



SCMS SCHOOL OF ENGINEERING AND TECHNOLOGY, KARUKUTTY

3.3.1 Number of research papers published per teacher in the Journals notified on UGC CARE list during 2020-2021

SL NO:	Title of paper	Name of the author/s	Department of the teacher	Name of journal	Calendar Year of publication	ISSN number	Link to the recognition in UGC enlistment of the Journal /Digital Object Identifier (doi) number		
							Link to website of the Journal	Link to article / paper / abstract of the article	Is it listed in UGC Care list
1	SDN Powered Humanoid with Edge Computing for Assisting Paralyzed Patients	Dr Varun G Menon	CSE	IEEE Internet of Things Journal	September 2020	ISSN 2327-4662	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI
2	The Hidden Role of Patriarchy in Malayalam Cinema: An Analysis of the movie 'Sufiyum Sujathayum'	Febini Joseph	Basic Science and Humanities	INTERNATIONAL JOURNAL OF ENGLISH LITERATURE AND SOCIAL SCIENCES	November 2020	ISSN: 2456-7620	https://ijels.com/	https://ijels.com/detail/the-hidden-role-	UGC CARE
3	"A novel spectrum sharing scheme using dynamic long short-term memory with CP-OFDMA in	Dr.Sunil Jacob	ECE	IEEE Transaction on Cognitive Communication	September 2020	ISSN: 2332-7731	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI
4	Service learning in engineering education: A study of student participatory survey for urban canal rejuvenation in Kochi, India	Dr Rathish Menon	CE	Procedia Computer Science	2020	ISSN: 1877-0509	https://www.sciencedirect.com/journal/procedia	http://dx.doi.org/10.1016/j.procs.2020.11.005	Scopus
5	Numerical Study on the Undrained Response of Silty Sands Under Static Triaxial Loading	Akhila M	CE	Advances in Computer Methods and Geomechanics, Lecture Notes in Civil Engineering	2020	ISSN: 2366-2557	https://link.springer.com/book/10.1007/978-3-030-54100-1_16	https://link.springer.com/chapter/10.1007/978-3-030-54100-1_16	Scopus
6	Effects of fines content and plasticity on liquefaction resistance of sands	Akhila M	CE	Proceedings of the Institution of Civil Engineers – Geotechnical Engineering	June 2020	ISSN 1353-2618	https://www.icevirtuallibrary.com/	https://doi.org/10.1680/jgeen.2020.13532618	SCI
7	PM10 source identification using the trajectory based potential source apportionment (TrapSA) toolkit at Kochi, India	Dr Rathish Menon	CE	Atmospheric Pollution Research	September 2020	ISSN: 1309-1042	https://www.sciencedirect.com/journal/atmospheric-pollution-research	https://www.sciencedirect.com/science/article/pii/S1309104220301000	SCI
8	Numerical Study on Cyclic Loading Effects on the Undrained Response of Silty Sand	Akhila M	CE	Geotechnical Characterization and Modelling, Lecture Notes in Civil Engineering, Springer	September 2020	ISSN: 2366-2557	https://link.springer.com/book/10.1007/978-3-030-54100-1_16	https://link.springer.com/chapter/10.1007/978-3-030-54100-1_16	Scopus
9	An efficient and adaptable multimedia system for converting PAL to VGA in real-time video processing	Dr Varun G Menon	CSE	Journal of Real Time Image Processing	December 2020	ISSN: 1861-8200	https://www.springer.com/journal/11097	https://doi.org/10.1007/s11554-020-00000-0	SCI
10	Dual-mode power reduction technique for real-time image and video processing board	Dr Varun G Menon	CSE	Journal of Real-Time Image Processing	June 2020	ISSN1861-8200	https://www.springer.com/journal/11097	https://link.springer.com/article/10.1007/s11554-020-00000-0	SCI
11	Secrecy Outage Probability of Relay Selection Based Cooperative NOMA for IoT Networks	Dr Varun G Menon	CSE	IEEE Access	December 2020	ISSN: 2169-3536	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI
12	Efficient Flow Processing in 5G-Envisioned SDN-Based Internet of Vehicles Using GPUs	Dr Varun G Menon	CSE	IEEE Transactions on Intelligent Transportation Systems	December 2020	ISSN 1558-0016	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI

13	A secure data deduplication system for integrated cloud-edge networks	Dr Varun G Menon	CSE	Journal of Cloud Computing	November 2020	ISSN: 2192-113X	https://journalofcloudcomputing.com	https://journalofcloudcomputing.com	SCI
14	Service offloading with deep Q-network for digital twinning empowered Internet of Vehicles in edge computing	Dr Varun G Menon	CSE	IEEE Transactions on Industrial Informatics	November 2020	ISSN: 1551-3203	https://ieeexplore.ieee.org	https://ieeexplore.ieee.org	SCI
15	SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network	Dr Varun G Menon	CSE	Journal of RealTime Image Processing	September 2020	ISSN1861-8200	https://www.springer.com/journal/11158	https://link.springer.com/article/10.1007/s11042-020-09512-3	SCI
16	I/Q Imbalance Aware Nonlinear Wireless-Powered Relaying of B5G Networks: Security and Reliability Analysis	Dr Varun G Menon	CSE	IEEE Transactions on Network Science and Engineering	September 2020	ISSN: 2327-4697	https://ieeexplore.ieee.org	https://ieeexplore.ieee.org	SCI
17	A Trust Analysis Scheme for Vehicular Networks within IoT-oriented Green City	Dr Varun G Menon	CSE	Environmental Technology & Innovation	November 2020	ISSN: 2352-1864	https://www.sciencedirect.com/journal/et&i	https://www.sciencedirect.com/science/article/pii/S2352186420300038	SCI
18	Enhancing the performance of flow classification in SDN-based intelligent vehicular networks	Dr Varun G Menon	CSE	IEEE Transactions on Intelligent Transportation Systems	August 2020	ISSN 1558-0016	https://ieeexplore.ieee.org	https://ieeexplore.ieee.org	SCI
19	SafeCity: Toward Safe and Secured Data Management Design for IoT-enabled Smart City Planning	Dr Varun G Menon	CSE	IEEE Access	August 2020	ISSN: 2169-3536	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI
20	Efficient equalisers for OFDM and DF-FT-OCDM multicarrier systems in mobile E-health video broadcasting with machine learning perspectives	Dr Varun G Menon	CSE	Physical Communication	October 2020	ISSN: 1874-4907	https://www.sciencedirect.com/journal/pc	https://www.sciencedirect.com/science/article/pii/S1874490720300038	SCI
21	Smart Sensing-enabled Decision Support System for Water Scheduling in Orange Orchard	Dr Varun G Menon	CSE	IEEE Sensors Journal	July 2020	ISSN: 1530-437X	https://ieeexplore.ieee.org	https://ieeexplore.ieee.org	SCI
22	High-performance flow classification using hybrid clusters in software defined mobile edge computing	Dr Varun G Menon	CSE	Computer Communications	July 2020	ISSN: 1873-703X	https://www.sciencedirect.com/journal/cc	https://www.sciencedirect.com/science/article/pii/S1873703X20300038	SCI
23	Into the World of Underwater Swarm Robotics: Architecture, Communication, Applications and Challenges	Dr Varun G Menon	CSE	Recent Patents on Computer Science	2020	ISSN: 2666-2566	https://benhamscience.com	https://www.curekjournals.com/article/94	SCOPUS
24	Failure mode effect and criticality analysis using dempster shafer theory and its comparison with fuzzy failure mode effect and criticality analysis: A case study	Nitty Rose Augustine	BSH	Process Safety and Environmental Protection	June 2020	ISSN: 1744-3598	https://www.sciencedirect.com/journal/psep	https://www.sciencedirect.com/science/article/pii/S1744359820300038	SCI
25	Energy-Efficient Transmission Range Optimization Model for WSN-Based Internet of Things	Dr.Varun G Menon	CSE	Computers, Materials and Continua	December 2020	ISSN: 1546-2225	https://www.techscience.com/journal/cm&c	https://www.techscience.com/cm&c/v6i7n	SCI
26	Exploring the narratives of Human Resilience in History & Highlighting their significance in Present Times as in The Diary of a Young Girl.	Divya M S	BSH	Galaxy: International Multidisciplinary Research Journal	2020	ISSN 2278-9529	https://www.galaxyintj.com/	https://www.the-criticon.com/V11i03/G	UGC CARE
27	Hardfacing and its effect on wear and corrosion performance of various ferrous welded mild steels	Suraj R	Mech	Materials Today Proceedings	January 2021	ISSN: 2214-7882	https://www.sciencedirect.com/journal/mtp	https://www.sciencedirect.com/science/article/pii/S2214788220300038	Scopus
28	A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System	Dr.Varun G Menon	CSE	Mobile Networks and Applications	January 2021	ISSN1572-8151	https://www.springer.com/journal/11158	https://link.springer.com/article/10.1007/s11042-020-09512-3	SCI
29	A Survey of Computational Intelligence for 6G: Key Technologies, Applications and Trends	Dr.Varun G Menon	CSE	IEEE Transactions on Industrial Informatics	January 2021	ISSN: 1941-0064	https://ieeexplore.ieee.org	https://ieeexplore.ieee.org	SCI
30	Hardware Impaired Ambient Backscatter NOMA Systems: Reliability and Security	Dr.Varun G Menon	CSE	IEEE Transactions on Communications	April 2021	ISSN: 1558-0844	https://www.comsoc.org/publicatio	https://ieeexplore.ieee.org	SCI

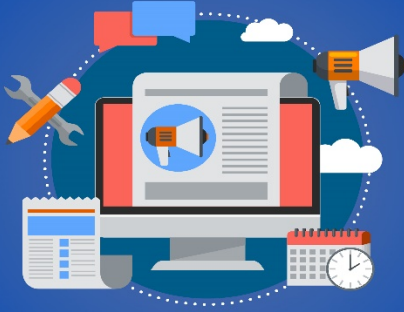
31	Learning based MIMO communications with imperfect channel state information for Internet of Things	Dr.Varun G Menon	CSE	Multimedia Tools and Applications	January 2021	ISSN1573-772	https://www.springer.com/journal/11	https://link.springer.com/article/10.1007/s00779-021-01016-1	SCI
32	Mechanical and tribological performance of Al-Fe-SiC-Zr hybrid composites produced through powder metallurgy process	Dr Vidyachandran	Mech	Materials Research Express	January 2021	ISSN: 2053-1591	https://iopscience.iop.org/	https://iopscience.iop.org/	SCI
33	Effect of nanofillers on the crystalline and mechanical properties of EVACO polymer nanocomposites	Dr Gibin George	Mech	Materials Today Proceedings	April 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/materials-today-proceedings	https://www.sciencedirect.com/science/article/pii/S2214785321000000	Scopus
34	Nature & Environmentalism: Post-Colonial Eco Critical rereading of selected Nigerian Poems.	Divya M.S	Basic Science	International Journal of Humanities and Arts	February 2021	ISSN 2664-7702	https://www.humanitiesjournals.net/	https://www.humanitiesjournals.net/	UGC CARE
35	A review on the use of ferrocement with stainless steel mesh as a rehabilitation technique	Anjana Susan John	CE	Materials Today Proceedings	February 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/materials-today-proceedings	https://dx.doi.org/10.1016/j.matpr.2021.02.001	Scopus
36	City scale water audit of a pilgrimage town in South India	Merin Mathew	CE	International Journal of Water Resources and Environmental Engineering	February 2021	ISSN 2141-6611	https://academicjournals.org/	https://academicjournals.org/	UGC CARE
37	Influence of multiwalled carbon nanotubes on the structure and properties of poly(ethylene-co-vinyl acetate-co-carbon monoxide) nanocomposites	Dr Gibin George	Mech	Polymer composites	May 2021	ISSN:0272-8399	https://4spublications.com/	https://4spublications.com/	SCI
38	IoT-powered deep learning brain network for assisting quadriplegic people	VINOJ P G	ECE	Computers and Electrical Engineering	March 2021	ISSN: 0045-7906	https://www.sciencedirect.com/journal/computer-electrical-engineering	https://www.sciencedirect.com/science/article/pii/S0045790621000000	SCI
39	Blockchain-based secure healthcare application for diabetic-cardio disease prediction in fog computing	Dr.Varun G Menon	CSE	IEEE Access	March 2021	ISSN: 2169-3535	https://ieeexplore.ieee.org/	https://ieeexplore.ieee.org/	SCI
40	Malware visualization and detection using DenseNets	Anandhi V	ECE	Personal and Ubiquitous Computing	May 2021	ISSN1617-4917	https://link.springer.com/article/10.1007/s00779-021-01016-1	https://doi.org/10.1007/s00779-021-01016-1	SCI
41	Automated Power Depiction System Using Iot Platform	Dr.Parvathy M	ECE	Turkish Online Journal of Qualitative Enquiry	April 2021	e-ISSN 1309-6591	https://www.tojqi.net/index	https://www.tojqi.net/index	Scopus
42	Hardware impaired modify-and-forward relaying with relay selection: Reliability and security	Dr.Varun G Menon	CSE	Physical Communication	April 2021	ISSN: 1876-3213	https://www.sciencedirect.com/journal/physical-communication	https://www.sciencedirect.com/science/article/pii/S1876321321000000	SCI
43	'Surface Modification of Tungsten Fillers for Application in Polymer Matrix Composites	Dr Jenson Joseph	Mech	Materials Today Proceedings	February 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/materials-today-proceedings	https://www.sciencedirect.com/science/article/pii/S2214785321000000	Scopus
44	An intelligent heart disease prediction system based on swarm-artificial neural network	Dr.Varun G Menon	CSE	Neural Computing and Applications	May 2021	ISSN1433-3055	https://www.springer.com/journal/10045	https://link.springer.com/article/10.1007/s00779-021-01016-1	SCI
45	Tribological and Corrosion analysis of Co-20Al-GNSA composites produced through powder metallurgy process	Dr Raghav G R	Mech	IOP Conf. Series: Materials Science and Engineering	2021	ISSN: 1757-899X	https://iopscience.iop.org/	https://iopscience.iop.org/	Scopus
46	Synthesis and characterization of Co-5Cr-RHA hybrid composite using Powder metallurgy	Dr Raghav G R	Mech	Materials Today Proceedings	April 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/materials-today-proceedings	https://www.sciencedirect.com/science/article/pii/S2214785321000000	Scopus
47	Mechanical and tribological performance of Al-Fe-SiC-Zr hybrid composites produced through powder metallurgy process	Dr Vidyachandran	Mech	Materials Research Express	January 2021	ISSN: 2053-1591	https://iopscience.iop.org/	https://iopscience.iop.org/	SCI
48	Impact of Ground Nut Shell Ash on Cobalt-Chromium metal matrix composites synthesized using Powder metallurgy process	Dr Raghav G R	Mech	IOP Conf. Series: Materials Science and Engineering	May 2021	ISSN: 1757-899X	https://iopscience.iop.org/	https://iopscience.iop.org/	Scopus

49	Dispersion analysis of nanofillers and its relationship to the properties of the nanocomposites	Dr Gibin George	Mech	Materials Today Proceedings	May 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/m	https://www.sciencedirect.com/science/ar	Scopus
50	Liquefaction resistance improvement of silty sands using cyclic preloading	Dr. Akhila M	CE	IOP Conference Series: Materials Science and Engineering	April 2021	ISSN: 1757-899X	https://iopscience.iop	https://iopscience.iop	Scopus
51	A review on the use of ferrocement with stainless steel mesh as a rehabilitation technique	Anjana Susan John	CE	Materials Today Proceedings	February 2021	ISSN: 2214-7853	https://www.sciencedirect.com/journal/m	http://dx.doi.org/10.1016/j.matpr	Scopus
52	Service Deployment Strategy for Predictive Analysis of FinTech IoT Applications in Edge Networks	Dr.Varun G Menon	CSE	IEEE Internet of Things Journal	May 2021	ISSN: 2327-4662	https://ieeexplore.ieee	https://ieeexplore.ieee	SCI
53	An intelligent heart disease prediction system based on swarm-artificial neural network	Dr.Varun G Menon	CSE	Neural Computing and Applications	May 2021	ISSN1433-3058	https://www.springer.com/journal/5	https://link.springer.com/article/10	SCI
54	Optimal Distribution of Workloads in Cloud-Fog Architecture in Intelligent Vehicular Networks	Dr.Varun G Menon	CSE	IEEE Transactions on Intelligent Transportation Systems	April 2021	ISSN: 1558-0016	https://ieeexplore.ieee	https://ieeexplore.ieee	SCI
55	Linked Data Processing for Human-in-the-Loop in Cyber-Physical Systems	Dr.Varun G Menon	CSE	IEEE Transactions on Computational Social Systems	April 2021	ISSN: 2329-924X	https://ieeexplore.ieee	https://ieeexplore.ieee	SCI
56	Blockchain-based secure healthcare application for diabetic-cardio disease prediction in fog computing	Dr.Varun G Menon	CSE	IEEE Access	March 2021	ISSN: 2169-3536	https://ieeaccess.ieee.org/	https://ieeexplore.ieee	SCI
57	Energy-Efficient Transmission Range Optimization Model for WSN-Based Internet of Things	Dr.Varun G Menon	CSE	Computers, Materials and Continua	December 2020	ISSN: 1546-2226(online)	https://www.techscience.com/journal/c	https://www.techscience.com/cm/v67n	SCI
58	IMPACT ANALYSIS OF THE COVID19 ON THE ATMOSPHERIC AIR QUALITY AND ELECTRICITY CONSUMPTION PER DAY IN	Dr.Sreelekha Menon	Basic Science and Humanities	Indian Journal of Applied Research	January 2021	ISSN:2249-555X	https://www.worldwidejournals	https://www.worldwidejournals	UGC CARE
59	An intelligent heart disease prediction system based on swarm-artificial neural network	Dr.Varun G Menon	CSE	Neural Computing and Applications	May 2021	0941-0643	https://link.springer.com/journal/5	https://link.springer.com/article/10	SCI
60	Energy-Efficient Transmission Range Optimization Model for WSN-Based Internet of Things	Dr.Varun G Menon	CSE	Computers, Materials and Continua	December 2020	ISSN: 1546-2226	https://www.techscience.com/journal/c	https://www.techscience.com/cm/v67n	SCI
61	Detection and robustness evaluation of android malware classifiers	Josna Philomina	CSE	Journal of Computer Virology and Hacking Techniques	May 2021	ISSN2197-9995	https://link.springer.com/journal/1	https://link.springer.com/article/10	SCI
Total number of research papers published per teacher in the Journals notified on UGC CARE list during 2020-2021									61



Joshi
PRINCIPAL
 SCMS SCHOOL OF ENGINEERING & TECHNOLOGY
 VIDYANAGAR, PALLISSERY, KARUKUTTY
 ERNAKULAM, KERALA-683 576

ISSN 0976 - 8165



THE CRITERION


AN INTERNATIONAL JOURNAL IN ENGLISH

11th Year of Open Access


**Bi-Monthly Refereed and Peer-Reviewed
Open Access e-Journal**

Vol. XI, Issue-3 (June 2020)

Editor-In-Chief : Dr. Vishwanath Bite
Managing Editor : Dr. Madhuri Bite



www.the-criterion.com



AboutUs: <http://www.the-criterion.com/about/>

Archive: <http://www.the-criterion.com/archive/>

ContactUs: <http://www.the-criterion.com/contact/>

EditorialBoard: <http://www.the-criterion.com/editorial-board/>

Submission: <http://www.the-criterion.com/submission/>

FAQ: <http://www.the-criterion.com/fa/>



ISSN 2278-9529

Galaxy: International Multidisciplinary Research Journal
www.galaxyimrj.com

Exploring the Narratives of Human Resilience in History and Highlighting their Significance in Present Times as in Anne Frank's *The Diary of a Young Girl*

Divya MS

Assistant Professor,

Department of English,

SCMS School of Engineering and Technology, Ernakulam.

Article History: Submitted-22/05/2020, Revised-30/06/2020, Accepted-01/07/2020, Published-10/07/2020.

Abstract:

“We realize the importance of our voice only when we are silenced”

Remembering the quotes of Malala Yousafzai, I would like to travel through the memoirs of holocaust faced by the Jews during Hitler's reign as marked in the writings of Anne Frank in her famous work “The Diary of a Young Girl”. It is indeed true that throughout the history, hundreds of thousands of individuals have undergone heart rending suffering and horrors beyond their worst dreams. Humans have, time and again exhibited extraordinary resilience in adapting to the situation. During these lock down days we do face a lot of discomfort and frustration to be confined into our own safe home. But just think about the thousands and millions who had gone on exile on fear of death and about the cruelties they had undergone. My attempt here is to pen down the agonies and resilience they had faced during these holocausts. Anne, a young Jewish girl is forced into hiding with her family and one other family in Nazi occupied Amsterdam. The inscriptions in the form of diary writing tells us about her feelings and experiences they had faced and also about her budding hopes to be free once again. Things began to change when the Nazis came to power. Their aim was to remove the Jews from German society even though they were less than 1% of the population. Nazi believed that Jews were the root of all the evils. Life was horrific for Jews and they began to flee from Germany. Nazis burned down the synagogues and Jewish owned shops and even burned their books. Jews were fleeing and tried to find shelter wherever they could. Nazis deported these people to forced labour camps, where they worked to produce supplies for the increasingly strained war economy. In most camps the prisoners were devoid of sufficient food, equipment, medicine and clothing. There was a complete disregard and their health was deteriorating day by day. As a result of these conditions, death rates in labour camps were extremely high.

Believing Holland was safe for Jews, Anne's family moved to Amsterdam in 1933. 'The Diary of a Young Girl' also known as 'The Diary of Anne Frank', a book of diary writings kept by Anne Frank while she was hiding for two years in the secret annex, with her family during the Nazi occupation of the Netherlands. The family was apprehended in 1944 and Anne Frank died of typhus in the Bergen-Belsen concentration camp in 1945. The diary was retrieved by Anne's father Mr. Otto Frank, the family's only known survivor after the war. The writings were from 14th June 1942 to 1st August 1944. Her father gifted her a red checked diary on her 13th birthday, June 12th 1942. It was not like a usual diary writing, she wrote as letters to her best friend that is, diary whom she addressed as Kitty. In August 1944, they were caught from the secret annex and were deported to Nazi concentration camps. Anne died when she was just fifteen years old. These letters were not just the experiences of a thirteen-year-old young girl, it gives us an insight into the most terrific inhumane situation that mankind had ever undergone. The wordings which she breathed became eternal and true.

"I want to go on living even after my death"

Keywords: Holocaust, Concentration Camps, Resilience.

"I want to bring out all kinds of things that lie buried deep in my heart".

As rightly said by Anne Frank, this is exactly what her writings were. A mixture of agonies, frustration, happiness, fear, realisations, relations, her first love and more over an account of what happened in the outside world. These letters were not just the experiences of a thirteen-year-old young girl, it gives us an insight into the most terrific inhumane situation that mankind had ever undergone. An account for her journals do tell us about how much Jews had suffered and deprived from the rest of the society. The restrictions imposed on them were even more harsh. They must always wear a yellow star and had to be indoors by eight o'clock and cannot even sit in their own gardens after that hour. They were forbidden to visit theatres, cinema halls and any other places of entertainment. Not allowed to take part in public sports, swimming baths, tennis courts, hockey fields and other sports grounds. They cannot visit any Christians and were allowed only to go to Jewish schools, many more such restrictions. A brief account of these imposed restrictions is clearly mentioned in the initial pages of the diary:

"The rest of our family, however, felt the full impact of Hitler's anti-Jewish laws, so life was filled with anxiety. After 1940 good times rapidly fled: first the war, then the

capitulation, followed by the arrival of the Germans, which is when the sufferings of us Jews really began. Anti-Jewish decrees followed each other in quick succession. Jews must wear a yellow star, Jews must hand in their bicycles, Jews are banned from trams and are forbidden to drive, Jews are only allowed to do their shopping between three and five o'clock and then only in shops which bear the placard "Jewish shop". (pg. 20, 21)

Amidst her class mates and friends she was all alone. She never had a real friend and was always in quarrel with her mother. She doesn't want to fit into the usual slot. Wished to have her own space and always voiced her own opinions. Father was her favourite and she used to say "I can understand my friends better than my own mother – too bad! (pg 53). She always longed for someone to be her best companion and to someone to whom she can express herself. And it is until when she received a diary as a birthday gift from father, she began to express all her emotions to her best friend "Kitty" – the diary. In her own words what the diary meant for her:

"In order to enhance in my mind's eye the picture of the friend for whom I have waited so long, I don't want to set down a series of bald facts in a diary like most people do, but I want this diary itself to be my friend, and I shall call my friend Kitty. No one will grasp what I am talking about if I begin my letters to Kitty just out of the blue, so, albeit unwillingly, I will start by sketching in brief the story of my life". (pg 20)

On July 5th 1942, Anne's elder sister Margot received an official summons to report to a Nazi work camp in Germany. On July 6th they went into hiding. They were later joined by Hermann Van Pels, Otto's business partner including his wife Auguste and their teenage son Peter. They hid in the sealed off upper rooms of the annex of Otto's company building in Amsterdam. The rooms they hid in were concealed behind a moveable book case, not easily noticeable. Mrs Van Pel's dentist Fritz Pfeffe, joined them four months later. They remained hidden there for two years and one month. Anne rightly called it as their "Secret Annex". They heard about the cruelties in camps and choose to be on exile than being caught. It is right what Otto Frank said before going into hiding, that "we don't want our belongings to be seized in by the Germans, but we certainly don't want to fall into their clutches ourselves. So we shall disappear of our own accord and not wait until they come and fetch us" (pg no. 31). All kinds of thoughts disturbed her as into where they are going to hide, "in a town or the country, in a house or a cottage, when, how, and where...?" (pg 33). It was quite natural to think of all such unique possibilities as all of a sudden when one is forced to go on exile. We cannot even

imagine of such a dreadful thing, all of a sudden to leave our belongings and to go somewhere we are not sure of, whether to live or to die. All of them wore two or three layers of dress and packed just enough to hold in a sachet and left their house with anxiety. On their way they received sympathetic looks from people and their face showed how sorry they were as they couldn't help because the gaudy yellow star spoke more than needed.

“Our many Jewish friends are being taken away by the dozen. These people are treated by the Gestapo without a shred of decency, being loaded into cattle trucks and sent to Westerbork, the big Jewish camp the big Jewish camp in Drente. Westerbork sounds terrible: only one washing cubicle for a hundred people and not nearly enough lavatories. There is no separate accommodation. Men, women and children all sleep together. One hears of frightful immorality because of this; and a lot of the women, and even girls, who stay there any length of time are expecting babies.” (pg 63)

It is impossible for them to escape, most of the people in the camp are branded as inmates by their shaven heads and many also by their Jewish appearance. If it is as bad as this in Holland whatever will it be like in the distant and barbarous regions they are sent to? We can assume that most of them are murdered. The English radio speaks of their being gassed. Perhaps that is quickest way to die. We feel helpless and sympathetic for them. Anne wrote an incident which really wets our eyes: “Just recently for instance, a poor old crippled Jewess was sitting on her doorstep; she had been told to wait there by the Gestapo, who had gone to fetch a car to take her away. The poor old thing was terrified by the guns that were shooting at English planes overhead, and by the glaring beams of the searchlights. No one would dare to take her in and to undergo such a risk.” (pg 64). The Germans strike without the slightest mercy. Prominent citizens and innocent people are thrown into prison to await their fate. If the saboteur can't be traced, the Gestapo simply put about five hostages against the wall. Announcements of their deaths appear in the papers frequently. These outrages are described as “fatal accidents” and countless people have gone to a terrible fate. Evening after evening the green and grey army lorries trundle past. The Germans ring at every front door to inquire if there are any Jews living in the house. No one has a chance of evading them unless one goes into hiding. Often, they go around with lists, and only ring when they know they can get a good haul. No one is spared not even the old people, babies, expectant mothers, the sick all join in the march towards death. Their nationality and even their very existence is being questioned. The fault is them is that they were born as Jews.

“Nice people, the Germans! To think that I was once one of them too! No, Hitler took away our nationality long ago. In fact, Germans and Jews are the greatest enemies in the world” (pg 65)

It is only on the second day of arrival Anne started writing her diary. It was about how she felt on hiding and the peculiar place and its ambience. “Then I had a chance, for the first time since our arrival, to tell you all about it, and at the same time to realize myself what had actually happened to me and what was still going to happen” (pg 40). There was no variety in thoughts or nothing new to be done. They go round and round like a roundabout – from Jews to food and from food to politics. It isn’t an easy task to go on hiding with such alerts outside. Day and night more of those poor miserable people are being dragged off, with nothing but a rucksack and a little money. On the way they are deprived even of these possessions. Families are torn apart, the men, women and children all being separated. Children coming from school find that their parents have disappeared. Women return from shopping to find their homes shut up and their families gone. Every night hundreds of planes fly over Holland and go to German towns, where the earth is ploughed up by their bombs, and every hour hundreds and thousands of people are killed in Russia and Africa. No one is able to keep out of it, the whole globe is waging war and although it is going better for the Allies, the end is not yet in sight. And as for us, we are fortunate. Yes, we are luckier than millions of people. The children here run about in just a thin blouse and clogs, no coat, no hat, no stockings and no one helps them. Their tummies are empty, they chew an old carrot to stay the pangs, go from their cold homes out into the cold street and when they get to school, find themselves in an even colder classroom. Countless children stop the passers-by and beg for a piece of bread. There is nothing we can do but wait as calmly as we can till the misery comes to an end. Jews and Christians wit, the whole earth waits, and there are many who wait for death. It was not easy to go on hiding because of the terrific things happened outside, an account of what was happening outside:

“We had a short circuit last evening, and on top of that the guns kept banging away all the time. I still haven’t got over my fear of everything connected with shooting and planes, and I creep into Daddy’s bed nearly every night for comfort. I know it is very childish but you don’t know what it is like. The A.A. guns roar so loudly that you can’t hear yourself speak. Mrs. Van Daan, the fatalist, was nearly crying, and said in a very timid little voice, “Oh, it is so unpleasant! Oh, they are shooting so hard,” by which she really means am so frightened.” (pg 100)

Apart from that there were frequent banging outside that feared everyone and Anne gathered all her belongings together. She packed a suitcase with the most necessary things for an escape. But as her mother rightly said “Where will you escape to?” (pg114). They are Jews, can't go anywhere. Even the nature, birds, animals are free to do as they like but not them. They are even degraded to that level. On exile they have only option to divert their mind books, reading and studying new things. “Ordinary people simply don't know what books mean to us, shut up here, reading, learning and the radio are our amusements.” (pg 121). With the little ration they receive from fake cards they moved on. Celebrated birthday's with whatever they find. One such poem written by Margot on Anne's birthday tells us how their daily life and thoughts have been immersed in fear and agonies.

“The first shot sounds at dead of night

Hush, look! A door creaks open wide,

A little girl glides into sight,

Clasping a pillow to her side” (pg 136)

Not just that Anne used to swallow Valerian pills every day against worry and depression, but that doesn't prevent her from being even more miserable the next day. She wrote “a good hearty laugh would help more than ten Valerian pills, but we have almost forgotten how to laugh. I feel afraid sometimes that from having to be so serious I will grow a long face and my mouth will droop at the corners”. (pg 150). She wrote about the ambience there to be so oppressive and sleepy and as heavy as lead. They can't hear a single bird singing outside and a deadly close silence hangs everywhere, catching hold of them as if it will drag them down deep into an underworld. She used to wander from one room to another, downstairs and up again, feeling like a song bird whose wings have been clipped and who is hurling himself in utter darkness against the bars of his cage. She longed to “Go outside, laugh, and take a breath of fresh air, a voice cries within me, but I don't even feel a response anymore; I go and lie on divan and sleep, to make the time pass more quickly and the stillness and the terrible fear, because there is no way of killing them” (pg 155). We could understand what she needs:

“When someone comes in from outside, with the wind in their clothes and the cold on their faces, then I could bury my head in the blankets to stop myself thinking: “When will we be granted the privilege of smelling fresh air?” And because I must not bury my head in the

blankets, but the thoughts will come. Believe me, if you have been shut up for a year and a half, it can get too much for you some days. In spite of all the justice and thankfulness you can't crush your feelings. Crying, dancing, whistling, looking out into the world, feeling young, to know that am free – that is what I long for, still I must not show it, because I sometimes think is all eight of a us began to pity ourselves or went about with discontented faces where would it lead us? I couldn't talk this to anyone but only van cry. Crying can bring such relief” (pg 168)

There are a number of organisations such as “The Free Netherlanda” which forge identity cards, supply money to people “underground”, find hiding places for people, and work for young men in hiding and it is amazing how much noble, unselfish work these people are doing, risking their own lives to help and save others. Our helpers are a very good example. They have pulled us through up till-now and we hope they will bring us safely to dry land. Otherwise, they will have to share the same fate as the many others who are being searched for. Never had they heard one word of the burden which they certainly must be to them, never has one of them complained of all the trouble we give. They put on the brightest possible faces, bring flowers and presents for birthdays and bank holidays are always ready to help and do all they can. That is something we must never forget; although the Germans our helpers display heroism in their cheerfulness and affection “(pg 195). Rauter, one of the German big shots, has made a speech. “All Jews must be out of the German occupied countries before July 1. Between April 1 and May 1 the province of Utrecht must be cleaned out (as if Jews are cockroaches). Between May 1 and June 1 the provinces of North and South Holland.”(pg 108). We cannot even imagine to face such a dreadful situation. On 29th march 1944, she heard a London radio broadcast made by the exiled Dutch minister for education, art and science Gerrit Bolkestein, calling for the preservation of “ordinary documents – a diary, letters....simple everyday material” to create an archive for posterity as testimony to the suffering of civilians during the Nazi occupation. That is when she began to write more seriously that someone may read it. In August 1944, they were discovered and deported to Nazi concentration camps and that is what she heard of her last. As she said she is living in many minds even after her death and too years and years apart. It is really relevant what she said, we also need to do that to keep our minds engaged during these lockdown days.

“I finally realized that I must do my school work to keep from being ignorant, to get on in life, to become a journalist, because that is what I want! I know I can write.....but it remains to be seen whether I really have talent.....”

Exploring the Narratives of Human Resilience in History and Highlighting their Significance in Present Times
as in Anne Frank's *The Diary of a Young Girl*

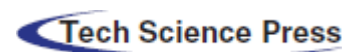
Works Cited:

Frank, Anne. *The Diary of a Young Girl*: Lexicon Books. 2011

Abrams, M.H. *A Glossary of Literary Terms*. Canada: Wadson Cengage Learning, 2009s

[\[BACK\]](#)*Computers, Materials & Continua*

DOI:10.32604/cmc.2021.015426

Article

Energy-Efficient Transmission Range Optimization Model for WSN-Based Internet of Things

Md. Jalil Piran¹, Sandeep Verma², Varun G. Menon³ and Doug Young Suh^{4,*}

¹Department of Computer Science and Engineering, Sejong University, Seoul, Korea

²Department of Electronics and Communication Engineering, D.B.R.A. National Institute of Technology, Jalandhar, India

³SCMS School of Engineering and Technology, Ernakulam, India

⁴Department of Electronics Engineer, Kyung Hee University, Yongin, Korea

*Corresponding Author: Doug Young Suh. Email: suh@khu.ac.kr

Received: 02 November 2020; Accepted: 10 December 2020

Abstract: With the explosive advancements in wireless communications and digital electronics, some tiny devices, sensors, became a part of our daily life in numerous fields. Wireless sensor networks (WSNs) is composed of tiny sensor devices. WSNs have emerged as a key technology enabling the realization of the Internet of Things (IoT). In particular, the sensor-based revolution of WSN-based IoT has led to considerable technological growth in nearly all circles of our life such as smart cities, smart homes, smart healthcare, security applications, environmental monitoring, etc. However, the limitations of energy, communication range, and computational resources are bottlenecks to the widespread applications of this technology. In order to tackle these issues, in this paper, we propose an Energy-efficient Transmission Range Optimized Model for IoT (ETROMI), which can optimize the transmission range of the sensor nodes to curb the hot-spot problem occurring in multi-hop communication. In particular, we maximize the transmission range by employing linear programming to alleviate the sensor nodes' energy consumption and considerably enhance the network longevity compared to that achievable using state-of-the-art algorithms. Through extensive simulation results, we demonstrate the superiority of the proposed model. ETROMI is expected to be extensively used for various smart city, smart home, and smart healthcare applications in which the transmission range of the sensor nodes is a key concern.

Keywords: Internet of Things; wireless sensor networks; routing; transmission range optimization; energy-efficiency; hot-spot problem; linear programming

1 Introduction

1.1 Background and Problem Statement

Data-driven wireless sensor networks (WSNs) are widely applied to enhance the Internet of Things (IoT) in terms of the data throughput, energy efficiency, and self-management [1]. WSN-based IoTs are composed of wireless sensor nodes, which realize data collection and communication [2,3]. In this framework, the sensor nodes are deployed in the physical environment to sense the phenomena and report their readings in a distributed manner to the sinks [4]. However, the sensor nodes exhibit certain limitations in terms of energy, computation resources, and communication range [5,6].

When a WSN-based IoT is deployed over a large application area, the nodes perform multihop communication due to the limited transmission range, and direct data transmission cannot be realized. Furthermore, it has been reported that a larger number of relay nodes on the path of data delivery to the sink corresponds to a higher probability of these nodes closer to the sink suffering from hot-spot

problem [7]. In such a scenario, the number of intermediate nodes should be reduced to decrease the emergence of a no-connection zone for distantly located nodes.

Moreover, the battery of the sensor nodes may not be able to be changed or recharged. Therefore, it is necessary to ensure efficient power consumption in a WSN-based IoT [8]. Furthermore, transmitting one kilobyte of data corresponds to the processing of three million instructions [9]. Therefore, data transmission in the WSNs should be minimized with regard to the distance between any two entities among sensor nodes, cluster heads (CHs), or sinks [10].

One solution is to maximize the transmission range between nodes. The key concept of transmission range maximization is that if a sensor initiates a data packet transmission to a sink located 1000 m away, the least number of relay sensors should be selected to forward the packet. The communication range of sensor nodes depends on their transmission power and the volume of the packet to be transmitted. Transmission over long distances requires a higher energy [11,12]. Therefore, it is necessary to determine the maximum possible distance (transmission range) to which the sensor nodes can transmit the data packets.

Many researchers have attempted to reduce the energy consumption by avoiding the hot-spot problem [13]. In particular, Verma et al. [14] proposed the multiple sink-based genetic algorithm-based optimized clustering (MS-GAOC) approach, in which four data collection sinks were incorporated outside the network. However, the cost of using four sinks may be prohibitive in various applications.

Moreover, researchers generally apply the corona-based model to avoid hot-spot problems. A survey of the various corona-based approaches has been presented in an existing study [15]. Nevertheless, even corona-based methods are not sufficiently reliable in mitigating the hot-spot problem. In fact, the literature review indicates that the concept of transmission range adjustment for the sensor nodes, to realize direct data transfer to the sink or transfer with the least possible number of intermediate nodes, has not been extensively investigated.

1.2 Motivation

The review pertaining to the mitigation of hot-spot problems indicated that the optimization-based approach can provide a balanced solution to specific problems. Therefore, in this work, we used linear programming (LP) to compute the maximum data transmission range [16,17]. In particular, LP exhibits remarkable exploration and exploitation capabilities, enabling fast convergence to the optimal solution. Moreover, LP is highly computationally efficient [16].

1.3 Our Contributions

In the context of the aforementioned problems, the key contributions of this work are as follow:

- a) We propose an energy-efficient transmission range optimized model for IoT (ETROMI) to optimize the transmission range of the sensor nodes to reduce the hot-spot problem in WSN-based IoT.
- b) The mathematical model and formulation using LP is presented.
- c) The simplex method is used to solve the defined problem.
- d) The proposed model's performance of the proposed model is analyzed in terms of various aspects, and the optimal solution is identified.

1.4 Paper Organization

The remaining paper is structured as follows. Section 2 presents the background of transmission range adjustment algorithms and describes the existing work pertaining to the hot-spot problem in WSNs. Section 3 describes the system model and explains the LP formulation. Section 4 describes the performance evaluation of ETROMI, which is used to compute the maximized distance corresponding to the transmission range of a node. The concluding remarks, along with the limitations and scope for future work, are presented in Section 5.

2 Related Work

In this section, we discuss the existing work focused on addressing the hot-spot problem through various state-of-the-art techniques and on realizing the transmission range adjustment of a sensor node.

2.1 Approaches to Solve the Hot-Spot Problem

In applications involving an extremely large network area, the sensor nodes inevitably perform multi-hop communication [18]. In this process, a hot-spot is created at the nodes located nearest to the sink. Several researchers have addressed this concern through various topology-based methods. Moreover, the many-to-one approach (many sensor nodes corresponding to one sink) has been widely implemented through corona-based structures [13]. Many researchers use the term “energy-hole,” which is equivalent in meaning to a hot-spot.

Elkamel et al. [19] proposed an unequal clustering method to overcome the hot-spot problem by placing the small and large clusters nearer to and farther from the sink, respectively. However, the proposed technique failed to eliminate the hot-spot problem, and the network’s energy consumption was high. Verma et al. [7,14] implemented multiple data sinks in a given network to mitigate the hot-spot problem. In their former and latter studies, the authors used the conventional approach and the genetic algorithm, respectively. However, the network incurred a higher financial cost owing to the use of multiple data sinks. The authors in [20] proposed a virtual-force-based energy-hole mitigation strategy to ensure sensor nodes’ uniform distribution. Moreover, the network was composed of various annuli, and virtual gravity was used to optimize the sensor node positions in each annulus. However, due to the multi-hop communication, the number of overheads in each annulus was extremely high, which increased the energy consumption in the network. Sharmin et al. [21] proposed a strategy in which the network was partitioned into several wedges, and residual energy was considered to combine the various wedges. The head node was selected based on the distance between the innermost corona and node. However, the inefficient selection of the head node led to the mediocre performance of this strategy.

In addition to the static network scenario, certain researchers introduced sink mobility to curb the hot-spot problem. Sahoo et al. [22] proposed a particle-swarm-optimization-based energy-efficient clustering and sink mobility (PSO-ECSM) technique, in which the sink mobility was used to alleviate the hot-spot problem. However, the mobility scenario was not efficiently utilized, and the slow convergence of the PSO degraded the performance of the proposed scheme. Furthermore, Kaur et al. [23] introduced dual sink mobility outside the network to target unattended applications. Although the authors implemented the PSO-based sink mobility, the use of the dual sink introduced overheads in the network, which increased the energy consumption. In addition, the data delivery was required to be synchronized when using the two sinks in the network. Certain other researchers also employed the sink mobility scenario to alleviate the hot-spot problem. However, it was observed that the use of sink mobility limited the applicability of the approaches in various real-time scenarios.

2.2 Transmission Range Adjustment Algorithms

In addition to the network topological changes associated with the introduction of the corona-based model, the characteristics of sensor nodes have been examined. The focus of the present study is to optimize the transmission range. Although certain researchers have attempted to adjust the transmission range to alleviate the hot-spot problem, the proposed approaches suffer from the inherent problems, which limit their relevance.

In an existing strategy [24] pertaining to the transmission range adjustment, the network was divided into various concentric sets termed as coronas. Every corona was assigned a transmission range level. Furthermore, the authors presented an ant colony optimization (ACO)-based transmission range adjustment strategy [24] to prolong the network lifetime. Liu [25] considered the energy consumption balancing (ECB) and energy consumption minimization (ECM) techniques to avoid the occurrence of energy holes. The authors exploited the short-trip moving scheme for the ACO, which helped in decreasing the complexity and in the amelioration of convergence speed. Furthermore, the authors considered a reference transmission distance to implement the ECB and ECM techniques. Xin et al. [26] were the first to attempt to solve the many-to-one data transmission problem, particularly in strip-based WSNs. The authors adjusted the transmission range based on the computation of the accurate distance. The objective was to prolong the network lifetime. However, the proposed algorithm was applicable only for strip-based WSNs, for example, railway track, bridge, and tunnel systems.

In summary, only a few studies have been focused on addressing the hot-spot problem through transmission range adjustment, and this approach exhibits considerable scope for improvement. Furthermore, the use of LP for energy-hole mitigation in the transmission range adjustment context is yet to be explored. Therefore, we implement these aspects in our proposed strategy.

3 System Model

In this section, we describe the network assumptions and the system model.

3.1 Network Assumptions

The following network assumptions are considered to implement ETROMI.

- The WSN is composed of one sink and several sensor nodes that collect data and transfer them to the sink.
- Each sensor has a unique ID.
- There is no dispute for medium access, and thus, proportional fair channel access is available to all the sensors.
- The minimum cost forwarding approach is employed as the multi-hop-routing protocol.
- The sensor nodes are homogeneous, i.e., all the nodes have the same configuration in terms of energy, computational resources, transmission range, etc.
- The entire network is static, including the CHs and the sink.
- The entire network has ideal conditions in terms of security, physical medium factors, reflection, refraction, splitting of signals, and presence of other obstacles.

3.2 Fundamental Principle of ETROMI

Assume a WSN with N sensor nodes, in which one of the sensor nodes initiates a data packet with the intent to transmit it to the sink (final receiver base station). In the conventional clustering method, a CH collects the data from the sensors node in the corresponding cluster and forwards the data toward the sink via the other CHs. However, this approach is not efficient because the CHs suffer from battery limitations, even more than the other data collecting elements, but must be involved in all transmissions.

In contrast, the lifetime of the WSNs depends on the remaining energy of the members, i.e., the sensor nodes. Therefore, the number of forwarding nodes must be minimized. In this study, we assume that instead of always selecting the CHs to receive and forward the data packets, the sensors select the farthest sensor node in their transmission range. In other words, the sensor nodes increase their transmission power to transmit a data packet over a longer distance. In this configuration, the number of nodes that are involved in a transmission are minimized, which can considerably improve the energy efficiency. In Fig. 1, the red dotted line represents the routing procedure in a clustering-based method.

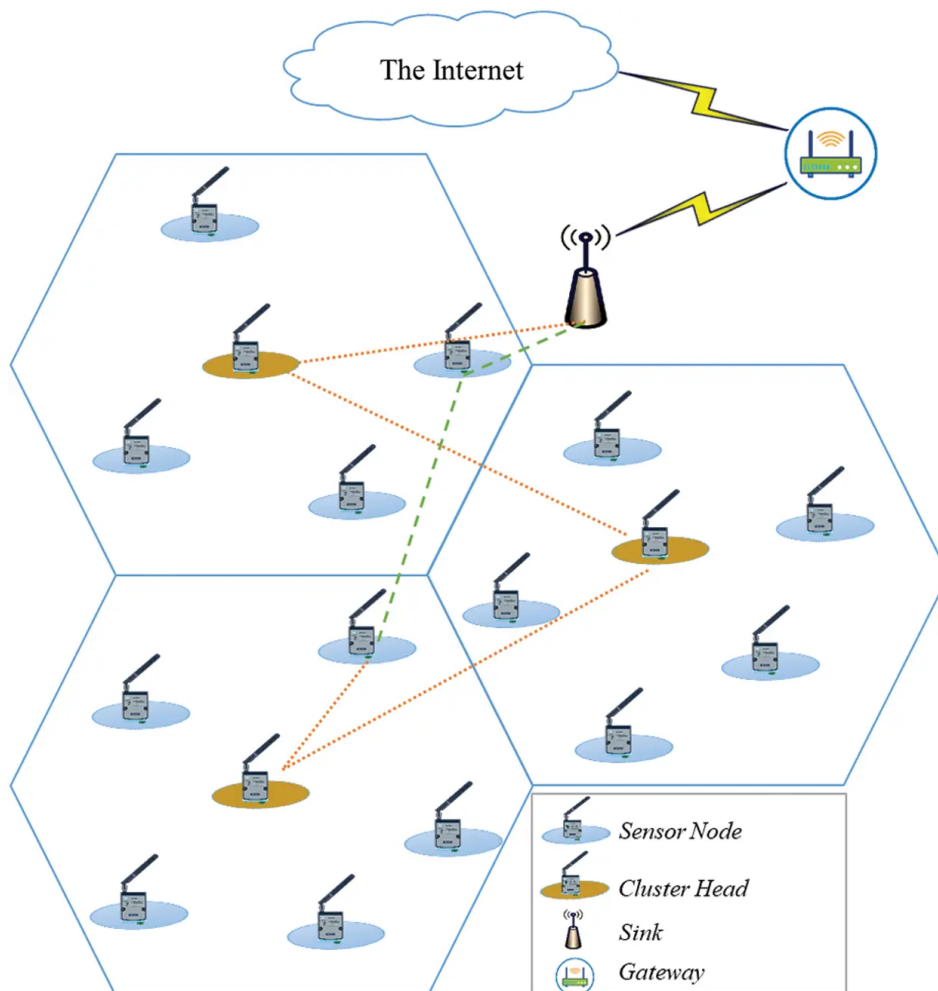


Figure 1: Routing procedure in WSNs

In this approach, the nodes send their packets to their corresponding CH, which then forwards the packet to the next CH and so on. Finally, the closest CH delivers the packet to the sink. In contrast, in the approach represented by the green dashed line, the node that initiates the packet sends the packet to the farthest node, and the receiver node follows the same principle and send the packet to the farthest node in its transmission range. Consequently, the number of nodes involved in the transmission procedure is less than that in the clustering-based method.

Consider a network involving 100 sensor nodes. Sensor 1 initiates a data packet and wants to send it to node 100. As mentioned earlier, in the WSNs, the topology is multi-hop. In other words, node 1 sends its packet to its neighbor, which receives the packet and forwards it to the neighboring nodes, excluding the node that the packet was received from. This process continues until node 100 receives the packet. The problem then is to determine the number of nodes in the transmission process that receive and forward the packet. This number of the relay nodes should be minimized to abate the energy consumption and, in turn, prolongs the network lifetime.

One solution is to increase the transmission range of the nodes involved in the transmission process from the source to the destination. In this case, a node selects a neighboring node, which is far from it, but in its range, i.e., on the edge of its transmission range, and the number of intermediate nodes is decreased. To this end, we consider the energy consumption accounted to transmission and reception of data packets and also the magnitude of data packets. The total required energy can be expressed as

$$E_{ij} = E^{TX} \times P_i \times d_{ij} + N \times E^{RX} + \sum_{i=1}^{N-1} D_i. \quad (1)$$

The list of main symbols used in this paper are listed in [Tab. 1](#).

Table 1: List of symbols

Symbol	Definition
D_i	The distance between node i to the sink
E^I	The initial power
E_{ij}	Total energy consumption of a link
E^{RX}	The receiving power
E^{TX}	Transmission power
P_i	Packet volume
d_{ij}	The distance between node i and j
N	The number of nodes

3.3 LP in ETROMI

Consider a WSN-based IoT represented by graph $G = (V, E)$, in which $V = v_1, v_2, v_3, \dots, v_n$ is the set of sensor nodes, and $E = e_1, e_2, e_3, \dots, e_n$ is the set of direct wireless links between the nodes, such that $E \subseteq V \times V$. Link (i, j) exists if and only if $j \in L_i$, where L_i is the set of all nodes that can be reached by sensor i directly with a certain transmission power level. Furthermore, each sensor i has the initial power E^I . The transmission energy consumed by node i to send a data packet to the neighboring sensor j is $E_{ij}^{TX} = \{e_{1j}^{TX}, e_{2j}^{TX}, e_{3j}^{TX}, \dots, e_{nj}^{TX}\}$; $E_{ij}^{RX} = \{e_{1j}^{RX}, e_{2j}^{RX}, e_{3j}^{RX}, \dots, e_{nj}^{RX}\}$ is the energy required for a node to receive a packet from node i ; and $P = \{p_1, p_2, p_3, \dots, p_n\}$ is the set of the packet volumes.

The objective function is to maximize the transmission distance with respect to the packet volume and the transmission and receiving energies, that is

$$\max_x \sum_{j=1}^n d_{ij}, \quad (2)$$

$$s.t. \sum_{i=1}^n p_i + \sum_{j=1}^n d_j \leq P, \quad (3)$$

$$\sum_{i=1}^n E_i^{TX} < E^{TX}, \quad (4)$$

$$\sum_{i=1}^n E_i^{RX} < E^{RX}, \quad (5)$$

$$E^{TX} + E^{RX} < E^I. \quad (6)$$

Constraint (3) specifies that the total number of packets received or transmitted to/from node i must be less than a threshold for a specific time slot. Constraints (4) and (5) control the maximum transmission and receiving energy consumptions, respectively. Constraint (6) ensures that the total consumed energy for transmission and receiving by node i , is not less than the initial energy of the node.

4 Performance Evaluation

To evaluate the performance of the technique, we consider that a data packet is to be sent from a sensor node to the sink via two intermediate sensor nodes. Therefore, four sensor nodes are involved in the process: one sender, one sink, and two relay nodes. The size of each packet is 50 bits, and the transmission and receiving power are 100 and 70 W, respectively.

4.1 Linear Problem

The linear problem can be expressed as

$$\max_x \quad x_1 + x_2 + x_3, \quad (7)$$

$$s.t. \quad -x_1 - 2x_2 + 4x_3 \leq 50, \quad (8)$$

$$-2x_1 + 5x_2 - 2x_3 \leq 100, \quad (9)$$

$$4x_1 - 2x_2 - x_3 \leq 70, \quad (10)$$

$$x_1, x_2, x_3 \geq 0. \quad (11)$$

By adding slack variables to constraints, the primal problem in standard format is represented as follows:

$$\max_x \quad x_1 + x_2 + x_3 + 0x_4 + 0x_5 + 0x_6, \quad (12)$$

$$s.t. \quad -x_1 - 2x_2 + 4x_3 + x_4 = 50, \quad (13)$$

$$-2x_1 + 5x_2 - 2x_3 + x_5 = 100, \quad (14)$$

$$4x_1 - 2x_2 - x_3 + x_6 = 70, \quad (15)$$

$$x_1, x_2, x_3, x_4, x_5, x_6 \geq 0. \quad (16)$$

4.2 Simplex Method

We use the simplex method to solve the problem. The simplex method is used to solve LP models by using slack variables, tableaus, and pivot variables to determine the optimal solution of an optimization problem [27]. To solve the optimization problem, the following steps are performed:

- a) Obtain the standard form,
- b) Introduce slack variables,
- c) Create the tableau,
- d) Identify the pivot variables,
- e) Create a new tableau,
- f) Check for optimality,
- g) Identify the optimal values.

The procedure starts with an initialization phase, followed by several iterations to determine the optimal solution.

The initialization step for our optimization problem is presented in Tab. 2. After the first step, x_6 is the leaving variable, x_1 is the entering variable, and 4 is the pivot element.

Table 2: Stating section

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	RHS	Θ
$0x_4$	-1	-2	4	1	0	0	50	-
$0x_5$	-2	5	-2	0	1	0	100	-
$0x_6$	4	-2	-1	0	0	1	70	70/4
$C_j - Z_j$	1	1	1	0	0	0	0	

Subsequently, we apply the first iteration, as indicated in Tab. 3. Upon completing this iteration, the leaving variable is x_5 , the entering variable is x_2 , and the pivot element is 4.

Table 3: Iteration I

$0x_4$	0	-5/2	15/4	1	0	1/4	270/4	-
$0x_5$	0	4	-5/2	0	1	1/2	135	135/4
$1x_1$	1	-1/2	-1/4	0	0	1/4	70/4	
$C_j - Z_j$	0	3/2	5/4	0	0	-1/4	70/4	

We continue by applying the second iteration as indicated in Tab. 4, which results in a leaving variable x_4 , an entering variable x_3 , and a pivot element, 35/16.

Table 4: Iteration II

$0x_4$	0	0	35/16	1	5/8	9/16	1215/8	486/7
$1x_2$	0	1	-5/8	0	1/4	1/8	135/4	-
$1x_1$	1	0	-9/16	0	1/8	5/16	275/8	-
$C_j - Z_j$	0	0	35/16	0	-3/8	-7/16	545/8	

We proceed to the third iteration, in which all $C_j - Z_j$ values are zero or negative; therefore, the simplex method is terminated at this step, as indicated in Tab. 5. The optimal solution for the defined problem is presented in Tab. 6.

Table 5: Iteration III

$1x_3$	0	0	1	16/35	2/7	9/35	486/7	
$1x_2$	0	1	0	2/7	3/7	2/7	540/7	
$1x_1$	1	0	0	9/35	2/7	16/35	514/7	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

Table 6: The optimal solution

Z	220
x_1	514/7
x_2	540/7
x_3	486/7

4.3 Duality

The duality refers to a specific relationship between an LP problem and another problem, both of which involve the same original data, albeit located differently [28]. The former and latter problems are referred to as the primal and dual problems, respectively. The feasible regions, optimal solutions, and optimal values of these problems must be strongly correlated. The duality and optimality conditions obtained from these aspects are a basis for the LP theory. Once either of the primal or dual problems is solved, both the problems can be solved owing to duality. To convert the primal problem to a dual problem, the following steps are performed:

- If the primal problem corresponds to “Maximize,” the dual problem corresponds to “Minimize.”
- The number of variables in the dual problem is equal to the number of constraints in the primal problem.
- The number of constraints in the dual problem, is equal to the number of variables in the primal problem.
- The coefficients of the objective function in the dual problem, are equal to the right-hand side (RHS) values in the primal problem.
- The RHS values in the dual problem are equal to the coefficients of the objective function in the primal problem.
- The coefficient variables in the constraints of the dual problem correspond to the transpose matrix of the coefficient variables in the primal problem.
- “ \leq ” constraints in the primal problem are “ \geq ” constraints in the dual problem, and vice versa.
- The variables in the dual problem are denoted as “ y ”.
- The objective function is denoted as “ w ”. The primal problem is as follows:

Our primal problem is as below:

$$\max \quad x_1 + x_2 + x_3, \quad (17)$$

$$s.t. \quad -x_1 - 2x_2 + 4x_3 \leq 50, \quad (18)$$

$$-2x_1 + 5x_2 - 2x_3 \leq 100, \quad (19)$$

$$4x_1 - 2x_2 - x_3 \leq 70, \quad (20)$$

$$x_1, x_2, x_3 \geq 0. \quad (21)$$

Coefficient matrix of basic variables in objective function is $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, Coefficient matrix of

basic variables in constraints is $\begin{bmatrix} -1 & -2 & 4 \\ -2 & 5 & -2 \\ 4 & -2 & -1 \end{bmatrix}$, and RHS matrix is $RHS = \begin{bmatrix} 50 \\ 100 \\ 70 \end{bmatrix}$. Based on

the aforementioned steps, our dual problem will be as bellow:

$$\max \quad 50y_1 + 100y_2 + 70y_3, \quad (22)$$

$$s.t. \quad -y_1 - 2y_2 + 4y_3 \geq 1, \quad (23)$$

$$-2y_1 + 5y_2 - 2y_3 \geq 1, \quad (24)$$

$$4y_1 - 2y_2 - y_3 \geq 1, \quad (25)$$

$$y_1, y_2, y_3 \geq 0. \quad (26)$$

By adding surplus variables, the dual problem is as follows:

$$\max \quad 50y_1 + 100y_2 + 70y_3 + 0y_4 + 0y_5 + 0y_6, \quad (27)$$

$$s.t. \quad -y_1 - 2y_2 + 4y_3 - y_4 = 1, \quad (28)$$

$$-2y_1 + 5y_2 - 2y_3 - y_5 = 1, \quad (29)$$

$$4y_1 - 2y_2 - y_3 - y_6 = 1, \tag{30}$$

$$y_1, y_2, y_3, y_4, y_5, y_6 \geq 0. \tag{31}$$

As indicated in the dual problem, no identity matrix exists for the coefficients of the variables in the constraints; therefore, artificial variables must be introduced. In this case, the dual problem in the standard format is:

$$\max \quad 50y_1 + 100y_2 + 70y_3 + 0y_4 + 0y_5 + 0y_6 - Ma_1 - Ma_2 - Ma_3, \tag{32}$$

$$s.t. \quad -y_1 - 2y_2 + 4y_3 - y_4 + a_1 = 1, \tag{33}$$

$$-2y_1 + 5y_2 - 2y_3 - y_5 + a_2 = 1, \tag{34}$$

$$4y_1 - 2y_2 - y_3 - y_6 + a_3 = 1, \tag{35}$$

$$y_1, y_2, y_3, y_4, y_5, y_6 \geq 0. \tag{36}$$

By adding artificial variables, an identity matrix can be generated, and the simplex method can be implemented.

As indicated in Tab. 7, after completing the initialization section, the leaving variable is a_3 , the entering variable is x_1 , and the pivot element is 4. Subsequently, we implement the first iteration, as indicated in Tab. 8. Upon completing iteration I, the leaving variable is a_2 , the entering variable is x_2 , and our pivot element is 4. We then proceed to the second iteration, as indicated in Tab. 9. After the second iteration, the leaving variable is a_1 , the entering variable is x_3 , and the pivot element is 35/16. We attempt to determine the optimal solution by using the two-phase simplex method. All the artificial variables are removed, and the problem can be solved through the other variables. We then apply the third iteration in two phases, as indicated in Tabs. 10 and 11.

Table 7: Starting section

Min	50	100	70	0	0	0	-M	-M	-M	RHS	Θ
	y_1	y_2	y_3	y_4	y_5	y_6	a_1	a_2	a_3		
-Ma ₁	-1	-2	4	-1	0	0	1	0	0	1	-
-Ma ₂	-2	5	-2	0	-1	0	0	1	0	1	-
-Ma ₃	4	-2	-1	0	0	-1	0	0	1	1	1/4
$C_j - W_j$	-1	-1	-1	1	1	1	0	0	0	3	

Table 8: Iteration I

-Ma ₁	0	-9/4	15/4	-1	0	-1/4	1	0	1/4	5/4	-
-Ma ₂	0	4	-9/4	0	-1	-1/2	0	1	1/2	3/2	3/8
50y ₁	1	-1/2	-1/4	0	0	-1/4	0	0	1/4	1/4	-
$C_j - W_j$	0	-3/2	-5/4	1	1	3/4	0	0	1/4	11/4	

Table 9: Iteration II

-Ma ₁	0	0	35/16	-1	-5/8	-9/16	1	5/8	9/16	35/16	1
100y ₂	0	1	-5/8	0	-1/4	-1/8	0	1/4	1/8	3/8	-
50y ₁	1	0	-9/16	0	-1/8	-5/16	0	1/8	5/16	7/16	-
$C_j - W_j$	0	0	-35/16	1	5/8	9/16	0	3/8	7/16	35/16	

Table 10: Phase I, Iteration III

70y ₃	0	0	1	-16/35	-2/7	-9/35	16/35	2/7	9/35	1
100y ₂	0	1	0	-2/7	-3/7	-2/7	2/7	3/7	2/7	1
50y ₁	1	0	0	-9/35	-2/7	-16/35	9/35	2/7	16/35	1
$C_j - W_j$	0	0	0	0	0	0	1	1	1	0

Table 11: Phase II, Iteration III

$70y_3$	0	0	1	$-16/35$	$-2/7$	$-9/35$	1
$100y_2$	0	1	0	$-2/7$	$-3/7$	$-2/7$	1
$1y_1$	1	0	0	$-9/35$	$-2/7$	$-16/35$	1
$C_j - W_j$	0	0	0	$514/7$	$540/7$	$486/7$	220

Finally, it is observed that the primal solution, presented in Tab. 12, is equal to the dual solution, presented in Tab. 13, that is $Z^* = W^*$.

Table 12: Primal optimal Solution

Z	220
x_1	$514/7$
x_2	$540/7$
x_3	$486/7$

Table 13: Dual optimal solution

W	220
y_1	1
y_2	1
y_3	1

4.4 Sensitivity Analysis

Sensitivity analysis is aimed at examining the influence of changes in the variables, such as the RHS, coefficients of the objective function, and constraints, on the solution. We start with Tab. 14 and make the some changes as explained in the next subsection.

Table 14: Simplex optimum tableau

Maximize	<u>1</u>	<u>1</u>	<u>1</u>	<u>0</u>	<u>0</u>	<u>0</u>	RHS
	x_1	x_2	x_3	x_4	x_5	x_6	
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$486/7$
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$540/7$
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$514/7$
$C_j - Z_j$	0	0	0	-1	-1	-1	220

4.4.1 Change in the Objective Function Coefficient for Non-Basic Variables

In the last iteration, no non-basic variables of the objective function exist. Therefore, if one of the coefficients is changed, the optimal solution is not influenced.

4.4.2 Change in the RHS Value

Suppose the intention is to change the first RHS to b_1 ; then we have $\begin{bmatrix} 50 \\ 100 \\ 70 \end{bmatrix}$. To calculate new RHS;

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} b_1 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} 16b_1 + 1630 \\ 2b_1 + 440 \\ 9b_1 + 2120 \end{bmatrix} \quad (37)$$

$$\therefore \begin{cases} \frac{16b_1 + 1630}{35} \geq 0 \Rightarrow b_1 \geq -\frac{815}{8} \\ \frac{2b_1 + 440}{7} \geq 0 \Rightarrow b_1 \geq -80 \\ \frac{9b_1 + 2120}{35} \geq 0 \Rightarrow b_1 \geq -\frac{725}{9} \end{cases} \quad (38)$$

Because, $b_1 \geq -815/8$. We suppose a b_1 value beyond the specified range; as an example -110 .

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} -110 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} 278 \\ 220 \\ 226 \\ 7 \end{bmatrix} \quad (39)$$

Now, we continue the tableau with new RHS values;

As indicated in Tab. 15, the primal solution is not feasible; therefore, we attempt to find the optimal solution through the dual problem.

Table 15: New RHS values

Maximize	1	1	1	0	0	0	RHS	⊖
	x_1	x_2	x_3	x_4	x_5	x_6		
1 x_3	0	0	1	16/35	2/7	9/35	278/7	
1 x_2	0	1	0	2/7	3/7	2/7	220/7	
1 x_1	1	0	0	9/35	2/7	16/35	226/7	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

The initialization step, as the first iteration, is presented in Tab. 16. After completing the initialization step, the leaving variable is x_1 , the entering variable is x_5 , and the pivot element is $3/7$.

Table 16: Initialization step

1 x_3	0	0	1	16/35	2/7	9/35	278/7
1 x_2	0	1	0	2/7	3/7	2/7	220/7
1 x_1	1	0	0	9/35	2/7	16/35	226/7
$C_j - Z_j$	0	0	0	-1	-1	-1	220
⊖	-	0	-	-7/2	-7/3	-7/2	

Subsequently, we implement the second iteration, as indicated in Tab. 17. After finishing the second iteration, it is noted that the primal is feasible; the leaving variable is x_5 , the entering variable is x_2 , and the pivot element is $7/3$.

Table 17: Iteration II

1 x_3	0	-2/3	1	28/105	0	7/105	394/21	-
0 x_5	0	7/3	0	2/3	1	2/3	220/3	220/7
1 x_1	1	-2/3	0	7/105	0	28/105	34/3	-
$C_j - Z_j$	0	7	0	-1/3	0	-1/3	30.1	

We then implement the third iteration as indicated in Tab. 18. All the values for $C_j - Z_j$ are zero or negative; therefore, the process is terminated at this step. The optimal solution is as presented in Tab. 19.

Table 18: Iteration III

$1x_3$	0	0	1	48/105	6/21	37/105	278/7
$1x_2$	0	1	0	2/7	3/7	2/7	220/7
$1x_1$	1	0	0	27/105	6/21	48/105	226/7
$C_j - Z_j$	0	0	0	-1	-1	-115/105	724/7

Table 19: Optimal solution

Z	724/7
x_1	226/7
x_2	220/7
x_3	278/7

4.4.3 Change in the Objective Function Coefficient for the Basic Variable

We consider the case in which the coefficient of x_1 changes. Suppose the coefficient of x_1 is c_1 .

Any change in the coefficient of the basic variables of the objective function affects the value of $C_j - Z_j$.

$$\text{For } x_4 \Rightarrow C_j - Z_j = 0 - \left[\left(1 \times \frac{16}{35}\right) + \left(1 \times \frac{2}{7}\right) + \left(c_1 \times \frac{9}{35}\right) \right] \Rightarrow \frac{-9c_1}{35} - \frac{26}{35} \tag{40}$$

$$\text{For } x_5 \Rightarrow C_j - Z_j = 0 - \left[\left(1 \times \frac{2}{7}\right) + \left(1 \times \frac{3}{7}\right) + \left(c_1 \times \frac{2}{7}\right) \right] \Rightarrow \frac{-2c_1}{7} - \frac{5}{7} \tag{41}$$

$$\text{For } x_6 \Rightarrow C_j - Z_j = 0 - \left[\left(1 \times \frac{9}{35}\right) + \left(1 \times \frac{2}{7}\right) + \left(c_1 \times \frac{16}{35}\right) \right] \Rightarrow \frac{-16c_1}{35} - \frac{19}{35} \tag{42}$$

If $C_j - Z_j \leq 0$ then the present solution remains optimal solution;

$$\frac{-9c_1}{35} - \frac{26}{35} \leq 0 \Rightarrow c_1 \leq \frac{-26}{9} \tag{43}$$

$$\frac{-2c_1}{7} - \frac{5}{7} \leq 0 \Rightarrow c_1 \leq \frac{-5}{2} \tag{44}$$

$$\frac{-16c_1}{35} - \frac{19}{35} \leq 0 \Rightarrow c_1 \leq \frac{-19}{16} \tag{45}$$

In this case, the range of c_1 is greater than $-26/9$. Thus, we assign c_1 beyond this range, for example $c_1 = -4$, and implement the first iteration, as indicated in Tab. 20.

Table 20: Iteration I

Maximize	-4	1	1	0	0	0	RHS	Θ
	x_1	x_2	x_3	x_4	x_5	x_6		
$1x_3$	0	0	1	16/35	2/7	9/35	486/7	270
$1x_2$	0	1	0	2/7	3/7	2/7	540/7	270
$-4x_1$	1	0	0	9/35	2/7	16/35	514/7	1285/8
$C_j - Z_j$	0	0	0	2/7	3/7	9/7	-1030/7	

Upon completing the first iteration, the leaving variable is x_3 , the entering variable is x_6 , and the pivot element is 9/35.

The second iteration is presented in Tab. 21. All the values for $C_j - Z_j$ are zero or negative; therefore, the process is terminated at this step. We conclude that the optimal solution is as follows: $x_1 = -50, x_2 = 0, x_3 = 270$ and $z = 470$.

Table 21: Iteration II

$1x_3$	0	0	$35/9$	$16/9$	$10/9$	1	270
$1x_2$	0	1	$-10/9$	$-2/9$	$1/9$	0	0
$-4x_1$	1	0	$-16/9$	$-7/9$	$-2/9$	0	-50
$C_j - Z_j$	0	0	$-80/9$	$-42/9$	$-19/9$	-1	470

4.4.4 Change in the Constraint Coefficient Corresponding to Non-basic Variables

In the last iteration, no non-basic variable of the objective function exists. Therefore, if one of the coefficients is changed, the optimal solution is not influenced.

4.4.5 Addition of a New Variable

Consider a new variable x_7 with coefficient $c_7 = 12$ and $P_7 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$, then;

$$\bar{P}_7 = B^{-1}P_7 \Rightarrow \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 638 \\ 105 \\ 7 \\ 61 \\ -35 \end{bmatrix} \tag{46}$$

In this case, we perform three iterations as indicated in Tabs. 22–24.

Table 22: Iteration I

Maximize	<u>1</u>	<u>1</u>	<u>1</u>	<u>0</u>	<u>0</u>	<u>0</u>	RHS	⊖
	x_1	x_2	x_3	x_4	x_5	x_6		
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$486/7$	
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$540/7$	
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$514/7$	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

Table 23: Iteration II

Maximize	<u>1</u>	<u>1</u>	<u>1</u>	<u>0</u>	<u>0</u>	<u>0</u>	12	RHS	⊖
	x_1	x_2	x_3	x_4	x_5	x_6	x_7		
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$212/35$	$486/7$	$104/9$
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$12/7$	$540/7$	45
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$61/35$	$514/7$	$2570/61$
$C_j - Z_j$	0	0	0	-1	-1	-1	$137/105$		

Table 24: Iteration III

Maximize	<u>1</u>	<u>1</u>	<u>1</u>	<u>0</u>	<u>0</u>	<u>0</u>	12	RHS	⊖
	x_1	x_2	x_3	x_4	x_5	x_6	x_7		
$12x_7$	0	0	$3/18$	$8/9$	$15/9$	$27/18$	1	$104/18$	
$1x_2$	1	0	$37/127$	$68/319$	$65/319$	$67/319$	0	$36/18$	
$1x_1$	0	1	$-18/63$	$-78/63$	$-153/63$	$-144/63$	0	$235/18$	
$C_j - Z_j$	0	0	-1	$-87/9$	$-160/9$	$-143/9$	0	115	

Upon completing the first iteration, the leaving variable is x_3 , the entering variable is x_7 , and the pivot element is $212/35$.

The third iteration is presented in Tab. 12, in which all the values for $C_j - Z_j$ are zero or negative and; therefore, the program is terminated at this step, and the optimal solution is as indicated in Tab. 25.

Table 25: Optimal solution

Z	115
x_1	$235/18$
x_2	$36/18$
x_7	$104/18$

4.4.6 Addition of a New Constraint

To examine the influence of the addition of a new constraint to the problem, we consider $x_3 \leq 40$:

As indicated in Tab. 26, the optimal solution is as follows: $x_1 = 514/7$, $x_2 = 540/7$, $x_3 = 486/7$, and $Z = 20$.

Table 26: Additional constraint

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	$\frac{0}{x_7}$	RHS
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	0	$486/7$
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	0	$540/7$
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	0	$514/7$
$0x_7$	0	0	1	0	0	0	1	40
$C_j - Z_j$	0	0	0	-1	-1	-1	0	220

5 Conclusion and Future Direction

The transmission range of a sensor node defines whether the communication mode is single-hop or multi-hop. In this paper, we proposed the use of ETROMI, which can determine the maximum distance to which a sensor node can transmit data with the least possible number of relay nodes. We presented an LP-based analytical model to determine the transmission range of the sensor node. Moreover, we explained the mathematical model associated with the ETROMI to reduce the energy consumption of WSN-based IoT. A key concern about the ETROMI is that it considers the ideal conditions involving no obstacles between the sensor nodes and the sink. Therefore, the model performance is specific to the circumstances. Furthermore, the network is assumed to be homogeneous, whereas homogeneity does not exist in an actual network due to the different factors associated with network deployment. In future work, we aim to extend our work to address the aforementioned scenarios.

Funding Statement: This research was supported by Korea Electric Power Corporation (Grant Number: R18XA02).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. S. Kumar and V. K. Chaurasiya. (2018). "A strategy for elimination of data redundancy in Internet of Things (IoT) based wireless sensor network (WSN)," *IEEE Systems Journal*, vol. 13, pp. 1650–1657.
2. P. Swarna, P. Maddikunta, M. Parimala, S. Koppu, T. Gadekallu et al. (2020). , "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," *Computer Communications*, vol. 160, pp. 139–149.
3. R. Vinayakumar, M. Alazab, S. Srinivasan, Q. Pham, S. Padannayil et al. (2020). , "A visualized bot net detection system based deep learning for the Internet of Things networks of smart cities," *IEEE*

Transactions on Industry Applications, vol. 56, no. 4, pp. 4436–4456.

4. M. Piran, Y. Cho, J. Yun and D. Y. Suh. (2014). “Cognitive radio-based vehicular ad hoc and sensor networks (CR-VASNET),” *International Journal of Distributed Sensor Networks*, vol. 2014, pp. 1–11.
5. T. M. Behera, S. K. Mohapatra, U. C. Samal and M. S. Khan. (2019). “Hybrid heterogeneous routing scheme for improved network performance in WSNs for animal tracking,” *Internet of Things*, vol. 6, pp. 1–9.
6. T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand et al. (2019). , “Residual energy-based cluster-head selection in WSNs for IoT application,” *IEEE Internet of Things Journal*, vol. 6, pp. 5132–5139.
7. S. Verma, N. Sood and A. K. Sharma. (2019). “A novelistic approach for energy efficient routing using single and multiple data sinks in heterogeneous wireless sensor network,” *Peer-to-Peer Networking and Applications*, vol. 12, pp. 1110–1136.
8. Y. Liu, C. Yang, L. Jiang, S. Xie and Y. Zhang. (2019). “Intelligent edge computing for IoT-based energy management in smart cities,” *IEEE Network*, vol. 33, pp. 111–117.
9. D. K. Gupta. (2013). “A review on wireless sensor networks,” *Network and Complex Systems*, vol. 3, no. 1, pp. 18–23.
10. L. Krishnasamy, R. K. Dhanaraj, G. D. Ganesh, G. Reddy, M. K. Aboudaif et al. (2020). , “A heuristic angular clustering framework for secured statistical data aggregation in sensor networks,” *Sensors*, vol. 20, pp. 1–15.
11. S. Bhattacharya, P. Maddikunta, S. Somayaji, K. Lakshmana, R. Kaluri et al. (2020). , “Load balancing of energy cloud using wind driven and firefly algorithms in internet of everything,” *Journal of Parallel and Distributed Computing*, vol. 142, pp. 16–26.
12. C. Iwendi, P. K. Maddikunta, T. R. Gadekallu, K. Lakshmana, A. K. Bashir et al. (2020). , “A metaheuristic optimization approach for energy efficiency in the IoT networks,” *Software: Practice and Experience*, vol. 22, no. 6, pp. 1–14.
13. H. Asharioun, H. Asadollahi, T. C. Wan and N. Gharaei. (2015). “A survey on analytical modeling and mitigation techniques for the energy hole problem in corona-based wireless sensor network,” *Wireless Personal Communications*, vol. 81, pp. 161–187.
14. S. Verma, N. Sood and A. K. Sharma. (2019). “Genetic algorithm-based optimized cluster head selection for single and multiple data sinks in heterogeneous wireless sensor network,” *Applied Soft Computing*, vol. 85, pp. 1–21.
15. A. U. Rahman, A. Alharby, H. Hasbullah and K. Almuzaini. (2016). “Corona based deployment strategies in wireless sensor network: A survey,” *Journal of Network and Computer Applications*, vol. 64, pp. 176–193.
16. D. Bertsimas and J. N. Tsitsiklis. (1997). *Introduction to Linear Optimization*, vol. 6. Belmont, MA: Athena Scientific.
17. V. Tabus, D. Moltchanov, Y. Koucheryavy, I. Tabus and J. Astola. (2015). “Energy efficient wireless sensor networks using linear-programming optimization of the communication schedule,” *Journal of Communications and Networks*, vol. 17, pp. 184–197.
18. V. Sandeep, N. Sood and A. K. Sharma. (2019). “QoS provisioning-based routing protocols using multiple data sink in IoT-based WSN,” *Modern Physics Letters*, vol. 34, pp. 1–36.
19. R. Elkamel, A. Messouadi and A. Cherif. (2019). “Extending the lifetime of wireless sensor networks through mitigating the hot spot problem,” *Journal of Parallel and Distributed Computing*, vol. 133, pp. 159–169.
20. C. Sha, C. Ren, R. Malekian, M. Wu, H. Huang et al. (2019). , “A type of virtual force-based energy-hole mitigation strategy for sensor networks,” *IEEE Sensors Journal*, vol. 20, pp. 1105–1119.
21. N. Sharmin, A. Karmaker, W. Lambert, M. Alam and M. Shawkat. (2020). “Minimizing the energy hole problem in wireless sensor networks: A wedge merging approach,” *Sensors*, vol. 20, pp. 1–25.
22. B. Sahoo, T. Amgoth and H. Pandey. (2020). “Particle swarm optimization based energy efficient clustering and sink mobility in heterogeneous wireless sensor network,” *Ad Hoc Networks*, vol. 106, pp. 1–21.
23. S. Kaur and V. Grewal. (2020). “A novel approach for particle swarm optimization-based clustering with dual sink mobility in wireless sensor network,” *International Journal of Communication Systems*, vol. 33, no. 16, pp. 1–
24. M. Liu and C. Song. (2012). “Ant-based transmission range assignment scheme for energy hole problem in wireless sensor networks,” *International Journal of Distributed Sensor Networks*, vol. 8, pp. 1–12.

25. X. Liu. (2016). "A novel transmission range adjustment strategy for energy hole avoiding in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 67, pp. 43–52.
26. H. Xin and X. Liu. (2017). "Energy-balanced transmission with accurate distances for strip-based wireless sensor networks," *IEEE Access*, vol. 5, pp. 16193–16204.
27. V. Zhadan. (2019). "Two-phase simplex method for linear semidefinite optimization," *Optimization Letters*, vol. 13, pp. 1969–1984.
28. S. Nasser and D. Darvishi. (2018). "Duality results on grey linear programming problems," *The Journal of Grey System*, vol. 30, pp. 127–142.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 Access through SCMS School of Engine... Purchase P... Access throug



Process Safety and Environmental Protection

Volume 138, June 2020, Pages 337-348



Article preview

Abstract


Introduction


Section snippets

References (52)

Cited by (25)

Failure mode effect and criticality analysis using dempster shafer theory and its comparison with fuzzy failure mode effect and criticality analysis: A case study applied to LNG storage facility

Manoj Jose Kalathil ^a , V.R. Renjith ^a , Nitty Rose Augustine ^b 

Show more 

+ Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.psep.2020.03.042>

[Get rights and content](#)

Abstract

Author

X

Nitty Rose Augustine

[View in Scopus](#)

SCMS, School of Engineering and Technology, Kerala, India

More documents by Nitty Rose Augustine

Provided by Scopus

[Failure mode effect and criticality analysis using dempster shafer...](#)

Process Safety and Environmental Protect...
Kalathil, M.J., ..., Augustine, N.R.

Activate Windows

Go to Settings to activate Windows.

FEEDBACK 

Editor-in-Chief >>

ISSN (Print): 2666-2558
ISSN (Online): 2666-2566

Back Journal Subscribe

INTO THE WORLD OF UNDERWATER SWARM ROBOTICS: ARCHITECTURE, COMMUNICATION, APPLICATIONS AND CHALLENGES

Author(s): Koyippilly Satheesh Keerthi, Bandana Mahapatra and Varun Girijan Menon*

Volume 13, Issue 2, 2020

Page: [110 - 119]

Pages: 10

DOI: 10.2174/2213275912666181129141638

Price: \$65

Purchase

PDF



Abstract

Background: With the curiosity of exploring the underwater world, science has devised various technologies and machines that can help them in performing activities like exploring, navigating and plunging into the unknown world of oceanography. Underwater Robot or vehicle can be claimed as an outcome of extensive research done by the scientists who aimed at discovering the unknown mysterious world of ocean and how it can benefit humanity. Swarm robotics is an entirely new section of robotics that has been developed based on

system.

Methods: A systematic review on state-of-the-art has been performed where contributions of various researchers was considered. The study emphasis on the concepts, technical background, architecture and communication medium along with its applicability in various fields that also include various issues and challenges faced while attaining them.

Results: The incorporation of swarm intelligence in underwater robotics provides a new angle altogether into the working pattern of underwater robotic system.

Conclusion: The article is a systematic presentation of swarm robot technologies, their mechanisms, conceived and designed communication medium with respect to adaptability of the vehicle to the versatile nature of water. The paper delineates the various conceptual and technical details and its beneficence to humanity.

Keywords: Applications, architect communication medium, AUV architecture, swarm intelligence, swarm robotics, issues and challenges, underwater swarms.

Graphical Abstract

Article Metrics



PDF

29



HTML

7



12

12 Total citations
3 Recent citations

3.74 Field Citation Ratio
n/a Relative Citation Ratio

Home About Publications Publish with us Marketing Opportunities Articles by Disease For Librarians For Authors & Editors More

Journal Information


- > About Journal
- > Editorial Board
- > Current Issue
- > Volumes /Issues

For Authors

For Editors



High-performance flow classification using hybrid clusters in software defined mobile edge computing

Mahdi Abbasi^a  , Azad Shokrollahi^a, Mohammad R. Khosravi^{b c}, Varun G. Menon^d

Show more 

 Outline |  Share  Cite

<https://doi.org/10.1016/j.comcom.2020.07.002> 

[Get rights and content](#) 

Abstract

Mobile Edge Computing (MEC) provides different storage and computing capabilities within the access range of mobile devices. This moderates the burden of offloading compute/storage-intensive processes of the mobile devices to the centralized cloud data centers. As a result, the network latency is reduced and the quality of service provided for the mobile end users is improved. Different applications benefit from the large-scale deployments of MEC servers. However, the considerable complexity of managing large scale deployments of the sheer number of applications for the millions of mobile devices is a challenge. Recently, Software Defined Networking (SDN) is leveraged to resolve the problem by providing unified and programmable interfaces for managing network devices. Most of the current SDN packet processing services are tightly dependent on the packet classification service. This primary service classifies any incoming packet based on matching a set of specific fields of its header against a flow table. Acceleration of this basic process considerably increases the performance of the SDN-based MEC. In this paper, the hierarchical tree algorithm, which is a packet classification method, is parallelized using popular platforms on a cluster of Graphics Processing Units (GPUs), a cluster of Central Processing Units (CPUs), and a hybrid cluster. The best scenario for the parallel implementation of this algorithm on the CPU cluster is that which combines OpenMP and MPI.

In this case, the throughput of the classifier is 4.2 million packets per second (MPPS). On the GPU cluster, two different scenarios have been used. In the first scenario, the global memory is used to store the rules and the Hierarchical-trie of the classifier while in the second scenario we break the filter set in a way that the resulting Hierarchical-trie of each subset could be stored in the shared memory of GPU. According to the results, although the first GPU cluster scenario achieves a throughput of 29.19 MPPS and a speedup 58 times as great as the serial mode, the second scenario is 12 times faster due to using the shared memory. The best performance, however, belongs to the hybrid cluster mode. The hybrid cluster achieves a throughput of 30.59 which is 1.4 MPPS more than the GPU cluster.

 Previous

Next 

Keywords

Flow classification; Graphics Processing Unit (GPU); GPU cluster; Mobile edge computing; Software Defined Networking; Software Defined Mobile Edge Computing (SDMEC)

1. Introduction

The advent of several mobile applications, such as intelligent transport systems[1], virtual reality[2], Human activity recognition, and control[3], [4], [5], [6], and smart environments[1], [7], [8] has implied the availability of a large pool of computing and storage resources. Hence, the considerable growth in data volume that comes from the massive number of devices enabled by 5G has made mobile edge computing more important than ever before[9], [10]. Beyond its abilities to reduce network traffic and improve user experience, edge computing also plays a critical role in enabling use cases for ultra-reliable low-latency communication in industrial manufacturing and a variety of other sectors[11]. Especially, facilitating cloud-like resources at the edge of the network is a challenging issue for the telecommunication sector[12]. By the introduction of the Mobile Edge Computing (MEC) in 2014, a sustainable business model is provided for mobile operators, service providers, and mobile subscribers[13], [14]. MEC aims to provide cloud capabilities within the Radio Access Network (RAN) in the area of mobile subscribers[15]. That is, it provides accelerated services, contents, and applications by increasing availability at the edge[16]. Recently, flexible and scalable solutions of SDN for a large set of challenges that are encountered within the traditional networking approach, have introduced it as a suitable collaborator for MEC[17]. When the intrinsic properties of SDN are considered, three practical models for fruitful collaboration of SDN and MEC in real-world scenarios would be exploited, including multi-tier edge computing architecture, service-centric access to the edge, and network function virtualization (NFV)[17], [18]. SDN has proficiencies of arranging the network, its services, and devices by hiding the complexities of the varied mobile environment from end-users. Thus, SDN can moderate the barriers and limits that multi-tier MEC infrastructure will meet. The SDN control mechanism can reduce the complexity of MEC by utilizing accessible resources more efficiently. SDN dynamically routes the traffic between MEC servers and cloud servers to deliver the highest quality of service to end-users. In the second model of collaboration, the NFV platform of SDN can be dedicated to MEC or shared with other network functions or applications. In this model, MEC can use NFV management and orchestration entities and interfaces. Finally, as the third form of collaboration, SDN provides high speeds in content delivery between the MEC and central cloud systems.

These collaboration models result in several benefits including high resolution & effective control, flexibility and low barrier on innovation, service-centric implementation, virtual machine mobility, adaptability, interoperability, low-cost solutions, and multiplicity of scope[17].

SDN uses the limited network resources optimally and enables flexible network management by separating the control operations from the data management[19]. For this purpose, the forwarding switch nodes cannot take decisions on their own, instead, a software-based controller that has a general view of the underlying network makes the forwarding and routing decisions. All operations of the SDN controller, especially its flexible communication with switches are carried out according to the OpenFlow protocol. The chief functionalities of the SDN controller include managing the flow tables on the forwarding nodes, collecting statistics, and populating them by editing the packet classification rules. To classify a packet appearing at an ingress port of an SDN switch, the switch performs a lookup on its flow table, according to a classification algorithm to find any matching rule. If found, the corresponding action on the packet. Otherwise, the switch requests the controller to figure out the most appropriate action. The decided action is applied to the packet, and the necessary classification rule is installed on the corresponding switch.

The performance of the classification engine of an SDN switch has a great impact on the overall performance of the system[20], [21]. There are different methods for parallel implementation of packet classification algorithms[1], [22], [23], [24], [25], [26]. Some of these methods, e.g. parallelization of algorithms on GPUs and multi-core CPUs, have been recently implemented. Due to the limitation of hardware resources, however, the throughput of these systems can hardly reach the cumulative throughput rates of current network systems. A common solution to overcome this limitation is CPU clusters and GPU clusters. The present study is the first attempt to parallelize the

Hierarchical-trie algorithm on a CPU cluster, a GPU cluster, and a hybrid cluster. The innovations of this study are as follows:

- For the first time, a cluster system has been used for packet classification.
- This study compares for the first time the performance of programming based on [Message Passing Interface \(MPI\)](#) with that of OpenMP-based programming in cluster computations.
- In all the scenarios, the effect of different types of memory in the hierarchy of GPU memory on the performance of concurrent processes of packet classification in a GPU cluster is investigated.
- The parameters of memory usage, speedup, and throughput are measured based on the results of our implementation of the scenarios which are combinations of MPI, CUDA, and OpenMP.

The paper is organized as follows. Section 2 discusses the Hierarchical-trie packet classification algorithm. Next, cluster systems and their programming will be described. In Section 3, the literature on different architectures of the cluster and parallel implementations of packet classification algorithms will be reviewed. Section 4 provides a description of the proposed scenarios for the parallelization of Hierarchical-trie algorithm on CPU clusters, GPU clusters, and hybrid clusters. In the next section, the implementation results will be analyzed and evaluated. The final section compares the results of our work with other findings in this field and proposes suggestions for further research.

2. Tools and algorithms

This section describes the structure of the Hierarchical-trie algorithm as well as how this algorithm classifies internet packets. Next, cluster computing and its related [parallelization](#) tools are discussed.

2.1. Hierarchical binary search tree

Decision-tree algorithms have an appropriate algorithmic structure for parallelization. A key decision-tree algorithm is the Hierarchical-trie algorithm which has a relatively low memory usage. In the Hierarchical-trie algorithm, the [decision tree](#) used for classification is constructed according to prefixes in source and destination IP addresses. For example, we can use the tree illustrated in [Fig. 1](#) based on the 9 filters listed in [Table 1](#).

To construct the tree, the bits of the prefix of the source IP address of the filters are read consecutively. Depending on whether a 0 or 1 is met, the left or right side of the tree is formed, respectively. A wildcard sign (*) denotes that the formation of the tree based on the source IP address of that rule is completed.

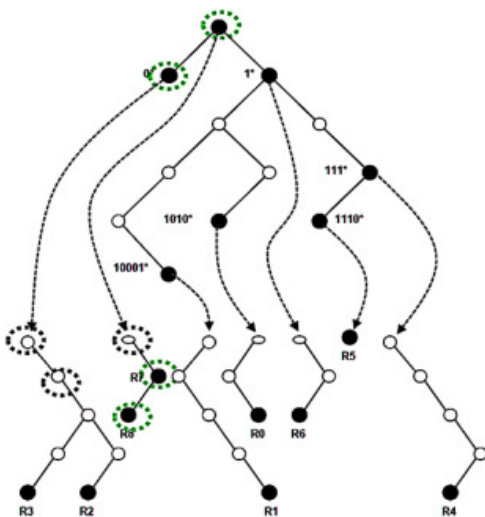
Table 1. Example of a filter set [27].

Filter	Destination IP	Source IP	Destination port	Source port	Protocol
R0	1010*	01*	0,65536	25,25	6
R1	10001*	0111*	53,53	443,443	4
R2	0*	1110*	53,53	1024,65535	17
R3	0*	1100*	53,53	443,443	4
R4	111*	1110*	53,53	25,25	4
R5	1110*	*	0,65535	2788,2788	17
R6	1*	10*	53,53	5632,5632	6
R7	*	1*	53,53	25,25	6
R8	*	10*	0,65535	2788,2788	17

In the next step, a pointer is used for reaching the root of the destination IP tree. In this tree, all the rules whose source IP prefix can be traversed from the root to a leaf of the source tree are inserted in a node of the destination tree based on their destination IP prefix. The procedure of creating the tree based on destination IP prefixes is the same as creating the tree based on source IP prefixes.

For classifying packets, the source and destination IP addresses, the source and destination port numbers, and the protocol are extracted from the header of every incoming packet. since the flow classification only requires the information extracted from the packet header, the packet size does not affect the performance.

The lookup is done by traversing the tree. The traversal begins at the root of the source tree and continuous based on the values of the bits of the source address field. During the traversal, the nodes that contain a pointer to the destination IP tree are inserted into a queue. At the end of the traversal, the trees corresponding to the nodes in the queue will be traversed according to the values of bits of the destination IP field of the packet. All the filters on the traversal path are stored and then searched linearly to find the best matching filter. The time complexity of Hierarchical-trie algorithm is $O(wd)$, where (w) represents the maximum prefix length of each field and d denotes the number of the fields in question. Also, $O(Ndw)$ represents the storage complexity of the algorithm, where N represents the number of classifying filters.



[Download : Download high-res image \(185KB\)](#)

[Download : Download full-size image](#)

Fig. 1. The Hierarchical-trie of the filters in Table 1 [27].

2.2. Cluster computing

A group of computer systems that consists of a set of independent systems is called a cluster. Systems in a cluster are closely interconnected. In fact, they can be viewed as a single system. A local network usually connects the components of a cluster. The number of these components should be two or more. In this structure, the systems require a message passing interface and a scheduler that controls the allocation of resources [8]. There is a variety of frameworks for scheduling and passing the messages in parallel computing and distributed computing platforms including, Spark, Flink, and MPI. Recently, it is shown that though MPI is harder to program, it outperforms existing frameworks [28], [29].

The first cluster system, which had a computation power of 1 Gigaflop, was called BEOWULF and was developed and implemented by NASA. Beowulf consisted of 16 PCs with 486 processors which were connected through a standard Ethernet. It is interesting to note that the total cost of the project amounted to only four percent of a single processor with the same computation power [30].

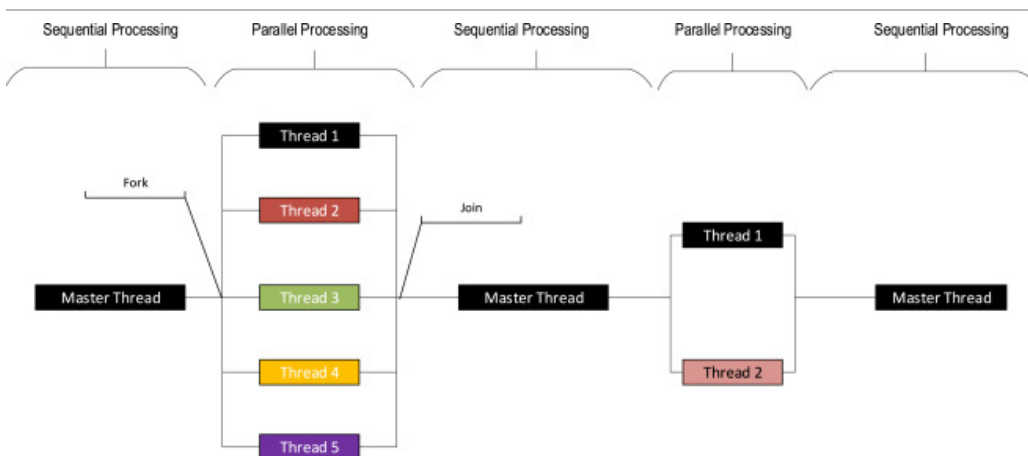
An important service provided by cluster systems is high-performance computing (HPC). In this mode, the software and hardware capacity of all the computers are used in a parallel manner for conducting a process. This will

remarkably reduce the time required for large amounts of processing [10]. In the following, we shall discuss different parallel models for the current programming interfaces.

2.2.1. OpenMP-based model

OpenMP is a shared memory programming interface in C, C++, and Fortran on all major compilers and platforms. The library includes parallelization patterns such as parallel blocks, loop, and tasks supported by basic concepts like data sharing and synchronization. Parallel algorithms can be expressed in simple, compact pieces of code to improve productivity. Also, OpenMP provides a more abstract interface than many common schedulers used in applied programming. This is a great advantage for implementation of scientific and analytical algorithms. The underlying concept of this library is the Fork/Join model for parallelization which is shown in Fig.2. The command used to interpret parallel blocks is `#pragma omp parallel` [31].

In this model, each parallel block has a master thread, several slave threads, and a specified operation. The master thread and the slave threads are collectively called a team. Each team has a certain size which is expressed in terms of the number of master and slave threads. At the beginning of each parallel block, there is a fork to synchronize the master and slave threads. After all the threads of the team have reached the fork, each member of the team begins to execute part of the team's operations that is assigned to it. These parts are assigned by the compiler at compile time. At the end of each parallel block, a join barrier is used to synchronize the master and slave threads so that the execution of the program continues sequentially through the master thread. Therefore, it can be said that processing in this model is divided into three main steps: the first stage is forking. At this step, the master thread first takes a team from the pool of defined teams or creates a new team. Then it sets the team size to a specific value and starts to run. The second step is execution in which the master thread processes along with slave threads that part of the team's operations which is assigned to it. The last step is joining. Here, the master thread creates a barrier and waits for all the slave threads to complete their work. Eventually, the team is collected; that is, it is either returned to the pool team or destroyed [31].



[Download : Download high-res image \(173KB\)](#)

[Download : Download full-size image](#)

Fig. 2. OpenMP parallelization mode.

2.2.2. MPI-based model

MPI is the most common technique of parallel programming. The message mediator is what controls the capability of an application programming interface to be exploited on shared-memory systems such as clusters of computing nodes. In MPI, the programmer explicitly communicated the data from the address space of one process to that of another through cooperative operations on each process. By caching data intended to be consumed by a consuming process before this inter-process communication, communication encounters minimum delay in passing the data. Therefore, in our proposed scenarios, we place the message data in a cache, bound, and make it ready to be consumed by the consuming process. The mediator provides this facility. The mediator of MPI is not a tool, but a

communication protocol that determines how parallel systems can communicate messages[32], [33], [34]. The chief advantage of MPI over other methods of message mediation is its lightness and high speed. Its high speed is due to the fact that it can be optimized while being executed on any hardware configuration. The most important feature of this method is that its functions can be called in different programming languages including C, C++, Fortran, Java, C#, and Python[34], [35], [36], [37], [38].

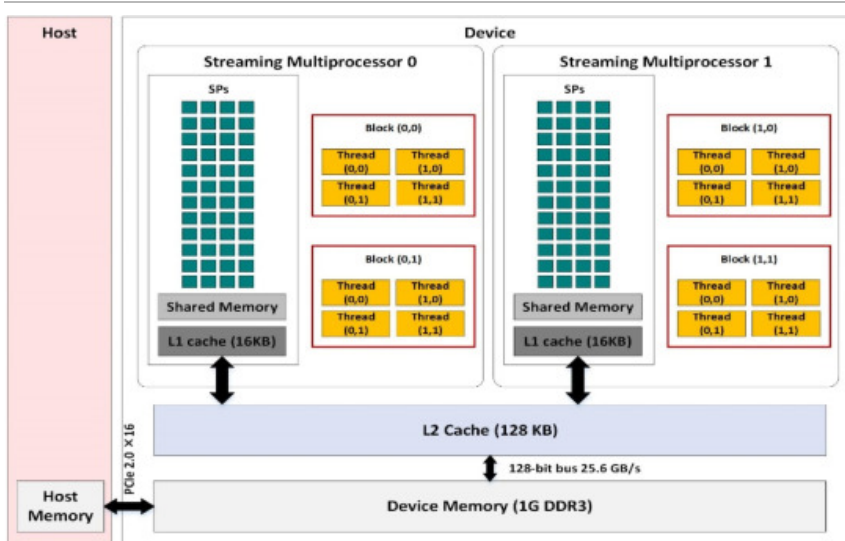
There are several implementations of MPI for different operating systems and configurations of hardware. One of these implementations is MPICH which is used as a dedicated open-source implementation of MPI for Linux. The chief advantage of parallel programs that use MPICH is that they can be executed on the most popular cluster architectures in the world. MPICH provides a set of libraries for Fortran, C, and C++ to use MPI-2. Hence, the MPICH has been used as the message mediator in our experiments.

2.2.3. CUDA-based model

GPU is a professional system to display graphic images in personal computers. Following the release of software development packages on this unit by great manufactures such as Nvidia[32] and ATI[39], the use of GPU was accepted as a powerful many-core computing hardware instead of central processing unit in accelerating technical computational. The main reason for this considerable revolution is that the architecture of GPU is specially designed for running compute-intensive and parallel programs (required for displaying graphic images). Hence, Nvidia provided a software platform called CUDA (Compute Unified Device Architecture) for performing nongraphic computations on graphic processors in 2006[32]. CUDA supports possibilities that could be used by technical programmers to have access to hardware capabilities of graphic processors in their nongraphic technical compute-intensive programs and increase the speed of running complex algorithms.

Several kinds of research including[33], [34] have attempted to study the use of the CUDA platform for parallelizing instruction-parallel or data-parallel network functions such as IP lookup in routing tables, aiming at having access to higher throughput. This tool has been used to parallelize evolutionary programs[35], convolutional neural networks[36] and in the other fields, as well[37], [38]. Also, the considerable capacity of parallel programming in the CUDA platform has been used in the field of cryptography for condensing databases and accelerating encryption and decryption in encryption algorithms[40], [41].

From a programming perspective, two CUDA processes are involved in parallel computations: host and device processes. The former runs on the central processing unit and executes the main program, whereas the latter is executed on GPU. Any program that is written on CUDA may be formed of several kernels. Each kernel is executed by a grid and each grid may be formed of several blocks. Each block is formed of several threads. Indeed, threads are responsible for performing programs.



[Download : Download high-res image \(464KB\)](#)

[Download : Download full-size image](#)

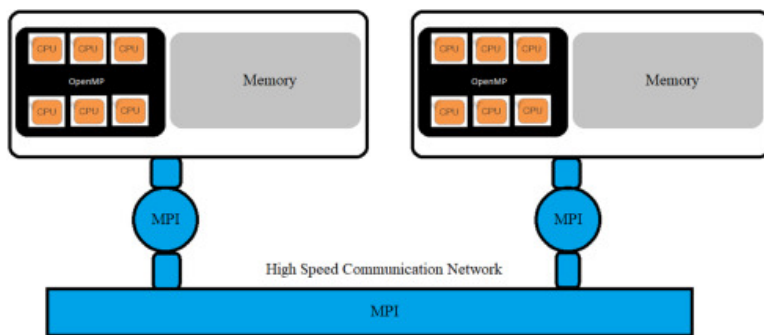
Fig. 3. Architecture of the NVIDIA GeForce 425M GPU[23].

One of the graphics processors used in this paper is GT 425M that is comprised of two streaming multiprocessors (SM), each consisting of 48 streaming processors (SP)[42]. Fig.3 shows the hardware structure of GT 425M GPU. This GPU has different types of memory including global, constant, texture, register, and shared memories. The CUDA grid in this figure includes four blocks that each of them consists of four threads.

2.2.4. Hybrid models

This method simultaneously uses OpenMP for cores with shared memory and MPI for those with separate memories. It should be noted, however, that MPI can also be used for the cores within the same system. The structure of hybrid architecture is illustrated in Fig.4. Another common hybrid architecture, which also takes into account the GPU, is the combination of CUDA and MPI. Fig.5 shows this architecture.

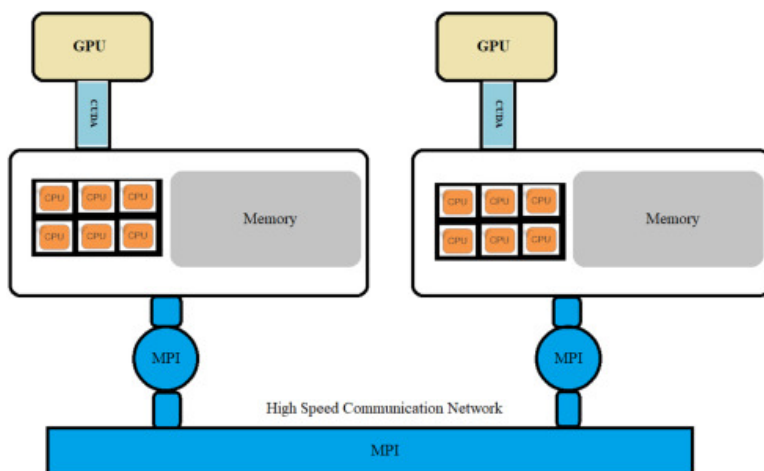
The most recent hybrid model is the combination of the above two models. By involving both GPU and CPU, this model seeks to use the maximum processing power of the systems on a network. It can provide a significantly high parallelization capacity. However, a major disadvantage of this model is the complicated nature of the task of combining MPI, OpenMP, and CUDA[16]. The present study makes use of all the three hybrid methods for the parallelization of Hierarchical-trie algorithm.



Download : [Download high-res image \(169KB\)](#)

Download : [Download full-size image](#)

Fig. 4. Combination of MPI and OpenMP programming models.



Download : [Download high-res image \(206KB\)](#)

Download : [Download full-size image](#)

Fig. 5. Combination of MPI and CUDA.

3. Review of literature

Zhou et al. conducted one of the preliminary research studies of the parallelization of packet classification algorithms on multi-core systems [43]. They parallelized linear search algorithm and area-based tree search algorithm using Pthread library. The maximum throughput of their parallel packet classifier was 11.5 Gbps. Another dominant study was conducted by Qu et al. in 2015. They parallelized the Bit-vector packet classification algorithm on multi-core processors using the OpenMP. The maximum throughput rate achieved in their study was 14/7 MPPS [44]. Recently, Tung et al. proposed a new algorithm for parallel packet classification on a processor with eight cores. According to their results, the throughput of their algorithm had been increased by 40 percent. By appropriately setting the parameter of the dependence of CPU on threads, they improved the productivity of cache memory [44].

GPU-based packet classification is one of the interesting fields in the literature that has emerged after the suggestion of GPU for parallel processing. Since the introduction of the GPUs for performing parallel algorithms, rare studies have been conducted on parallelization of the packet classification algorithms on such many-core machines. An important study in this domain is done by Nottingham et al. [45], which theoretically studies the possibility of parallel execution of packet classification algorithms on highly-threaded many-core GPUs. Though they introduce the concept of parallel packet classification on CUDA and OpenCL platforms, their work lacks experimental evaluation. Hung et al. [46] assess two parallel packet classification algorithms, namely RFC and BPF, according to four and eight different scenarios of using GPU memory hierarchy, correspondingly. According to their experimental results, for both algorithms, the most effective scenario is one in which the classification filters and test packets are stored in constant and global memory, respectively. They use an impractical test dataset, which includes only three filters as the classification filter set. Certainly, the experimental results of this unrealistic study are not consistent.

Deng et al. [47] proposed a hybrid method that utilizes both the CPU and GPU for the parallelizing linear packet classification algorithm. In this study, only the slow global memory of the GPU was used. Thus, their hybrid method may not be efficient in seamlessly exploiting the capacity of the hierarchical memory system of GPUs. Kang et al. [44] introduce a meta-program model for GPU-based packet classification. By using this technique, the rules are compiled into instructions and consequently the expensive latency of memory accesses would be avoided.

Zhou et al. investigate the impact of efficiently exploiting the various types of memory on the performance of packet classification algorithms on GPU-like many-core machines [48]. They study GPU's characteristics in terms of thread parallelization and memory access in parallelizing the Bit-vector packet classifier using the CUDA platform. Their parallel classification that has only one CUDA block of 32 threads (one warp) classifies each packet. Their parallel algorithm tries to store all of the filters in the shared memory of the block. Hence, the performance of this kernel decays with an increase in the number of filters. In such a case, the shared memory cannot hold the whole dataset and the remaining part of data is stored in the slow global memory of GPU. As a result, the access time of threads increases.

In their pioneering research, Varvello et al. propose a kernel model for effective parallel packet classification on GPUs [49]. Their kernel model tries to maximize the parallelization capacity of threads by splitting the filter set among several blocks, so that each block is responsible for checking the incoming packets only against a specific part of the filter set. Also, the authors assess the effect of some parameters including the size of the filter set, the number of blocks as well as the threads per block on the overall efficiency of the parallel kernel model. However, their study lacks an efficient analysis method to estimate the effect of the algorithm complexity and the different hardware characteristics of GPU on the overall performance of the classifier.

Unfortunately, recent studies like [39], [50] have only inspected the effect of tuning parameters including the number of threads and blocks on the overall kernel runtime. In all, all of the reviewed studies in the field of GPU-based packet classification have obscurely parallelized some packet classification algorithms. None of them has studied the parallel processing models of the kernel threads and the different types of memory modules.

Recently, Abbasi et al. have proposed a novel complexity analysis method that analytically estimates the efficiency of parallelization methods considering the threads, the type of memory used for storing the data structure, and

memory access latency. Their method provides an inclusive analysis framework that helps in designing efficient algorithms for parallel packet classification on GPU-like many-core systems. Their work also introduces applied models for parallel packet classification on GPUs.

Abbasi et al. [24] assessed different scenarios for parallel packet classification using OpenMP, MPI, and their combination on a simple system with a single multi-core processor as well as on complex multi-core number processor clusters. Their results show the speed of packet classification is linearly related to the number of cores. Also, though MPI uses more memory than OpenMP, it provides a higher rate of packet classification. Moreover, in the case of using a combined MPI-OpenMP interface, the maximum speed of packet classification is reached when the number of processes and threads is equal to the number of processor cores.

None of the above works has made simultaneous use of CPU clusters and GPU clusters for the parallelization of packet classification. In what follows, we shall have a brief look at some major works that have utilized CPU clusters to solve the problems of various fields of engineering.

Henty et al. proposed a combination of OpenMP and MPI on a CPU cluster [19]. Their work showed that a combined method has a lower performance than the MPI-alone method due to its processing load. They also showed that the OpenMP-alone method is more efficient than the combined method in small-scale parallel problems. They also concluded that the efficiency is highly depended on the structure of program code.

J. Hutter et al. used a combination of the shared-memory and MPI in a 1024-core cluster. Their results showed that the communication delay is an important factor in the overall efficiency of the cluster [48]. Cappello et al. used two methods to solve numerical simulation problems of aerodynamics, i.e., a combined method and an MPI-alone method. They found that the efficiency of computation in a cluster depends on several parameters such as memory access pattern and hardware performance [31]. In another work in 2018, M. Ferretti et al. implemented Cross Motif search in a parallel form on a CPU cluster. Their results showed that the use of MPI alone is more efficient than the combination of MPI and OpenMP [49]. Q. Zhao et al. showed in 2019 that large-scale numerical simulation for the analysis of discrete spherical forms is extremely long. The results of exploiting OpenMP, MPI, and their combination showed that the combined method may result in a better performance than the other two methods [50].

The first GPU cluster called Lincoln was implemented by D. A. Jacobsen and his colleagues at Illinois University [28]. They focused on hardware resources, power consumption, and the price of the graphics cards used in the cluster and their main aim was to find the best graphics card for their cluster.

In [29] the architecture, resource sharing, and programming models of GPU clusters are discussed as far as HPC is concerned. This study reports the main challenges such as resource management, task scheduling, and security in two well-known GPU clusters, namely, Lincoln and AC. These two clusters are based on the NVIDIA S1070 graphics card.

J. Song et al. studied the combination of MPI, CUDA, and OpenMP and showed that, for applications that use the CUDA+MPI model, a compromise among different performance factors is necessary [30].

This review shows that hybrid clusters provide a higher capacity for the parallelization of algorithms using different combinations of programming models.

For this reason, the present study combines CPU and GPU clusters for packet classification. Hybrid clustering could significantly increase the classification speed. Nevertheless, it is obvious that there are always limitations on hardware capacity and the number of CPUs and GPUs on a system cannot be freely determined. This is why parallel computing has come to the help of computer clusters. In contrast, hardware clustering could provide higher extension capacity and any new node can be easily added to the system, thereby increasing the number of CPUs and GPUs. In addition to combining CPUs and GPUs, therefore, the present study also uses clustering for the task of packet classification.

It is the first time that the Hierarchical-trie algorithm is parallelized on a combination of CPU and GPU clusters. This significantly increases the speedup and throughput rates. Different scenarios have been created and compared here

by using several programming models on a combined CPU and GPU cluster.

4. Proposed method

In this section, we shall first describe the cluster used for the implementation. Next, we shall explain the parallelization of Hierarchical-trie algorithm on a CPU cluster, a GPU cluster, and a hybrid cluster. As [Table 2](#) predicts, the hybrid cluster can reveal the highest performance level. Also, it is expected that the complexity of a hybrid [cluster algorithm](#) is more than that of any parallel algorithm.

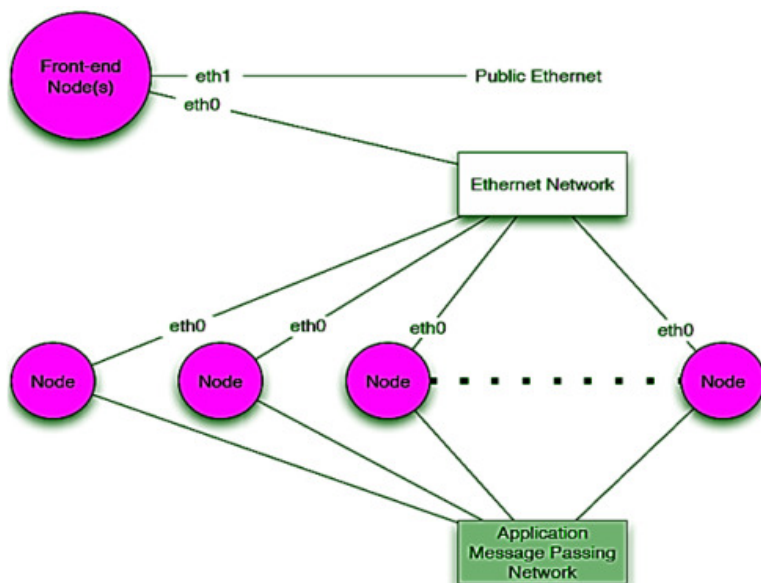
Table 2. Comparing the expected performance and complexity level of parallelization methods.

Method	Performance level	Complexity level
CPU	Very low	Very low
GPU	Low	Low
Hybrid CPU–GPU	Average	Average
CPU cluster	High	High
GPU cluster	High	High
Hybrid cluster	Extremely high	Very high

4.1. Specifications of the implemented cluster

Implementation of our cluster was based on the Rocks distributed operating system which is based on Centos operating. The Rocks reduce the complexity of processing, development, and management as well as to improve performance in a parallel cluster system. To install Rocks on a front-end node, two network adapters, one for internal communications (eth0), and the other for external communications (eth1) are required. The operating system is first installed on the front-end node system, and next on other computers. Most of the required packages including MPICH and OpenMP are installed by default with Rocks[\[51\]](#).

The nodes (systems) within a cluster communicate through switches. The switch used in this study was D-Link 10/100. Three systems were used in this study, but only two of them, which were homogeneous and had a quad-core CPU and similar GPUs, were used for computing. [Fig.6](#) illustrates the architecture of the systems as well as how they are connected (see [Fig.7](#)).

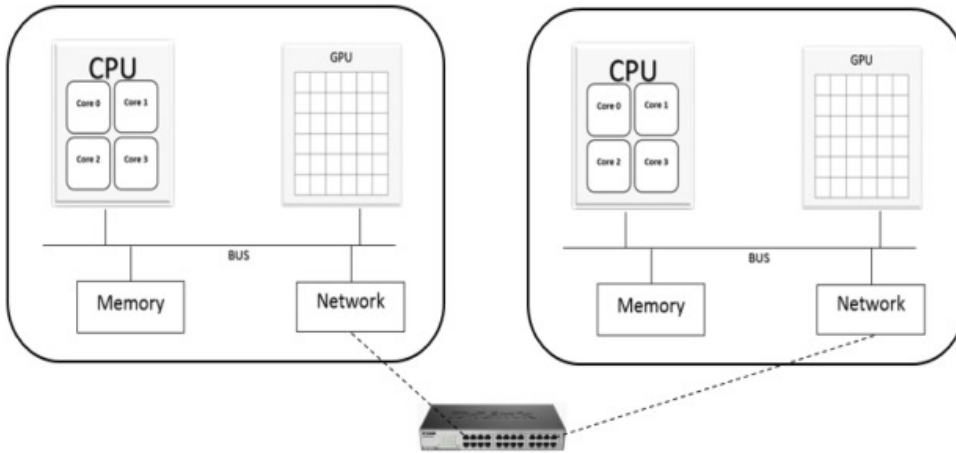


[Download : Download high-res image \(361KB\)](#)

[Download : Download full-size image](#)

Fig. 6. Clustering with Rocks[24].

As can be seen in the figure, each system was equipped with a separate CPU, GPU, and memory (i.e. symmetric multiprocessing or SM). The specifications of the quad-core CPU used in the systems are listed in [Table 3](#).



[Download : Download high-res image \(201KB\)](#)

[Download : Download full-size image](#)

Fig. 7. The structure of the connections among the systems.

The graphics card used was GeForce GTX 960. Its specifications are shown in [Table 4](#). Given its Maxwell architecture, the eight SMs of GeForce GTX 960 provide a CUDA computation capacity of 5.2. The size of the shared memory for each SM and each block in this card is 96KB and 48KB, respectively, and the maximum number of static threads for each SM is 2048. Since the maximum number of threads in each block cannot exceed 1024, two blocks can be defined in each SM. Given that there are eight SMs, a total number of 16 blocks can be defined on each system for the task of packet parallelization (see [Fig. 8](#)).

Table 3. Specifications of the CPU.

Processor number	Q6600
Status	End of interactive support
Launch date	Q1'07
Lithography	65nm
# of cores	4
Processor base frequency	2.40GHz
Cache	8 MB L2
Bus speed	1066 MHz FSB
FSB parity	NO
Cache size	4096 KB
TDP	105 W
VID voltage range	0.8500V–1.500V

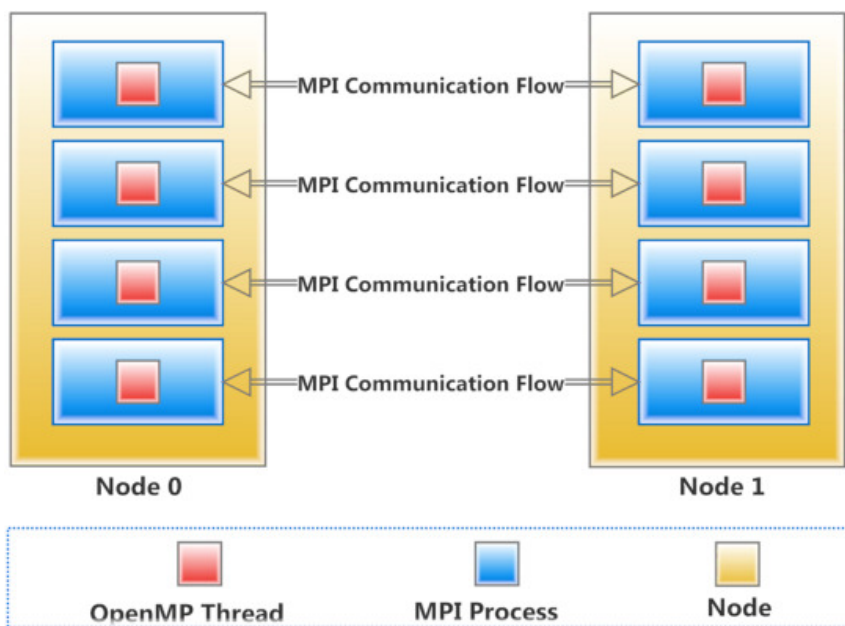
Table 4. Specifications of the graphics card.

GeForce GTX 960	
Streaming multiprocessors	8
CUDA cores	1024
Base clock	1126
Memory clock	7.0 Gbps
MemoryBandwidth (GB/sec)	112
Architecture	Maxwell
Bus support	PCI Express 3.0
Standard memory config	2 GB

4.2. Parallelization of hierarchical-trie algorithm on a CPU cluster

4.2.1. The first scenario

As fully explained in [24], this scenario only makes use of MPI. Given that the total number of CPU cores on the two systems is eight, eight MPI processes are defined. When the processes are produced, a file is given to each process to be executed. Also, a separate address space in the cluster memory is allocated to each process. Therefore, there are eight different address space in this scenario, each of which contains the memory, filters, results array, tree structure, and packets of its corresponding process (H_i).



[Download : Download high-res image \(432KB\)](#)

[Download : Download full-size image](#)

Fig. 8. The configuration discussed.

MPI distributes the execution of the file simultaneously on all the cores. In doing so, the packets corresponding to each process can be separated by referring to the rank of the process, which is a unique ID. In this study, this technique is used to distribute the packets among different processes. According to Algorithm 1, the total number of processes (S), the packets (H), and the rank of each process are given to the function. Using the unique rank of each process, the function determines and returns the start index ($start_i$) and the end index (end_i) of the packets which should be classified by the i th process. H_i denotes the number of packets related to the i th process.

Algorithm 1 The pseudocode for the distribution of packets among the processes.

Input: MPIsize S , Headers H , rank R
Output: Headers H_i , $start_i$, end_i

```

1  if  $(|H| \% S) = 0$  then
2:    $|H_i| \leftarrow |H| / S$ 
3:  else
4:    $|H_i| \leftarrow (|H| / S) + 1$ 
5:    $start_i \leftarrow R * |H_i|$ 
6:    $end_i \leftarrow ((R + 1) * |H_i|) - 1$ 
7:  end if
8:  if  $(R = (S - 1) \ \&\& \ (end_i > H))$  then
9:    $end_i \leftarrow |H|$ 
10: end if

```

[Download : Download high-res image \(171KB\)](#)

[Download : Download full-size image](#)

After distributing the packets among the processes and obtaining the indexes (end_i , $start_i$) as well as the number of the packets belonging to each process (H_i), these values along with the filter set R and the tree structure T are given as arguments to the [classification algorithm](#). As can be seen in Algorithm 2, the classification algorithm stores the filter that best matches the i th process in *ruleIndexArray* $_i$, which is in the range [$start_i$, end_i], and returns it as the output. This is done simultaneously for all the processes.

Algorithm 2 The pseudocode for packet classification on a CPU cluster.

Input: rules R , H-trie T , Headers H_i , $start_i$, end_i
Output: *ruleIndexArray* $_i$

```

1: for all  $i \in [start_i, end_i]$ 
2:    $P \leftarrow$  Read Packet ( $i$ )
3:    $rIdx \leftarrow$  Classify ( $P, T, R$ )
4:   if  $rIdx \neq$  Null then
5:     ruleIndexArray ( $i, rIdx$ )
6:   end if
7:    $i \leftarrow i + 1$ 
8: end for

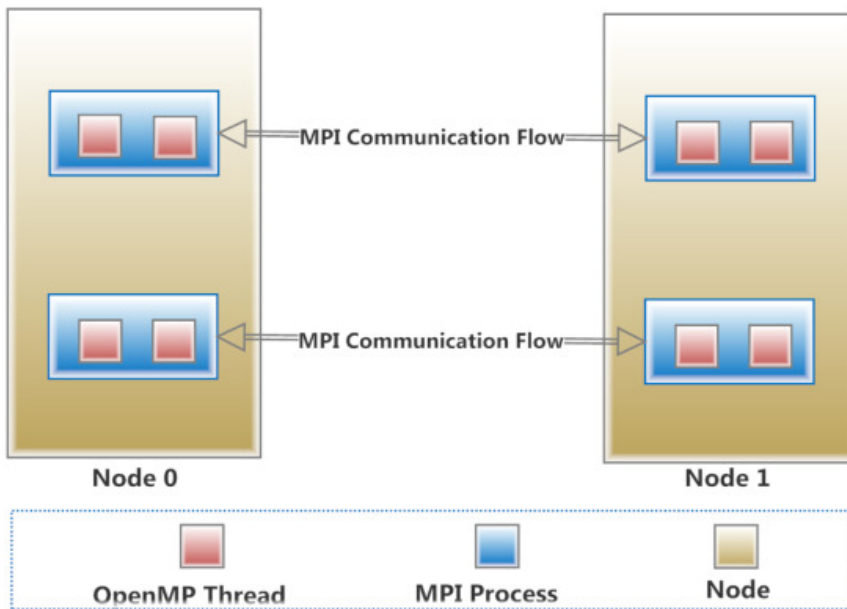
```

[Download : Download high-res image \(162KB\)](#)

[Download : Download full-size image](#)

4.2.2. The second scenario

In addition to MPI, this scenario also uses OpenMP. The number of MPI processes and OpenMP threads are determined as four and two, respectively. Since four processes are defined and there are two systems, two processes are allocated to each system. Each process becomes responsible for parallel handling of two threads. As in the previous scenario, the filters, the results array, the tree structure, and the number of packets assigned to each process (H_i) are copied into the allocated space of processes. Since the number of OpenMP threads in this scenario is two, H_i allocated to each system is partitioned into two separate parts, each being classified by one thread. The advantage of this method is that it reduces the address space required by the processes almost by half in comparison with the previous scenario. This is shown in [Fig.9](#).



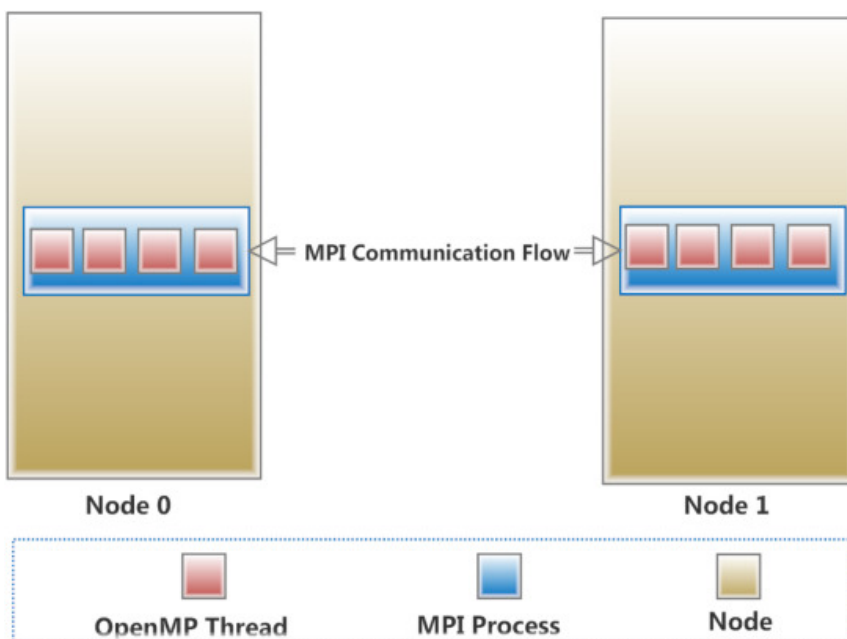
[Download : Download high-res image \(342KB\)](#)

[Download : Download full-size image](#)

Fig. 9. The configuration of the second scenario for CPU cluster.

4.2.3. The third scenario

In contrast to the previous scenario, two processes are used here. We need four OpenMP threads per process to involve all the processing cores. The rest of the scenario resembles the last two scenarios. This is shown in Fig. 10.



[Download : Download high-res image \(300KB\)](#)

[Download : Download full-size image](#)

Fig. 10. The configuration of the third scenario for CPU cluster.

4.3. Parallelization of the h-trie algorithm on a GPU cluster

4.3.1. The first scenario

Using MPI, two different processes are generated and the packets are divided into two equal groups, i.e. H_1 and H_2 , by means of the pseudocode in Algorithm 3.

Algorithm 3 The pseudocode for packet classification in the global memory of a GPU cluster.

Input: rules R, H-trie T, Headers H, gc, ruleIndexArray, np
Output: ruleIndexArray
1: **if** $|H| \% gc = 0$
2: $b \leftarrow |H| / gc$
3: **else**
4: $b \leftarrow (|H| / gc) + 1$
5: **end else if**
6: $H_{cpu} \leftarrow (np - 1) * b$
7: $|H_{gpu}| \leftarrow |H| - |H_{cpu}|$
8: **if** rank = 0
9: ruleIndexArray \leftarrow GPU (T, H_{gpu} , R)
10: **else**
11: ruleIndexArray \leftarrow CPU (T, H_{cpu} , R)
12: **end else if**

[Download : Download high-res image \(205KB\)](#)

[Download : Download full-size image](#)

Fig. 11 illustrates the distribution of packets among the processes on the GPU cluster. After sending the packet groups to two systems, as shown in the pseudocode, each system receives the filters (R), the tree structure (T), and the H_i as arguments. In the first line of the pseudocode, these values are transferred from the host's memory to the global memory of the GPU. In line 2, the total number of packets is divided by the total number of processing threads. If the number of packets is greater than the number of threads, E packets are given to each thread. In line 4, the index of the packet in question is obtained using the thread index in the block, the block index in the grid, the dimension size of the block, and the number of packets allocated to each thread (E). If the obtained index is less than the total number of packets, the thread retrieves one packet from the global memory and classifies it using the tree structure T and the filter set R, which are both located in the global memory. Since it is possible for the incoming packet to match several filters, the index of the best matching filter is stored in ruleIndexArray and returned as the output. In fact, the algorithm stores the index of the filter that best matches the i th process in ruleIndexArray which is in the range $[start_i, end_i]$.

4.3.2. The second scenario

In this scenario, the filter sets are divided into trees with a size that can be stored in the shared memory of the block. The aim of this division is to make maximum use of the shared memory. The main reason is that access to the shared memory is 100 times as quick as the global memory of the GPU. As mentioned earlier, the size of the shared memory is 48KB. Given that each node occupies 40B, we can create a tree with a maximum of 1200 nodes that could be stored in the shared memory. This scenario can have 199 filters, and needs a corresponding tree with 1193 nodes. The input to this algorithm consists of the small filter set, the packets, and the tree corresponding to the small filter set. After transferring these data to the global memory, the small tree is copied into the shared memories of the blocks in both systems and the packets are classified based on this tree.

4.4. Parallelizing the H-trie algorithm on a hybrid cluster

In the CPU cluster mode, when the CPU is in use, the GPU is idling and no packet is assigned to it to be classified. Similarly, using the GPU means that the CPU is idling. In the hybrid cluster mode, however, attempts are made to utilize the capacity of both processors simultaneously.

In the implemented cluster described in this study, the graphics card (GTX 960) has 1024 computing cores. On the other hand, the CPU (Q6600) has only four cores, which means a low parallelization capacity. Therefore, the packets should be distributed proportionately between the CPU and the GPU. For this purpose, a variable called gc is defined which represents the ratio of the speeds of the GPU and the CPU for packet classification. The value of this variable was calculated by testing the algorithm with different numbers of incoming packets. The gc value, the filter set, the

packets, and the results array are given as input to the algorithm. The pseudocode for this scenario is presented in Algorithm 4. In lines 1–5, the variable b is initialized. It is calculated by dividing the number of packets by gc . Using b and np (the number of packets), the number of packets that should be allocated to the CPU can be determined. The number of packets that should go to the GPU is also easily calculated by subtracting the number of CPU packets from the total number of packets.

Algorithm 4 The pseudocode for dividing the packets between the CPU and the GPU.

```

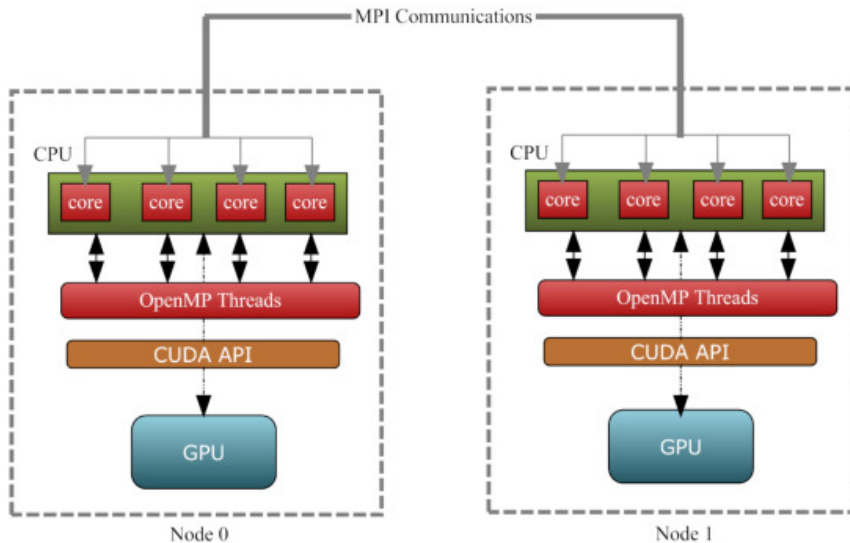
Input: rules R, H-trie T, Headers H, gc, ruleIndexArray, np
Output: ruleIndexArray
1: if  $|H| \% gc = 0$ 
2:    $b \leftarrow |H| / gc$ 
3: else
4:    $b \leftarrow (|H| / gc) + 1$ 
5: end else if
6:  $H_{cpu} \leftarrow (np - 1) * b$ 
7:  $|H_{gpu}| \leftarrow |H| - |H_{cpu}|$ 
8: if rank = 0
9:   ruleIndexArray  $\leftarrow$  GPU (T,  $H_{gpu}$ , R)
10: else
11:   ruleIndexArray  $\leftarrow$  CPU (T,  $H_{cpu}$ , R)
12: end else if

```

[Download : Download high-res image \(201KB\)](#)

[Download : Download full-size image](#)

To sum up this scenario, when two processes has been defined according to the pseudocode in Algorithm 4 and the packets have been divided equally between the two systems, the number of packets that should go to the CPU and the GPU are determined according to the pseudocode in Algorithm 4. Finally, the packets are classified according to this scheme. In this scenario, two GPUs and eight cores from two CPUs are involved in classification.



[Download : Download high-res image \(289KB\)](#)

[Download : Download full-size image](#)

Fig. 11. Distribution of packets among the processes on the GPU cluster.

5. Implementation and evaluation

In this section, we will first describe a software suite for generating the filter sets of experimental headers. Next, we will discuss our criteria and evaluate the results from the scenarios described in Section 4.

5.1. ClassBench

ClassBench is a simulator based on C language and [Linux platform](#) which is used to generate experimental filters and their corresponding headers. This suite creates rule sets that are similar to the actual rule sets in the classification process. It produces filters based on input parameters. ClassBench uses two modules. The first module generates rule sets with any desirable number of rules. It can generate three kinds of rule set: Firewall (FW), Access Control (ACL), and Chain (IPC). The second module produces a set of random packets based on the statistical features of the rule sets generated by the first module. In the following we briefly introduce abovementioned types of ruleset.

5.1.1. ACL ruleset

The ACL represents the standard format for security and [network address translation](#) (NAT) rules. It is a set of rules that define how to forward or block a packet at the router's interface. By applying an ACL on a routing device for a specific interface, all the packets flowing through it, are compared with the [ACL rules](#), which can either block or allow them.

5.1.2. FW ruleset

The FW is the registered format for specifying security rule-sets for firewall devices. FW rules determine what traffic your firewall allows and what is blocked. They also inspect the control information in individual packets, and either block or permit them according to the defined conditions. As a result, these rules dictate the firewalls on how to protect the network from [malicious programs](#) and illegal access.

5.1.3. IPC ruleset

IPC allows a network administrator to manage rules in the kernel [packet filtering](#) area, by adding or deleting rules from various chains. It also proposes a [decision tree](#) format for security rules. A set of rules that direct the traversal according to the header information of a packet toward a specific node specifies a chain of the decision-tree. The end-node of the chain contains a set of commands, which must be applied to the packet. There are three types of chain including the input chain, the forward chain, and the output chain. The input chain checks the permissions of the incoming packet. After this, the routing decision is applied, and the packet goes through a forward chain or an output chain regarding whether it is an internal packet or not.

A review of the literature on packet classification shows that the majority of studies[51], [52] have used ClassBench to generate filters. The main reason is that this suite uses a random method and produces datasets that are very close to authentic data.

In the present study, we also used this suite to generate packets and filters corresponding to IPC2. 1k filters as well as 32k, 64k, 128k, 256k, 512k, and 1024k incoming packets were generated for the evaluation of our scenarios. The algorithm was executed ten times for each number of packets and the mean of the results was recorded as the final result of each classification. It should be noted that, due to space limitations, we shall only present part of the results in this paper.

5.2. Evaluation criteria

In this section, the performance criteria for packet [classifier systems](#) are defined.

Throughput: Throughput is defined as the number of packets which can be classified in the unit of time. It is measured in packets per second (PPS).

Classification time: It refers to the duration of the processing of packets in CPU or GPU. The unit of this measure is milliseconds and it is denoted by $T_{\text{classification}}$. This time can be calculated in two ways for each process. The first method considers the maximum time among the times of all processes. The reason is that it takes this amount of time to complete all classifications. Eq.(1) represents how the time is calculated in this method.

$$T_{max} = \max \{ t_{\text{classification}_1}, t_{\text{classification}_2}, \dots, t_{\text{classification}_m} \} \quad (1)$$

In the second method, the average of the times is taken into account instead of the maximum time. The average classification time can be calculated using Eq.(2).

$$T_{avg} = \frac{\sum_{i=0}^{np-1} t_{classification,i}}{np} \quad (2)$$

Speedup: Speedup is obtained by dividing the classification time in the serial mode by the classification time in the parallel mode.

$$S_p = \frac{T(1)}{T(p)} \quad (3)$$

In Eq.(3), $T(p)$ denotes the time required for parallel execution of the algorithm, and $T(1)$ denotes the time required for serial execution of the algorithm.

Transfer time: It refers to the time it takes to copy the required data structure from the CPU memory into the global memory of the GPU. Transfer time is measured in milliseconds.

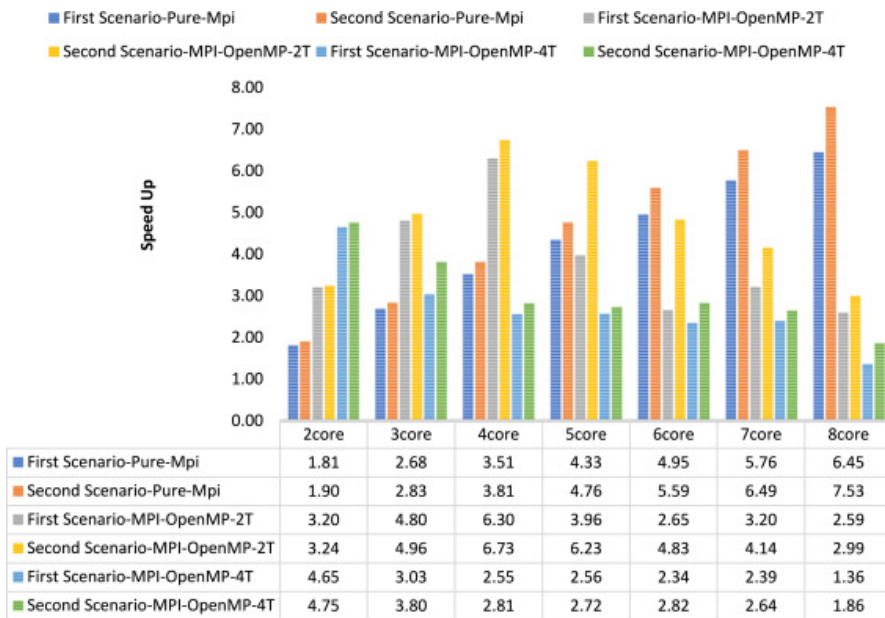
Memory usage: An important criterion for evaluating the performance of classification algorithms is memory usage. Thus, an algorithm with lower memory usage has a better performance.

In this study, IPC filters are used. The total number of IPC filters are 634. The Hierarchical-trie corresponding to this filter set contains 7215 nodes. Each node needs 40 bytes to be stored on a Linux platform, which means a total amount of 288 KB of memory. In addition, some memory should also be allocated to filters, packets, and an output array for displaying the results. *Memory_i* represents the total amount of memory needed for any classification process. Given that processes have separate address spaces, this amount of memory should be multiplied by np in order to obtain the memory usage of np processes.

$$Memory_{Total} = np * Memory_i \quad (4)$$

5.3. CPU cluster

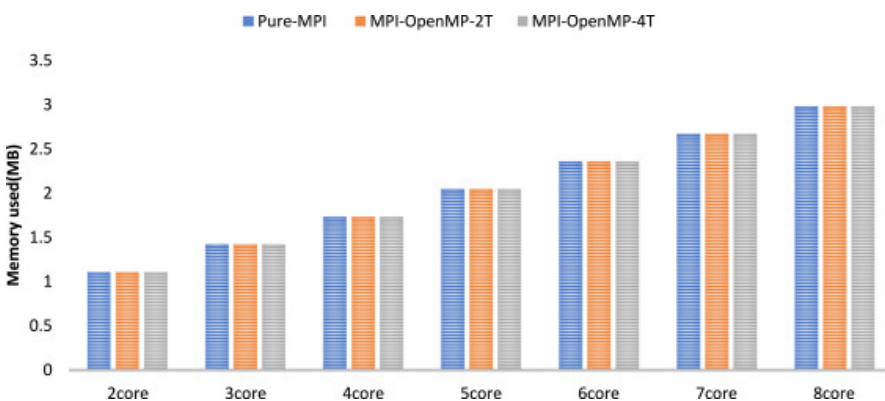
Classification time for different numbers of packets was calculated. The speedup rate for these classification times was also calculated for different numbers of packets. For example, Fig.12 shows the graph for the speedup rate of the CPU cluster for 32k packets in all three scenarios obtained by the average method and the maximum method. The first two columns show the speedup rate with two cores involved and in a mode that uses MPI alone for packet classification. In the first column, classification time was calculated with the maximum method, and in the second column, it was calculated with the average method. The second and third two columns show the same results for the second scenario. Also, the third and fourth two columns show these results for the third scenario. As the number of cores is eight, in this scenario, the cores were increased from 1 to 8, and the speedup rate was calculated after the addition of each new core. The highest speedup achieved in the first scenario belongs to the case where all cores are involved in the computation. In this case, the speedup rate calculated by the first and the second method is 6.4 and 7.5, respectively. In fact, with the addition of new cores, the speedup rate will increase. When eight processes are used, this rate will achieve its maximum.



[Download : Download high-res image \(648KB\)](#)

[Download : Download full-size image](#)

Fig. 12. Speedup rate of CPU cluster for 32k packets in all three scenarios with different numbers of threads assigned to each core (denoted by the coefficient of T).



[Download : Download high-res image \(388KB\)](#)

[Download : Download full-size image](#)

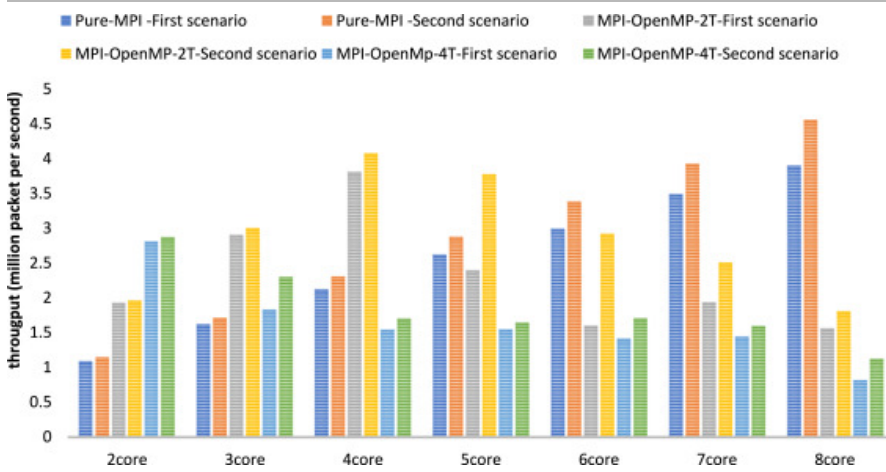
Fig. 13. Memory required for classification of 32k packets by the CPU cluster in the three scenarios.

Fig. 13 shows memory usage for the classification of 32k packets in the three scenarios. According to the graphs in this figure, the increased number of cores will increase memory usage due to using a separate address space. Note that, as one process is executed per core, any increase in the number of cores will increase the processes.

As the threads share an address space in the second and third scenarios, memory usage does not differ from the MPI-alone mode. As can be seen in the graphs, the maximum memory required for the classification of 32k packets is 2.98 MB and belongs to the case where eight processes are defined.

Throughput: Fig. 14 illustrates the throughput of all three scenarios in the classification of 32k packets as calculated by the maximum method and the average method. The maximum throughput, which is 4.5 Mega PPS, belongs to the case where the eight processes are running in eight cores. This is the maximum throughput among the three scenarios. The maximum throughput in the second scenario, which is 4.07 Mega PPS, is achieved when two processes are involved in classification. Finally, the maximum throughput in the third scenario, i.e., 2.87 Mega PPS, is achieved when two cores are involved in classification.

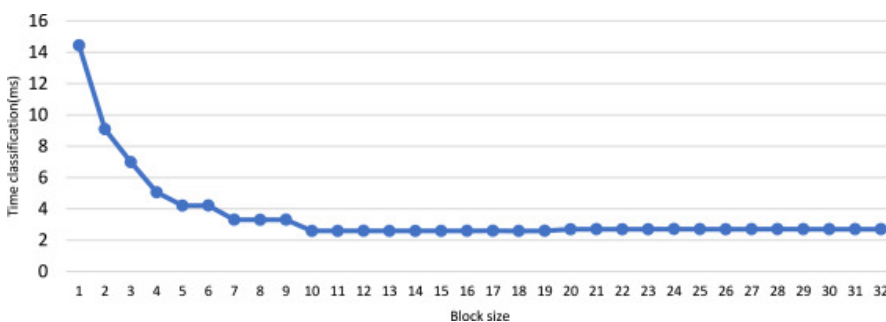
From the results of the CPU cluster scenarios we can infer that, to achieve the maximum speedup rate, the total number of MPI processes plus the number of OpenMP threads must be equal to the total number of computing cores. In addition, the larger the number of processes and the smaller the number of threads, the higher the speedup rate is. Using more processes, however, means using more memory. In fact, the best scenario for a CPU cluster is to define one process per system and to determine the number of OpenMP threads equal to the number of the cores in each system so that the largest number of packets could be classified in the unit of time with the lowest memory usage.



[Download : Download high-res image \(542KB\)](#)

[Download : Download full-size image](#)

Fig. 14. Throughput of CPU cluster for 32k packets in all three scenarios.



[Download : Download high-res image \(134KB\)](#)

[Download : Download full-size image](#)

Fig. 15. The effect of the number of blocks on the classification time with 32K packets.

5.4. GPU cluster

The first scenario

According to the results of [23], the number of blocks in the GPU should be determined as 16 for the optimum use of processing threads. In our evaluations, we sought to examine whether a smaller or larger number of blocks would change the classification time. For this purpose, the number of blocks was increased from 1 to 32 and classification was performed for different numbers of blocks. The performance of the classifier was measured in each run for 32K to 1024K packets.

Fig. 15 shows the classification time with 32K experimental packets for different numbers of blocks. It can be seen that this time is 14.4ms for one block. This time is significantly reduced and becomes 9ms for two blocks. The more the number of blocks, the less the slope of the curve of time reduction. Our results suggest that this reduction

continues up to 16 blocks and, from this point on, the increase in the number of blocks has no effect on the classification time.

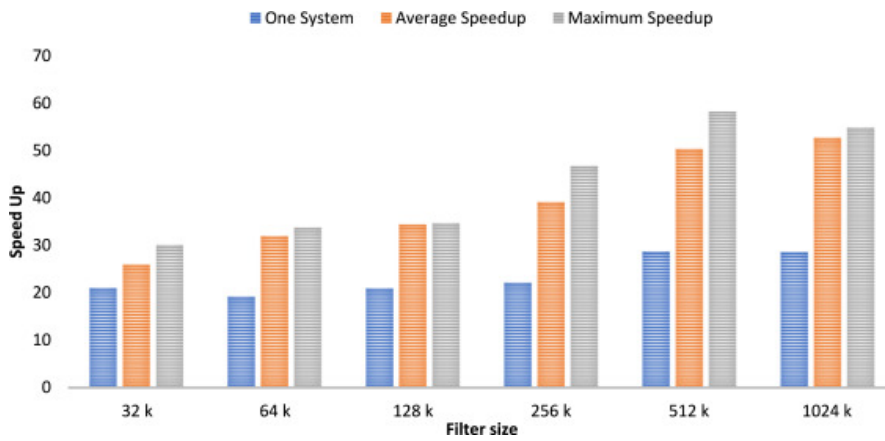
After obtaining a fixed number for the CUDA blocks, the classification algorithm was executed on the GPU cluster and the classification time, the transfer time, and the sum of both times were calculated for different numbers of packets on both systems. The results are shown in Table 5. By comparing the classification times of the two systems, we find out that they are almost equal, which may be explained by the fact that the graphics cards are the same and the numbers of packets are equal. We should however note that the times are not exactly equal because the packets allocated to each process as well as the execution conditions may be different. With 512K packets, for example, the classification times of systems 1 and 2 are 19.66ms and 14.31 ms, respectively. The 5-millisecond difference is mostly due to the difference in the allocated packets. In other words, the classification of packets with different headers may require that different path be traversed on the filter tree with different numbers of accesses.

In the previous section, two methods, naming average and maximum methods were described for calculating the classification time in cluster systems. Using these two methods, the speedup graph for different numbers of packets can be drawn as in Fig. 16. It shows the speedup rate in the one-system mode and in the first scenario of the GPU cluster mode. In the one-system mode where we used one GPU on a single system for packet classification, the average speedup rate achieved for packets less than 512K is 20 times as much as the serial packet classification.

Table 5. The classification time for different numbers of packets on the two GPU cluster systems.

Number of packets	System 1			System 2		
	Computing time of the core	Transfer time	Total time	Computing time of the core	Transfer time	Total time
32K	1.507	0.503	2.010	2.088	0.498	2.587
64K	3.354	0.689	4.043	3.737	0.644	4.381
128K	6.487	1.044	7.532	6.589	1.063	7.652
256K	8.327	1.815	10.142	12.327	1.838	14.166
512K	19.662	3.368	23.03	14.319	3.325	17.644
1024K	34.483	6.366	40.849	37.346	6.327	43.673

For packets more than 512K, the speedup is almost 30 times. This shows that with the increase in the number of packets, the parallelization capacity of the GPU becomes more manifest. In the classification of 32K packets by the GPU as in the first scenario, the average speedup is 25 times and the maximum speedup is 30 times. With increasing the number of packets, the speedup begins to rise until, for 512K packets, the average and maximum speedup values become 50 and 58, respectively. These values are 1.78 and 2 times as great as the speedup values achieved with a single GPU. This is what we expected from the outset. In fact, by increasing the number of processors from one to two, the speedup also doubles. It can be seen that the GPU could significantly increase the speedup value in comparison with the serial classification. Also, the ratio of the speedup, particularly the maximum speed up, in the GPU cluster mode to the speedup in the one-system mode increases with the number of packets and almost reaches 2 for 512K or more packets.



[Download : Download high-res image \(350KB\)](#)

[Download : Download full-size image](#)

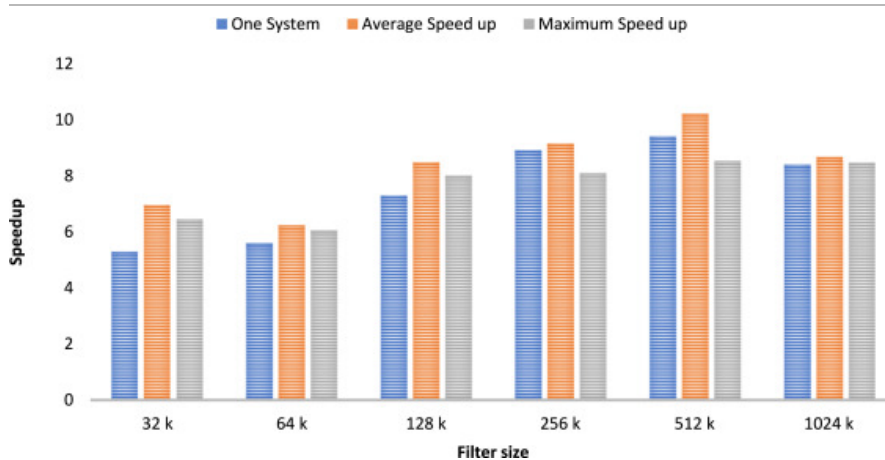
Fig. 16. The speedup achieved in classifying different numbers of packets on the GPU cluster.

The second scenario

Using the classification time in the global and shared memory of the GPU, the ratio of the speedup in using the shared memory to the speedup in using the global memory was calculated for both the one-system and the two-system modes. Fig. 17 shows the speedup in the classification of different numbers of packets.

In the one-system mode, the speedup is less than in the two-system mode. For example, for the classification of 128K packets by one system, the speedup in the second scenario is 7.3 times as much as the speedup in the first scenario and the ratio of the average and maximum classification time in the first scenario to the average and maximum time in the second scenario on the GPU cluster is 8.74 and 8.00, respectively. An interesting point in this graph is the remarkable speedup resulting from using the shared memory of the GPU in comparison with the first scenario.

According to Fig. 17, the ratio of the speedup in the second scenario to the speedup in the first scenario for different numbers of packets is almost 5. The speedup for 512K packets is around 10. This increase in speedup shows that the shared memory of the GPU can provide a remarkable speedup in packet classification. The reason is that the lifetime of the data in the shared memory amounts to the lifetime of its corresponding block and, in addition, this memory has a very short access time; therefore, the data in it can be accessed very rapidly.



[Download : Download high-res image \(435KB\)](#)

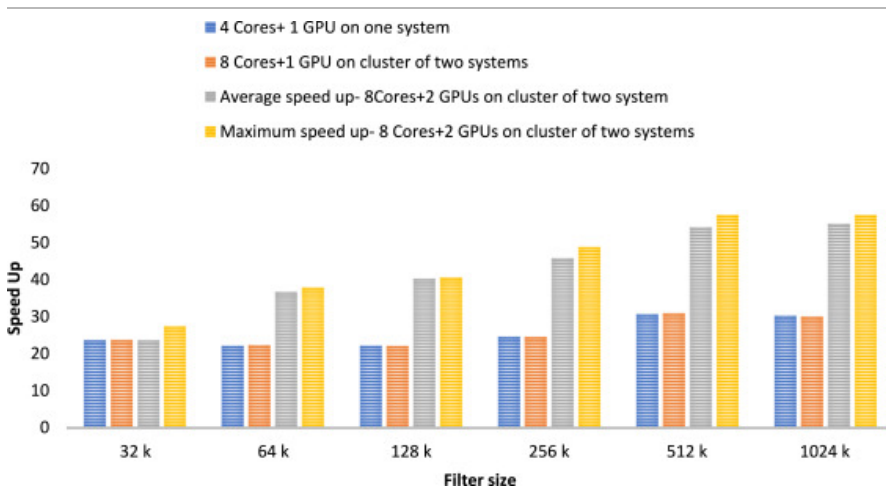
[Download : Download full-size image](#)

Fig. 17. Comparison of the speedup in the first and the second scenarios of GPU cluster.

5.5. Hybrid cluster

Under the circumstances described in Section 4.4, we measured the classification time for different numbers of packets on the hybrid cluster. The speedup graph is shown in Fig. 18. In this graph, the left column represents the speedup for each number of packets in the one-system mode. In this mode, the four CPU cores and all the GPU cores of a single system were involved in packet classification. The second column from left shows the same results with the difference that the GPU of one system is used along with the CPUs of both systems in the cluster. The motivation behind this experiment was that it might be the case that the GPU of one of the nodes in the GPU cluster would not be ready for processing. According to the results, there is not any significant difference in speedup between these two modes. The third and fourth columns respectively represent the maximum and average speedup achieved by the hybrid cluster when all cores of the CPUs and GPUs were engaged in classification. In this mode, the maximum capacity of the cluster was utilized. The ratio of the speedup of the hybrid cluster to the speedup in the one-system mode for 1024K, 512K, and 256K packets is 1.89, 1.85, and 1.95, respectively. These values indicate that the classification speed in the two-system mode is almost twice as much as the speed in the one-system mode. A comparison with the speedup graph of the GPU cluster would show that the speedup achieved by the hybrid cluster is not significant because the CPU of the cluster nodes has a high computation power.

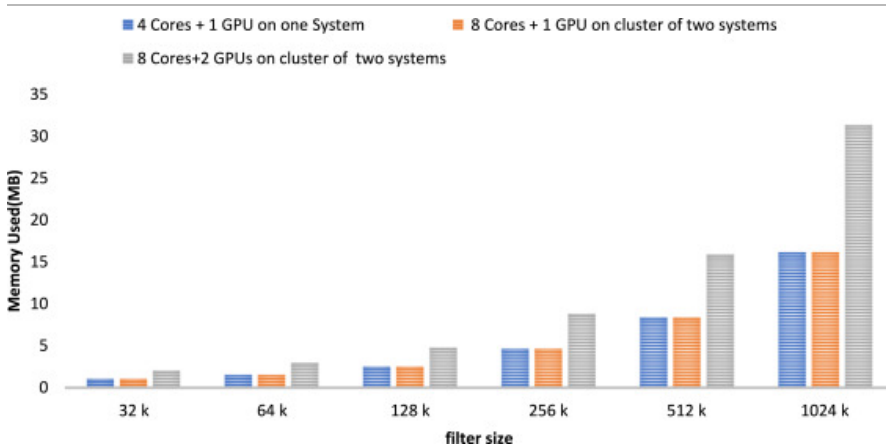
Fig. 19 shows the memory usage of the hybrid cluster. The first column represents the one-system mode where one CPU and one GPU are used. The second column represents the memory usage of the two-system cluster. In the latter, the CPUs of one system and the GPU of the other system were used. The amounts of memory used in both modes are equal. Also, the memory usage of a perfect hybrid cluster is twice as much as in the one-system mode.



[Download : Download high-res image \(392KB\)](#)

[Download : Download full-size image](#)

Fig. 18. The speedup of the one-system and two-system hybrid clusters for different numbers of packets.



[Download : Download high-res image \(208KB\)](#)

[Download : Download full-size image](#)

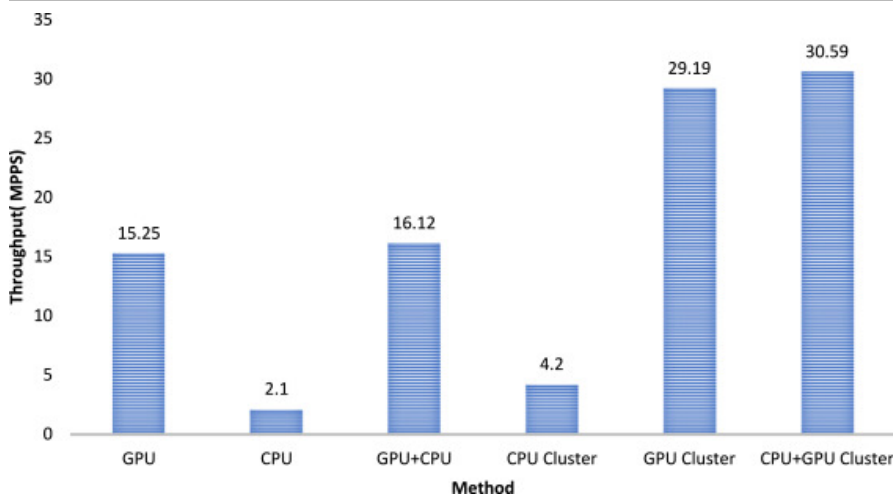
Fig. 19. The memory usage of the one-system and two-system hybrid clusters for different numbers of packets.

5.6. Comparing proposed scenarios

In this section, we compare the throughput achieved by the CPU cluster, the GPU cluster, and the one-system and two-system hybrid clusters. For this comparison, the results of the optimum scenarios of each cluster are used.

Fig.20 illustrates the throughput achieved by all of the clusters for 1024K packets. It can be seen that when only one CPU (with four active cores) was used, the throughput reached 2.1 MPPS. Due to the larger number of cores and its higher parallelization capacity, the GPU has classified 13.15 million packets more than the CPU in the unit of time. When both the CPU and the GPU of a system were used, the number of packets classified in the unit of time increased from 15.25 to 16.12 million. Also, the CPU cluster has reached a throughput of 4.2 MPPS which is twice as much as the throughput of the one-system mode. The highest throughput was 30.59 MPPS and belongs to the hybrid cluster.

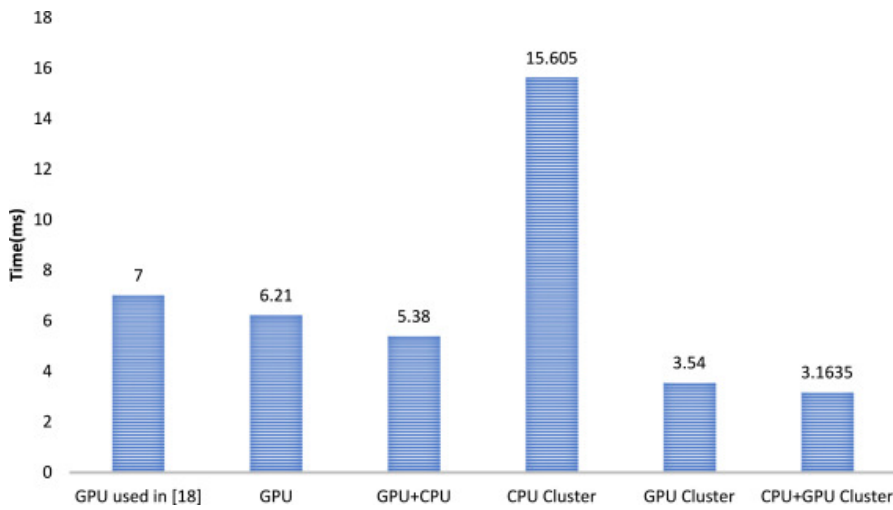
A similar work has been carried out by Abbasi et al.[23]. The hardware specifications used in that study are listed in Table6. Fig.21 compares the classification time of their best scenario with all the scenarios of the present study for an IPC filter set. The number of packets is assumed as 64K. The reason for selecting this number is that the maximum number of packets used in [23] is 64K.



[Download : Download high-res image \(299KB\)](#)

[Download : Download full-size image](#)

Fig. 20. The throughput of the three clusters for 1024K packets.



[Download : Download high-res image \(275KB\)](#)

[Download : Download full-size image](#)

Fig. 21. Comparing the running time of three methods used in [23] in classifying 64K packets.

In their best scenario, the minimum time for the classification of a packet using the GPU is 7ms. The GPU in the present study has reduced this time down to 6.22ms. This reduction is explained by the higher processing power of our GPU. By using engaging the CPU in the hybrid mode, this time has been reduced by 0.83ms to reach 5.38ms. In the GPU cluster mode and the hybrid mode, the classification time has increased by 3.44ms and 3.84ms, respectively, in comparison with the results of Abbasi et al. In the CPU cluster mode, the classification time is comparatively lower because the CPU used in our study has lower processing power.

Table 6. System specifications in [23].

CPU	Intel(R) Core™ i7Q 740 @ 1.73GHz	
RAM	4 GB	
OS	Windows 7 Ultimate, 64-bit	
GPU	Model (Nvidia)	GTX 750
	Architecture	Maxwell
	CUDA Cores	512
	Core clock	1020MHz
	Memory Clock	1250 MHz
	Processing Power (GFLOPS)	1044
	Memory bandwidth	80 GB/s
	SMS	4
	Bus width (bit)	128

6. Conclusion

MEC is a new accelerated trend in ubiquitous computing where the computational resources are being brought nearer to the mobile users. SDN, can serve as an enabler to reduce the complexity barriers involved and let the real potential of MEC be achieved. That is, the complexities resulted from deploying the cloud-like resources and related services at the edge of the mobile network can be solved by a control mechanism that can orchestrate the

distributed environment. All data flow management, service orchestration, and other management tasks are accomplished by the central SDN controller that is transparent to the end-user.

The undeniable efficiency of SDN in managing the sheer number of flows in MEC owes to packet classification. Packet classification is a process whereby network packets are divided into flows. Central to this process is the search for the best matching packet through matching the packet headers against the filters.

A desirable classification algorithm is one that could perform classification with the least memory usage and in the shortest time possible. To reduce the classification time, the packet classification algorithm can be parallelized. With this aim, the present study attempted for the first time to parallelize a typical trie-based packet classification algorithm through different scenarios on a CPU cluster, a GPU cluster, and a hybrid cluster. To evaluate these scenarios, an IPC filter set was used to classify different numbers of packets.

Two major conclusions can be drawn concerning the use of CPU clusters. First, the optimum results are obtained when the total number of MPI processes plus the number of OpenMP threads are equal to the total number of processing cores. Second, MPI has a higher performance than OpenMP in all scenarios but its memory consumption is also higher.

Two scenarios were considered for the GPU cluster mode. In the first scenario in which the filter set and the packets were stored in the global memory, we calculated the optimum number of blocks and achieved a speedup 58 times as large as in the serial mode. In the second scenario, we tried to decompose the filter set down to a point in which the sub-tree corresponding to each subset of the filters could be stored in the shared memory of the CPU. We concluded that the scenario using the shared memory was 12 times as fast as the scenario using the global memory.

In the hybrid method, the entire computing capacity of both CPU and GPU was utilized. The results show that a combination of MPI, CUDA, and OpenMP programming models is the optimum mode. In this mode, the maximum capacity of the cluster is used and all the cores of the CPU and the GPU are involved in packet classification. In comparison with GPU cluster mode, this mode increased the throughput by 1.4 MPPS.

A powerful technology that can enhance GPU performance is GPU Direct. It requires only a single pre-processing time. In other words, all CPUs are involved in the task of pre-processing; rather, after receiving and reading the information, each system can transmit it to its GPU without engaging the CPU of other systems.

CRedit authorship contribution statement

Mahdi Abbasi: Supervision, Conceptualization, Methodology, Formal analysis, Visualization, Writing - review & editing. **Azad Shokrollahi:** Data curation, Software, Writing - original draft. **Mohammad R. Khosravi:** Conceptualization, Validation, Writing - review & editing. **Varun G. Menon:** Methodology, Validation, Writing - review & editing.

Declaration of Competing Interest




The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[Special issue articles](#) [Recommended articles](#)

References

- [1] Abbasi M., Rafiee M., Khosravi M.R., Jolfaei A., Menon V.G., Koushyar J.M.
An efficient parallel genetic algorithm solution for vehicle routing problem in cloud implementation of the intelligent transportation systems
J. Cloud Comput., 9 (2020), p. 6
2020/02/03

[View in Scopus](#) ↗ [Google Scholar](#) ↗

- [2] Lai Z., Hu Y.C., Cui Y., Sun L., Dai N., Lee H.-S.
Furion: Engineering high-quality immersive virtual reality on today's mobile devices
IEEE Trans. Mob. Comput. (2019)
[Google Scholar](#) ↗
- [3] Wan S., Qi L., Xu X., Tong C., Gu Z.
Deep learning models for real-time human activity recognition with smartphones
Mob. Netw. Appl. (2019), pp. 1-13
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [4] Menon V.G., Jacob S., Joseph S., Almagrabi A.O.
SDN powered humanoid with edge computing for assisting paralyzed patients
IEEE Internet Things J. (2019), p. 1
[Google Scholar](#) ↗
- [5] Wan S., Gu Z., Ni Q.
Cognitive computing and wireless communications on the edge for healthcare service robots
Comput. Commun., 149 (2020), pp. 99-106
2020/01/01
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [6] Wan S., Goudos S.
Faster R-CNN for multi-class fruit detection using a robotic vision system
Comput. Netw., 168 (2020), Article 107036
2020/02/26
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [7] Gomez C., Chessa S., Fleury A., Roussos G., Preuveneers D.
Internet of Things for enabling smart environments: A technology-centric perspective
J. Ambient Intell. Smart Environ., 11 (2019), pp. 23-43
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [8] Sourì A., Hussien A., Hoseyninezhad M., Norouzi M.
A systematic review of IoT communication strategies for an efficient smart environment
Trans. Emerg. Telecommun. Technol. (2019), Article e3736
[Google Scholar](#) ↗
- [9] Jiang C., Fan T., Gao H., Shi W., Liu L., Cérin C., *et al.*
Energy aware edge computing: A survey
Comput. Commun., 151 (2020), pp. 556-580
2020/02/01
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [10] Taleb T., Dutta S., Ksentini A., Iqbal M., Flinck H.
Mobile edge computing potential in making cities smarter
IEEE Commun. Mag., 55 (2017), pp. 38-43
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [11] Mao Y., You C., Zhang J., Huang K., Letaief K.B.
A survey on mobile edge computing: The communication perspective
IEEE Commun. Surv. Tutor., 19 (2017), pp. 2322-2358

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[12] Menon V.G., Prathap J.

Vehicular fog computing: challenges applications and future directions

Int. J. Veh. Telemat. Infotain. Syst., 1 (2017), pp. 15-23

[Google Scholar](#) ↗

[13] Mach P., Becvar Z.

Mobile edge computing: A survey on architecture and computation offloading

IEEE Commun. Surv. Tutor., 19 (2017), pp. 1628-1656

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[14] Safavat S., Sapavath N.N., Rawat D.B.

Recent advances in mobile edge computing and content caching

Digit. Commun. Netw. (2019)

2019/09/04

[Google Scholar](#) ↗

[15] Wan S., Zhao Y., Wang T., Gu Z., Abbasi Q.H., Choo K.-K.R.

Multi-dimensional data indexing and range query processing via Voronoi diagram for internet of things

Future Gener. Comput. Syst., 91 (2019), pp. 382-391

2019/02/01

 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗

[16] Wan S., Li X., Xue Y., Lin W., Xu X.

Efficient computation offloading for Internet of Vehicles in edge computing-assisted 5G networks

J. Supercomput. (2019), pp. 1-30

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[17] Chen M., Hao Y.

Task offloading for mobile edge computing in software defined ultra-dense network

IEEE J. Sel. Areas Commun., 36 (2018), pp. 587-597

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

[18] Baktir A.C., Ozgovde A., Ersoy C.

How can edge computing benefit from software-defined networking: A survey, use cases, and future directions

IEEE Commun. Surv. Tutor., 19 (2017), pp. 2359-2391

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[19] Nugroho H.P., Irfan M., Faruq A.

Software defined networks: a comparative study and quality of services evaluation

Sci. J. Inform., 6 (2019), p. 43

[Google Scholar](#) ↗

[20] Saraswat S., Agarwal V., Gupta H.P., Mishra R., Gupta A., Dutta T.

Challenges and solutions in software defined networking: A survey

J. Netw. Comput. Appl., 141 (2019), pp. 23-58

 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗

[21] Zaw H.T., Maw A.H.

Traffic management with elephant flow detection in software defined networks (SDN)

Int. J. Electr. Comput. Eng., 9 (2019), pp. 2088-8708

[Google Scholar](#) ↗

[22] Abbasi M., Fazel S.V., Rafiee M.

MBitCuts: optimal bit-level cutting in geometric space packet classification

J. Supercomput. (2019), pp. 1-24

[CrossRef](#) ↗ [Google Scholar](#) ↗

[23] Abbasi M., Rafiee M.

A calibrated asymptotic framework for analyzing packet classification algorithms on GPUs

J. Supercomput., 75 (2019), pp. 6574-6611

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

[24] Abbasi M., Shokrollahi A.

Enhancing the performance of decision tree-based packet classification algorithms using CPU cluster

Cluster Comput. (2020)

2020/03/16

[Google Scholar](#) ↗

[25] Abbasi M., Vakilian S., Fanian A., Khosravi M.R.

Ingredients to enhance the performance of two-stage TCAM-based packet classifiers in internet of things: greedy layering, bit auctioning and range encoding

EURASIP J. Wireless Commun. Networking, 2019 (2019), pp. 1-15

[CrossRef](#) ↗ [Google Scholar](#) ↗

[26] Rafiee M., Abbasi M., Nassiri M.

An efficient method for parallel implementation of H-trie packet classification algorithm on GPU

Tabriz J. Electr. Eng., 46 (2016), pp. 181-196

[Google Scholar](#) ↗

[27] L. Hyesook, Survey and proposal on packet classification algorithms, in: 2010 International Conference on High Performance Switching and Routing, 2010, pp. 1–134.

[Google Scholar](#) ↗

[28] Kamburugamuve S., Wickramasinghe P., Ekanayake S., Fox G.C.

Anatomy of machine learning algorithm implementations in MPI, Spark, and Flink

Int. J. High Perform. Comput. Appl., 32 (2017), pp. 61-73

2018/01/01

[Google Scholar](#) ↗

[29] Liang F., Lu X.

Accelerating iterative big data computing through MPI

J. Comput. Sci. Tech., 30 (2015), pp. 283-294

2015/03/01




[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

[30] Daydé M., Marques O., Nakajima K.

High performance computing for computational science–VECPAR 2012

(2005)

[Google Scholar](#) ↗

- [31] F. Wolf, I. Psaroudakis, N. May, A. Ailamaki, K.-U. Sattler, Extending database task schedulers for multi-threaded application code, in: Proceedings of the 27th International Conference on Scientific and Statistical Database Management, 2015, p. 25.
[Google Scholar](#) ↗
- [32] NVIDIA
NVIDIA CUDA (Compute unified device architecture) programming guide
(2018)
Available: http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf ↗, (Accessed July 2018)
[Google Scholar](#) ↗
- [33] Y. Li, D. Zhang, A.X. Liu, J. Zheng, GAMT: a fast and scalable IP lookup engine for GPU-based software routers, in: Proceedings of the Ninth ACM/IEEE Symposium on Architectures for Networking and Communications Systems, 2013, pp. 1–12.
[Google Scholar](#) ↗
- [34] Lin F., Wang G., Zhou J., Zhang S., Yao X.
High-performance IPv6 address lookup in GPU-accelerated software routers
J. Netw. Comput. Appl., 74 (2016), pp. 1-10
 [View PDF](#) [View article](#) [Google Scholar](#) ↗
- [35] Zhao Y., Chen L., Xie G., Zhao J., Ding J.
GPU implementation of a cellular genetic algorithm for scheduling dependent tasks of physical system simulation programs
J. Comb. Optim., 35 (2018), pp. 293-317
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [36] Gong T., Fan T., Guo J., Cai Z.
GPU-based parallel optimization of immune convolutional neural network and embedded system
Eng. Appl. Artif. Intell., 62 (2017), pp. 384-395
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [37] Ghidouche K., Sider A., Couturier R., Guyeux C.
Efficient high degree polynomial root finding using GPU
J. Comput. Sci., 18 (2017), pp. 46-56
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [38] Fernández J.L., Ferreiro-Ferreiro A.M., García-Rodríguez J.A., Vázquez C.
GPU parallel implementation for asset–liability management in insurance companies
J. Comput. Sci., 24 (2018), pp. 232-254
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [39] AMD: Global provider of innovative graphics, processors ...
(2018)
Available: <http://www.amd.com> ↗, (Accessed July 2018)
[Google Scholar](#) ↗
- [40] Przymus P., Kaczmarski K.
Dynamic compression strategy for time series database using GPU
New Trends in Databases and Information Systems, Springer (2014), pp. 235-244
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

- [41] G. Vasiliadis, E. Athanasopoulos, M. Polychronakis, S. Ioannidis, PixelVault: Using GPUs for securing cryptographic operations, in: Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, 2014, pp. 1131–1142.
[Google Scholar](#) ↗
- [42] Specifications of the NVIDIA Geforce GT 425M graphics card (2018)
Available: <http://www.geforce.com/hardware/notebook-gpus/geforce-gt-425m/specifications> ↗, (Accessed July 2018)
[Google Scholar](#) ↗
- [43] Zhou S., Qu Y.R., Prasanna V.K.
Multi-core implementation of decomposition-based packet classification algorithms
Parallel Computing Technologies, Springer (2013), pp. 105-119
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [44] K. Kang, Y.S. Deng, Scalable packet classification via GPU metaprogramming, in: Design, Automation & Test in Europe Conference & Exhibition, DATE, 2011, pp. 1–4.
[Google Scholar](#) ↗
- [45] A. Nottingham, B. Irwin, GPU packet classification using OpenCL: a consideration of viable classification methods, in: Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists, 2009, pp. 160–169.
[Google Scholar](#) ↗
- [46] Hung C.-L., Lin Y.-L., Li K.-C., Wang H.-H., Guo S.-W.
Efficient GPGPU-based parallel packet classification
Trust, Security and Privacy in Computing and Communications, TrustCom (2011), pp. 1367-1374
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [47] Deng Y., Jiao X., Mu S., Kang K., Zhu Y.
NPGPU: Network processing on graphics processing units
Theoretical and Mathematical Foundations of Computer Science, Springer (2011), pp. 313-321
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [48] S. Zhou, S.G. Singapura, V.K. Prasanna, High-performance packet classification on GPU, in: High Performance Extreme Computing Conference, HPEC, 2014, pp. 1–6.
[Google Scholar](#) ↗
- [49] Varvello M., Laufer R., Zhang F., Lakshman T.
Multilayer packet classification with graphics processing units
IEEE/ACM Trans. Netw., 24 (2016), pp. 2728-2741
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [50] Zheng J., Zhang D., Li Y., Li G.
Accelerate packet classification using GPU: A case study on hicuts
Computer Science and Its Applications, Springer (2015), pp. 231-238
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [51] Zheng J., Zhang D., Li Y., Li G.
Accelerate packet classification using GPU: A case study on hicuts
Park J.J., Stojmenovic I., Jeong H.Y., Yi G. (Eds.), Computer Science and Its Applications: Ubiquitous Information Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg (2015), pp. 231-238

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

[52] S. Zhou, S.G. Singapura, V.K. Prasanna, High-performance packet classification on gpu, in: High Performance Extreme Computing Conference (HPEC), 2014 IEEE, 2014, pp. 1–6.

[Google Scholar](#) ↗

Cited by (20)

[Multi-view clustering via matrix factorization assisted k-means](#)

2023, Neurocomputing

[Show abstract](#) ▼

[Reliability and robust resource allocation for Cache-enabled HetNets: QoS-aware mobile edge computing](#)

2022, Reliability Engineering and System Safety

Citation Excerpt :

...According to the previous studies, some frameworks have been developed for dynamic power optimization in cache-enabled systems and power-sharing, respectively [9–11]. The aforementioned works in [12] are mainly concentrated on the performance analysis, rate maximization, power minimization problem for NOMA-based mobile edge computing and cooperative multi-server cloud systems. At present, worldwide is advocating wireless technologies as backhauling to obtain green communication, aiming to increase energy efficiency, and providing higher bit rate services....

[Show abstract](#) ▼

[Special Issue on Optimization of Cross-layer Collaborative Resource Allocation for Mobile Edge Computing, Caching and Communication](#)

2022, Computer Communications

[Intelligent workload allocation in IoT–Fog–cloud architecture towards mobile edge computing](#)

2021, Computer Communications

Citation Excerpt :

...Also, AT&T networks use 200 petabytes of bandwidth each year. Sending all the data to the cloud requires a large amount of bandwidth used for sophisticated methods of high-performance flow processing [13,17,18]. Many IoT devices have limited resources and are unable to fulfill their own computation needs [8,15]....

[Show abstract](#) ▼

[Leveraging crowd knowledge to curate documentation for agile software industry using deep learning and expert ranking](#)

2023, Multimedia Systems

[A Flow Table Compression Algorithm for Improving the Storage Capacity of Software-Defined Network Switches](#)

2022, Hsi-An Chiao Tung Ta Hsueh/Journal of Xi'an Jiaotong University

[↗](#) View all citing articles on Scopus

[View Abstract](#)

© 2020 Elsevier B.V. All rights reserved.



Copyright © 2023 Elsevier B.V. or its licensors or contributors.
ScienceDirect® is a registered trademark of Elsevier B.V.

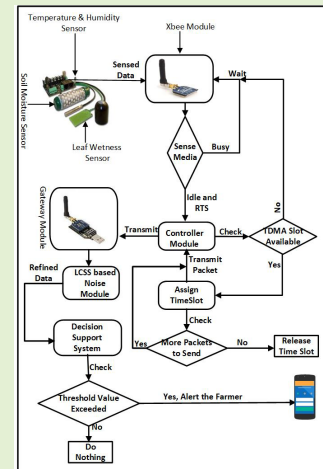


Smart Sensing-Enabled Decision Support System for Water Scheduling in Orange Orchard

Rahim Khan^{ID}, Muhammad Zakarya^{ID}, Member, IEEE, Venki Balasubramanian^{ID}, Member, IEEE, Mian Ahmad Jan^{ID}, Senior Member, IEEE, and Varun G. Menon^{ID}, Senior Member, IEEE

Abstract—The scarcity of water resources throughout the world demands its optimum utilization in various sectors. Smart Sensing-enabled irrigation management systems are the ideal solutions to ensure the optimum utilization of water resources in the agriculture sector. This paper presents a wireless sensor network-enabled Decision Support System (DSS) for developing a need-based irrigation schedule for the orange orchard. For efficient monitoring of various in-field parameters, our proposed approach uses the latest smart sensing technology such as soil moisture, leaf-wetness, temperature and humidity. The proposed smart sensing-enabled test-bed was deployed in the orange orchard of our institute for approximately one year and successfully adjusted its irrigation schedule according to the needs and demands of the plants. Moreover, a modified Longest Common SubSequence (LCSS) mechanism is integrated with the proposed DSS for distinguishing multi-valued noise from the abrupt changing scenarios. To resolve the concurrent communication problem of two or more wasp-mote sensor boards with a common receiver, an enhanced RTS/CTS handshake mechanism is presented. Our proposed DSS compares the most recently refined data with pre-defined threshold values for efficient water management in the orchard. Irrigation activity is scheduled if water deficit criterion is met and the farmer is informed accordingly. Both the experimental and simulation results show that the proposed scheme performs better in comparison to the existing schemes.

Index Terms—Wireless sensor network, precision agriculture, irrigation management systems, DSS, RTS/CTS, LCSS.



I. INTRODUCTION

WORLDWIDE, water is a scarce resource and it requires considerable attention from the research community and industry to ensure its maximum utilization. Agriculture sector is one the water's main consumer because it consumes approximately 70% of the available water to fulfill the food requirements of the world fast growing population [1]. Generally, irrigation schedules are based on farmers experience, crop requirements, environmental conditions, and

soil properties. However, these traditional irrigation procedures are not efficient from the resource utilization perspective as a considerable amount of water is wasted. Due to the recent technological advancements, particularly sensors and actuators, it is possible to develop a smart sensing-enabled automated Decision Support System (DSS) that has the ability to identify water-deficit locations and irrigate those areas on priority basis if needed [2].

Manuscript received June 24, 2020; revised July 19, 2020; accepted July 24, 2020. Date of publication July 28, 2020; date of current version August 13, 2021. The associate editor coordinating the review of this article and approving it for publication was Dr. Hari P. Gupta. (Corresponding author: Mian Ahmad Jan.)

Rahim Khan, Muhammad Zakarya, and Mian Ahmad Jan are with the Department of Computer Science, Abdul Wali Khan University Mardan, Mardan 23200, Pakistan (e-mail: rahimkhan@awkum.edu.pk; mohd.zakarya@awkum.edu.pk; mianjan@awkum.edu.pk).

Venki Balasubramanian is with the School of Engineering, Information Technology and Physical Sciences, Federation University Australia at Mount Helen Campus, Ballarat, VIC 3350, Australia.

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India.

Digital Object Identifier 10.1109/JSEN.2020.3012511

A network of smart sensing devices has the potentials to collect the real-time data for developing an automated Irrigation Management System (IMS) or DSS, that is also known as precision agriculture [3]. The DSS aims to provide the right resources at the right time, which has a direct correlation with the yield improvement of various crops. To realize this objective, smart sensing devices are deployed in agricultural fields where specialized sensors probe their surrounding phenomena such as soil moisture, soil temperature, pH, humidity, and leaf wetness, etc. These gathered phenomena are thoroughly observed by the DSS on a centralized device. Various agriculture-related activities are subject to these observations. There exist numerous studies in this context. In [4], a WSN-based remote water management system for agriculture

sector was implemented in Thailand. The proposed approach was deployed for five months in agricultural fields and various data mining techniques were used to analyze the captured data. However, the proposed approach completely neglected outliers, i.e., noisy data. A DSS-enabled irrigation prediction system was presented in [5] for optimizing the water scheduling of orange orchard. This system has the ability to predict the soil moisture of a particular region within the orchard by applying a hybrid of Support Vector Regression (SVR) and K-mean clustering algorithm on the gathered data. The proposed DSS emphasized on the refinement of captured data before it is being processed by the DSS. However, this system fails to distinguish multivariate noise from abrupt changing scenarios [6]. In [7], a WSN-enabled water management system was deployed in adjacent agricultural fields where different crops were grown. The proposed test-bed utilized both the historical data and variations in the climate values to devise an effective irrigation schedule. An Internet of Things (IoT) and neural network-based DSS was developed in [8] to predict the water-deficit locations. The proposed DSS is capable to precisely detect the required amount of water for irrigation. However, this approach does not consider the outliers in the gathered data.

In precision agriculture, outliers or noise is defined as unwanted data that severely affect the performance of an operational DSS. Usually, this problem occurs due to malfunctioning sensor nodes, interference, collision of packets, circuit failure, extreme pressure, high temperature, and other environmental conditions. As a result, refinement of sensed data prior to its processing by the concern DSS is a challenging issue. Moreover, in the case of shared media, collision of Request To Send / Clear To Send (RTS/CTS) packets and data packets is another challenging issue. Therefore, the development of a precise and accurate technology-assisted DSS is desperately needed to ensure maximum utilization of water in agriculture sector.

In this paper, a smart sensing-enabled DSS for the orange orchard is presented to resolve the aforementioned issues. In our proposed approach, the sensed data by the various smart sensing devices/nodes is processed using a refinement module to ensure accuracy and integrity of data at the destination. Moreover, every node is bounded to transmit its data only if the medium of communication is free. The main contributions of this paper are:

- 1) A smart sensing-enabled DSS is presented for proper management of the irrigation activities in agriculture sector. The proposed agricultural DSS is a need-based system that provides water to a particular area only if it is identified as water-deficit.
- 2) A modified LCSS mechanism is proposed that enables the proposed DSS to differentiate multivariate noise from the abrupt changing scenario.
- 3) An enhanced RTS/CTS mechanism is proposed to resolve the collision issue associated with concurrent communications. Before any transmission activity, every sensor node is bounded to use the classifier-based mechanism that ensures a collision-free communication between two or more operational devices.

The rest of the paper is organized as follows. In Section II, an overview of the literature is presented. In Section III, a detailed description of the proposed smart sensing-enabled DSS module and its deployment for orange orchard are presented. In Section IV, both experimental and simulation results are discussed in detail. Finally, concluding remarks are provided in Section V.

II. LITERATURE REVIEW

Precision agriculture is the technology-assisted farming, which is based on sensor-enabled monitoring, measurement and response generation via the DSS. The responses are generated based on the varying conditions of crops [9]. It enables the farmer to provide the right resources at the right time and right place to any crop [10]. In agriculture, water is an essential resource that is needed to bring forth the maximum potential of the agricultural fields. Moreover, it enables crops to make full use of other yield enhancing production factors [11]. In this section, a brief overview of various test-beds which are related to the proposed work is presented.

Keswani *et al.* [8] have presented an optimal Internet of things (IoTs) and neural network-based irrigation management system that has a one-hour prior prediction capability of water-deficit location(s). For this purpose, various sensors were deployed such as soil moisture, temperature, CO₂, light intensity, and humidity sensors. A hybrid DSS was proposed by Viani *et al.* [12] which was based on the fuzzy logic and farmer's experiences. Likewise, WSNs and General Packet Radio Services (GPRS) were used to form an optimal irrigation management system. Soil moisture sensors with controller modules were deployed in agricultural field(s) [13]. A machine learning and agronomist's encysted knowledge-based irrigation prediction system was proposed that concluded that Gradient Boosted Regression Trees (GBRT) was the best regression model with approximately 93% irrigation prediction accuracy [14]. Dursun and Ozden [15] presented an automatic drip irrigation management system for the cherry trees. A low-cost IoT system for smart irrigation was proposed by NK. Nawandar and Satpute [16]. The system capabilities include the estimation of irrigation schedule, neural-based decisions and remote monitoring. Similarly, a WSN-based irrigation management platform was presented that has the capacity to calculate the quantity of water needed for irrigating a specific area [17]. A novel watering management system, which is based on low-cost IoT components was presented by Khoa *et al.* [18]. Additionally, LoRa LPWAN technology was used for the transmission to ensure the best performance of the proposed system. In FLOW-AID project, WSNs were used to identify water-deficit locations, a situation where plants need water desperately [19]. In 2011, the Information and Communication Technology unit of Commonwealth Scientific and Industrial Research Organization (CSIRO) used various sensor nodes to recover the ecological integrity of Queensland's National Park [20]. Pardossi *et al.* [21] described a methodology of integrating Root Zone sensors with WSNs which is used to identify water deficit locations in agricultural fields. Harun *et al.* [22] described WSNs as an efficient tool to resolve both the decision-making and resource optimization

issues associated with technology-assisted farming. A need-based irrigation practice was presented by Abrishambaf *et al.* [23]. This system has the capacity to schedule the irrigation activity for lowest cost period by using various parameters to temperature, soil moisture, wind, precipitation forecast, and soil calculation. An IoT-based irrigation system was developed to automate the irrigation activity of crops using soil and environmental data [24]. Dong *et al.* [25] presented a pivot-based irrigation procedure to optimize the irrigation activity via wireless underground sensor networks.

III. PROPOSED LCSS-BASED DATA REFINEMENT MECHANISM

Data fusion or refinement of the WSNs capture data has become a dominant research area; as the majority of our daily-life activities are either partially or completely dependent on these DSS-based networks. In this section, a space free LCSS-based data fusion scheme is presented to enhance accuracy and precision level of the proposed agricultural DSS. The captured data of every device C_i , that is wasp-mote agricultural board in the proposed test-bed, is passed through the LCSS-based noise detection module that ensures the accuracy of the refined data. Moreover, the proposed fusion scheme has the capacity to distinguish outliers or noisy data from the abrupt changing scenarios, i.e., an abrupt change occurs if water directly interacts with soil moisture or leaf wetness sensors.

A. Sequence Matching: Definitions and Preliminaries

A sequence SQ_i is defined as a collection of related values, $a_1, a_2, \dots, a_n \in SQ_n$ where n represents length of the data set. The longest common subsequence (LCSS) is represented by $LCSS(k)$, where k defines length of the LCSS. Concatenation process in LCSS is defined as appending any two symbols X and Y to form a subsequence XY such that $X \& Y \in SQ_a$.

Definition 1: Any two values $X \in SQ_a$ and $Y \in SQ_b$ are considered as equal **iff** $\text{distance}(X, Y) \leq 0.05$ or $X \approx Y$.

definitions $LCSS(0, 0)$ is used to describe an empty LCSS.

Definition 2: A value $a_1 \in SQ_a$ is considered as a predecessor of another value $a_2 \in SQ_a$ in $LCSS$ **iff** $\text{index}(a_2 > a_1)$ and for both values \exists (a value $b_m \in SQ_b$ such that $a_1 \approx b_m$ and $a_2 \approx b_m$).

Definition 3: A value $a_1 \in SQ_a$ is not considered as a predecessor of another value $a_2 \in SQ_a$, that is $a_1 \notin LCSS_{\text{matched}}$, **iff** $\text{index}(a_2 < a_1)$ that is $1 > 2$. Although, for both values $a_1 \& a_2 \in SQ_a \exists$ ($b_y \& b_z \in SQ_b$ such that $a_1 \approx b_y \& a_2 \approx b_z \forall$ where y and z represent indexes information).

Definition 4: Similarity index of any two sequences or data sets $SQ_a \& SQ_b$ are higher **iff** length of their $LCSS_{\text{matched}} > \text{length}(SQ_{3 \times n/2} \& SQ_{3 \times m/2})$

Definition 5: The $LCSS(k)$ represents the LCSS of SQ_a and SQ_b **iff** $\forall (a_n \in SQ_a)$ there exist a $b_m \in SQ_b$ such that $a_n \approx b_m$ and $\exists (n' < n \text{ and } m' < m \text{ such that } LCSS(k - n', l - m')$ is generated by n' and m').

B. Computation of the LCSS

The proposed approach uses two different sequences of the same size n that is 10 in this case i.e., SQ_a for storing

current data values and SQ_b for previously transmitted data where $a = 1, 2, 3 \dots n$ and $b = 1, 2, 3 \dots m$. Every device $C_i \in WSN$ is bounded to store the collect data in SQ_a until $a = n$. Initial values for SQ_b is defined manually, once at the deployment stage of WSNs, and are updated according to collected data of sensor(s). For example, soil moisture sensor values are set according to the average values of three different soil moisture sensors which were deployed in dried soil i.e., 250Hz to 260Hz. Once, the network becomes fully operational i.e., sensors begin to probe the phenomena after the defined interval of time, that is 30 second in the deployed WSN infrastructure. In the proposed test-bed, every wasp-mote board C_i is bounded to store their captured data temporarily in sequence SQ_a until value of $a = 10$ and then send it to the gateway.

To refine this data, the proposed test-bed uses a modified form of LCSS and gap-free LCSS. LCSS is used to find the similarity indexes of the currently received data SQ_a and existing data SQ_b . Initially, a matching window control parameter δ is defined that is used to limit the matching window of a value in sequence SQ_a , i.e., $a = 1$, with another sequence SQ_b . In the proposed test-bed, the value of δ is set to three (3) which means that the first element of SQ_a is matched with at-most three elements of SQ_b **iff** these elements are not matched. In phase-I, the first element of SQ_a , i.e., $SQ_1 \in SQ_a$, is matched with element(s) of SQ_b , i.e., $b_1, b_2, \dots, b_\delta \in SQ_b$, such that either a match is found or maximum limit δ is reached. If first element $a_1 \in SQ_a$ matches with any element $b_{1 \text{ to } \delta} \in SQ_b$ then b_m is stored in class $LCSS_{\text{matched}}$ with its position information. However, if a_1 does not match with any element of SQ_b within the defined window δ then a_1 is ignored and subsequent element $a_2 \in SQ_a$ is processed. Likewise, second value $a_2 \in SQ_a$ is matched using similar approach with a slight modification that is its matching criteria with the SQ_b is subjected to the following conditions.

- 1) $a_2 \in SQ_a$ is matched with the first value of SQ_b **iff** $a_1 \in SQ_a$ does not have a matching value in the defined window, i.e., δ .
- 2) $a_2 \in SQ_a$ is matched with value of SQ_b that is stored after previously matched value i.e., $b_1 \in SQ_b$ **iff** $a_1 \in SQ_a \approx b_1 \in SQ_b$.

If $a_2 \in SQ_a$ has a match in SQ_b then matching value $b_{2 \text{ to } \delta} \in SQ_b$ is stored in class $LCSS_{\text{matched}}$ with its position information. For the remaining values of SQ_a , this process is repeatedly applied to compute their LCSS.

In phase-II, first value of SQ_a , i.e., $a_1 \in SQ_a$, is ignored **iff** $a_1 \in LCSS_{\text{matched}}$ and the required LCSS is not computed yet. The remaining values, i.e., $a_2, a_3, \dots, a_n \in SQ_a$, is considered as a refined data set which has nine values. Then, the aforementioned process, i.e., finding $LCSS_{\text{matched}}$ of SQ_a and SQ_b , is repeated. Both LCSSs, i.e., current $LCSS_{\text{matched}}$ and previous $LCSS_{\text{matched}}$, are compared and $LCSS$ with the maximum length is selected whereas other is discarded. This process is repeated until the required LCSS.

To understand this idea, consider two sequences SQ_a and SQ_b which contain data generated by the temperature sensor(s) i.e., $SS_n = 30 \ 34 \ 31 \ 30 \ 33 \ 35 \ 34 \ 30 \ 34 \ 32$ and SQ_m

= 33 30 32 34 30 33 34 30 34 32 where $n=m=10$ and $\delta = 3$. First value $30 \in SQ_a$ is matched with every value of SQ_b within the defined window $\delta = 3$; starting with the first, i.e., $32 \in SQ_b$. A match is encountered at the 2^{nd} position in SQ_b i.e., $30 = 30$. Value 30 is stored in $LCSS_{matched}$ with its position information. Second value $34 \in SQ_a$ is then matched with every value of SQ_b starting from the position 3^{rd} i.e., 31 in this case. However, 31 does not have a matching value in SQ_b within δ . Therefore, it is neglected and the subsequent value $31 \in SQ_b$ is processed which is matched with 3^{rd} value in SQ_b . 31 is stored with its location information in $LCSS_{matched}$. For the remaining values of SQ_a , this process is repeatedly applied until their LCSS is found. It is to be noted that phase-II is applicable only if the computed LCSS length is less than the length of $SQ_b/2$.

Lemma 1: $LCSS(p)$ represents the longest common subsequence of SQ_a & SQ_b of length n and m respectively **iff** $k \geq 1$ and $\exists (a_1 \dots p$ such that $a_1 \dots p \cong b_{1 \dots p} \therefore a_{1 \dots p} \in SQ_a$ and $b_{1 \dots p} \in SQ_b$) where $p \leq n$ & m .

Proof: Applying mathematical induction i.e., for $k = 1$ The $length(LCSS(1))$ is equal to 1 **iff** $a_1 \cong b_1$ where $a_1 \in SQ_a$ and $b_1 \in SQ_b$. (According to **Definition-4**). Hence, the lemma is true for $k = 1$.

Suppose that the lemma is true for $k - 1$ values. We need to prove that the lemma is true for k values. If $LCSS(n, m)$ represents the LCSS of SQ_a & SQ_b then $\exists (n' < n$ and $m' < m$ such that $LCSS(n', m')$ is the LCSS of SQ'_a & SQ'_b for $k - 1$ value).

According to our assumption,

$LCSS(n', m') = p'_1, p'_2, p'_3, \dots, p'_k$ such that $p' < p$ and $p'_1, p'_2, p'_3, \dots, p'_k \in SQ_a$ & $SQ_b \therefore a_n = b_m \therefore n' < n$ and $m' < m$.

Therefore, LCSS of SQ_a and SQ_b is of length k . $\therefore length(LCSS(n', m')) + length(LCSS(n, m)) = k$. Hence, it proves that k is the length of required ($LCSS(p)$).

Conversely, if $length(LCSS(n, m)) \geq k$ and $a_n = b_n$, where $a_n \in SQ_a$ and $b_m \in SQ_b$ then $\exists (n' < n$ and $m' < m$ such that $a_{n'} \cong b_{m'}$. Moreover, $length(LCSS(n', m')) = length(LCSS(n, m)) - 1 \geq k - 1$. Therefore, $LCSS(n', m')$ is the LCSS of $k - 1$ length data sets (By inductive Hypothesis). Hence, the proof i.e., $LCSS(p)$ represents the longest common subsequence of SQ_a & SQ_b . \square

C. Proposed Methodology: Classifier-Based Based RTS/CTS Handshake

To resolve one of the aforementioned issue, i.e., collision of RTS/CTS packets, a classifier-based scheduled RTS/CTS mechanism is presented. Every device $C_i \in IoTs$ shares its communication schedule T_s with the neighboring devices via a smaller scheduled-frame preferably after the deployment phase. The proposed communication scheme consists of two phases i.e., hop-count and classifier-based optimal neighbors discovery phases.

1) Hop-Count Discovery Phase: The base station module S_j broadcasts a scheduled-frame which contains a transmission schedule (T_s), hop-count (H_c) and back-off timer T_b . Moreover, the hop-count value is set to zero as base station is the ultimate destination for every device $C_i \in IoTs$ and

T_b value is set to infinity which distinguishes S_j from the ordinary devices. Active devices C_i which reside in the closed proximity of S_j receive this frame and update it according to their stored information, i.e., $H_c = 1$, T_b and T_s are set according to the equation. 2 & 3 respectively. Moreover, every device C_i maintains a schedule table where valuable information about neighboring nodes is stored i.e., H_c , T_s , residual energy E_r that is calculated using equation 1.

$$E_r = E_i - E_c \quad (1)$$

where E_i and E_c represent the initial and consumed energies respectively.

Back-off timer T_b is computed using equation 2. The idea of adding an H_c value or δ with the generated random number is to minimize the collision probability of neighboring nodes as, usually, these nodes have different H_c values. However, if back-off timer T_b of the two neighboring devices are similar then these devices should recompute their T_b . δ is an infrastructure dependent parameter i.e., for flat networks its value ranges from 5-15 whereas in hierarchical networks its value ranges from 2-5.

$$T_b(C_i) = rand(0 - 1000) + min\left(\frac{T_p(C_i)}{H_c(C_i)}, \delta\right) \quad (2)$$

Transmission schedule T_s is computed using equation 3.

$$T_s(C_i) = T_b + T_p + \gamma \quad (3)$$

where T_p is the average propagation time of C_i 's first hop neighbors which includes both the transmission and processing delays. γ represents the sampling rate of a particular device which will be similar for every $C_i \in WSNs$.

Once a first hop neighboring node C_i updates the schedule-frame, it doesn't broadcast the frame immediately rather it waits for T_b . When T_b expires, C_i broadcasts an updated version of the scheduled-frame which is received by devices reside in vicinity. C_i 's neighboring devices are divided into two groups i.e., Group-I which consists of devices such that their $H_c \leq H_c(C_i)$ whereas Group-II has devices with $H_c > H_c(C_i)$. When a device $C_{i+1} \in Group - I$ receives a scheduled-frame from a neighboring device C_i it updates the schedule table entries according to the message contents and discard it. C_{i+1} discards the received scheduled-frame because it has either transmitted a scheduled-frame or waiting for its T_b to expires as it has already received a similar message from the base station module S_j . Conversely, if the scheduled-frame is received by a device $C_{i+2} \in Group - II$ then it updates the scheduled table information particularly about C_i such as H_c , T_b and T_s . Moreover, C_{i+2} computes its back-off timer using equation 2 and updates the scheduled-frame by replacing H_c , T_b and schedule time T_s with its own. When T_b of C_{i+2} expires it broadcasts the updated scheduled-frame. This process is repeatedly applied until every device $C_i \in WSNs$ in an operational network has a defined H_c value and information about neighboring node's transmission schedules T_s . Additionally, C_i 's transmission schedule is not affected even if it serves as a relaying device, that is forwarding packets of neighboring devices, in addition to its own duties.

2) *Classifier-Based Mechanism to Mitigate the Collisions Ratio of RTS/CTS and Data:* In the proposed scheme, every device C_i maintains a schedule table which contains information about neighboring devices. This information is very useful in both scenarios i.e., minimizing the collision probability and finding an optimal device.

- 1) Where multiple devices initiate a request-to-send message (RTS) at the same time and are interested to start a communication process with a shared device i.e., base station or cluster head (CH) or neighboring node.
- 2) Where a single device C_i has multiple recipient and needs to start communication with a reliable and optimal device.

In scenario-I, without a proper schedule information of neighboring devices, particularly first hop neighbors, collisions will occur and retransmission is mandatory which is not only time-consuming but power consuming too. However, if these devices are bounded to store sufficient information about neighboring devices such as T_s , T_b and H_c then packets collisions are minimized or even avoided. The proposed scheme uses a classifier-based mechanism to resolve the collision issue associated with devices interested in communication with a shared base station or other entity. Since, every neighboring device C_i has a unique back-off timer T_b , hence, the collision probability is zero even if two neighboring devices initiate the RTS process simultaneously.

In scenario-II, if a device C_i is interested to initiate a communication session with a reliable and optimal neighboring device or CH or base station then this device needs a simplified classification mechanism which identifies an optimal device. The proposed classifier-based mechanism uses various parameters such as T_s , T_b , E_r and H_c values to find an optimal neighbor. Neighboring devices are classified using equation 4.

$$C_{opt} = (W_1 * T_b + w_2 * T_s)C_i \quad (4)$$

where $W_1 = 50\%$, $W_2 = 50\%$ represent different weight-ages assigned to these parameters. A neighboring device C_i with minimum value of C_{opt} is an ideal and reliable candidate. However, if H_c and E_r of neighboring devices are not considered by our classifier then it is possible that either the transmitted packets may propagate in opposite directions or forwards to a device with minimum residual energy. In both cases the results are not favorable specifically in resource limited infrastructures, therefore, once the classifier described in equation 4 identifies the optimal neighbors then the two most optimal devices are passed to another classifier as described in equation $\xi_{reliable} = W_3 * H_c + W_4 * E_r(C_i)$ (5)

where $W_3 = 40\%$, $W_4 = 60\%$ are weight-ages assigned to the residual energy and hop-count parameters. A neighboring device C_i with maximum value of $C_{reliable}$ is considered as optimal and reliable device.

D. Implementation of the Proposed Scheme in Agricultural Environment: A Case Study

A precise and accurate DSS (preferably technology-assisted) is subjected to the selection of appropriate devices or sensors C_i , parameters to be sensed, data refinement and communication mechanisms. To accomplish this, wasp-mote

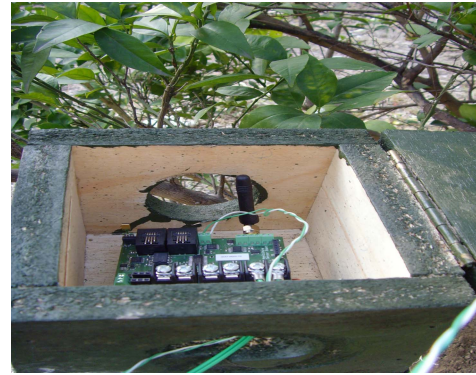


Fig. 1. Deployment of the wasp-mote agriculture boards with humidity and temperature sensors.



Fig. 2. Deployment of leaf wetness sensor in orange orchard.

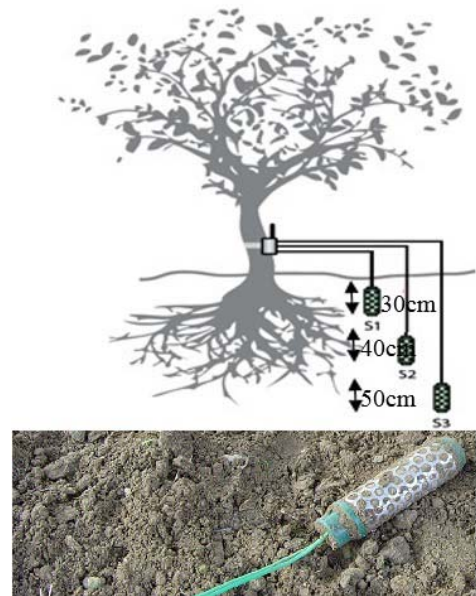


Fig. 3. Deployment of soil moisture sensor in orange orchard.

agricultural boards with a gate way were deployed in the orange orchard of our institute for approximately one year to form an automatic irrigation management system as shown in Fig. 1, Fig. 2 and Fig. 3, respectively. These boards were equipped with soil moisture, temperature, humidity and leaf wetness sensors to collect real time data continuously after a defined interval of time i.e., 30 seconds.

In the proposed DSS, soil moisture parameter is considered due to its vital role in the development of a precise watering

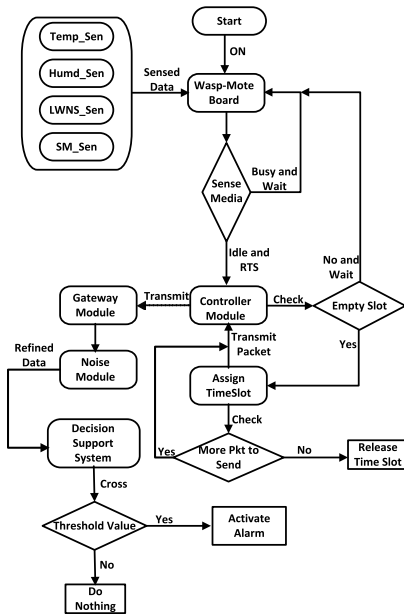


Fig. 4. Data flow diagram of the proposed WSN's based DSS for agriculture.

schedule. For example, if the sensed value is below the threshold value, then that particular area is needed to irrigated on priority basis. To further precise the proposed DSS, soil moisture sensors were deployed at three different levels in the agricultural field, as shown in Fig. 3. Likewise, atmospheric moisture exerts drastic effects on the watering schedules of various crops. Therefore, leaf wetness sensors were deployed in closed proximity to the orange leaves as shown in Fig. 2. Moreover, Temperature and humidity sensors were integrated with the wasp-mote boards to further enhance accuracy of the proposed DSS as shown in Fig. 1. The gateway module is directly connected to a computer via a USB serial port (port-6 in this case) to receive data.

In our previous data collection and communication infrastructure [6], a simplified approach was used to resolve the collision issue associated with simultaneous transmission of two or more devices C_i to a common destination. However, the system enters to deadlock if more than two devices are simultaneously transmitting to a common destination. In this paper, a modified version of the RTS/CTS handshake approach is used to resolve this issue. A detailed description of the proposed WSNs-based DSS for agriculture sector is presented in Fig. 4.

Every wasp-mote board collects real time data from various sensors i.e., temperature, humidity, soil moisture and leaf wetness. This data, say packet-x, is sent to the gateway either directly or through the relaying nodes. In both cases, the transceiver module uses the RTS/CTS handshake approach to avoid collision of packet(s). In the proposed experimental setup, the gateway module is directly connected to a central computer via USB cable specifically through port-6 and the received data is (temporarily) stored automatically using Cool Term software. Before DSS, packet-x is passed through the noise detection module to get the refined data let say packet-y. The DSS module of the proposed system checks packet-y against the threshold value, that is 250Hz for soil moisture sensor, and if the threshold value is crossed then the alarming

TABLE I
WSN'S SIMULATION PARAMETERS SETUP AND THEIR VALUES

Parameters	values
WSN Deployment Area	1000m * 1000m
Sensor Node	50, 100, 500, 1000
Base Station	One
Initial Energy (E_S)	52000 mAh
Residual Energy (E_r)	$E_S - E_c$
Packet Transmission Power Consumption (P_{Tx})	91.4 mW
Channel Delay (Ch_{delay})	10 milliseconds
Packet Receiving Power Consumption (P_{Rx})	59.1 mW
Idle Mode Power Consumption	1.27 mW
Sleep Mode Power Consumption	15.4 μ W
Transceiver Energy (T_i)	1 mW
Transmission Range (T_r)	500m
Receiving Power Threshold (RTS_n)	1024 bits
Packet Size (P_{size})	128 bytes
Initial Hop Count (H_c) of Sensor Nodes	∞
Maximum Distance between Nodes	300-450m
Sampling Rate of sensor nodes	10 to 30 seconds
Topological Infrastructure	Static and Random
Traffic Type	CBR and UDP

unit is activated along with a text message to the farmer on his mobile or LAN. If data is within the defined threshold then it is stored permanently.

IV. EXPERIMENTAL AND SIMULATION RESULTS

In this section, a detail description of both experimental and simulation results are presented to verify the exceptional performance of the proposed system against the existing schemes in terms of computational time, decisions accuracy, packet collision ratio and packet loss ratio. These algorithms were implemented in OMNET++, which is an open source simulation tool specifically designed for the resource limited networks. Initially, a static topological infrastructure, which was later on changed to the random, with a fixed propagation delay was used to mimic the real deployment process of WSNs in general and our deployment infrastructure in particular. Additionally, other networks related parameters such as interference and path-loss ratio were kept constant. A detail description of various simulation related parameters are presented in table I.

Initially, the real-time data set, that is collected through the deployment of various wasp-mote boards based tested in the orange orchard, is used as a testing tool to check the performance of these algorithms particularly in terms of computational time and decision accuracy of the underlined DSS. In terms of computational cost, the performance of these algorithms is presented in Fig. 5, which clearly depicts that the proposed algorithm performance is better than existing algorithms except the noise evading algorithm. However, a common problem associated with NE algorithm is its vulnerability to multi-valued noise, which is quite common in WSNs. Moreover, NE does not differentiate multi-valued noise from an abrupt change scenario. Similarly, the proposed scheme performance is not affected by changing data set size either statically or dynamically because it always uses a fixed sliding window. Therefore, the proposed algorithm is suitable for both scenarios, i.e., static and dynamic datasets. Additionally, these algorithms were evaluated on different static versions of the real-time dataset obtained through our deployed test bed, that

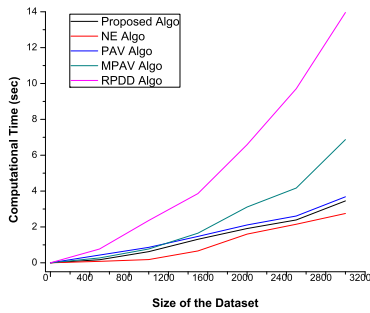


Fig. 5. Performance of DSS in terms of computational time.

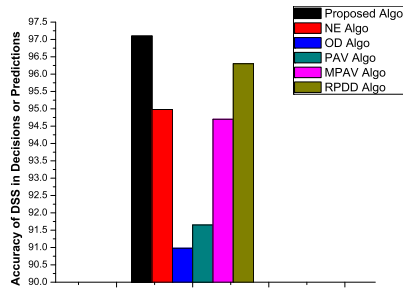


Fig. 6. Accuracy of the DSS in terms of decisions or predictions.

TABLE II
ANALYSIS OF THE PROPOSED & EXISTING ALGORITHMS ON BENCHMARK DATA SETS IN TERMS OF COMPUTATIONAL TIME

Data Set Benchmark	Proposed Algo	NE Algo [6]	PAV Algo [26]	MPAV Algo[26]	RPDD Algo[27]
50words	2.6825	2.3280	3.6090	2.8590	3.0780
B-Cancer	2.4257	2.4220	3.0780	2.7800	2.9060
Two Pattern	2.1865	2.1700	3.1880	2.7130	2.2960
Yoga	2.2958	2.1250	2.8750	2.3900	2.7030
Fish	2.3827	2.0775	2.5630	2.4850	2.5620
Mote Strain	2.8964	2.7340	3.7350	2.9370	3.0160
Diatom-Red	2.0571	2.0180	2.8600	2.0938	2.6980
Amex	2.2145	2.1090	3.4840	2.2500	3.000
Hobo Link	2.3982	2.3120	3.3590	2.4380	2.9530
Face UCR	2.6789	2.0158	3.7340	2.6400	3.4840

was approximately collected in a month or two. The proposed scheme performance is intact as shown in Fig.5.

In agriculture sector, the farmer’s attraction to the technology based infrastructures or DSS will be increased **iff** majority of their decisions or predictions are accurate. Therefore, the proposed algorithm is eager to improve accuracy of the agricultural DSS with the available computational resources and minimum cost. The decision accuracy of the proposed and existing algorithms based DSS is depicted in Fig. 6 which shows the exceptional performance of the proposed algorithm based DSS than existing algorithms. Moreover, it is evident from Fig. 6 that the NEA based DSS has a high probability of errors or wrong decision(s).

The claims of an algorithm is considered as authentic **iff** it is tested on publicly available benchmark datasets. Therefore, these algorithms were tested on various publicly available benchmark datasets as shown in Table-II. The computational time of the proposed algorithm is less than that of existing algorithms except NE which has other issues as described above. We have observed that the computational time of the proposed scheme is inversely proportional to the similarity indexes of the datasets or matching windows i.e., if similarity

TABLE III
COMPARATIVE ANALYSIS OF THE PROPOSED & EXISTING ALGORITHMS ON BENCHMARK DATA SETS IN TERMS OF ACCURACY

Data Set Benchmark	Proposed Algo	NE Algo [6]	RPDD Algo[27]	PAV Algo [26]
50words	96.21	90.40	94.66	96.59
B-Cancer	96.33	89.29	93.48	96.30
Two Pattern	96.10	89.95	93.12	95.83
Yoga	96.79	90.62	94.50	96.74
Fish	95.78	90.18	94.31	95.69
Mote Strain	95.48	89.63	93.85	95.35
Diatom-Red	96.67	90.06	94.19	96.56
Amex	96.14	87.54	93.65	95.91
Hobo Link	96.89	91.05	94.76	96.82
Face UCR	96.99	91.98	94.38	96.98

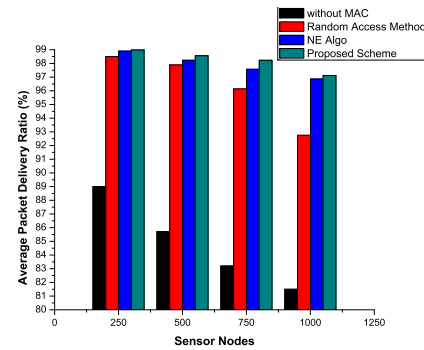


Fig. 7. Comparison of the average packet delivery ratio (simulated results).

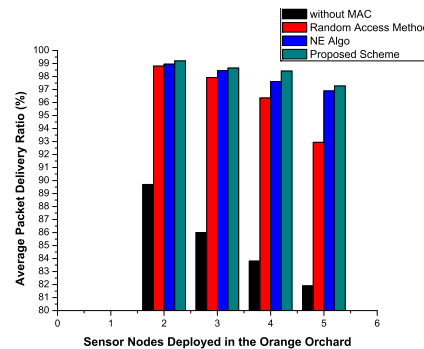


Fig. 8. Average packet delivery ratio (experimental results).

index is high the computational time will be small and vice versa. Due to the high similarity indexes of the benchmark datasets B-Cancer and Two Patterns, the computational time of the proposed algorithms is approximately equal to that of NE algorithm as shown in Table II.

Accuracy of the proposed and existing algorithms based DSS on various publicly available benchmark datasets are depicted in Table III. It is evident from Table III that the proposed mechanism is an ideal candidate for the design of an accurate and precise technology based DSS for the agriculture sector.

Packet delivery ratio is the ratio of successfully delivered packets, particularly at the destination module, to the transmitted one in an operational network. We have observed that the proposed scheme has the maximum packet delivery ratio for both real-time and simulated data against its rival schemes as shown in Fig. 7 and Fig. 8; as packet delivery ratio is inversely proportional to the packet loss ratio which is mostly due to

the packet collision. In proposed scheme, the collision issue is resolved by utilizing the RTS/CTS handshaking scheme.

V. CONCLUSION

Smart sensing-enabled networks, such as WSNs, have the ability to predict when and where the irrigation activities need to be performed. These networks enable the farmers to evaluate the required amount of water for irrigation purposes based on the data sensed by various nodes. In this paper, a real-time smart sensing-enabled Decision Support System (DSS) was presented for optimizing the water schedules for orange orchard. Smart sensing-enabled devices were deployed in different regions for approximately one year to collect soil moisture, temperature, humidity and leaf-wetness of the orchard. The gathered raw data were refined by passing it through a noise module for outliers detection. The DSS module matches the refined data against the threshold values using a modified LCSS mechanism. If these data are below the threshold value, e.g., less than 250Hz for the soil moisture sensor, then irrigation activity is scheduled in that region and the farmer is notified via a text message. Moreover, a modified version of the RTS/CTS handshake mechanism was presented to ensure the successful delivery of packets and collision avoidance. Both the experimental and simulation results showed the exceptional performance of our proposed scheme against the existing schemes for outliers detection and successful delivery of packets.

REFERENCES


- [1] G. Husak and K. Grace, "In search of a global model of cultivation: Using remote sensing to examine the characteristics and constraints of agricultural production in the developing world," *Food Secur.*, vol. 8, no. 1, pp. 167–177, Feb. 2016.
- [2] J. J. Estrada-López, A. A. Castillo-Atoche, and E. Sanchez-Sinencio, "Design and fabrication of a 3-D printed concentrating solar thermo-electric generator for energy harvesting based wireless sensor nodes," *IEEE Sensors Lett.*, vol. 3, no. 11, pp. 1–4, Nov. 2019.
- [3] H. M. Jawad *et al.*, "Accurate empirical path-loss model based on particle swarm optimization for wireless sensor networks in smart agriculture," *IEEE Sensors J.*, vol. 20, no. 1, pp. 552–561, Jan. 2020.
- [4] J. Muangprathub, N. Boonnam, S. Kajornkasirat, N. Lekbangpong, A. Wanichsombat, and P. Nillaor, "IoT and agriculture data analysis for smart farm," *Comput. Electron. Agricult.*, vol. 156, pp. 467–474, Jan. 2019.
- [5] A. Goap, D. Sharma, A. K. Shukla, and C. R. Krishna, "An IoT based smart irrigation management system using machine learning and open source technologies," *Comput. Electron. Agricult.*, vol. 155, pp. 41–49, Dec. 2018.
- [6] R. Khan, I. Ali, M. Zakarya, M. Ahmad, M. Imran, and M. Shoaib, "Technology-assisted decision support system for efficient water utilization: A real-time testbed for irrigation using wireless sensor networks," *IEEE Access*, vol. 6, pp. 25686–25697, 2018.
- [7] S. A. Nikolidakis, D. Kandris, D. D. Vergados, and C. Douligieris, "Energy efficient automated control of irrigation in agriculture by using wireless sensor networks," *Comput. Electron. Agricult.*, vol. 113, pp. 154–163, Apr. 2015.
- [8] B. Keswani *et al.*, "Adapting weather conditions based IoT enabled smart irrigation technique in precision agriculture mechanisms," *Neural Comput. Appl.*, vol. 31, no. S1, pp. 277–292, Jan. 2019.
- [9] D. K. Shannon, D. E. Clay, and N. R. Kitchen, *Precision Agriculture Basics*, vol. 176. Hoboken, NJ, USA: Wiley, 2020.
- [10] D. Shadrin, A. Menshchikov, D. Ermilov, and A. Somov, "Designing future precision agriculture: Detection of seeds germination using artificial intelligence on a low-power embedded system," *IEEE Sensors J.*, vol. 19, no. 23, pp. 11573–11582, Dec. 2019.
- [11] H. Panda, H. Mohapatra, and A. K. Rath, "WSN-based water channelization: An approach of smart water," in *Smart Cities—Opportunities and Challenges*. Singapore: Springer, 2020, pp. 157–166.
- [12] F. Viani, M. Bertolli, M. Salucci, and A. Polo, "Low-cost wireless monitoring and decision support for water saving in agriculture," *IEEE Sensors J.*, vol. 17, no. 13, pp. 4299–4309, Jul. 2017.
- [13] J. Gutierrez, J. F. Villa-Medina, A. Nieto-Garibay, and M. A. Porta-Gandara, "Automated irrigation system using a wireless sensor network and GPRS module," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 1, pp. 166–176, Jan. 2014.
- [14] A. Goldstein, L. Fink, A. Meitin, S. Bohadana, O. Lutenberg, and G. Ravid, "Applying machine learning on sensor data for irrigation recommendations: Revealing the agronomist's tacit knowledge," *Precis. Agricult.*, vol. 19, no. 3, pp. 421–444, Jun. 2018.
- [15] M. Dursun and S. Ozden, "A wireless application of drip irrigation automation supported by soil moisture sensors," *Sci. Res. Essays*, vol. 6, no. 7, pp. 1573–1582, 2011.
- [16] N. K. Nawandar and V. R. Satpute, "IoT based low cost and intelligent module for smart irrigation system," *Comput. Electron. Agricult.*, vol. 162, pp. 979–990, Jul. 2019.
- [17] A. Dahane, B. Kechar, Y. Meddah, and O. Benabdellah, "Automated irrigation management platform using a wireless sensor network," in *Proc. 6th Int. Conf. Internet Things, Syst., Manage. Secur. (IOTSMS)*, Oct. 2019, pp. 610–615.
- [18] T. A. Khoa, M. M. Man, T.-Y. Nguyen, V. Nguyen, and N. H. Nam, "Smart agriculture using IoT multi-sensors: A novel watering management system," *J. Sens. Actuator Netw.*, vol. 8, no. 3, p. 45, Aug. 2019.
- [19] J. Balendonck, J. Hemming, B. V. Tuijl, L. Incrocci, A. Pardossi, and A. Marzaletti, "Sensors and wireless sensor networks for irrigation management under deficit conditions (FLOW-AID)," *AgEng2008*, Crete, Greece, Tech. Rep., 2008. [Online]. Available: <https://research.wur.nl/en/publications/sensors-and-wireless-sensor-networks-for-irrigation-management-un>
- [20] L. P. Shoo *et al.*, "Moving beyond the conceptual: Specificity in regional climate change adaptation actions for biodiversity in South East Queensland, Australia," *Regional Environ. Change*, vol. 14, no. 2, pp. 435–447, Apr. 2014.
- [21] A. Pardossi *et al.*, "Root zone sensors for irrigation management in intensive agriculture," *Sensors*, vol. 9, no. 4, pp. 2809–2835, Apr. 2009.
- [22] A. N. Harun, M. R. M. Kassim, I. Mat, and S. SarahRamli, "Precision irrigation using wireless sensor network," in *Proc. Int. Conf. Smart Sensors Appl. (ICSSA)*, 2015, pp. 71–75.
- [23] O. Abrishambaf, P. Faria, L. Gomes, and Z. Vale, "Agricultural irrigation scheduling for a crop management system considering water and energy use optimization," *Energy Rep.*, vol. 6, pp. 133–139, Feb. 2020.
- [24] J. Boobalan, V. Jacintha, J. Nagarajan, K. Thangayogesh, and S. Tamilarasu, "An IoT based agriculture monitoring system," in *Proc. Int. Conf. Commun. Signal Process. (ICCCSP)*, Apr. 2018, pp. 0594–0598.
- [25] X. Dong, M. C. Vuran, and S. Irmak, "Autonomous precision agriculture through integration of wireless underground sensor networks with center pivot irrigation systems," *Ad Hoc Netw.*, vol. 11, no. 7, pp. 1975–1987, Sep. 2013.
- [26] X.-Y. Chen and Y.-Y. Zhan, "Multi-scale anomaly detection algorithm based on infrequent pattern of time series," *J. Comput. Appl. Math.*, vol. 214, no. 1, pp. 227–237, Apr. 2008.
- [27] D. T. J. Huang, Y. S. Koh, G. Dobbie, and R. Pears, "Detecting changes in rare patterns from data streams," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Cham, Switzerland: Springer, 2014, pp. 437–448.



Full length article

Efficient equalisers for OFDM and DFrFT-OCDM multicarrier systems in mobile E-health video broadcasting with machine learning perspectives

Hani H. Attar^a , Ahmad A.A. Solyman^b , Abd-Elnaser Fawzy Mohamed^c , Mohammad R. Khosravi^{d e} ,
Varun G. Menon^f  , Ali Kashif Bashir^g , Pooya Tavallali^h 

[Show more](#) [Outline](#) | [Share](#)  [Cite](#) <https://doi.org/10.1016/j.phycom.2020.101173> [Get rights and content](#) 

Abstract

Recently, the orthogonal frequency-division multiplexing (OFDM) system has become an appropriate technique to be applied on the physical layer in various requests, mainly in wireless communication standards, which is the reason to use OFDM within mobile wireless medical applications. The OFDM with cyclic prefix (CP) can compensate lacks for the time-invariant multi-path channel effects using a single tap equaliser. However, for mobile wireless communication, such as the use of OFDM in ambulances, the Doppler shift is expected, which produces a doubly dispersive communication channel where a complex equaliser is needed. This paper proposes a low-complexity band LDL^H factorisation equaliser to be applied in mobile medical communication systems. Moreover, the discrete fractional Fourier transform (DFrFT) is used to improve the communication system's performance over the OFDM. The proposed low-complexity equaliser could improve the OFDM, and the DFrFT-orthogonal chirp-division multiplexing (DFrFT-OCDM) system's performance, as illustrated in the simulation results. This proves that the recommended system outperforms the standard benchmark system, which is an essential factor as it is to be applied within mobile medical systems.



Keywords

Factorisation equaliser; Doppler shift; Cyclic prefix; Doubly dispersive channel; Channel model

1. Introduction

Mobile wireless communication systems for E-health applications have received more attention recently with the goal to achieve a mobile hospital and patient monitoring system. Accordingly, the mobile wireless communication system, including video broadcasting features, is urgently needed in such applications. The orthogonal frequency-division multiplexing (OFDM) is the base for several communication systems such as, European digital video broadcasting systems like Digital Video Broadcasting-Terrestrial (DVB-T), DVB-Second Generation Terrestrial (DVB-

T2), DVB Handheld (DVB-H), Long-Term Evolution (LTE), and fifth generation (5G) mobile communication systems. The popularity of OFDM systems is based on its ability to compensate the effect for the time-invariant channel matrix. However, the OFDM loses its optimality against intercarrier interference (ICI) due to Doppler shift (doubly dispersive channel) or carrier frequency offset; accordingly, the system will be in need of sophisticated equalisers [1], [2]. In [3], [4], the discrete Fourier transform (DFT) was replaced by the discrete fractional Fourier transform (DFrFT) for multicarrier systems, which resulted in decreasing the Doppler frequency spread's effect, benefiting from the DFrFT subcarrier's chirped nature that mitigates the Doppler shift. As such, the ICI was reduced.

While DFrFT gives a more improved performance than DFT under the doubly dispersive fading channel, there is still a need for a complex equaliser [5]. Simple equalisers were proposed in [6], [7], [8], wherein [6] least-squares problems (LSQR) algorithm was offered to solve the matrix inversion iteratively; accordingly, no matrix inversion is needed, which simplifies the equaliser. In [7], the equaliser was simplified by using a banded matrix, while in [8], a new approach is proposed, which is based on both a banded matrix and the LDL^H factorisation algorithm.

Unlike the aforementioned simple equalisers which were applied with OFDM, this paper proposes the Orthogonal Chirp Division Multiplexing (DFrFT-OCDM) systems, and then combines the simple suggested equaliser under a time-variant multi-path channel, which is deemed to be suitable for a medical mobile video broadcasting system. Furthermore, the DFrFT-OCDM system was introduced in detail, including the way it is able to replace the DFT on the OFDM, and primarily, the simple equaliser has been added with DFrFT-OCDM to fit the mobile medical applications. Moreover, the doubly dispersive channel details, with their effects on the OFDM and the DFrFT-OCDM system's performance, will be outlined. The equalisation challenges will be specified, then an assessment between some known complex equalisers will be delivered to evaluate the equalisers' behaviour when improving the systems. The suggested simple equalisation methods based on the LDL^H factorisation algorithm is explained and presented as a practical solution for mobile medical applications.

802.11-WLAN video streaming was investigated in [9] over m-health claims, where a medical channel-adaptive fair allocation scheme was proposed to enhance the Quality of service (QoS) for IEEE 802.11 (WLAN). More recent work against real-time medical applications were explained in [10], where an adaptive video encoder compared to a real-time medical use is investigated to maximise the encoded video's quality, improve encoding rate, and to minimise the bit rate demands. In [11], an experimental set was introduced to provide mobile WiMAX video streaming performance analysis for Bandwidth on demand (BOND) services. More recent research and proposed wireless medical applications can be found in [12], [13], [14], [15]. Comprehensive knowledge regarding the structure of health monitoring and machine learning can be found in the well-cited reference book [16], where the theory and the demonstration of the health monitoring structure were presented. Recently, a lot of research has been carried out in this field, including, [17], where the limitations of machine learning approaches have been investigated, and future clinical translations defined. Specific application for using machine learning within health monitoring is rapidly increasing, for example, in [18] where this technique was proposed for the early prediction of asthma attacks.

On the other hand, [19] investigates the scenario of E-health applications that apply the multi-service stream network, which concludes that the mathematical model class $G / G / 1$ – in its general case of a single-channel system – is regarded as an appropriate technique to be implemented within the E-health applications. Indeed, the short time delay and the jitter are practically suitable for E-health primarily. Moreover, the packet losses and the error rates are also considered to be suitable within E-health.

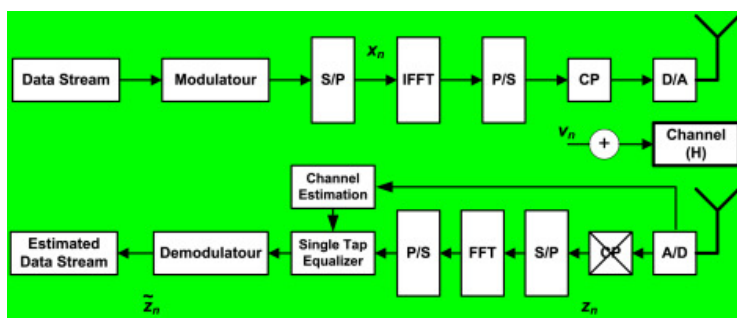
The paper is organised as follows: In Section 2, the background of the OFDM system equalisation is explained to equip the reader with a more comprehensive understanding of the research work presented. The preliminaries for this research are also stated in this section. The proposed approach is presented in Section 3. Section 4 details the performance analysis of the proposed method, technical discussion, and deep computing/machine learning perspectives. Finally, conclusions are presented in the last section.

2. Background and preliminaries

The OFDM allows high data rates to be reliably and efficiently transmitted over the delay-dispersive channels. By dividing the transmitted signal into several narrow bandwidth sub-carriers, OFDM can mitigate the undesirable multi-path effects, mainly, the inter symbol interference (ISI) quantity in long symbol time systems. Moreover, at the beginning of each symbol, a guard period is added – termed cyclic prefix (CP) – to eliminate the expected effects of ISI over the multi-path signals' delay. The multi-path effect tolerant resulting from CP makes OFDM more suitable for high data rate transmission over terrestrial locations rather than single carrier transmissions.

The CP has significant influence over the OFDM system equalisation, as a result of inserting it at the first part of the OFDM symbol. This transfers the multi-path frequency fading channel matrix into a circulant matrix that can be diagonalised by the FFT at the receiver side. The diagonalised channel matrix can be compensated using a single tap equaliser, which can be considered as a simple multiplication in the gain and phase components.

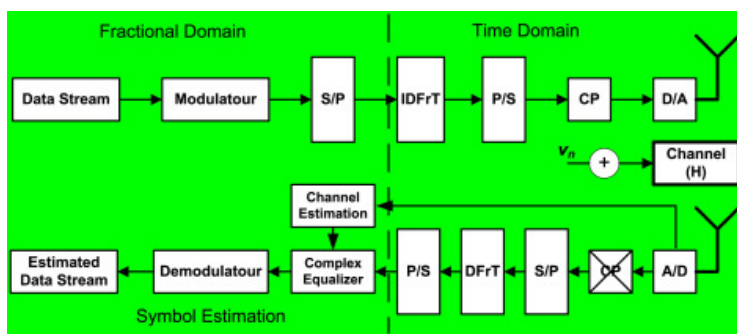
The basic block diagram for a baseband OFDM transmission and reception system is shown in [Fig. 1](#). The OFDM signal is corrupted by passing through the channel. Taking into consideration that the receiver's mission is to obtain the useful information from the corrupted message correctly, accordingly, the receiver converts the received signal into its original form depending on a single tap equaliser.



[Download : Download high-res image \(193KB\)](#)

[Download : Download full-size image](#)

Fig. 1. Basic baseband OFDM transmission and reception system.



[Download : Download high-res image \(235KB\)](#)

[Download : Download full-size image](#)

Fig. 2. The Basic baseband DFrFT-MCM transmission and reception system with a complicated equaliser.

Let us consider the communication system in [Fig. 1](#) in its sequence processing steps over a noisy, frequency fading channel. The received symbols are shown in (1):

$$\mathbf{z}_n = \mathbf{H}\mathbf{F}^H \mathbf{x}_n + \mathbf{v}_n \quad (1)$$

where \mathbf{z}_n is the received signal, \mathbf{H} is the $N \times N$ frequency fading channel matrix, N is the number of subcarriers, \mathbf{F}^H is the inverse discrete Fourier transform (IDFT) matrix of DFT, \mathbf{x}_n is the data vector transmitted in the n th OFDM symbol, and \mathbf{v}_n is the time domain of the white Gaussian noise (WGN). After demodulation and using DFT,

the received vector can be calculated as:

$$\tilde{\mathbf{z}}_n = \mathbf{F}\mathbf{H}\mathbf{F}^H \mathbf{x}_n + \mathbf{F}\mathbf{v}_n \quad (2)$$

Where \mathbf{H} is a circulant matrix (resulting from CP), $\mathbf{F}\mathbf{H}\mathbf{F}^H$ becomes a diagonal matrix[20], and we can equalise the received signal by simple adjustment of the amplitude and phase for the received sequence[21]. This property is one of the key advantages of OFDM as it reduces the complexity of the equalisation process in a multi-path fading channel, which is a harsh environment requiring complex equalisers[22]. However, this property is valid only in time-invariant frequency-selective multi-path channels[23].

When the channel is doubly selective or the receiver induces a frequency offset,[24], the channel matrix is no longer circulant, the DFT cannot diagonalise and ICI appears. In this event, OFDM needs a complicated equaliser[8], [22], [25], such as the minimum mean square error (MMSE) equaliser. The fractional Fourier transform (FrFT) proposed a new base for OFDM[3], [4] that can enhance multicarrier modulation (MCM) systems' performance under a doubly dispersive fading channel because of its ability to cope with the Doppler shifts and to compensate its effects.

2.1. Discrete Fractional Fourier Transform (DFrFT)

The FrFT was presented as a new idea in 1929[26] as a generalisation of the FT. Namias reintroduced FrFT in mathematics for applications in quantum mechanics in 1980[27]. The DFrFT appeared after many groups of researchers reinvented FrFT[28]. Later, low-complexity representations, computational cost, and applications for the DFrFT were investigated in[29], [30]. Nowadays, DFrFT is being used in various requests, such as in optics, image processing, and signal processing.

One of the FrFT definitions is that:

"A fractional Fourier transform is a rotation operation on the time-frequency distribution by angle α " [28].

For $\alpha = 0$ when DFrFT has no effect, for $\alpha = \pi / 2$ when DFrFT returns to FT, and for any value of α in between 0 to $\pi / 2$; the DFrFT substitutes the time-frequency distribution based on the value of α .

The transformation kernel of the continuous FrFT is according to[30]:

$$K_\alpha(t, y) = A_\alpha e^{j\pi(t^2+y^2)\cot\alpha - j2\pi ty \csc\alpha} \quad (3)$$

where α is the rotation angle for the transformation process and

$$A_\alpha = \frac{e^{-j\pi \operatorname{sign}(\sin\alpha)/4 + j\alpha/2}}{\sqrt{|\sin\alpha|}} \quad (4)$$

The FrFT becomes:

$$f_\alpha\{d(t)\}(y) = X_\alpha(y) = \int_{-\infty}^{\infty} d(t)K_\alpha(t, y)dt \quad (5)$$

$$d(t) = \int_{-\infty}^{\infty} D_\alpha(y)K_{-\alpha}(t, y)dy \quad (6)$$

The fractional Fourier signal domain is defined by the α angle for $0 < |\alpha| < \pi$. Fourier transform can be obtained using $\alpha = \pi / 2$. There are several DFrFT algorithms with various properties and accuracies. The DFrFT algorithm proposed in[31] is used in this work. Suppose that the input and output functions of the DFrFT $f(t)$ and $F_\alpha(y)$ respectively have the chirp period of order p with the period $T_p = N\Delta t$, $F_p = M\Delta y$, and the sampled signals are bound between the interval Δt and Δy as:

$$d(n) = f(n\Delta t), D_\alpha(m) = F_\alpha(m\Delta y) \quad (7)$$

where $n = 0, 1 \dots N - 1$ and $m = 0, 1 \dots M - 1$. When $\alpha \neq X \cdot \pi$ (X is an integer), (5) can be converted to:

$$D_\alpha(m) = A_\alpha \Delta t e^{j\frac{\alpha}{2} \cot\alpha m^2 \Delta y^2} \sum_{n=0}^{N-1} e^{j\frac{\alpha}{2} \cot\alpha n^2 \Delta t^2} e^{j \csc\alpha n \cdot m \cdot \Delta t \cdot \Delta y} d(n) \quad (8)$$

when $M = N$ the transformation is reversible, with the condition:

$$\Delta t \Delta y = 2\pi \sin \alpha / M \quad (9)$$

Eq.(8) may also be written in a multiplication of matrix and vector form,

$$\mathbf{D} = \mathbf{F}_\alpha \mathbf{d} \quad (10)$$

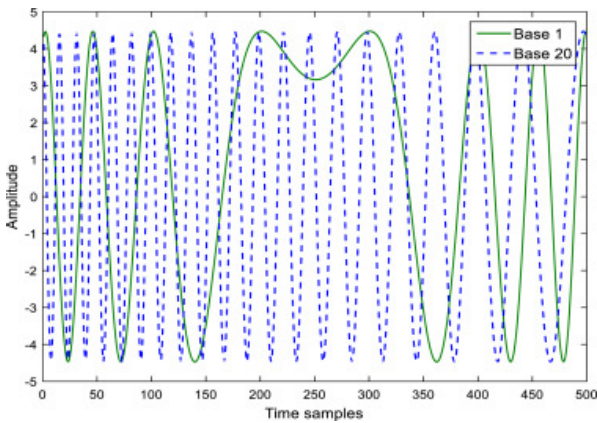
where $\mathbf{D} = [D\alpha(0), D\alpha(1), \dots, D\alpha(N-1)]^T$, $\mathbf{d} = [d(0), d(1), \dots, d(N-1)]^T$, and \mathbf{F}_α is an $N * N$ matrix in the same way, the IDFrFT may be written as:

$$\mathbf{d} = \mathbf{F}_{-\alpha} \mathbf{D} \quad (11)$$

where $\mathbf{F}_{-\alpha} = \mathbf{F}_\alpha^H$.

The DFrFT-MCM system shown in Fig.2 is a system based on block data transfer, and the subcarriers are orthogonal to each other where each subcarrier is a different chirp signal, so we can call it DFrFT-OCDM. Two bases for the DFrFT-OCDM system are shown in Fig.3 and the spectral energy distribution for the two bases are shown in Fig.4.

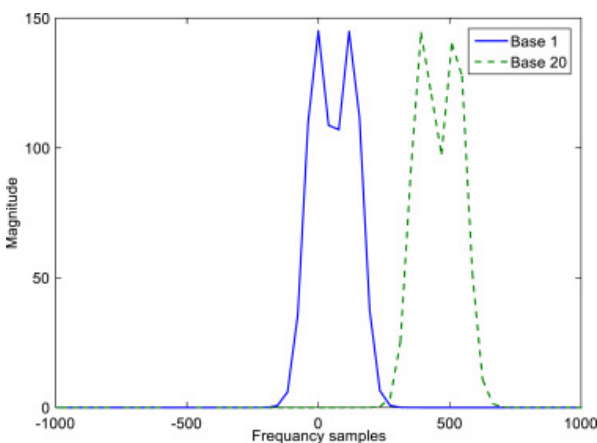
The Wigner distribution in time and frequency domain for the 1st basis signal and the 20th basis signal with $\alpha = 0.7$ are shown in Fig.5. The figure shows that the DFrFT bases are frequency varying with time, which is a property of the DFrFT transformation.



[Download : Download high-res image \(506KB\)](#)

[Download : Download full-size image](#)

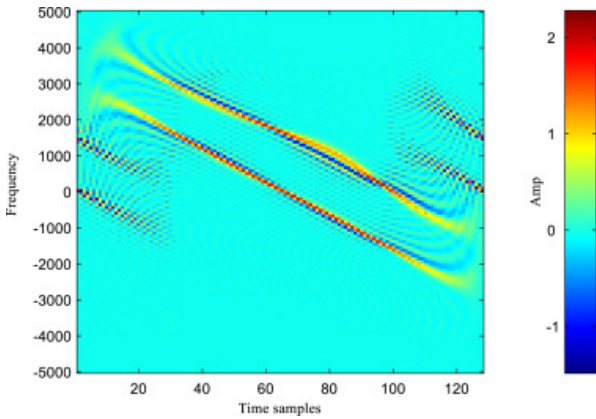
Fig. 3. DFrFT-OCDM basis for the 1st and the 20th basis signals.



[Download : Download high-res image \(140KB\)](#)

[Download : Download full-size image](#)

Fig. 4. Spectral Energy Distribution for the 1st and the 20th basis signals.



Download : [Download high-res image \(459KB\)](#)

Download : [Download full-size image](#)

Fig. 5. The Wigner distribution for the 1st basis signal and the 20th basis signal.

The transmitted DFrFT-OCDM signal is a combination of many blocks, starting with a CP to eliminate the ISI. However, the need for fast mobile communications with significant high data rates and long symbols introduces larger ICI, which forces the use of complicated equalisers.

The DFrFT-OCDM system complexity is nearly equivalent to the traditional OFDM system [3], and both systems exhibit almost the same performance when the channel is time-invariant. However, neither of them can diagonalise the time-variant channel matrix; the DFrFT-OCDM can compress it towards the diagonal much more effectively than OFDM, which is the main advantage of the DFrFT-OCDM system. This enables it to achieve a better performance than OFDM and provides the opportunity to use low-complexity equalisers while maintaining this better performance.

2.2. Doubly dispersive channel

The channels of the mobile-radio applications have time-variant behaviour, due to the transmitter and/or the receiver movement that results in the continual changing of the propagation paths. The changing pace of the propagation circumstances is causally related to the fading rapidity, i.e. the speed of the changing rate of fading environments.

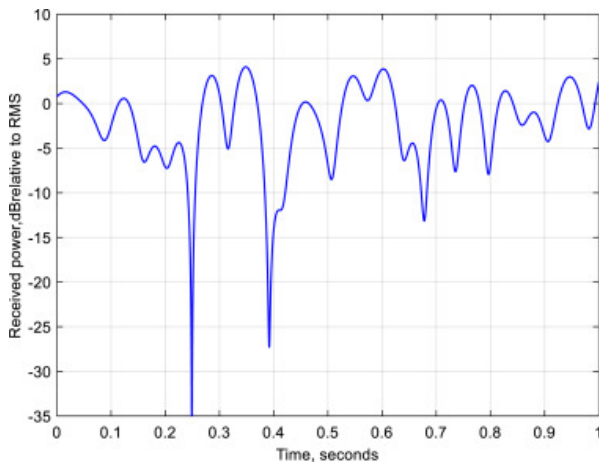
In the environment of proposed multimedia mobile communications for medical applications, multiple copies of the transmitted signal are received with different delays and phases at the receiver. This creates the phenomenon of multi-path, resulting in a random frequency modulation on each of its multi-path components due to the Doppler shifts. Hence, the resultant received signal may suffer from severe attenuation and interference that can lead to errors at the receiver and system performance degradation.

The variation in the fading channel frequency response with time due to the Doppler shift in fading channels is called a doubly dispersive fading channel. The Doppler shift calculations are given by:

$$f_d = (\Delta u / C) f_C \quad (12)$$

where f_d is the Doppler shift frequency, Δu is the velocity difference between the transmitter and the receiver, C is the speed of light, and f_C is the signal carrier frequency.

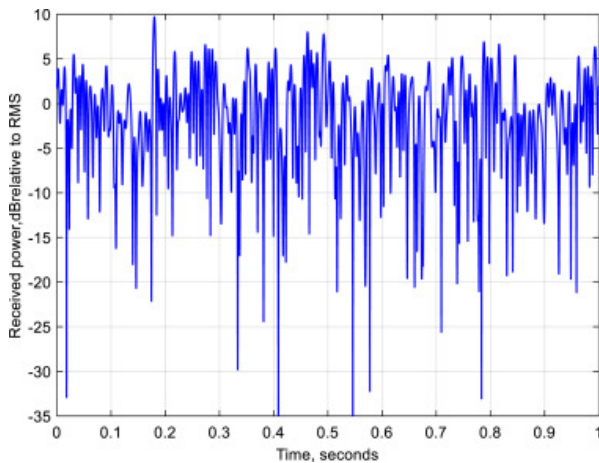
Fig. 6, Fig. 7 show the Rayleigh fading channel mutual shape for two different maximum Doppler shift frequencies of 10Hz and 100Hz, respectively. These Doppler shifts are measured respectively for 6 km/h and 60 km/h velocities at 1800 MHz, which is one of the working frequencies for GSM mobile networks. The intensity of the signal can fall by several thousand factors or 30–40dB in some “deep fades”.



[Download : Download high-res image \(204KB\)](#)

[Download : Download full-size image](#)

Fig. 6. Rayleigh fading channel corresponding to a Doppler shift of 10Hz.



[Download : Download high-res image \(446KB\)](#)

[Download : Download full-size image](#)

Fig. 7. Rayleigh fading channel corresponding to a Doppler shift of 100Hz.

2.3. Communication channel equalisation

The majority of errors at the receiver side are caused by the channel distortion, whilst the most effective method to compensate for this channel distortion effect, in order to recover the original signal's shape, is the equalisation, as shown, for example, in Fig.8[32]. The most fundamental method used by the equalisation is to select the correct receiver's filter to compensate for the selectivity of the radio channel frequency completely. This could be accomplished by choosing the receiver's filter impulse response that satisfies the relation in (13):

$$\mathbf{W} \otimes \mathbf{h} = \mathbf{1} \quad (13)$$

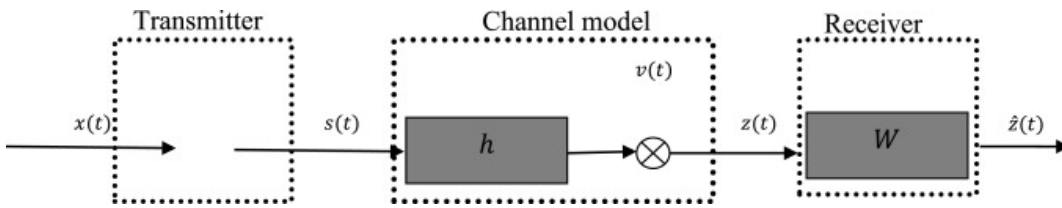
where \mathbf{W} is the equaliser impulse response, \mathbf{h} is the channel impulse response, and " \otimes " represents the linear convolution. This algorithm of equalisation is called "zero-forcing (ZF) equalisation" and, according to [32], [33], [34], ZF can provide the complete removal of any frequency selectivity in the radio channel. As a result, destruction-free and corruption-free signals can be achieved. However, ZF equalisation may become a great source of noise amplification that occurs after the filtering process. This may have an undesirable effect in that it may cause severe degradation in the system's performance.

An alternative suggestion for the ZF equalisation is to build a filter that provides a compromise between the noise/interference level and the signal distortion level based on the level of the radio-channel frequency selectivity. This might be accomplished by the MMSE equaliser that selects the filter to minimise the mean-square error (ϵ) between the transmitted signal and the equaliser output:

$$\epsilon = E\left\{|\hat{z}(t) - z(t)|^2\right\} \quad (14)$$

where $\hat{d}(t)$ is the estimated signal, and $d(t)$ is the actual transmitted signal.

It was proved that using a single carrier modulation system in frequency fading channel is inefficient due to the time equalisation complexity. On the other hand, the OFDM systems give an instant solution to this problem using a single tap equaliser in the frequency domain. As the OFDM cannot deal with doubly dispersive channels, there is a motivation to search for other bases that can match the channel frequency variation with time like DFrFT.



[Download : Download high-res image \(146KB\)](#)

[Download : Download full-size image](#)

Fig. 8. General time-domain linear equaliser.

3. Proposed method

Fig.9 shows the OFDM system data flow, $\mathbf{x}_n = [x_0, x_1 \dots x_{N_a-1}]^T$ is the data vector transmitted in the n th OFDM symbol, whilst its samples are permuted by the binary matrix $\mathbf{P} \in \mathbb{Z}^{N \times N_a}$ in the frequency domain, which allocates a data vector $\mathbf{x}_n \in \mathbb{C}^{N_a}$ to N subcarriers, with only N_a active:

$$\mathbf{P} = [0_{N_a \times (N-N_a)/2} \mathbf{I}_{N_a} 0_{N_a \times (N-N_a)/2}] \quad (15)$$

\mathbf{I}_{N_a} is an identity matrix with $N_a \times N_a$ dimensions. The vector $\mathbf{s}_n = [s_0 s_1 \dots s_N]^T$ is calculated from:

$$\mathbf{s}_n = \mathbf{F}^H \mathbf{P} \mathbf{x}_n \quad (16)$$

where \mathbf{F}^H is the N -point IDFT matrix.

The time and frequency fading channel can be demonstrated by the time-variant discrete impulse response : $h(n, u)$, where n is the time instant, and u is the time delay. The justification of this model with further details can be found in [1], [35], [36] that could be stated in the formula of (time-variant, or circular) convolution matrix by:

$$[\mathbf{H}]_{n,u} := h(n, \langle n - u \rangle_N) \quad (17)$$

Supposing the causal channel and the maximum delay spread N_h were shorter than the CP $N_h \leq L$; after removing the CP, the n th OFDM received symbol can be specified by:

$$\mathbf{z}_n = \mathbf{H}_n \mathbf{x}_n + \mathbf{v}_n \quad (18)$$

where \mathbf{v}_n are the samples of the additive white Gaussian noise (AWGN) with variance of σ^2 . In static setting, \mathbf{H}_n is circulant and the DFT matrix can be decoupled. The received subcarriers are demodulated by DFT:

$$\mathbf{y} = \mathbf{F} \mathbf{z}_n \quad (19)$$

where \mathbf{F} is the DFT matrix. The equaliser matrix $\mathbf{W}_n \in \mathbb{C}^{N_a \times N_a}$ deals with the input:

$$\tilde{\mathbf{z}}_n = \mathbf{P}^H \mathbf{F} \mathbf{H}_n \mathbf{F}^H \mathbf{P} \mathbf{x}_n + \mathbf{P}^H \mathbf{F} \mathbf{v} = \mathbf{U}_n \mathbf{x}_n + \tilde{\mathbf{v}}_n \quad (20)$$

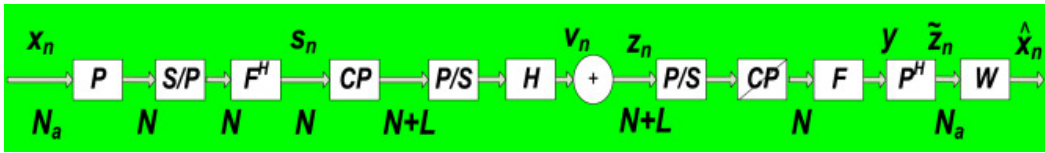
with a system matrix $\mathbf{U}_n \in \mathbb{C}^{N_a \times N_a}$, where $\mathbf{U}_n = \mathbf{P}^H \mathbf{F} \mathbf{H}_n \mathbf{F}^H \mathbf{P}$. \mathbf{P} is a binary matrix used to remove the components that may appear in the lower left and upper right corners in \mathbf{U}_n [37], and to support reducing the out-of-band emissions. The estimated data vector after using equaliser is specified by:

$$\hat{\mathbf{x}}_n = \mathbf{W} \tilde{\mathbf{z}}_n \quad (21)$$

It is easy to show that $[\tilde{\mathbf{H}}]_{m,k} = \tilde{h}(m-k, k)$, where

$$\tilde{h}(m, k) = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{u=0}^{N-1} h(n, u) e^{-j2\pi(uk+mn)/N} \quad (22)$$

From (22), it is clear that $\{\tilde{h}(0, :)\}$ is on the main diagonal of $[\tilde{\mathbf{H}}]_{m,k}$, $\{\tilde{h}(1, :)\}$ and $\{\tilde{h}(-1, :)\}$ is on the first sub-diagonal and the first super-diagonal respectively. It is obvious that $\tilde{h}(m, k)$ is the response of the frequency-domain, at subcarrier $k+m$, to a frequency-domain impulse centred at subcarrier k . In $\tilde{h}(m, k)$, k is the frequency index and m known as Doppler index. In $h(n, u)$, n is known as the time index and u as the lag index.



Download : [Download high-res image \(157KB\)](#)

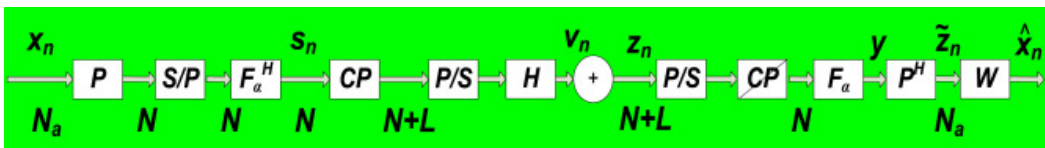
Download : [Download full-size image](#)

Fig. 9. OFDM data flow block diagram.

The DFrFT-OCDM system [data flow diagram](#) is shown in Fig. 10, which illustrates the difference from the OFDM system by using inverse fractional Fourier transform (IDFrFT) and the DFrFT for modulation and demodulation respectively. Using identical sequences for the data vector in the transmitter and the receiver, it can be shown that the equaliser $\mathbf{W}_n \in \mathbb{C}^{N_a \times N_a}$ function is to estimate the transmitted data using Eq.(20):

$$\tilde{\mathbf{z}}_n = \mathbf{P}^H \mathbf{F}_\alpha \mathbf{H}_n \mathbf{F}_{-\alpha} \mathbf{P} \mathbf{x}_n + \mathbf{P}^H \mathbf{F}_\alpha \mathbf{v} = \mathbf{U}_{n,\alpha} \mathbf{x}_n + \tilde{\mathbf{v}}_n \quad (23)$$

where \mathbf{F}_α and $\mathbf{F}_{-\alpha}$ represent the DFrFT matrix and the IDFrFT matrix respectively, with fractional angle $= \alpha$. The channel matrix and the noise vector in the [fractional domain](#) are given by $\tilde{\mathbf{H}}_\alpha = \mathbf{F}_\alpha \mathbf{H} \mathbf{F}_{-\alpha}$ and $\tilde{\mathbf{v}} = \mathbf{F}_\alpha \mathbf{v}$, respectively.



Download : [Download high-res image \(155KB\)](#)

Download : [Download full-size image](#)

Fig. 10. DFrFT-OCDM System data flow block diagram.

$\tilde{\mathbf{H}}$ and $\tilde{\mathbf{H}}_\alpha$ introduce ICI because they are non-diagonal subcarrier channel matrices, which is the case in a doubly dispersive fading channel, accordingly, the process of the symbol estimation will be complicated and as a result, it is necessary to use a complex equaliser.

3.1. Zero forcing and MMSE block equalisers

The ZF and MMSE equalisers can estimate the transmitted data by minimising $E\{\|\mathbf{x}_n - \mathbf{W}\tilde{\mathbf{z}}_n\|\}$ [37]:

$$\hat{\mathbf{x}}_{ZF} = \tilde{\mathbf{H}}_\alpha^+ \tilde{\mathbf{z}}_n = \tilde{\mathbf{H}}_\alpha^H \left(\tilde{\mathbf{H}}_\alpha \tilde{\mathbf{H}}_\alpha^H \right)^{-1} \tilde{\mathbf{z}}_n \quad (24)$$

$$\hat{\mathbf{x}}_{MMSE} = \tilde{\mathbf{H}}_\alpha^H \left(\tilde{\mathbf{H}}_\alpha \tilde{\mathbf{H}}_\alpha^H + \gamma^{-1} \mathbf{I}_{N_a} \right)^{-1} \tilde{\mathbf{z}}_n \quad (25)$$

when $\alpha = \pi / 2$, the fractional domain channel matrix $\tilde{\mathbf{H}}_\alpha$ can be reduced to the frequency domain channel matrix $\tilde{\mathbf{H}}$, $\hat{\mathbf{x}}_{ZF}$, and $\hat{\mathbf{x}}_{MMSE}$ is the estimated data using the ZF and the MMSE equalisers respectively, γ is the signal-to-noise ratio (SNR), and $\tilde{\mathbf{H}}_\alpha^+$ is the fractional domain Moore–Penrose pseudo-inverse of the channel matrix[38]. In (24), (25), complete information of the channel matrix \mathbf{H}_α is presumed thanks to the channel estimation, even when the guard subcarriers are not used by the equaliser. Moreover, it is presumed that:

$$E\{\mathbf{x}_n\} = E\{\tilde{\mathbf{v}}_n\} = 0, E\{\mathbf{x}_n \mathbf{x}_n^H\} = \mathbf{I}, E\{\mathbf{d}_n \tilde{\mathbf{v}}_n^H\} = 0, \text{ and } E\{\tilde{\mathbf{v}}_n \tilde{\mathbf{v}}_n^H\} = \sigma^2 \mathbf{I}$$

The ZF equaliser enhances the noise so its performance is poor, whilst the performance of the MMSE equaliser is the best in all linear equalisers[5]. However, it is the most complicated because it needs channel matrix inversion that involves $\mathcal{O}(N_a^3)$ complex processes[39]. For high values of N_a like DVB-T, DVB-H and WiMAX, it is not practical.

3.2. Main contribution

Reduced complexity MMSE equalisers are proposed in[8], [22], [25], [37], [40], [41], [42], [43]. In[37], a sequential MMSE equaliser is suggested and banded equalisers were presented in[8]. As identified in[37], a nearly-banded channel matrix produced in the frequency and fractional domains under doubly dispersive channels based on these conditions adapting LDL^H factorisation, can reduce the MMSE equaliser complexity[8], [20]. All the low-complexity equalisers are itemised for the OFDM systems alone; this is not the case for the DFrFT-OCDM systems.

LDL^H factorisation equaliser (Linear equaliser) was proposed in[42] as a low-complexity equaliser for the OFDM systems, benefiting from the banded properties of the frequency domain channel matrix $\tilde{\mathbf{H}}$. In like manner, the DFrFT-OCDM systems can use the LDL^H factorisation equaliser because the system matrix in the fractional domain is almost banded more than the system matrix in the frequency domain[3].

The calculation of the equaliser matrix \mathbf{W}_n is restricted to the first Q sub- and super-diagonals of $\tilde{\mathbf{H}}_\alpha$ by applying the binary masking matrix \mathbf{M} with elements:

$$\mathbf{M}(m, n) = \begin{cases} 1 & 0 \leq |m - n| \leq Q \\ 0 & Q < |m - n| < N_a \end{cases} \quad (26)$$

where the masked matrix:

$$\mathbf{B}_n = \mathbf{M} \odot \tilde{\mathbf{H}}_\alpha \quad (27)$$

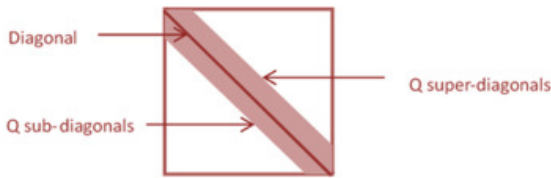
which is shown in Fig. 11, where \odot denotes to the element-wise multiplication. The MMSE equaliser can be defined according to[42] as:

$$\mathbf{W}_{n,MMSE} = \mathbf{B}_n^H (\mathbf{B}_n \mathbf{B}_n^H + \gamma^{-1} \mathbf{I}_{N_a})^{-1} \quad (28)$$

where \mathbf{B}_n is a banded matrix with Q off-diagonal terms below and above the diagonal, which corresponds to a band structure for $(\mathbf{B}_n \mathbf{B}_n^H)$, only the first $2Q$ off-diagonal terms above and below the diagonal enclosed elements can reduce the calculation complexity of the MMSE equaliser in (28). In the meantime, (28) is regarded as time-dependent where $\hat{\mathbf{x}}_n = \mathbf{W}_{n,MMSE} \tilde{\mathbf{z}}_n$ can be deduced with no need for explicitly determining $\mathbf{W}_{n,MMSE}$.

The LDL^H factorisation of the Hermitian band matrix $\mathbf{B}_n \mathbf{B}_n^H + \gamma^{-1} \mathbf{I}_{N_a} = \mathbf{LDL}^H$ can then be directly calculated[39], reaching to:

$$\hat{\mathbf{x}}_n = \mathbf{B}_n^H (\mathbf{LDL}^H)^{-1} \tilde{\mathbf{z}}_n = \mathbf{B}_n^H \mathbf{d}_n \quad (29)$$



Download : [Download high-res image \(57KB\)](#)

Download : [Download full-size image](#)

Fig. 11. The desired structure of the band matrix B inside the whole matrix \tilde{H}_α .

As an alternative option to calculate the inverse in (29), the system can be resolved by forwarding the substitution to obtain $\mathbf{d}_{2,n}$ via the lower left triangular matrix \mathbf{L} and a rescaling by the diagonal matrix \mathbf{D}^{-1} to calculate $\mathbf{d}_{1,n}$. Finally, back substitution with the upper right triangular \mathbf{L}^H yields \mathbf{x}_n , which can be inserted into (29) to determine $\hat{\mathbf{d}}_n$:

$$(\mathbf{L}\mathbf{D}\mathbf{L}^H)^{-1}\tilde{\mathbf{z}}_n = \mathbf{d}_n \quad (30)$$

$$\tilde{\mathbf{z}}_n = (\mathbf{L}\mathbf{D}\mathbf{L}^H)\mathbf{d}_n \quad (31)$$

$$\tilde{\mathbf{z}}_n = \underbrace{\mathbf{L}\mathbf{D}\mathbf{L}^H}_{\mathbf{d}_{2,n}} \mathbf{d}_n \quad (32)$$

The overall complexity for obtaining $\hat{\mathbf{x}}_n$ is $(8Q^2 + 22Q + 4)N_a$ complex operations [42]. The choice of the parameter Q is a trade-off between performance and sophistication. So, for example, a larger Q produces a slight estimation error, resulting in performance enhancement. On the other hand, the calculations' complexity increases as a consequence of the higher bandwidth of \mathbf{B} .

4. Results and discussion

In the following channel environments, the performance of uncoded bit error rate (BER) for the conventional OFDM and DFrFT-OCDM systems is studied:

- 1- Time-invariant channel.
- 2- Time-variant channel.

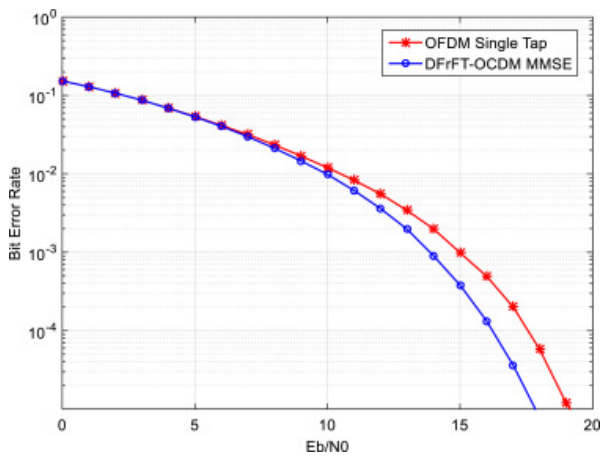
The QPSK modulated OFDM system under investigation has the following parameters:

$L = 8$, $N = 128$, and $N_a = 96$. The communication channel simulated in this proposed work is the Rayleigh fading channel that has an exponential power delay profile and a root-mean-square delay spread of 3. The adopted carrier frequency is chosen to be ultra-high frequency based on the suggested application investigated in this paper, therefore, the subcarrier spacing is $\Delta f = 20$ kHz and $f_C = 10$ GHz. This Doppler frequency corresponds to a high mobile speed $\mathbf{V} = 324$ Km/h. Simulation is carried over 10^5 continuous channels and different OFDM symbols, which means $10^5 * 96 * 4$ data bits.

4.1. Time-invariant channel

Doppler frequency is equal to zero in the time invariant channel environment ($f_d = 0$). The OFDM system uses the single tap equaliser, and the DFrFT-OCDM system uses the MMSE equaliser. Fig. 13 shows the BER performance for both systems. The OFDM system performance is compared to the work carried out in [42], and it was found to match.

From Fig. 12, although the DFrFT-OCDM system has a superior performance, the OFDM system with the single tap equaliser has a very competitive performance with much less complexity. As a result, there is a recommendation to use the OFDM in the time-invariant fading channel scenarios.



[Download : Download high-res image \(359KB\)](#)

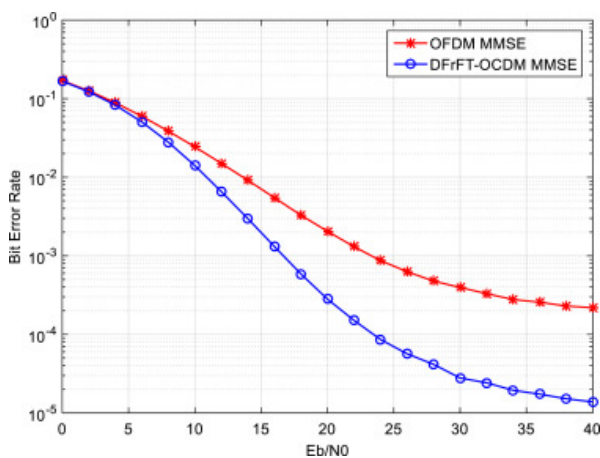
[Download : Download full-size image](#)

Fig. 12. OFDM and DFrFT-OCDM BER comparison in time invariant channel environment.

4.2. Time-variant channel

In the time-variant channel environment, we consider the maximum Doppler frequency $f_d = 0.15\Delta f$. The MMSE equaliser was used for the OFDM and the DFrFT-OCDM system.

From Fig. 13, the DFrFT-OCDM system has a superior performance when compared to the OFDM system with the same MMSE equaliser, and with the same complexity. As a result, the DFrFT-OCDM is regarded as a better choice in time-variant fading channel scenarios.



[Download : Download high-res image \(381KB\)](#)

[Download : Download full-size image](#)

Fig. 13. OFDM and DFrFT-OCDM BER comparison in time-variant channel environment.

4.3. LDL^H factorisation equaliser simulation

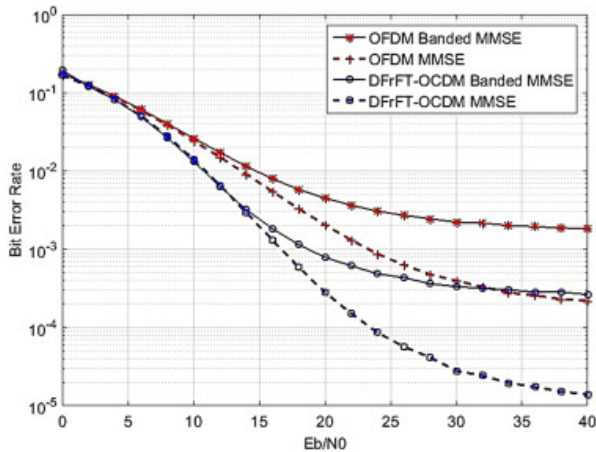
To measure the performance of the proposed OFDM system, simulation was run for an OFDM transmission with $N = 128$ subcarriers, where only $N_a = 96$ are active subcarriers, with CP of length $L = 8$, and QPSK modulation. The channel model is the same as the one used in [8], which adapts an exponential power delay profile with an RMS delay spread of 3 sampling periods, with a maximum Doppler spread f_d equal to 15% of the carrier spacing over a group of 10^5 Rayleigh fading channels.

A comparison between the block MMSE equaliser and the LDL^H low-complexity equaliser (Banded MMSE) is shown in Fig. 14 for the OFDM system and the DFrFT-OCDM system ($\alpha = 0.2\pi / 2$). The low-complexity equaliser

functions with $Q = [5, 96]$, where 96 corresponds to the regular block MMSE equaliser. Performance results are shown in Fig. 14 in terms of BER. The OFDM curves correspond with those stated in [42]. The DFrFT-OCDM system with $Q = 5$ shows a slight degradation over the full MMSE OFDM system at high SNR, but needs only 3.4% [42] of the calculation's cost in terms of complex operations, and still outperforms OFDM.

To investigate and determine the masking level Q 's influence, the power components in $\widetilde{\mathbf{H}}_\alpha$ and the after-masking process by \mathbf{M} reduced to \mathbf{B}_n , should be compared. We take into consideration that the ensemble trace operator $\text{tr}\{\cdot\}$'s average power ratio is given in the following relation:

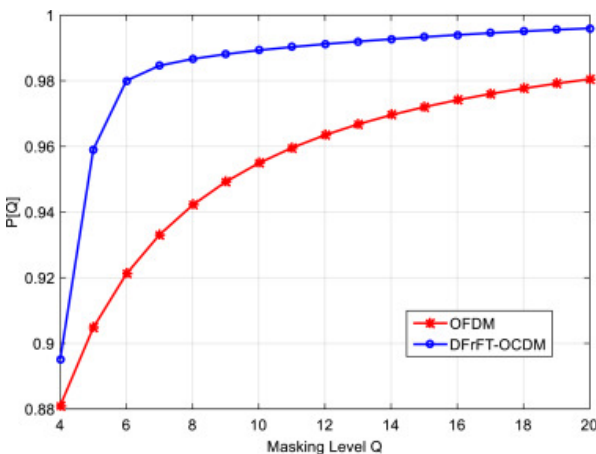
$$\rho[Q] = \varepsilon \left\{ \frac{\text{tr}\{\mathbf{B}_n \mathbf{B}_n^H\}}{\text{tr}\{\widetilde{\mathbf{H}}_\alpha \widetilde{\mathbf{H}}_\alpha^H\}} \right\}, \quad 0 \leq \rho[Q] \leq 1 \quad (33)$$



Download : [Download high-res image \(374KB\)](#)

Download : [Download full-size image](#)

Fig. 14. BER for MMSE equalisation with a block of $Q = N_\alpha$ (96), and low-complexity at $Q = 5$ approaches for DFrFT-OCDM at $\alpha = 0.2\pi / 2$ and OFDM.



Download : [Download high-res image \(190KB\)](#)

Download : [Download full-size image](#)

Fig. 15. Comparing OFDM and DFrFT-OCDM at $\alpha = 0.2\pi / 2$ for different percentage values of $\widetilde{\mathbf{H}}_\alpha$ power restricted in \mathbf{B}_n , determined by $[Q]$ that depends on the number of off-diagonal elements Q that are measured by \mathbf{M} .

Fig. 15 clearly illustrates that OFDM suffers from the effect of Doppler fading that causes energy to spread away from the main diagonal. The results show that the spread of energy is not even limited to nearby off-diagonals, which makes it essential to imply a high value of Q to collect a large amount of the power contained in $\widetilde{\mathbf{H}}$. A similar effect

of Doppler fading can be noted with the DFrFT-OCM because of its inability to diagonalise $\tilde{\mathbf{H}}_{\alpha}$. However, unlike OFDM, DFrFT-OCM's leaked power exists adjacent to the off-diagonal element. As a result, a much lower value of Q can be applied to collect the required power of the components of $\tilde{\mathbf{H}}_{\alpha}$ in \mathbf{B}_n , which justifies the performance improvement achieved with even less difficulty, making the proposal of this system for mobile medical application with machine learning perspectives preferable.

4.4. Deep computing perspectives

Although various concepts of machine learning and big data applications are explored in mobile health, there has been little attention paid to the usage of machine learning and optimisation techniques for coding at the physical layer. At present, the machine learning and deep learning fields have overcome issues concerning compression of neural networks and neural auto-encoders. Consequently, there is adequate opportunity for applying such techniques in the mobile health field; indeed, it could be used in encoding the data prior to broadcasting and then decoding it after passing through the channel. This opportunity is regarded as essential in the sense that a robust and compressed neural net is needed. At the same time, such a model must be able to decrease the load of transmitted data through its auto-encoding structure. Other opportunities consist of the application's design using intelligent models that are specialised for specific medical tasks. They can also communicate through the channel both efficiently and in a compressed manner.

There are various methods that can apply machine learning models for the specific tasks of medical images or video processing [43]. Such models can be quickly and easily learned through different optimisation techniques, such as greedy growing and pruning for trees. These can be easily used for fast auto-encoder tasks because they are compact and simple to interpret. Other types of models, such as neural nets, can be learned through some gradient-based methods. However, for neural nets, due to the high complexity of the model, the issue of compressing and decreasing inference complexity must be tackled for specific tasks of data transmitting. As an extension to the current work, one can further design a model that is suitable for the previously mentioned tasks whilst achieving optimal efficiency in data transmission, encoding, and bandwidth usage.

5. Conclusions

In this paper, a mobile medical video streaming broadcasting system was proposed. The time and frequency fading channel with its effects on OFDM system performance were investigated. The DFrFT-OCM MCM system was studied as an alternative MCM system that can enhance the overall MCM system performance. It was demonstrated that using simple equalisers with MCM systems was in high demand within the medical video broadcasting system because of its large symbol length, despite its simplicity. The DFrFT-OCM was found to be a good alternative for the OFDM in a doubly dispersive channel environment, dependent upon changing the traditional OFDM basis with a chirp basis using the DFrFT, which can cope with the channel variations.

Low-complexity equalisers based on LDL^H factorisation were proposed with the DFrFT-OCM system, and it was demonstrated that this new combination shows improved performance when compared to OFDM using the same equalisers. This justifies the reason this system is recommended to be applied within mobile medical video broadcasting. Future work will be implemented to incorporate this proposed system with other techniques that could reduce the complexity or the power consumption, such as searching for new low-complexity equalisers, searching for alternative bases that can improve the MCM system's performance under doubly dispersive fading channels in other applications like social media, introducing Network Coding in [15], [44], [45], [46] with MCM systems to improve the overall system performance, or with searching for the optimum number of paths for realisation of multi-path routing as in [47].

CRedit authorship contribution statement

Hani H. Attar: Conceptualization, Methodology, Writing - original draft. **Ahmad A.A. Solyman:** Conceptualization, Methodology, Writing - original draft, Investigation, Supervision. **Abd-Elnaser Fawzy Mohamed:** Investigation, Visualization. **Mohammad R. Khosravi:** Resources, Supervision, Writing - review & editing. **Varun G. Menon:**

Software, Resources, Writing - review & editing. **Ali Kashif Bashir:** Software, Resources, Writing - review & editing.
Pooya Tavallali: Resources, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

[Special issue articles](#) [Recommended articles](#)

References

- [1] Tiejun W., *et al.*
Performance degradation of OFDM systems due to doppler spreading
IEEE Trans. Wirel. Commun., 5 (2006), pp. 1422-1432
[Google Scholar](#) ↗
- [2] Q. Huang, *et al.* A novel OFDM equalizer for large doppler shift channel through deep learning, in: 2019 IEEE 90th Vehicular Technology Conference, VTC2019-Fall, 2019, pp. 1–5.
[Google Scholar](#) ↗
- [3] Martone M.
A multicarrier system based on the fractional fourier transform for time-frequency-selective channels
IEEE Trans. Commun., 49 (2001), pp. 1011-1020
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [4] Nassiri M., Baghersalimi G.
Comparative performance assessment between FFT-based and FRFT-based MIMO-OFDM systems in underwater acoustic communications
IET Commun., 12 (2018), pp. 719-726
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [5] Yang-Seok C., *et al.*
On channel estimation and detection for multicarrier signals in fast and selective Rayleigh fading channels
IEEE Trans. Commun., 49 (2001), pp. 1375-1387
[Google Scholar](#) ↗
- [6] G. Taubock, *et al.* LSQR-based ICI equalization for multicarrier communications in strongly dispersive and highly mobile environments, in: Signal Processing Advances in Wireless Communications, 2007. SPAWC 2007. IEEE 8th Workshop on, 2007, pp. 1–5.
[Google Scholar](#) ↗
- [7] L. Guanghui, *et al.* Simple equalization of OFDM signal over doubly selective channels, in: Intelligent Signal Processing and Communication Systems (ISPACS), 2010 International Symposium on, 2010, pp. 1–4.
[Google Scholar](#) ↗
- [8] Luca Rugini P.B., Leus Geert
Low-complexity banded equalizers for OFDM systems in doppler spread channels
EURASIP J. Appl. Signal Process., 2006 (2006), p. 67404
p. 13

[Google Scholar](#) ↗

- [9] Y.E. Tan, et al. Fragility Issues of Medical Video Streaming over 802.11e-WLAN m-health Environments, in: 2006 International Conference of the IEEE Engineering in Medicine and Biology Society, 2006, pp. 6316–6319.
[Google Scholar](#) ↗
- [10] Antoniou Z.C., *et al.*
Real-time adaptation to time-varying constraints for medical video communications
IEEE J. Biomed. Health Inf., 22 (2018), pp. 1177–1188
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [11] A. Alinejad, et al. Performance analysis of medical video streaming over mobile WiMAX, in: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, 2010, pp. 3471–3474.
[Google Scholar](#) ↗
- [12] Attar H., Khosravi M.R., Igorovich S.S., Georgievan K.N., Alhihi M.
Review and performance evaluation of FIFO, PQ, CQ, FQ, and WFQ algorithms in multimedia wireless sensor networks
Int. J. Distrib. Sens. Netw. (2020), [10.1177/1550147720912950](#) ↗
[Google Scholar](#) ↗
- [13] IEEE Health informatics–point-of-care medical device communication part 10207: Domain information and service model for service-oriented point-of-care medical device communication
IEEE Std 11073-10207-2017
(2018), pp. 1-437
[Google Scholar](#) ↗
- [14] Y. Li, et al. The reliability of Bluetooth data transmission in Mobile Medical information acquisition system, in: 2016 IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference, IMCEC, 2016, pp. 498–505.
[Google Scholar](#) ↗
- [15] Attar H., *et al.*
Cooperative network-coding system for wireless sensor networks
IET Commun., 6 (2012), pp. 344–352
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [16] Farrar C.R., Worden K.
Structural Health Monitoring: A Machine Learning Perspective
John Wiley & Sons (2012)
[Google Scholar](#) ↗
- [17] Dwyer D.B., *et al.*
Machine learning approaches for clinical psychology and psychiatry
Annu. Rev. Clin. Psychol., 14 (2018), pp. 91–118
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [18] Finkelstein J., Cheol Jeong I.
Machine learning approaches to personalize early prediction of asthma exacerbations
Ann. New York Acad. Sci., 1387 (2017), p. 153
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [19] Attar H., *et al.*

E-health communication system with multiservice data traffic evaluation based on a $G / G / 1$ analysis method

Curr. Signal Transduction Therapy, 15 (2020)

[Google Scholar](#) ↗

- [20] Golub G.H., Van Loan C.F.
Matrix Computations
(third ed.), Johns Hopkins Univ (1996)

[Google Scholar](#) ↗

- [21] P. Robertson, S. Kaiser, The effects of Doppler spreads in OFDM(A) mobile radio systems, in: Vehicular Technology Conference, 1999. VTC 1999 - Fall. IEEE VTS 50th, vol. 1, 1999, pp. 329–333.

[Google Scholar](#) ↗

- [22] L. Rugini, P. Banelli, Performance analysis of banded equalizers for OFDM systems in time-varying channels, in: Signal Processing Advances in Wireless Communications, 2007. SPAWC 2007. IEEE 8th Workshop on, 2007, pp. 1–5.

[Google Scholar](#) ↗

- [23] Burchill W., Leung C.
Matched filter bound for OFDM on Rayleigh fading channels
Electron. Lett., 31 (1995), pp. 1716-1717

[View in Scopus](#) ↗ [Google Scholar](#) ↗

- [24] Han J., *et al.*
Low-complexity equalization of orthogonal signal-division multiplexing in doubly-selective channels
IEEE Trans. Signal Process., 67 (2019), pp. 915-929

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

- [25] Dogan H., *et al.*
Low-complexity joint data detection and channel equalisation for highly mobile orthogonal frequency division multiplexing systems
IET Commun., 4 (2010), pp. 1000-1011

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

- [26] Saxena R., Singh K.
Fractional Fourier transform: A novel tool for signal processing
J. Indian Inst. Sci., 58 (2005), pp. 11-26

[View in Scopus](#) ↗ [Google Scholar](#) ↗

- [27] Namias V.
The fractional order Fourier transform and its application to quantum mechanics
IMA J. Appl. Math., 25 (1980), pp. 241-265

[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

- [28] Almeida L.B.
The fractional Fourier transform and time-frequency representations
IEEE Trans. Signal Process., 42 (1994), pp. 3084-3091

[View in Scopus](#) ↗ [Google Scholar](#) ↗

- [29] Bultheel A., Martínez Sulbara H.E.
Computation of the fractional Fourier transform

(2004)

[Google Scholar](#) ↗

[30] Candan C., *et al.*

The discrete fractional fourier transform

IEEE Trans. Signal Process., 48 (2000), pp. 1329-1337

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[31] Ozaktas H.M., *et al.*

Digital computation of the fractional Fourier transform

IEEE Trans. Signal Process., 44 (1996), pp. 2141-2150

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[32] Simon M.K., Alouini M.-S.

Digital Communication over Fading Channels, vol. 86

Wiley-IEEE Press (2004)

[Google Scholar](#) ↗

[33] RAPPAPORT

Wireless Communication Systems

(1996)

[Google Scholar](#) ↗

[34] Wireless Communications

John Wiley & Sons, Ltd (2005)

[Google Scholar](#) ↗

[35] Xiaodong C., Giannakis G.B.

Bounding performance and suppressing intercarrier interference in wireless mobile OFDM

IEEE Trans. Commun., 51 (2003), pp. 2047-2056

[Google Scholar](#) ↗

[36] Ye L., Cimini Jr. L.J.

Bounds on the interchannel interference of OFDM in time-varying impairments

IEEE Trans. Commun., 49 (2001), pp. 401-404

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[37] Schniter P.

Low-complexity equalization of OFDM in doubly selective channels

IEEE Trans. Signal Process., 52 (2004), pp. 1002-1011

[View in Scopus](#) ↗ [Google Scholar](#) ↗

[38] A.A.A. Solyman, *et al.* Low-complexity LSMR equalisation of FrFT-based multicarrier systems in doubly dispersive channels, in: Signal Processing and Information Technology (ISSPIT), 2011 IEEE International Symposium on, 2011, pp. 461–465.

[Google Scholar](#) ↗

[39] Golub G.H., Van Loan C.F.

Matrix Computations

Johns Hopkins University Press (1996)

[Google Scholar](#) ↗

- [40] S. Ahmed, et al. Low complexity iterative method of equalization for ofdm in doubly selective channels, in: Signals, Systems and Computers, 2005. Conference Record of the Thirty-Ninth Asilomar Conference on, 2005, pp. 687–691.
[Google Scholar](#) ↗
- [41] T. Hrycak, G. Matz, Low-complexity time-domain ICI equalization for ofdm communications over rapidly varying channels, in: Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference on, 2006, pp. 1767–1771.
[Google Scholar](#) ↗
- [42] Rugini L., et al.
Simple equalization of time-varying channels for OFDM
IEEE Commun. Lett., 9 (2005), pp. 619-621
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [43] Solyman A.A.A., et al.
A low-complexity equalizer for video broadcasting in cyber-physical social systems through handheld mobile devices
IEEE Access, 8 (2020), pp. 67591-67602
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [44] Nazir S., et al.
Relay-assisted rateless layered multiple description video delivery
IEEE J. Sel. Areas Commun., 31 (2013), pp. 1629-1637
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [45] H. Attar, et al. Deterministic network coding over long term evaluation advance communication system, in: 2014 Fourth International Conference on Digital Information and Communication Technology and its Applications, DICTAP, 2014, pp. 56–61.
[Google Scholar](#) ↗
- [46] Alhihi M., et al.
Network Coding Cooperation Performance Analysis in Wireless Network over a Lossy Channel, M Users and a Destination Scenario
(2016)
[Google Scholar](#) ↗
- [47] Alhihi M., et al.
Determining the optimum number of paths for realization of multi-path routing in MPLS-TE networks
TELKOMNIKA Indonesian J. Electr. Eng., 15 (2017), pp. 1701-1709
[CrossRef](#) ↗ [View in Scopus](#) ↗ [Google Scholar](#) ↗

Cited by (16)

[EMMM: Energy-efficient mobility management model for context-aware transactions over mobile communication](#)

2021, Sustainable Computing: Informatics and Systems

Citation Excerpt :

...The geographical region is partitioned in small service area to efficient use of frequency, which is known as a cell. Fixed host (FH), base station (BS), mobile support station (MSS) and mobile host (MH) are the pillar of mobile communication [1,2]. FH used as permanent shareable data repositories and linked via high speed wired network infrastructure....

[Show abstract](#) 

OPTIMIZED MIMO BASED ENHANCED OFDM FOR MULTI CARRIER SYSTEM WITH 5G WAVEFORMS

2023, Economic Computation and Economic Cybernetics Studies and Research

Single-Frequency Network Terrestrial Broadcasting with 5G NR Numerology Using Recurrent Neural Network

2022, Electronics (Switzerland)

Mobile multimedia computing in cyber-physical surveillance services through UAV-borne Video-SAR: A taxonomy of intelligent data processing for IoMT-enabled radar sensor networks

2022, Tsinghua Science and Technology

Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to eHealth and Patient Data Monitoring

2022, Anomaly Detection and Complex Event Processing Over IoT Data Streams: With Application to eHealth and Patient Data Monitoring

Distributed and Big Health Data Processing for Remote and Ubiquitous Healthcare Services Using Blind Statistical Computing: Review and Trends on Blindness for Internet of Artificially Intelligent Medical Things

2022, Intelligent Healthcare: Infrastructure, Algorithms and Management



[View all citing articles on Scopus](#)



Hani H. Attar received his Ph.D. from the Department of Electrical and Electronic Engineering, University of Strathclyde, United Kingdom in 2011. Since 2011, he has been working as a researcher of electrical engineering and energy systems. Dr Attar is now a university lecturer at Zarqa University, Jordan. His research interests include Network Coding, Wireless Sensor Networks, and Wireless Communications.



Ahmad A. A. Solyman graduated from the University of Strathclyde, United Kingdom in 2013. His Ph.D. researches lie in Multimedia Services over Wireless Networks Using OFDM. He is currently a lecturer at the Department of Electrical and Electronics Engineering, Istanbul Gelisim University, Turkey. His research interests contain Wireless Communication Networks, and MIMO Communication Systems.



Abd-Elnaser Fawzy Mohamed is currently a researcher with Bilbeis Higher Institute for Engineering, Egypt. He received his Ph.D. in the area of Electrical and Computer Engineering from Cairo University, Egypt.



Mohammad R. Khosravi is with the Department of Electrical and Electronic Engineering, Shiraz University of Technology, Iran, and Department of Computer Engineering, Persian Gulf University, Iran. His main interests include statistical signal and image Processing, medical bioinformatics, radar imaging and satellite remote sensing, computer communications, industrial wireless sensor networks, underwater acoustic communications, information science and scientometrics.



Varun G. Menon is currently Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. His research interests include sensors, IoT, fog computing and underwater acoustic sensor networks. He has completed his Ph. D in Computer Science and Engineering from Sathyabama University, India in 2017. He is also a Distinguished Speaker of ACM.



Ali K. Bashir is a Senior Lecturer at the Department of Computing and Mathematics, Manchester Metropolitan University, United Kingdom. His past assignments include Associate Professor of Information and Communication Technologies, Faculty of Science and Technology, University of the Faroe Islands, Denmark; Osaka University, Japan; Nara National College of Technology, Japan; the National Fusion Research Institute, South Korea; Southern Power Company Ltd., South Korea, and the Seoul Metropolitan Government, South Korea. He received his Ph.D. in computer science and engineering from Korea University, South Korea. MS from Ajou University, South Korea and BS from University of Management and Technology, Pakistan. He is supervising/co-supervising several graduate (MS and Ph.D.) students. His research interests include internet of things, wireless networks, distributed systems, network/cyber-security, network function virtualization, etc. He has authored over 80 peer-reviewed articles. He has served as a chair (programme, publicity, and track) on top conferences and workshops. He has delivered over 20 invited and keynote talks in seven countries. He is a Distinguished Speaker, ACM; Senior Member of IEEE; Member, ACM; Member, IEEE Young Professionals; Member, International Association of Educators and Researchers, UK. He is serving as the Editor-in-chief of the IEEE FUTURE DIRECTIONS NEWSLETTER. He is advising several start-ups in the field of STEM-based education, robotics, internet of things, and Blockchain.



Pooya Tavallali received the B.Sc. and the M.Sc. degrees in Electrical Engineering (Communication Systems) from the Department of Electrical and Electronic Engineering, Shiraz University, Shiraz, Iran, in 2013 and 2016, respectively. Since 2016, he has been a Ph.D. scholar at the Department of Electrical Engineering and Computer Science, University of California, Merced, USA. His scientific interest consists of machine learning, statistical signal and image processing, neural networks, statistical pattern recognition and optimisation algorithms.

[View Abstract](#)

© 2020 Elsevier B.V. All rights reserved.



Copyright © 2023 Elsevier B.V. or its licensors or contributors.
ScienceDirect® is a registered trademark of Elsevier B.V.



Received May 7, 2020, accepted July 29, 2020, date of publication August 6, 2020, date of current version August 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014622

SafeCity: Toward Safe and Secured Data Management Design for IoT-Enabled Smart City Planning

HUI ZHANG^{1,2}, MUHAMMAD BABAR³, MUHAMMAD USMAN TARIQ³,
MIAN AHMAD JAN^{4,5}, VARUN G. MENON⁶, (Senior Member, IEEE),
AND XINGWANG LI⁷, (Senior Member, IEEE)

¹School of Energy Science and Engineering, Henan Polytechnic University, Jiaozuo 454003, China

²Coal Mining and Design Branch, China Coal Research Institute, Beijing 100013, China

³Department of Management, Abu Dhabi School of Management, Abu Dhabi, United Arab Emirates

⁴Informetrics Research Group, Ton Duc Thang University, Ho Chi Minh City, Vietnam

⁵Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam

⁶Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India

⁷School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China

Corresponding author: Mian Ahmad Jan (mianjan@tdtu.edu.vn)

This work was supported in part by the National Natural Science Foundation General Fund under Project 51874109, in part by the Key Scientific and Technological Projects in Henan Province under Grant 182102310005, in part by the Science and Technology Support Plan of Guizhou Province (Science Support of Guizhou Province) under Grant [2019] 2861, in part by the Science and Technology Project for Outing and Young Talents of Guizhou (Talents of Science Platform in Guizhou) under Grant [2019] 5674, in part by the Key Scientific Research Projects of Higher Education Institutions in Henan Province under Grant 20A510007, and in part by the Fundamental Research Funds for the Universities of Henan Province under Grant NSFRF180309.

ABSTRACT The interaction among different Internet of Things (IoT) sensors and devices become massive and insecure over the Internet as we probe to smart cities. These heterogeneous devices produce an enormous amount of data that is vulnerable to various malicious threats. The generated data need to be processed and analyzed in a secure fashion to make smart decisions. The smart urban planning is becoming a reality through the mass information generated by the Internet of Things (IoT). This paper exhibits a novel architecture, SafeCity, that limelight the ecosystem of smart cities consists of cameras, sensors, and other real-world physical devices. SafeCity is a three-layer architecture, i.e., a data security layer, a data computational layer, and a decision-making layer. At the first layer, payload-based symmetric encryption is used to secure the data from intruders by exchanging only the authentic data among the physical devices. The second layer is used for the computation of secured data. Finally, the third layer extracts visions from data. The secured exchange of data is ensured by using Raspberry Pi boards while the computation of data is tested on trustworthy datasets, using the Hadoop platform. The assessments disclose that SafeCity presents precious insights into a secured smart city in the context of sensors based IoT environment.

INDEX TERMS Internet of Things, smart city, symmetric encryption, data management design, data analytics, data mining.

I. INTRODUCTION

Currently, 55% population of the world is in the cities that are expected to grow up to 67% by the year 2050 [1], [2]. The gradual increase in the urbanization poses various encounters for the decision-makers in proposing different facilities to the inhabitants of these cities. The ICT (Information and Communication Technologies) are used to make the cities smart enough by deploying and promoting sustainable devel-

The associate editor coordinating the review of this manuscript and approving it for publication was Jesús Hamilton Ortiz.

opment practices for addressing the growing challenges of urbanization. A solid foundation is offered for the Internet of Things (IoT) with an advancement in the field of smart cities' sensors by enabling them to interconnect [3]. Technology in the shape of smartphones, sensors, and other devices is playing a pivotal role in bringing the era of ubiquitous computing. In 2017, Gartner predicted that the number of interconnected devices will increase by 31% in 2017 by getting 8.5 billion and exceeded 20+ billion by the year 2020.

The IoT-enabled environment is a pattern where the processing of information is connected with every encountered

activity [4]. A huge number of real-world physical devices in a ubiquitous environment will generate voluminous data containing a variety of information that needs new forms of computation to facilitate enhanced decision making. The vast amount of data generated by the ubiquitous devices will add veracity, value, and variability to the Internet [5]. Advancement in the ubiquitous computing is causing in a large-scale valuable data or information, and with the assistance of Big Data tools and proficient machine learning methods, there is a great potential of analytical amenities to the smart cities [6]–[9]. A number of proposals are found to process and analyze the data generated by heterogeneous devices to perform efficient decision making.

Smart city data computation and pervasive intelligence expose the networks to security attacks, malware, and other cyber breaches. The inter-connectivity requirements of everyday physical devices would probably add numerous groundbreaking and resourceful malicious prototypes to IoT data computing [10]. The presence of malicious intruders may generate fabricated data to manipulate the sensed information of legitimate devices. The intruders may adversely affect the services and decision making in a ubiquitous environment. Furthermore, these malicious entities may liftoff attacks like denial-of-service by disrupting the transmission, and sensing of a ubiquitous environment to reduce the eminence of smart services [11].

Security provisioning in a ubiquitous environment is an intricate work since every machine possesses its identifiable unique characteristics and the uniqueness to be verified when connected to the Internet. The solutions for these ubiquitous devices in the marketplace lack the secured characteristics and are exposed to an extensive kind of adversarial attacks [12]. Besides, the existing privacy-preserving and authentication algorithms for smart ubiquitous environments involve complex and resource-intensive operations that require an abundance of resources. Most of these algorithms are not suitable for delay-sensitive and priority-based traffic generated in these environments.

In this article, we propose a safe and secured data management design for smart city planning using ubiquitous computing. The key contributions of the proposed architecture are as follows.

1. Payload-based symmetric encryption is proposed for a smart ubiquitous environment that is simple, lightweight, robust, and resilient against various malicious threats. The proposed approach uses 128-bit security primitives for secured exchange of data among the real-world physical devices.
2. A customized utility is proposed for the efficient loading of secured data into Hadoop. The proposed loading utility is efficient in terms of time and storage. The default HDFS (Hadoop Distributed File System) architecture is customized to achieve effective data storage. Our customized HDFS reduces storage consumption along with the network overhead.
3. The traditional YARN (Yet Another Resource Negotiator) Hadoop definition is customized for efficient data

computation. This is accomplished by introducing the concept of dynamic scheduling into the Hadoop YARN definition.

The remaining paper is ordered as follows. In Section 2, we spotlight the existing studies. In Section 3, we spotlight our proposed SafeCity framework for an IoT sensors based environment. In Section 4, the experimental results for secured data transmission and processing are presented. Finally, the paper is concluded in Section 5.

II. LITERATURE REVIEW

In this section, first we highlight the current works about the secure transmission of ubiquitous data collected from the smart cities, followed by their processing to extract valuable features.

A. SECURED TRANSMISSION OF DATA

Over the last decade, a lot of hype has been witnessed around building the concept of smart cities. Finally, the presence of sensor-embedded Internet of Things (IoT) platforms, ubiquitous connectivity, and cloud and data analytics has turned this concept into a reality. Although cities around the globe are seeking to become smarter, the applications of smart cities face a plethora of challenges in terms of security and privacy. These applications need to secure the gathered data from unauthorized access, disruption, annihilation, modification, inspection, and various other malevolent activities. In literature, numerous studies exist to protect the voluminous data traffic of smart ubiquitous cities from malicious entities. The error-prone communication channels used by the resource-starving sensors of smart cities limit the usage of TLS (Transport Layer Security) for seamless traffic flow [13]. As a result, most of the sensor nodes in smart ubiquitous environments rely on DTLS (Datagram Transport Layer Security) for the secured transmission of their data [14]. Nonetheless, the record layers of DTLS and handshake have a collective overhead of 25 bytes in each datagram header. The DTLS needs to be stripped of the resource-intensive operations to suit the resource-starving sensor nodes of smart cities [15].

In [16], the authors proposed an extremely lightweight encryption approach for the secured establishment of a unicast communication system in smart cities. The authors claimed that their model decreases the energy consumption and computational time of the sensor nodes. However, they did not provide any experimental and analytical results to verify their claim. In [17], the authors studied the use of DTLS for secured communication in a smart ubiquitous environment. They argued that the streaming applications of smart cities require an abundance of memory space and the use of DTLS is not feasible for them. The authors emphasized the use of compressed IPSec to offer security at the network layer for streaming applications.

A robust and resilient secured scheme for ubiquitous applications of smart cities was proposed in [18]. An RSA-based DTLS implementation was used for the secured exchange of ubiquitous data. However, both the RSA and DTLS have higher computational overheads due to resource-intensive

handshake mechanisms. The presence of complex cipher suites of RSA incurs a higher energy consumption and computational overhead for the ubiquitous operation of sensor nodes. The performance of the DTLS handshake was evaluated for ubiquitous smart devices using the Elliptic Curve Cryptography (ECC) [19].

In [20], a DTLS implementation for smartphones was proposed using the Constrained Application Protocol (CoAP). The proposed scheme involves computationally difficult encryption suites, requires ample processing and power memory, and is not suitable for sensor nodes of the smart cities. In [21], a lightweight encryption approach was proposed for ubiquitous communication in a smart city environment. Prior to establishing a secured session, the proposed approach validates the identities of clients and servers. For authentication, symmetric encryption with 128-bit security primitives were used. However, the proposed scheme is not validated experimentally to verify its efficiency, robustness, and resilience.

B. DATA PROCESSING AND FEATURE EXTRACTION

In this section, the challenges and issues in the existing works for smart city planning utilizing the Big Data analytical techniques are presented. In [22], the authors designed a model to compute Big Data generated in the IoT-based smart health setting. It involves the separation of vigorous data into subclasses that are based on hypothetical simulation of data fusion to improve computational effectiveness. The key issues underlined in this model are the use of customary MapReduce Cluster management for Apache Hadoop server, insufficient data loading to Hadoop, a conceptual framework, and the utilization of only healthcare datasets.

A Big Data analytics framework comprised of various tiers was proposed for urban planning in [23]. Each tier of the framework is responsible for different activities of the Big Data analytics to have efficient modularization of the overall process. Although, it is a complete framework from data generation and collection to application and usage of the analyzed data, it causes significant delay in processing and the use of classical MapReduce deteriorates the performance [24]. Moreover, prior to data loading, the authors focused on data aggregation while overlooking the data loading competence.

An IoT-enabled framework using Hadoop-based Big Data analytics was proposed in [25] for a smart city application. The proposed framework has different layers from data acquisition to the application. The main problem of this framework is that the data loading efficiency was ignored.

A proposal based on the analysis of Big Data that endorses the perception of SCC (smart and connected societies) for smart cities was proposed in [26]. The SCC model is a conceptual framework that was not implemented. A similar model was proposed for the ubiquitous smart city application in [27]. However, this model was not implemented as well. Moreover, [26] and [27] overlooked the data loading and ingestion into a distributed ubiquitous smart city environment. In addition, many solutions have been proposed to treat similar problems of Big Data analytics in smart ubiquitous 145258

environments [28], [29]. Vecular fog computing may also be utilized for smart city planning [30]. However, a critical issue in the design of these methods is the deployment of a traditional cluster resource management scheme and insufficient data loading to the Hadoop server.

A graph-oriented architecture to analyze the Big Data in a smart ubiquitous transportation system was proposed in [31]. This graph-based solution is more scalable and efficient, but it incurs additional delay due to graph processing. In addition to processing delay, the proposed solution was tested only for the transportation dataset, and loading the Big Data to the Hadoop server and its efficiency was overlooked. The proposed architecture was tested only for a healthcare dataset. The authors proposed a multi-level data processing scheme, based on parallel processing, for Big Data analysis. However, a YARN-enabled solution was provided but the data ingestion efficacy was ignored.

III. A SAFE AND SECURED DATA MANAGEMENT FRAMEWORK

For a smart and safe city to perform intelligent and secure decisions, the ubiquitous data collected by the devices are processed using different approaches. In SafeCity, the data analysis and machine learning approaches are applied to the data generated and acquired in a ubiquitous environment. The acquisition is carried out by systems that convert the analog information into digital. The cellular technology, i.e. 4G/LTE, is used as a bridging technology between the users, devices, and the system, as shown in Figure 1.

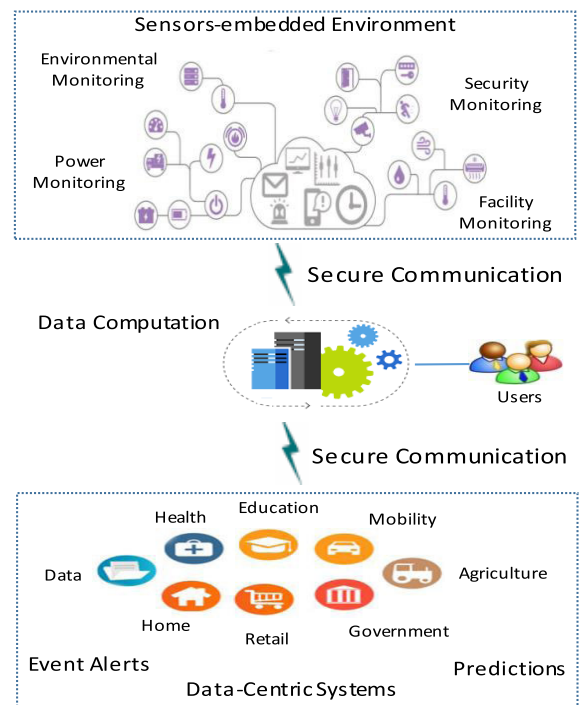


FIGURE 1. Overview of the proposed system.

To design a ubiquitous environment, numerous surveillance cameras, wired and wireless sensors, and device-

mounted sensors are deployed. Data sensing, acquisition, and collection are performed in this environment. Digital loggers and digital data acquisition systems are used to detect and collect data from devices and disseminate them with the help of the Internet. The produced ubiquitous data are secured before forwarding to a computational unit for safe and secured processing and transmission. Afterward, the decisions are made on the secured ubiquitous data. The proposed system is a three-layer architecture, i.e., a ubiquitous data security layer, a ubiquitous data computation layer, and a decision-making layer. A payload-based authentication approach is utilized in the first layer to make the ubiquitous data secured from adversaries.

This layer ensures that only secured data is forwarded. The second layer is accountable for the resource-intensive processing of secured ubiquitous data at the conventional computing platforms. Finally, the third layer provides insights from the ubiquitous data and makes smart decisions. The proposed architecture is shown in Figure 2. The comprehensive description of each layer is given in the following subsections.

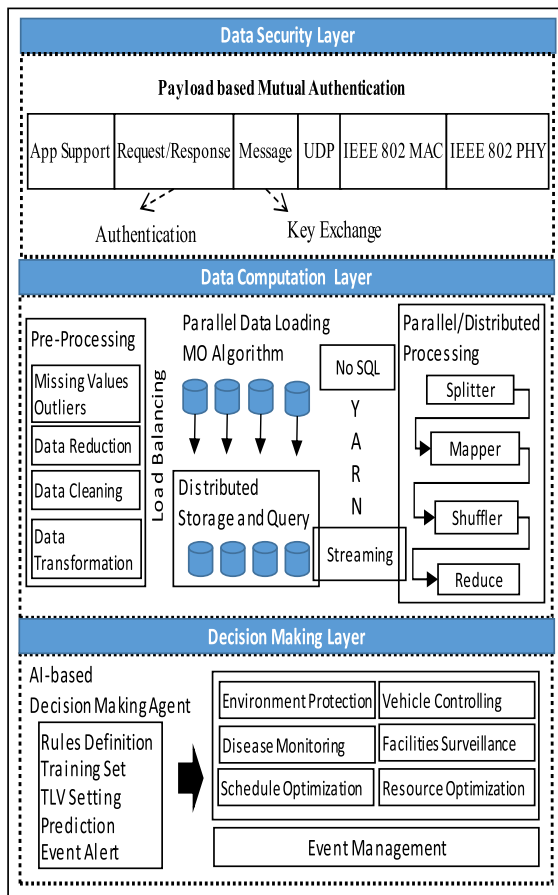


FIGURE 2. System architecture of SafeCity.

A. DATA SECURITY LAYER

This layer of SafeCity is linked to the data sources. The data received from the sensors are in the form of messages. At this layer, message identification and authentication

are performed using a simple payload-based authentication scheme. The proposed scheme uses the CoAP protocol [32] for message exchange and authentication at the application layer of each data source. In ubiquitous environments, most of the CoAP-based solutions are relied on the use of DTLS to ensure the protected transfer of resources between the devices. However, the DTLS-enabled CoAP stack incurs an excessive computational and communication overhead. Furthermore, the use of DTLS in combination with CoAP adds an extra layer of protocol header for security provisioning. In our approach, the security of data messages is not compromised while transferred between clients and servers. The session key is transmitted within the payload messages while authentication is achieved at the request-response communication, as shown by the top layer in Figure 2. In SafeCity, CoAP is equipped with secured features for authentication, efficiency, robustness, and defense against a number of malevolent threats.

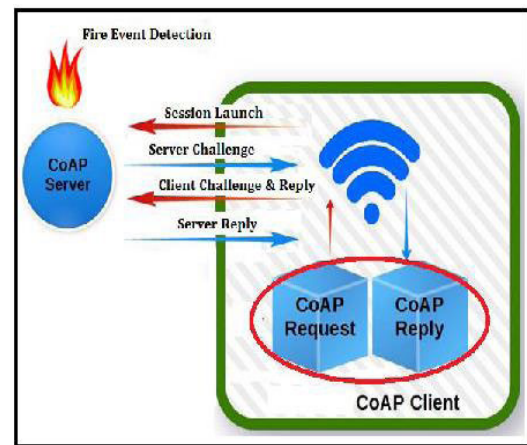


FIGURE 3. Mutual authentication.

During the authentication process, the resource-constrained clients communicate with a server to verify each other identities. As an example, the ubiquitous clients of Figure 1 observe various events such as, temperature, humidity, pollution, and fire eruption, at the server. For a server to provide access to the residing resources, both the parties need to be mutually authenticated. In SafeCity, the authentication is accomplished using four handshake messages. A maximum of 256-bits is used within the payload of each message. The four handshake messages are session launch, server challenge, client challenge and reply, and server reply, as shown in Figure 3. The session launch is headed by a provisioning stage where the clients share a secret key with the server. The server conserves a trace of keys, based on an associated unique identifier (ID). The exchange of a session key between the client and server takes place upon successful authentication. For each client, a session key is implanted on the device at the manufacturing time. If an impostor strives to raze the client, a specific alarm is spawned to notify the crack. To encode the payload of authentication information, the Advanced Encryption Standard (AES) is utilized.

During the session launch, a secret key λ_i is shared with the server, where λ_i is 128-bit long. The λ_i is identified only by client_{*i*} (it belongs) and server, where $i \in \{1,2,3,\dots,I\}$. Each i has a unique identifier that helps the server to execute a look-up table for verification of identity. The session launch is similar to a Hello message and its payload consists of CoAP options fields, i.e., **Auth** and **Auth-Msg-Type**, to indicate the type of operations performed between the client and a server. After the session launch, the next step is the server challenge, in which the server creates a challenge for the client. The encounter containing a pseudo-random nonce η_r and a session key μ produced by the server. The following equations are used to create a challenge.

$$\vartheta = \lambda_i \oplus \mu \tag{1}$$

$$C_r = \text{AES}\{\lambda_i, (\vartheta|\eta_r)\} \tag{2}$$

where, i is the ID of a client, ϑ is the intermediate value generated by the server, and C_r is the challenge sent to i . In the client challenge and reply message, the client retrieves η_r and λ from the server challenge and creates a challenge in response using the following equations.

$$\vartheta' = \eta_r \oplus \lambda_i \tag{3}$$

$$C_i = \text{AES}\{\mu, (\vartheta'|\eta_i)\} \tag{4}$$

where η_i is the pseudo-random nonce and ϑ' is the intermediate value generated by the client, and C_i is the challenge sent to the server. Upon receiving, the server tries to retrieve η_r from the client's challenge. If this nonce is present, the status of i changes to **Authenticated**, and the server responds to the client's challenge to complete the authentication process, using the following equation.

$$C_r = \text{AES}\{\lambda_i, (\eta_r|\mu)\} \tag{5}$$

B. DATA PROCESSING AND COMPUTATION LAYER

Versatile analysis and intelligent processing on huge data streams can be unrealistic and infeasible if the data streams are not properly pre-processed. Data pre-processing are performed prior to the core computation and processing. The pre-processing steps involve the reduction to realize the reduced data with similar properties, data transformation to standardize data to an appropriate arrangement for processing, and data cleansing. These activities are carried out using machine learning approaches. The objective is to dig out the data about various sets of an IoT domain, based on its characteristics. Next, the data loading is performed using multiple attribute criteria model (MACM) in the context of the Hadoop ecosystem. The MACM includes parallel data loading using the customized utility. The HDFS saves the huge files in small chunks that are customized to avoid too much data and metadata, that would otherwise create the overhead.

In HDFS, a replication method is to replicate the original chunk of data which is a time-consuming task. As a result, customized replication is proposed in this paper. Moreover, the Sqoop utility is used due to the parallel loading of data

using the map method. The proposed scheme utilizes Sqoop that offers connectivity to the external databases. The utilization of Sqoop brings a variety of features in SafeCity, such as loading with increments, complete import, parallel import, and corresponding export, compression, easy movement, enterprise independence, and auto-generation of tedious user side's code. Data processing and analytics are carried out using the MapReduce programming paradigm. Hadoop divides input dataset into small blocks of same size files, known as input splits. The size of the split is usually identical to the block or chunk size. One specific task (known as map task) is formed for each split that performs the function of the map, defined by the programmer, for each row (a record). A RecordReader is used to arrange the rows as a pair (key-value). The MapReduce process is depicted in Figure 4. The outputs of the map are not stored in HDFS, these results are stored in the local storage. Results from a number of mappers are the input for the reduce task. Reduce tasks do not include the advantage of data locality characteristic. Therefore, the stored map results have to transfer crossway the system to that specific location, where the job of reducing is performing. The of the reducer result is stored on HDFS.

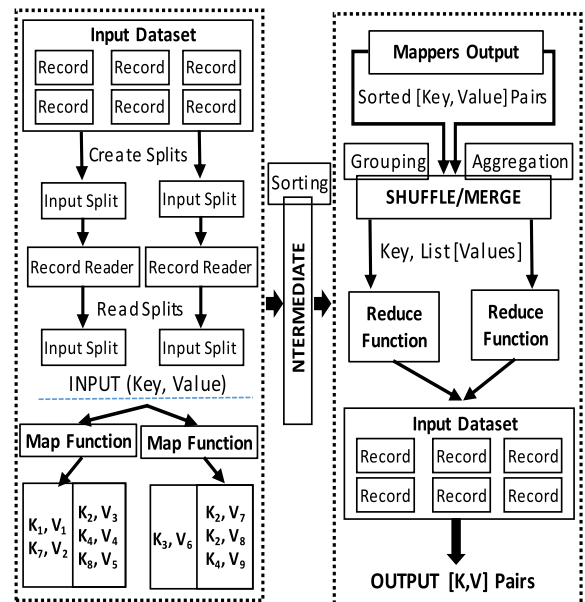


FIGURE 4. MapReduce paradigm.

Our projected scheme is grounded on the up-to-date depiction of Apache Hadoop framework which is embedded with Yet Another Resource Negotiator (YARN) and is accountable for data computation and cluster management. Unlike conventional MapReduce, the computation elements and resource management is separated by YARN. The YARN-enabled model is not limited to the MapReduce classical mechanism. The YARN is preferred due to limitations of classical MapReduce that are mostly associated with scalability and workload support. In the proposed architecture,

YARN has a ResourceManager that runs as a master daemon by managing the accessible cluster resources among a wide range of competing and contending applications. The ResourceManager keeps track of the available resources and live nodes on the cluster. As it is the solo process having this information, so it coordinates the resource allocation and scheduling between the submitted applications. The allocation decisions are made in a secured, multi-tenant, and shared way, e.g. based on queuing capacity, data locality, an application priority, etc. On the submission of an application, a lightweight process instance, also known as ApplicationMaster, is initiated that is responsible for the execution of all the tasks within an application. It is comprised of tasks monitoring, restarting failed tasks, and calculating the overall values of the used application counters. In the existing literature, the classical MapReduce framework is utilized where a single JobTracker is responsible to take care of these responsibilities for all the jobs. Utilizing a single JobTracker in huge clusters exposes them to the scalability bottleneck.

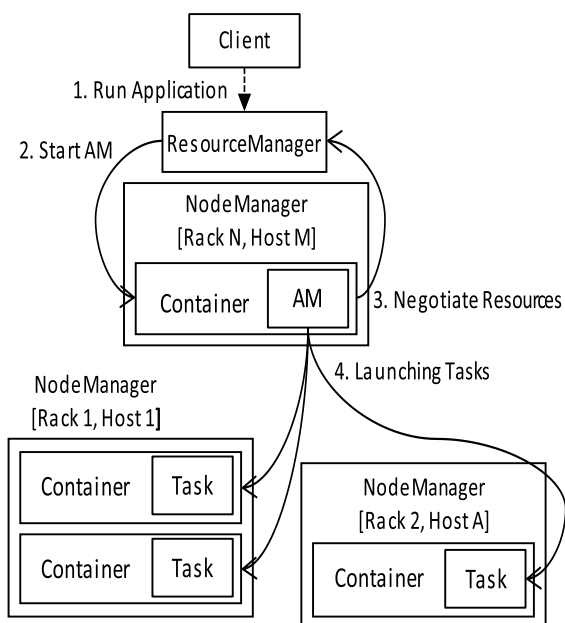


FIGURE 5. Yet another resource negotiator (YARN).

Different tasks associated with a particular application and an ApplicationMaster are controlled, monitored, and managed by the corresponding NodeManagers. Unlike the TaskTracker of a classical MapReduce framework, NodeManager is an efficient and more generic version of the TaskTracker. The NodeManager has many resource containers that are created dynamically, rather than having a defined number of slots (maps and reduces). All the components of the YARN such as ResourceManager, NodeManagers, ApplicationMaster, and containers cooperate with each other in a specific way upon the submission of an application in the cluster of YARN. This interaction of different parts of a YARN framework is shown in Figure 5.

The application is submitted using the Hadoop jar command in CLI or using Java IDE to RM, in a similar way to classical MR. A complete list of running jobs on the Hadoop cluster and all the available and accessible resources on every NM (live) are maintained by RM. The RM needs to decide which application is the next to acquire a piece of cluster resource. A number of constraints are taken into consideration while taking this decision such as fairness and capacity of the queue. The RM employs a scheduler that focuses mainly on scheduling activities. It deals with accessing the resources of a cluster and decides when and who will access them. Within an application, the task monitoring is not carried out by the scheduler and it never tries to restart a failed task. When the submission of a new application is accepted by ResourceManager, first the scheduler decides to select a container where ApplicationMaster will be started and run.

The ApplicationMaster will be in charge of the entire life cycle of the application when it starts. Primarily, ApplicationMaster would be requesting for various resources to the overall manager (ResourceManager) in order to inquire for different containers that are required to execute tasks of a particular application. A request for a particular resource is just a demand for several containers to assure various resource necessities, i.e., a number of resources. For example, CPU share, MB memory, preferred location, e.g. rack name, hostname or if no preference is required then * is used, and priority inside the current application.

The ResourceManager grants a container, whenever possible, that satisfies the request made by an ApplicationMaster. On a specific host, the application is permitted by the container to utilize specified resources. ApplicationMaster requests the NodeManager to launch an application-specific task to utilize these resources after a container is granted.

Please recall that the NodeManager is responsible to manage the host on which a particular container is assigned. The application-specific task could be any particular task written in any framework, e.g. MapReduce. The NodeManager only monitors and examines the resource usage in the containers. It does not monitor the tasks and destroys them if they use more than the allocated memory.

The ApplicationMaster is responsible for monitoring the restarting tasks in fresh containers that are failed, the progress of tasks and its application, and provides the progress back to a client. The ApplicationMaster closes itself and releases its container on completion of the application. Nevertheless, the RM does not check the tasks inside an application at all. It only confirms the health of the ApplicationMasters. In this paper, a flowchart is proposed using the MapReduce programming paradigm that is applied to a water dataset. This flowchart is used to collect the values/quantity of water consumption against different houses to govern the level of water and its demand. The pictorial illustration of recommended MapReduce is depicted in Figure 6.

The mapper gets the offset of a line as a specific key and the entire row is considered as a value. The time parameter (timestamp) and associate values are produced as output

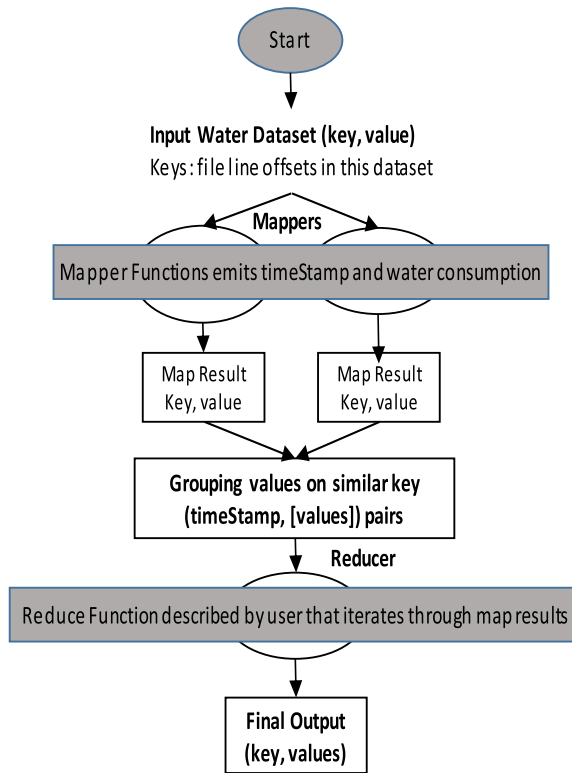


FIGURE 6. MapReduce flowchart for water dataset.

Algorithm 1 Mapper for Water Dataset

```

BEGIN
Input:
  key: line-offset
  value: = row
Output:
  key: facilityID
  value: LOTLINK
  //containing water consumption measurement

  // line splitting
  facilityID, LOTLINK: = line.split('\t')
  key: = facilityID
  value: = LOTLINK
  emit (key, value)
END
  
```

by the mapper. The reducer clusters the necessary associate values alongside every timeStamps and relates with the TLV (threshold limit value). Information with regard to the water consumption of different houses is obtained with the help of such algorithms. As the MapReduce executes various jobs in 2 phases, i.e., Map phase and Reduce phase, therefore, a separate Map function and a Reduce function is proposed for the flowchart of Figure 6. In Algorithm 1, we present the mapper for the water dataset and in Algorithm 2, we present the reducer for the same dataset.

Algorithm 2 Reducer for Water Dataset

```

BEGIN
Input:
  key: facilityID
  value: LOTLINK
Output:
  key: facilityID
  value: LOTLINK greater than threshold
initialize threshold
final []
FOR each (LOTLINK) at facilityID DO
IF (LOTLINK > threshold)
Begin
  final.append (LOTLINK)
  key: = facilityID
  value: = final
  emit (key, vaue)
End IF
END
  
```

C. DECISION-MKING LAYER

The intelligent decision making is the key to our SafeCity framework that includes the prediction, creation of training sets, thresholds setting, rules definition, and event management. It acts as the moderator between the end-users and it is carried out by the decision-making agent, based on AI approaches. Various limits are defined and several rules are set for the assessment of different datasets. The processing of data is carried out using these rules according to proposed algorithms. The TLV (Threshold Limit Value) is a precise value set for each dataset also known as threshold or limit which is the base for event generation and decision making. Likewise, several rules are set centered on corresponding limits in the form of if/then statements that are utilized for decision making. The notification and event alert component determines the specific recipient of a generated event. Hence, it notifies the operator with the generated event for further actions.

IV. SYSTEM EVALUATION AND ANALYSIS

The detailed analysis and discussion of results achieved using SafeCity discuss in this segment. The secured data authentication is realized using Raspberry Pi boards for the client-server interface model. The Libcoap library is used for Raspbian operation system that provides basic communication among the ubiquitous devices. The analysis is carried out on a dataset that is realistic to evaluate the SafeCity scheme using the premeditated algorithms. The implementation of our ubiquitous data computation layer is carried out using the Hadoop cluster on Ubuntu OS along with Sqoop. Moreover, Java is used for the MapReduce implementation by utilizing the pre-defined classes (mapper and reducer). The data is received from diverse but trustworthy sources that

are authentic. These datasets contain the transportation data, i.e., vehicles on roads in Aarhus city, Denmark. The water dataset homes are gained from the houses in Surrey, Canada.

A. SYSTEM EVALUATION FOR SAFETY AND SECURITY

The experimental results concerning the ubiquitous data security layer are illustrated here. A comparison of our payload-based authentication for SafeCity and CoAP-based DTLS implementation for smartphones is provided in Figure 7. DTLS+ denotes a smartphone (ubiquitous device) operating as a server and a workstation as a client. On the other hand, DTLS* denotes the handshake between a smartphone and a workstation, where the smartphone operates as a client and the workstation as a server. As the figure shows, SafeCity has a much lower handshake duration and standard deviation in comparison to DTLS* and DTLS+.

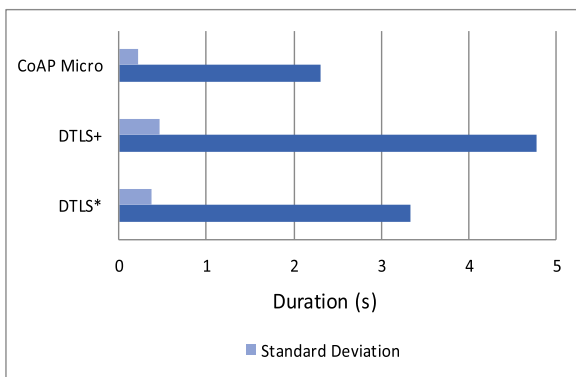


FIGURE 7. Handshake duration.

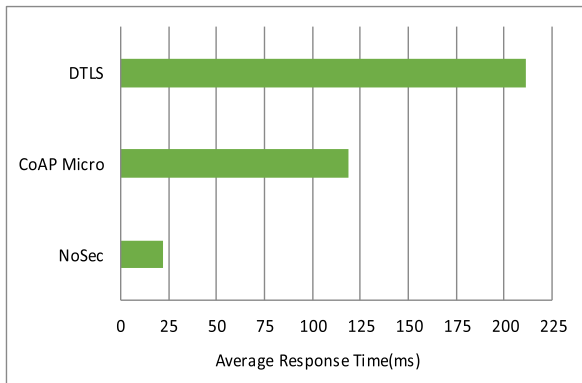


FIGURE 8. Average response time.

Similarly, SafeCity focuses on asynchronous communication of CoAP messages over the UDP sockets. A record of transferred Confirmable (CON) requests is maintained by every client. The mean reaction time for one CON request message of 1 byte is compared with DTLS exchange and the CoAP protocol with no added security, in Figure 8. SafeCity has a much lower average response time in comparison to DTLS because the latter involves computationally complex cipher suites and a resource-intensive record layer. CoAP with no added security has a slower response time but it is prone to various malicious and adversarial attacks.

TABLE 1. Average consumption (kb).

CoAP Micro	HTTP	HTTP/U DP	CoAPBlip	TinyCoAP
207	802	4009	7160	8498

The memory utilization of a CON request is evaluated at the compile time in Table 1. The proposed SafeCity is compared with the existing schemes for a CON message of minimum 500 bytes, as depicted in Figure 9 too. Among the current schemes, CoAPBlip [33] allocates considerable storage to messages at the compile time of the message. TinyCoAP [34] is a variation of the standard libraries of C that need the TinyOS element for its installation on a ubiquitous device. HTTP has a short foot-print of memory as it doesn't offer a trustworthiness method or correlation of a request/response. Both TinyCoAP and CoAPBlip use resource-consuming libraries and have a much higher memory consumption.

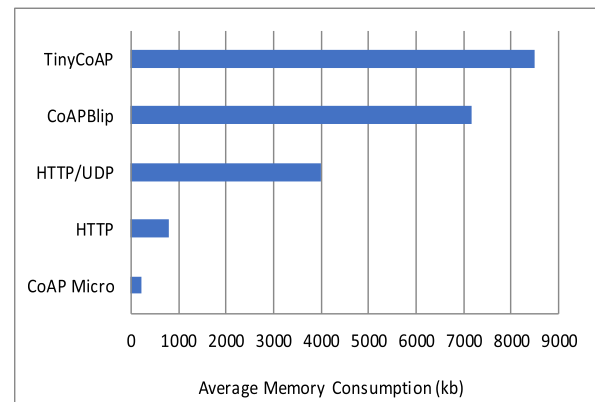


FIGURE 9. Average memory consumption.

B. SYSTEM EVALUATION FOR DATA PROCESSING AND COMPUTATION

Our SafeCity architecture generates alerts in real-time for a particular ubiquitous environment. In this section, we evaluate SafeCity in terms of efficiency by considering the execution time and throughput. To examine the system performance in real-time, various datasets, such as vehicular and water, are replayed to our Hadoop-based YARN framework of SafeCity. The throughput is assessed using datasets by increasing the data size. The efficiency concerning throughput is measured as shown in Figure 10. It can be observed that with the growth in size, the processing speed is reduced. The system throughput of Yarn-based framework is considerably higher in comparison to the existing classical MR-based solution.

Table 2 reveals the processing time, also known as the execution time proposed framework in the context of data volume. The execution time is evaluated for different sizes

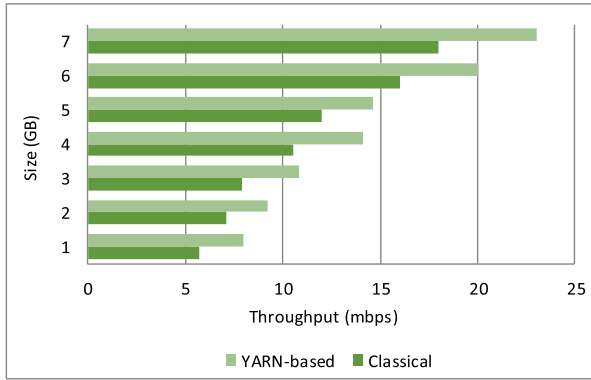


FIGURE 10. System throughput.

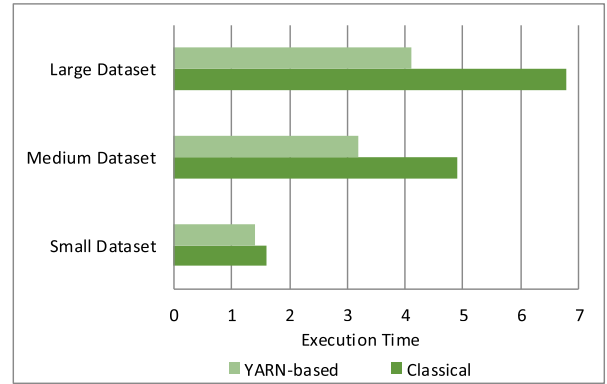


FIGURE 11. Execution time (s).

TABLE 2. Processing time of proposed framework.

Size (GB)	Time (ms)
1	67
2	78
3	96
4	119
5	133.5
6	150.9
7	168.3
8	185.7
9	203.1
10	220.5
11	237.9
12	255.3
13	270

TABLE 3. Average consumption (kb).

	Minor	Average	Huge
Traditional	1.4	4.9	6.8
Proposed	1.6	3.2	4.1

of data. The data size is started from 500MB and experienced up to 13 GB of data.

Table 3 determines the processing time in comparison to the classical structure. The time is calculated for minor, average, and huge datasets. It is observed that the processing time improves when the dataset size is increased. Figure 11 demonstrates the execution time of jobs using our Yarn-based framework in comparison to the existing scheme. The execution time is evaluated for small, medium, and large datasets. It is observed that the processing time improves when the dataset size is increased. It is mostly because of the data loading efficiency and improvement.

C. DATA ANALYSIS

The time difference of data loading is not perceptible when the size is smaller. The data ingestion time is pretty evident when the bulk of a dataset is larger due to the

replication approach. The query that arises is the threshold data, to discover the TLV size, the data loading performance is measured by testing the different sizes of data.

The TLV size is the point where the time difference becomes positive (greater than 0) which means a significant change occurs. The TLVs for various attributes are set using the outputs of similar trials. Taking into account the data ingestion tool experiments, the TLV size is 900MB (size of data). At this value, the effect of the data ingestion period is experienced as shown in Figure 12. This figure demonstrates that 1GB of size does not generate any change even if the automated ingestion is practice. The productivity is attained when dataset size is greater than 900 MB at least.

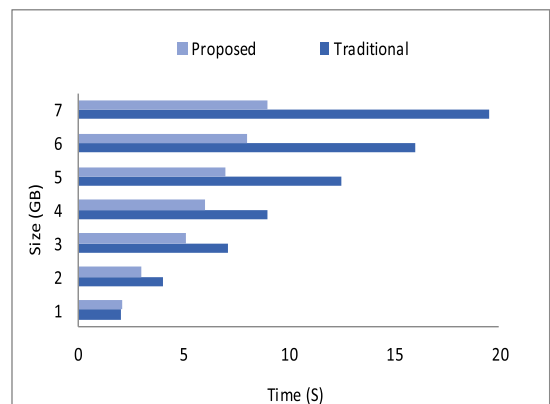


FIGURE 12. Data loading efficiency.

The water consumption is evaluated to achieve sustainable water management in the city due to the inconsistent consumption of water could be a disaster in the future. The data utilized in our research contains information about the city of Surrey, Canada. It comprises of the water intake of the houses in Surrey that is processed using our proposed algorithms. The results are demonstrated in Figure 13.

It shows the houses consumed more than 82000 liters each month. The defined TLV is 82000 found from the rule engine. The water usage higher than the TLV is particularly highlighted in this figure and this can cause frightening

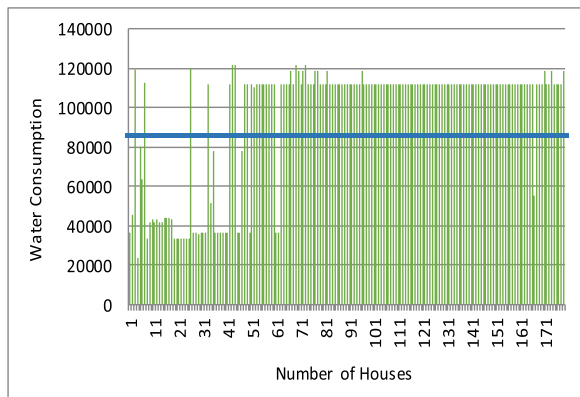


FIGURE 13. Water consumption.

situations for the authorities. It is observed that almost 50% of the consumers consumed more than the threshold limit. Most of the consumers, above the TLV limit, consumed water between 110000 to 120000 liter, which is quite alarming. Up-to-date fabrication methods could be industrialized to control the issues of the consumers in a city.

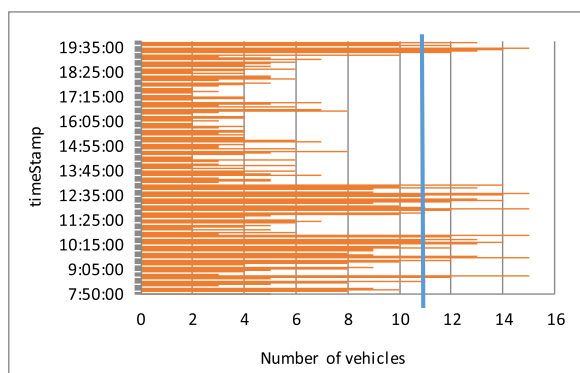


FIGURE 14. Number of vehicles on the road.

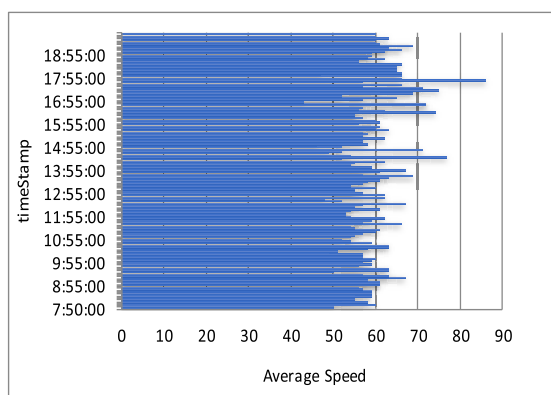


FIGURE 15. Average speed of automobiles.

Regarding traffic management, we consider the traffic data about road congestion. The data is intelligently processed using the SafeCity framework to overcome the traffic issues when the vehicles on roads surpass TLV. Figure 14 reveals

the vehicles and the corresponding TLV. It depicts vehicles at a different time on the roads. It is observed that due to schooling hours, there are more cars between 8:05-12:15 PM due to school and office timing in the city.

Furthermore, the average speed of vehicles is revealed in Figure 15. It is noticed that the average speed of the vehicles is quite alike all day, except from 13:00 to 18:00, when there are few vehicles.

V. CONCLUSION

This paper has envisioned the vital role of safety and security in IoT-enabled data computation and communication to achieve safe and secure decisions. The data generated by IoT sensors exploit the association between various features of data and enables the meaning of a safe city. We have suggested the conception of SafeCity and proven its applicability using apache and Hadoop, via cautious investigation and assessment of the presence of residents in the evolving smart cities. SafeCity carefully controls the encounter of security and computation faced by the ubiquitous data. It is a layered architecture that is composed of a data security layer, data computation layer, and decision-making layer. A payload-based authentication approach is utilized at the ubiquitous data security layer to secure the ubiquitous data from malevolent entities.

The data computation layer is liable for the processing of secured data. Finally, the decision-making layer extracts insights for making smart decisions. The ubiquitous data security is evaluated using the Raspberry Pi boards while the ubiquitous data computation is tested on trustworthy datasets, using Hadoop. In association with the current methods, SafeCity is trivial about handshake duration, response time, and average memory consumption. Furthermore, it attains a lesser processing time, greater throughput, and efficient about massive data ingestion.

REFERENCES

- [1] P. Bocquier, "World urbanization prospects: An alternative to the UN model of projection compatible with the mobility transition theory," *Demograph. Res.*, vol. 12, pp. 197–236, May 2005.
- [2] J. L. Hernández, R. García, J. Schonowski, D. Atlan, G. Chanson, and T. Ruohomäki, "Interoperable open specifications framework for the implementation of standardized urban platforms," *Sensors*, vol. 20, no. 8, p. 2402, Apr. 2020.
- [3] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 4, pp. 2347–2376, 4th Quart., 2015.
- [4] M. Weber and I. P. Žarko, "A regulatory view on smart city services," *Sensors*, vol. 19, no. 2, p. 415, 2019.
- [5] A. Entezami, H. Sarmadi, B. Behkamal, and S. Mariani, "Big data analytics and structural health monitoring: A statistical pattern recognition-based approach," *Sensors*, vol. 20, no. 8, p. 2328, Apr. 2020.
- [6] M. Babar and F. Arif, "Smart urban planning using big data analytics based Internet of Things," in *Proc. ACM Int. Joint Conf. Pervas. Ubiquitous Comput., ACM Int. Symp. Wearable Comput.*, Sep. 2017, pp. 397–402.
- [7] M. Babar and F. Arif, "Smart urban planning using big data analytics to contend with the interoperability in Internet of Things," *Future Gener. Comput. Syst.*, vol. 77, pp. 65–76, Dec. 2017.
- [8] M. Babar and F. Arif, "Real-time data processing scheme using big data analytics in Internet of Things based smart transportation environment," *J. Ambient Intell. Hum. Comput.*, vol. 10, no. 10, pp. 4167–4177, Oct. 2019.

- [9] M. Babar, A. Rahman, F. Arif, and G. Jeon, "Energy-harvesting based on Internet of Things and big data analytics for smart health monitoring," *Sustain. Comput., Informat. Syst.*, vol. 20, pp. 155–164, Dec. 2018.
- [10] J. Granjal, E. Monteiro, and J. S. Silva, "Security for the Internet of Things: A survey of existing protocols and open research issues," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1294–1312, 3rd Quart., 2015.
- [11] S. Rajesh, V. Paul, V. Menon, and M. Khosravi, "A secure and efficient lightweight symmetric encryption scheme for transfer of text files between embedded IoT devices," *Symmetry*, vol. 11, no. 2, p. 293, Feb. 2019.
- [12] M. A. Jan, F. Khan, M. Alam, and M. Usman, "A payload-based mutual authentication scheme for Internet of Things," *Future Gener. Comput. Syst.*, vol. 92, pp. 1028–1039, Mar. 2019.
- [13] A. Vėnčauskas, N. Morkevičius, V. Jukavičius, R. Damaševičius, J. Toldinas, and Š. Grigaliūnas, "An edge-fog secure self-authenticable data transfer protocol," *Sensors*, vol. 19, no. 16, p. 3612, Aug. 2019.
- [14] S. L. Keoh, S. S. Kumar, and H. Tschofenig, "Securing the Internet of Things: A standardization perspective," *IEEE Internet Things J.*, vol. 1, no. 3, pp. 265–275, Jun. 2014.
- [15] S. Raza, L. Seitz, D. Sitenkov, and G. Selander, "S3K: Scalable security with symmetric keys—DTLS key establishment for the Internet of Things," *IEEE Trans. Autom. Sci. Eng.*, vol. 13, no. 3, pp. 1270–1280, Jul. 2016.
- [16] A. Bhattacharyya, A. Ukil, T. Bose, and A. Pal. *Lightweight Mutual Authentication for CoAP (WIP)*. Accessed: Mar. 3, 2014. [Online]. Available: <https://draft-bhattacharyya-core-coap-lite-auth-00>
- [17] J. Granjal, E. Monteiro, and J. S. Silva, "On the feasibility of secure application-layer communications on the Web of things," in *Proc. 37th Annu. IEEE Conf. Local Comput. Netw.*, Oct. 2012, pp. 228–231.
- [18] T. Kothmayr, C. Schmitt, W. Hu, M. Brünig, and G. Carle, "DTLS based security and two-way authentication for the Internet of Things," *Ad Hoc Netw.*, vol. 11, no. 8, pp. 2710–2723, Nov. 2013.
- [19] J. Granjal, E. Monteiro, and J. S. Silva, "On the effectiveness of end-to-end security for Internet-integrated sensing applications," in *Proc. IEEE Int. Conf. Green Comput. Commun. (Green-Com)*, Nov. 2012, pp. 87–93.
- [20] D. Tralbalza, S. Raza, and T. Voigt, "Indigo: Secure coap for smartphones," in *Wireless Sensor Networks for Developing Countries*. Berlin, Germany: Springer, 2013, pp. 108–119.
- [21] M. A. Jan, P. Nanda, X. He, Z. Tan, and R. P. Liu, "A robust authentication scheme for observing resources in the Internet of Things environment," in *Proc. IEEE 13th Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Sep. 2014, pp. 205–211.
- [22] S. Din, H. Ghayvat, A. Paul, A. Ahmad, M. M. Rathore, and I. Shafi, "An architecture to analyze big data in the Internet of Things," in *Proc. 9th Int. Conf. Sens. Technol. (ICST)*, Dec. 2015, pp. 677–682.
- [23] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, "Urban planning and building smart cities based on the Internet of Things using big data analytics," *Comput. Netw.*, vol. 101, pp. 63–80, Jun. 2016.
- [24] M. M. Rathore, A. Paul, A. Ahmad, M. Anisetti, and G. Jeon, "Hadoop-based intelligent care system (HICS): Analytical approach for big data in IoT," *ACM Trans. Internet Technol.*, vol. 18, no. 1, p. 8, Dec. 2017.
- [25] B. N. Silva, M. Khan, C. Jung, J. Seo, Y. Yoon, J. Kim, S. Jin, J. Kang, and K. Han, "Planning of smart cities: Performance improvement using big data analytics approach," in *Proc. 4th Int. Conf. Adv. Comput., Electron. Commun. Inst. Res. Eng. Doctors*, 2016, pp. 51–55.
- [26] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of Things and big data analytics for smart and connected communities," *IEEE Access*, vol. 4, pp. 766–773, 2016.
- [27] R. Tönjes, M. I. Ali, P. Barnaghi, S. Ganea, F. Ganz, M. Haushwirth, B. Kjærgaard, D. Kümper, A. Mileo, S. Nechifor, A. Sheth, V. Tsiatsis, and L. Vestergaard, "Real time iot stream processing and large-scale data analytics for smart city applications," in *Proc. Eur. Conf. Netw. Commun.*, 2014, pp. 1–5.
- [28] B. Cheng, S. Longo, F. Cirillo, M. Bauer, and E. Kovacs, "Building a big data platform for smart cities: Experience and lessons from santander," in *Proc. IEEE Int. Congr. Big Data*, Jun. 2015, pp. 592–599.
- [29] M. M. Rathore, A. Paul, A. Ahmad, and G. Jeon, "IoT-based big data: From smart city towards next generation super city planning," *Int. J. Semantic Web Inf. Syst.*, vol. 13, no. 1, pp. 28–47, 2017.
- [30] V. G. Menon and J. Prathap, "Vehicular fog computing: Challenges applications and future directions," *Int. J. Veh. Telematics Inf. Syst.*, vol. 1, no. 2, pp. 15–23, 2017.
- [31] M. M. Rathore, A. Ahmad, A. Paul, and G. Jeon, "Efficient graph-oriented smart transportation using Internet of Things generated big data," in *Proc. 11th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2015, pp. 512–519.
- [32] Z. Shelby, K. Hartke, and C. Bormann, *The Constrained Application Protocol (CoAP)*, document RFC 7252, 2014.
- [33] K. Kuladinithi, O. Bergmann, T. Pötsch, M. Becker, and C. Görg, "Implementation of coap and its application in transport logistics," in *Proc. IP+SN*, Chicago, IL, USA, 2011, pp. 1–6.
- [34] A. Ludovici, P. Moreno, and A. Calveras, "TinyCoAP: A novel constrained application protocol (CoAP) implementation for embedding RESTful Web services in wireless sensor networks based on TinyOS," *J. Sens. Actuator Netw.*, vol. 2, no. 2, pp. 288–315, May 2013.



HUI ZHANG received the B.Sc. degree in communication engineering from Henan Polytechnic University, China, in 2007, the M.Sc. degree from the School of Energy Science and Engineering, Henan Polytechnic University, in 2010, and the Ph.D. degree from the School of Energy and Mining Engineering, China University of Mining and Technology, in 2013. He is currently an Associate Professor with the School of Energy Science and Engineering, Henan Polytechnic University.

He has authored several articles in journal and conferences, and holds several patents. His research interests include mining communication and smart mine.



MUHAMMAD BABAR received the bachelor's degree (Hons.) in computer sciences from the University of Peshawar, Pakistan, in 2008, and the Master of Science and Ph.D. degrees in computer software engineering from National University of Sciences and Technology (NUST), Islamabad, Pakistan, in 2012. He is currently with Iqra University, Islamabad. He has published his research work in various IEEE and ACM/Springer international conferences and journals. His research interests include big data analytics, the Internet of Things (IoT), smart city design and planning, and Social Web of Things (SWOT). He is an active Reviewer and a guest editor in the reputed journals.



MUHAMMAD USMAN TARIQ received the bachelor's and Master of Science degrees in computing, with a specialization in software engineering, and the Ph.D. degree (Hons.) in management from Calsouthern, USA. He has more than 13 years' experience in industry and academia. He has a passion for learning and development, project management, and training that made him achieve four patents. His research interests include management, the IoT, six sigma, knowledge management, information technology, economics, organizational change, facial recognition, biomedical devices, and computer science.



MIAN AHMAD JAN received the Ph.D. degree in computer systems from University of Technology Sydney (UTS), Australia, in 2016. He is currently a Researcher with Ton Duc Thang University, Vietnam. His research has been published in various prestigious the IEEE TRANSACTIONS and Elsevier Journals. His research interests include security and privacy in the Internet of Things, and wireless sensor networks. He was a recipient of various prestigious scholarship during his studies, notably the International Research Scholarship (IRS) at the UTS, and the Commonwealth Scientific Industrial Research Organization (CSIRO) scholarships. He has been received the Best Researcher awarded for the year 2014 at the UTS, Australia. He has been the general Co-Chair of Springer/EAI 2nd International Conference on Future Intelligent Vehicular Technologies, in 2017. He has been a guest editor of numerous special issues in various prestigious journals, such as the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, *Future Generation Computer Systems* (Elsevier), *Mobile Networks and Applications* (MONET) (Springer), *Ad Hoc & Sensor Wireless Networks*, and *MDPI Information*.



VARUN G. MENON (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. His research interests include the Internet of Things, fog computing and networking, underwater acoustic sensor networks, cyber psychology, hijacked journals, ad-hoc networks, and wireless sensor networks. He is a Distinguished Speaker of ACM Distinguished Speaker. He is currently a Guest Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE SENSORS JOURNAL, the *IEEE Internet of Things Magazine*, and the *Journal of Supercomputing*. He is an Associate Editor of *IET Quantum Communications*. He is also an Editorial Board Member of the IEEE Future Directions: Technology Policy and Ethics.



XINGWANG LI (Senior Member, IEEE) received the B.Sc. degree from Henan Polytechnic University, Jiaozuo, China, in 2007, the M.Sc. degree from the University of Electronic Science and Technology of China, in 2010, and the Ph.D. degree from the Beijing University of Posts and Telecommunications, in 2015. From 2010 to 2012, he was working as an Engineer with Comba Telecom Ltd., Guangzhou, China. From 2016 to 2018, he was also a Visiting Scholar with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. From 2017 to 2018, he was a Visiting Scholar with Queen's University Belfast, Belfast, U.K. He is currently an Associate Professor with the School of Physics and Electronic Information Engineering, Henan Polytechnic University. His research interests include MIMO communication, cooperative communication, hardware constrained communication, non-orthogonal multiple access, physical layer security, unmanned aerial vehicles, and the Internet-of-Things. He has served as many TPC members, such as the IEEE/CIC International Conference on Communications in China (ICCC'2019) and the IEEE Global Communications Conference 2018 (Globecom'18). He is also an Editor on the Editorial Board of IEEE ACCESS, *Computer Communications*, and *KSI Transactions on Internet and Information Systems*. He is also the Lead Guest Editor for the Special Issue on Recent Advances in Physical Layer Technologies for the 5G-Enabled Internet of Things of Wireless Communications and Mobile Computing and the Lead Guest Editor for the Special Issue on Recent Advances in Multiple Access for 5G-enabled IoT.

...

Enhancing the Performance of Flow Classification in SDN-Based Intelligent Vehicular Networks

Mahdi Abbasi¹, Hajar Rezaei, Varun G. Menon², *Senior Member, IEEE*,
Lianyong Qi³, *Member, IEEE*, and Mohammad R. Khosravi⁴

Abstract—Intelligent vehicular networks converged with software-defined networking provides several flow-based surveillance services to mobile applications on vehicular nodes. But, as the scale of such networks grows exponentially, a substantial delay in processing tremendous flows emerges. The delay can be reduced by accelerating the packet classification methods, which are nowadays exploited in software-defined vehicular networks. Fast packet classification lets firewalls to inspect each incoming packet at wire speed. One of the well-known packet classification methods is the KD-tree algorithm. This paper presents an enhanced version of this algorithm that uses the geometric space to display different fields and increases search speed by recursive decomposition of the search space. Also, the enhanced KD-tree is integrated with a leaf-pushing technique, which enhances the performance of KD-tree search during classification. The proposed algorithm is implemented using a bloom filter data structure and a hash table. Experimental results show that the proposed leaf-pushed KD-tree algorithm improves packet classification speed up to 24 times in comparison with the conventional KD-tree. Moreover, the proposed algorithm can significantly reduce the classification time in comparison with state-of-the-art tree-based algorithms.

Index Terms—Intelligent vehicular network, flow classification, KD-tree algorithm, leaf-pushing, performance, software-defined-networking (SDN).

I. INTRODUCTION

INTELLIGENT Vehicular Network (IVN) is one of the world-evolving technologies that help enhance road safety and efficient traffic control in smart cities [1]. This technology uses various communication technologies to provide organized routes to high mobility vehicular nodes [2], [3]. Although recently exploited high-speed communication technologies can provide dependable and universal mobile coverage [4], several

prominent features of novel deployments of IVN lead new challenges, such as unbalanced traffic flow in a multi-path topology and inefficient network utilization [5]–[7]. Thus, flexible and programmable architectures like software-defined-networking (SDN) have been recently proposed as a key solution for IVNs. The network programmability feature of the SDN, when added to IVN lets external applications to simply reconfigure the equipment and wireless devices [8]. That is, the SDN provides considerable flexibility in evolving vehicular network infrastructure [9], [10]. For this purpose, flow classification rules are configured and assigned to switches dynamically according to the network conditions and the requirements for applications on IVN [11]. Flow classification enables an SDN controller to provide several on-demand IVN surveillance services. Each SDN controller manages a dynamic set of packet classification rules, each of which corresponds to a data stream to/from a specific vehicular node [11], [12]. An essential prerequisite for classifying data into specific flows is the packet classification [13]–[17]. Packet classification refers to the process of classifying network packets into flows in routers and switches. Various methods have been so far developed for this purpose which are different in terms of classification time and memory usage. The methods are either software-based or hardware-based. Major hardware-based methods make use of Field-Programmable Gate Array (FPGA) and Ternary Content-Addressable Memory (TCAM) [18]. In general, although hardware-based classifiers achieve high speeds and throughput rates up to 100 MPPS (million packets per second), they cannot be easily developed and customized due to the limited resources on the chip [19], [20]. Moreover, these systems carry high costs and have a low efficiency-to-cost ratio. This is why software-based methods have become the focus of attention in recent years [19]–[24].

In spite of their extensibility, software-based classifiers do not function efficiently in networks with high bandwidth due to the low speed of the serial processing of instructions in CPUs. The challenge of accelerating the software-based classifiers of IP packets, therefore, has resulted in considerable research with the aim of developing methods to increase the speed of classification algorithms. In this study, we seek to use the leaf pushing technique to enhance the performance of KD-trees. The KD-tree is a decision-tree packet classification algorithm. Decision tree-based algorithms are considered as an important class of software-based classification methods. In this type of classification, the rule sets are stored in the search tree based

Manuscript received February 28, 2020; revised July 8, 2020; accepted July 30, 2020. Date of publication August 13, 2020; date of current version July 12, 2021. This work was supported by Bu-Ali Sina University. The Associate Editor for this article was S. Mumtaz. (*Corresponding author: Mahdi Abbasi.*)

Mahdi Abbasi and Hajar Rezaei are with the Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan 6516738695, Iran (e-mail: abbasi@basu.ac.ir; rezaei@eng.basu.ir).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683582, India (e-mail: varunmenon@scmsgroup.org).

Lianyong Qi is with the School of Information Science and Engineering, Qufu Normal University, Jining 273165, China (e-mail: lianyongqi@gmail.com).

Mohammad R. Khosravi is with the Department of Computer Engineering, Persian Gulf University, Bushehr 75169-13817, Iran, and also with the Telecommunications Group, Shiraz University of Technology, Shiraz 71555-313, Iran (e-mail: mohammadkhosravi@acm.org).

Digital Object Identifier 10.1109/TITS.2020.3014044

on binary patterns in the rule fields. Hence, to find the rule that best matches the incoming packet, the tree is traversed based on the binary content of the fields in question [25]. Various tree-based algorithms such as AQT [26], HiCuts [27], and Hyper-Cuts [28] have been so far developed. These algorithms first, seek to obtain efficient search methods by using the geometric representation of rules, and then construct the corresponding decision tree.

As the main contribution, we propose a classification method that makes use of leaf-pushing to allocate search space in a KD-tree. In this method, the nodes in each path that contain rules are reduced to one leaf node. The rules to be compared with each packet are confined to the rules stored in the leaf node and the process of searching the tree is completely separated from the process of rule comparison. As a result of optimizing the KD-tree and using leaf pushing technique, both memory usage and access to off-chip memory are reduced.

The paper is organized as follows. Section II reviews the related literature. Next, the proposed method is described in Section III and evaluated in Section IV. The final section concludes the discussion and shows the direction of further research.

II. RELATED WORK

In this section, the Area-based Quad Tree (AQT) algorithm and the other relevant methods are briefly explained. Next, the key idea behind leaf-pushing is fully explained.

A. Area-Based Quad Tree

In this algorithm, each packet is represented as a point in the geometric space. Space decomposition algorithms provide a search technique that uses a tree or tree-like structure to find a rule that covers the packet. An area-based quad tree (AQT) has a search area that consists of the source prefix address on the X axis and the destination prefix address on the Y axis. Each rule is represented as a square formed by the source and destination prefix addresses [26].

B. Other Algorithms

Linear search compares the rules sequentially with the incoming packet and has a low performance in terms of time. Characteristic of space decomposition algorithms is their geometric approach. In fact, the space of the classification problem is represented as a d-dimensional geometric space in which separators are shown as rectangles. While the rules are stored only once in AQT, other space decomposition algorithms allow their repetition to increase the efficiency of packet classification. Hierarchical Intelligent Cutting algorithm (HiCuts), for example, produces a decision tree by recursive decomposition of the search space. On each node of the tree, one decision is applied to decompose the current search space into several subsets so that each subset would specify a child. Each internal node keeps the information about the divisions performed in the node including the field used in the cutting, the number of cuttings, and the pointers to its children. Each leaf node keeps the rules relating to the space covered by

the node. In grid-of-tries structure, pointers are used instead of rule repetition to relate the nodes. This contributes to the reduction of memory usage. This method does not require recursive traversals; rather, it only traces the pointers back to the node. The algorithm's update time is so long that it is better to recreate the data structure from scratch for addition or omission of a rule. Therefore, this algorithm is appropriate for static packet classifiers in two dimensions, but it cannot be easily extended to multidimensional modes. An algorithm that is suitable for multidimensional modes is Cross-product. In this algorithm, for any given packet P, the best match for each header field is found and all of the results are finally combined to find the best match [27].

Another algorithm is Recursive Flow Classification. This algorithm works by mapping the packet header information onto a smaller number of bits in several phases according to the features of actual classifiers. It is suitable for large numbers of fields and provides a relatively high speed of access, but it has low scalability because it changes the structure of classification fields by adding a new field and requires hardware implementation which is usually difficult to modify [27].

C. Leaf Pushing

A leaf-pushed tree pushes all the prefixes in the internal nodes downward into the leaves, thus storing prefixes only in its leaves [29], [30].

None of the algorithms so far proposed have been able to compromise between classification time and memory usage. In other words, each of these algorithms is optimal either in terms of classification time or in terms of the memory used by its data structure. Therefore, we need a classification algorithm that would be efficient concerning both criteria. With this aim, the next section proposes such a method by making use of the best features of previous algorithms.

III. THE PROPOSED METHOD

In this section, first, we explain the basic KD-tree algorithm and its related data structure using a sample ruleset. Next, we explain how our proposed method applies the leaf-pushing technique on the sample KD-tree. Finally, a bloom-filter based implementation of the leaf-pushed KD-tree is completely explained.

A. KD-Tree Structure

In this algorithm each packet is represented as a point in the geometric space. Space decomposition algorithms provide a search technique that uses a tree structure to find a rule that covers the packet. In a tree structure, all children of a node share an identical prefix that is inherited from the parent node. For example, the children of a parent node that begins with "0" will begin with "0".

Fig. 1 shows an example of a space decomposed by the two fields F1 and F2 which represent the source and destination prefix addresses from Table I, respectively. The wild card state, represented by *, means that the rest of the bits can be 0 or 1.

TABLE I
EXAMPLE OF A RULE SET [30]

Rule No	Source Prefix	Destination Prefix	Source Port	Destination Port	Protocol Type
R0	010*	011*	0,65535	1704,1704	6
R1	01100*	0110*	161,161	1711,1711	6
R2	0110*	1001*	1024,1024	1521,1521	6
R3	1010*	1101*	119,119	1717,1717	6
R4	1*	10*	53,53	2110,2110	6
R5	00*	0*	1024,1024	1717,1717	6
R6	*	110*	80,80	1221,1221	6
R7	000*	*	0,65535	0,65535	6
R8	001*	00*	0,65535	0,65535	*
R9	00*	111*	0,65535	0,65535	*

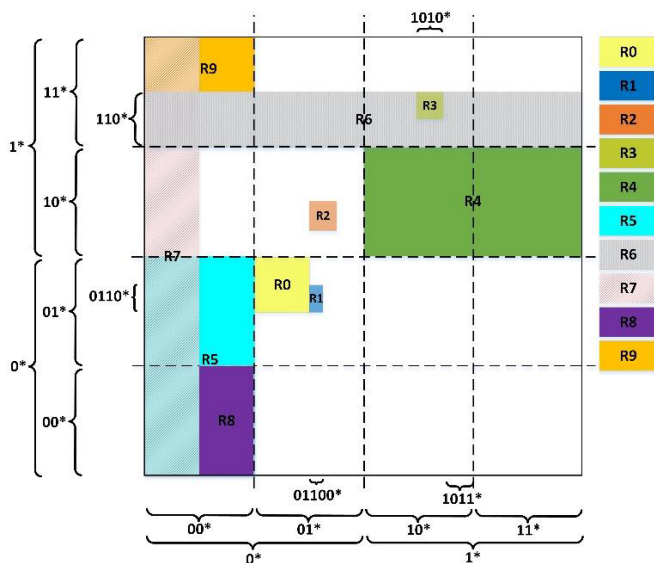


Fig. 1. The rules from Table I as represented in the geometric decomposition space of the KD-tree.

The space covered by a prefix on one axis is inversely related to the prefix length; that is, a shorter prefix covers a larger space. The length of the wild card state, for example, is always 0 and covers all the input spaces on the axis.

The partitioning of the geometric space is as following. The search space is recursively decomposed into two equal partitions based on F1 and F2. At the first level, this is done through F1 which is the first dimension and, at the second level, this is done through F2 which is the second dimension. Thus, if one of the corners of the square space of a rule crosses the boundary of its partition, the rule is considered as part of the Crossing Filter Set (CFS) of that partition.

A KD-tree makes combines recursive decomposition and tree-like structure. In fact, it provides two-dimensional packet classification for binary trees with the aim of searching an IP address.

As shown in Fig. 2, a KD-tree is built by the source and destination prefix address of a rule. Each level of the tree in the search space is divided into two parts based on one of the prefixes.

We begin by the root node which covers the entire search space. Partitioning at this level is based on the source prefix

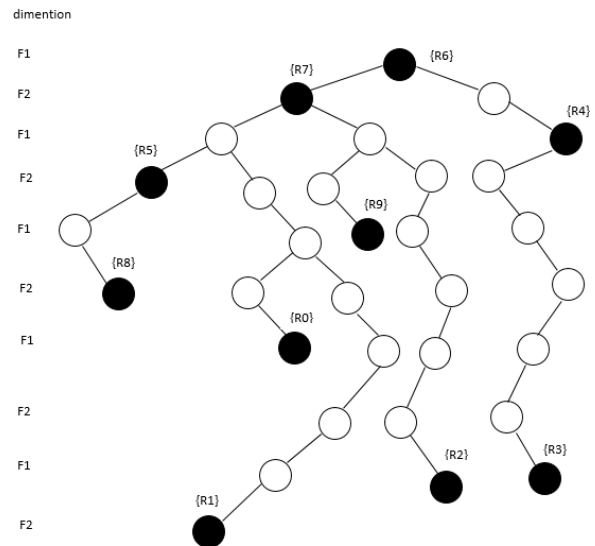


Fig. 2. The binary KD-tree of the geometric space represented in Fig. 1.

code. In this way, all the rules with a source prefix code of 0 (in the left-hand partition of the geometric space) are inserted on the left side of the root and the rules with a source prefix code of 1 (in the right-hand partition of the geometric space) are inserted on the right side. At the next level, partitioning is done on the basis of the destination prefix address. This will continue until every rule has been placed in a node. The inserted into the CFS of a partition have identical prefixes which are derived from the shortest prefix of each rule. They are inserted into a node where the area and path correspond to the sum of the lengths of source and destination prefix addresses and the value of the source and destination prefix codes, respectively. In this method, the rules are stored once without any repetition.

Note that the shortest length of two prefixes determines the area in which the rule is stored. In other words, the KD-tree does not exactly represent the decomposed space. This will increase the number of nodes on each path from the root to a leaf and decrease the efficiency of search. The reason is that the code used in the generation of a KD-tree is produced based on the length of the shortest prefix field of the rule and the rest of the length of longer fields is not used.

B. Leaf-Pushed KD-Tree

A leaf-pushed tree pushes all the prefixes in the internal nodes downward into the leaves. Therefore, prefixes are only stored in the leaves Fig. 3 represents the implementation of the leaf-pushing on the tree of Fig. 2. The prefixes in a leaf-pushed tree are joined, which optimizes the IP address search. Each leaf node in the leaf-pushed tree corresponds to the joined range of coverage and stores the prefixes which the range covers. The leaf-pushing technique used here differs from that utilized in IP address search problems. In IP address search where the longest prefix matching is at stake, only the longest prefixes are pushed into the leaves while here we push all prefixes that cover the same range into the leaves with the aim of solving packet classification problems.

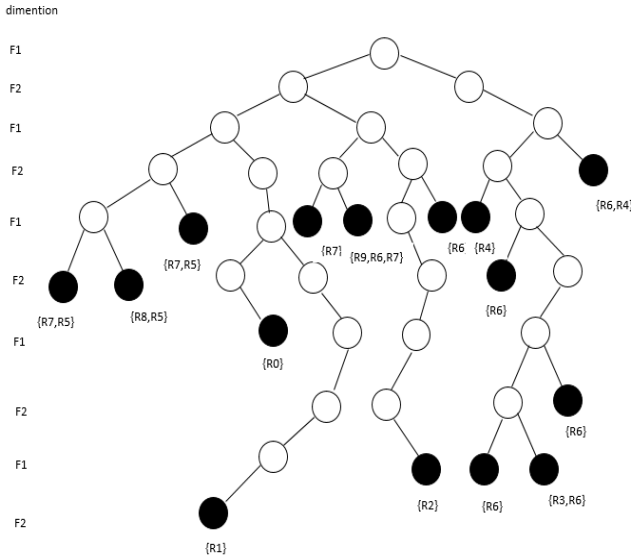


Fig. 3. The leaf-pushed tree of the geometric space represented in Fig. 1.

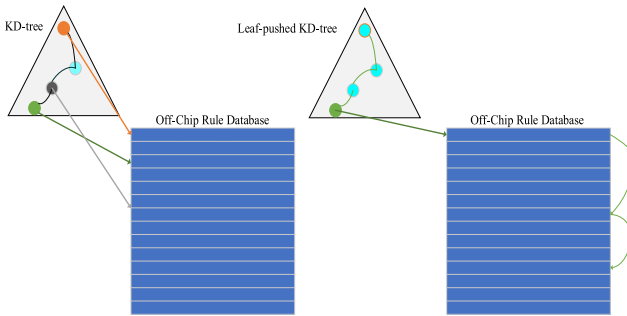


Fig. 4. Comparison of the architecture of conventional KD-tree and leaf-pushing tree.

In what follows, we seek to turn a KD-tree into a leaf-pushed tree. The leaf-pushed tree is created as following. In the example in Fig. 2, the rules stored in the internal nodes include R4, R5, R6, and R7. Let us examine the leaf-pushing process for R4 ($1^* 10^*$). This rule is in the first dimension. Since there is no other prefix in this dimension, the rule can cover both the left and the right child. Therefore, if we extend the prefix address of the first dimension, which is the starting point, we will obtain 10^* and 11^* and the rule R4 will be transferred to its two child nodes. As the right node is a leaf, further extension on this side is not necessary. On the left side, as R4 still lies in an internal node, it should be further extended. Since this rule still has a prefix code on the second dimension (i.e. the destination), this bit will be used. The bit is 0. Therefore, we move to the left side of the node and stop further extension on arriving at a leaf node. This process will continue for all rules in the internal nodes. In fact, further extension of rules should stop with the end of their nested relations because, although further extension will increase search efficiency, the required memory will also increase due to the repetition of rules in the nodes.

Algorithm 1 shows the pseudo code for searching the leaf-pushed KD-tree. The input to this function is the input that was assumed for explaining the search process in this tree, i.e.

Algorithm 1 The Pseudo Code for Searching the Leaf-Pushed KD-Tree

```

Input: packet in_pkt
Output: rule R
1: function SearchLeafPushingKdtree(in_pkt)
2:   BMR = default
3:   next_node = root; i = 0
4:   while (next_node != NULL) do
5:     node = next_node
6:     if ((node.type =
           = RuleNode)&&(BMR
           > node.pri)) then
7:       BMR = linearSearch(in_pkt)
8:       break
9:     else
10:      if (node.dimension == 0) then
11:        | next_node = node.ptr(in_pkt.srcA[i])
12:      else if (node.dimension == 1) then
13:        | next_node = node.ptr(in_pkt.dstA[i])
14:        | i++
15:      end if
16:    if end
17:  end while
18:  //search for wildcard rules
19:  if (BMR > wild.threshold) then
20:    | BMR = linearSearch(in_pkt);
21:  if end
22:  return BMR
23: function end

```

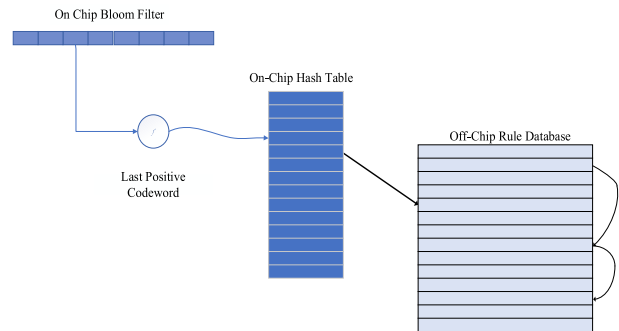


Fig. 5. Searching a tree by means of a Bloom filter and a hash table.

(6, 1711, 161, 01100, 01100). The output of the function is the best matching rule (BMR). First, a default value is determined for cases where the packet does not match any of the rules in the database (line 2). The default value is usually the wild-card state. Lines 4 to 17 traverse the tree. The traversal begins at the root and the first dimension. The algorithm takes the first bit of the source prefix code, which is 0, and moves to the left child (line 10). At the next level, it takes the first bit of the destination prefix code and moves to the left child (line 12). Then it takes the second bit of the source prefix code, which is 1, and moves to the right child. As the process continues and a leaf node containing R1 is achieved (line 6), the search is finished. R1 is compared with the packet header and, if they

TABLE II
COMPARISON OF THE BEHAVIOR OF CONVENTIONAL KD-TREE AND THE PROPOSED LEAF-PUSHING TREE

Size Rule set Type	KD-tree				Leaf-pushed KD-tree			
	5K	10K	50K	100K	5K	10K	50K	100K
ACL	4834	9835	49220	97450	4834	9835	49220	97450
	4834	9835	49220	97450	36838	46649	79361	145263
	14842	21789	8616	9107	25825	37003	10619	10921
IPC	4731	9533	32111	57104	4731	9533	32211	57104
	4731	9533	32211	57104	26265	62125	393463	655118
	22970	46349	4146	4399	40839	82333	4869	5011
FW	4710	9387	32578	44828	4710	9387	32578	44828
	4710	9387	32578	44828	194160	364919	439512	881245
	20052	40476	11742	1454	3143	7013	20571	1745

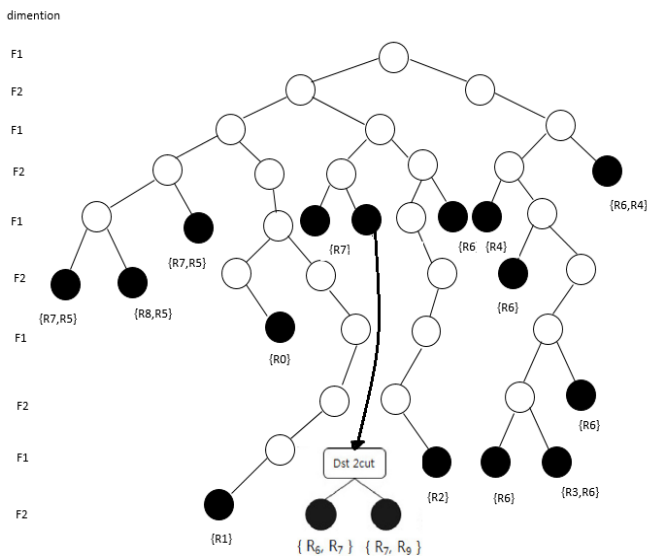


Fig. 6. The structure of the leaf-pushing KD-tree in Figure (3) as modified by means of HiCuts.

match, it is returned as the best matching rule (line 7). In this step, if there are several rules in the node, they are searched linearly to find the best matching rule. In this example, the search is finished only by comparing one rule. Comparison with R6 and R7 is avoided because they lie in the leaves. Lines 18 to 22 are executed when none of the rules in the tree match the packet. In this case, the packet is matched linearly against a list of rules in which both input fields have wild-card values and which have already been ordered by priority.

Fig. 4 compares the architecture of conventional KD-tree and leaf-pushed tree. It should be noted that we keep the

KD-tree in the on-chip memory and the database in the off-chip memory due to its large size. When a node containing a rule is observed in a KD-tree, the algorithm is referred to the memory whereas, in a leaf-pushed tree, the entire search process is performed within the on-chip memory. The pointer obtained in this search is used to access the off-chip memory which keeps the classifier’s database.

C. Generating a Leaf-Pushed KD-Tree by Using a Bloom Filter

In this section, we shall introduce a useful method for implementing a leaf-pushing KD-tree. Characteristic of this tree is that all the nodes that contain rules lie at the last level. Thus, an efficient search method is to use a Bloom filter and a hash table. Fig. 5 illustrates the proposed method which makes use of a Bloom filter, a hash table, and a rule database.

The Bloom filter is responsible for determining whether or not each input substring has a corresponding node in the tree. Therefore, the Bloom filter should be applied to all the nodes that contain rules in a leaf-pushing KD-tree.

First, the length of prefixes in the tree is sorted in a descending order and represented using vectors. Then a substring with the same length as the longest prefix in the tree is retrieved from the source and destination address prefixes of the packet and a query is sent to the Bloom filter. If the result is positive, the node with this prefix length contains a rule that matches the input. As a Bloom filter never produces false negatives, a negative result means that there is no node with the current length. Afterwards, further queries will be sent to the Bloom filter as the length of the input substring is being reduced down to smaller lengths in the prefix vector. This will continue until

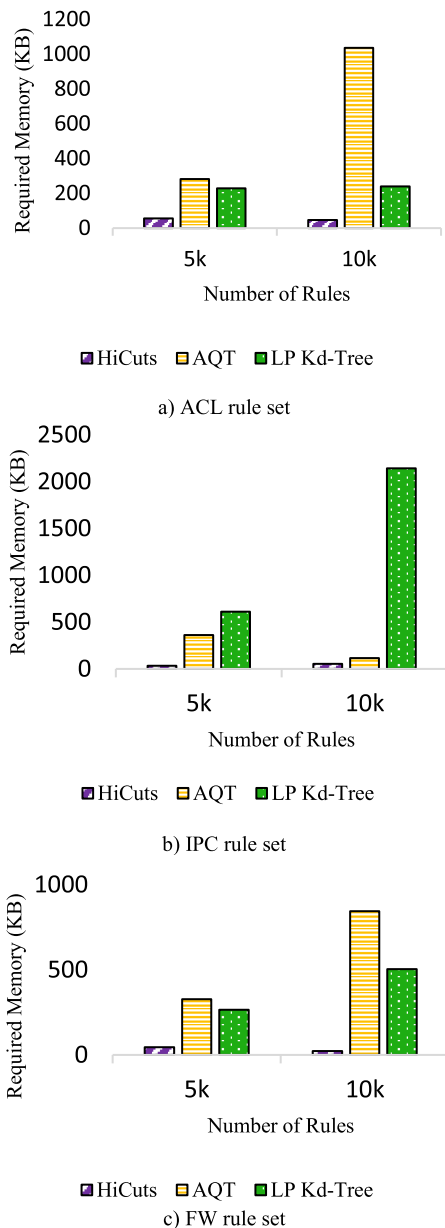


Fig. 7. Comparison of memory access among the proposed algorithm, HiCuts, and AQT.

a positive result is obtained. In this way, the search proceeds only by querying the Bloom filter. The role of the hash table is to provide a pointer to possibly matching rules in the database. For this purpose, every rule node must be stored in the hash table.

For example, let us assume the input packet (01100, 01100, 161, 1711, 6). In the tree in Fig. 3, the vector of prefix lengths is <3, 4, 5, 6, 7, 8, and 9>. The pseudo code for Bloom filter search is shown in Algorithm 2. The input to this function is our example packet. The output of the function is the best matching rule (BMR). The Bloom filter programmed according to the nodes of the tree in Fig. 3 will return a positive result (line 3 in Algorithm 2) for the substring 001111000*. Suppose that the probability of false positive results is sufficiently small. Using the substring 001111000

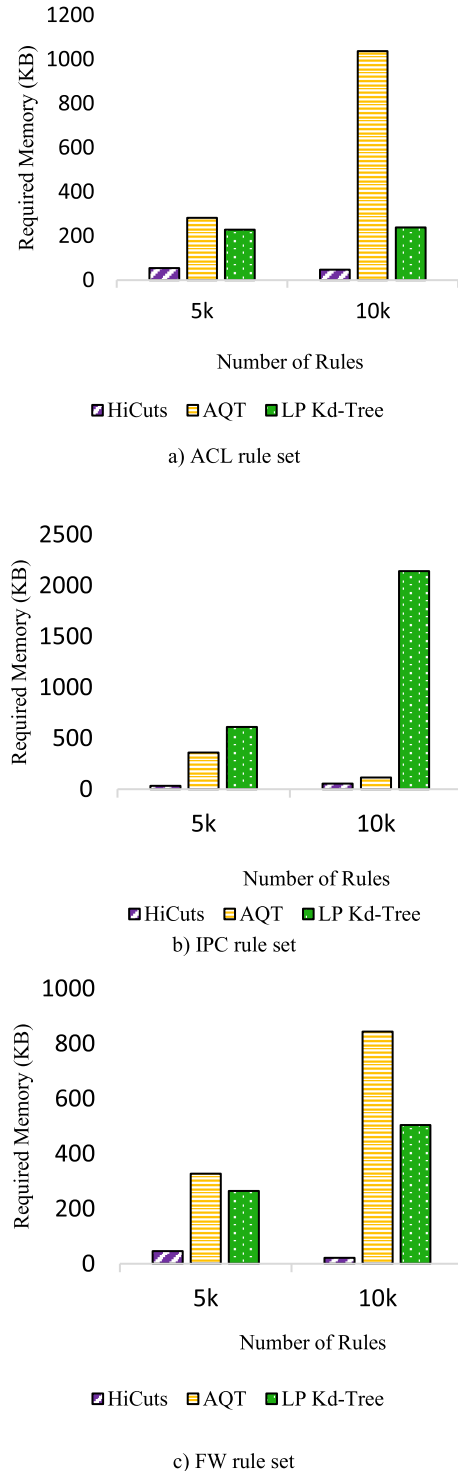


Fig. 8. Comparison of memory usage among the proposed algorithm, HiCuts, and AQT.

(which is a positive substring) as a hash key, the hash table is accessed (line 4). By obtaining a pointer from the hash table, R1 is accessed. Next, R1 is compared with the packet header and, if they match, it is returned as the best matching rule (line 6).

D. Modification of the Algorithm

Space decomposition algorithms such as HyperCuts [28], HiCuts [27], and BC [31] are controlled by a predetermined

TABLE III
COMPARISON OF THE MEMORY USED BY KD-TREE AND THE PROPOSED LEAF-PUSHED KD-TREE

Dataset	# Rules		KD-tree		Leaf-pushed KD-tree			
			On-chip	Off-chip	On-chip		Off-chip	
			M_i (KB)	M_r (MB)	M_b (KB)	M_h (KB)	Total (KB)	M_r (MB)
ACL	5K	4834	76	0.09	64	143	207	3.4
	10K	9835	119	0.185	128	186	314	5.42
	50K	49220	47	0.938	32	106	138	1.52
	100K	97450	51.13	1.87	32	111	142	2.08
IPC	5K	4731	123.37	0.08	128	220	348	4.28
	10K	9533	265.34	0.179	256	320	576	5.365
	50K	32211	21.25	0.61	16	50	66	7.64
	100K	57104	23.9	1.089	16	52	67	12.8
FW	5K	4710	93	0.088	128	392	519	2.9
	10K	9387	232.22	0.176	256	811	1066	3.96
	50K	32578	63	0.617	64	210	274	8.54
	100K	44828	6.8	0.855	4	18	22	8.54

Algorithm 2 The Pseudo Code for Searching by the Proposed Algorithm Using a Bloom Filter

Input: *packet in_pkt*

Output: *rules R*

```

1: function SearchWithBF (n_pkt)
2:   BMR = default
3:   BML = SearchBF (n_pkt)
4:   rulePtr = SearchHASH (BML)
5:   //rule database search
6:   BMR = linearSearch(in_pkt, rulePtr)
7:   //search for wildcard rules
8:   if (BMR > wild.threshold) then
9:     | BMR = linearSearch(in_pkt);
10:  end if
11:  return BMR
12: end function

```

proposed in the previous section is presented here so as to reduce the number of memory accesses. In this modified structure, a HiCuts tree is produced for each leaf node in which the number of rules is greater than the value of *binth*.

In other words, the space covered by each node in the leaf-pushing KD-tree is partitioned to make the number of rules in a decomposition space equal to or smaller than the value of *binth*. Fig. 6 represents the modified structure of the leaf-pushing KD-tree in Fig. 3, with *binth* set to 2. The space separated by the node 0101* contain three rules, which is greater than *binth*. As a result, this space is partitioned by HiCuts.

IV. IMPLEMENTATION AND EVALUATION

The proposed algorithm was implemented using C++ and Classbench Suite [32]. Two of the most important criteria used in the evaluation of packet classification algorithms include search time (which is directly related to the number of memory accesses) and memory usage. In our discussion, *N* denotes the number of rules in the database and *W* denotes the maximum prefix length in the rule database. Every rule has *d* dimensions.

A. Classbench

Classbench [33] is a simulator for generating rule sets with any distribution along with headers corresponding to the rules. This software suite can also produce the required packets. It performs this task by using the control information and the input parameters called ‘seed’ which are given to it through a text file. This simulator fulfills the need of the developers of packet classification algorithms for authentic, heterogeneous rules that are found in firewalls, IP chains, and Access Control Lists. In this study, we used three filter sets corresponding to the parameters *Acl2*, *Fw2*, and *Ipc2* with the number of rules being 5k, 10k, 50k, and 100k.

B. Metrics

In this section, the efficiency of the suggested algorithm is studied from different aspects such as memory required for storing the data structure, complexity of algorithm, and

TABLE IV
COMPARISON OF THE PROPOSED ALGORITHM WITH OTHER ALGORITHMS

Evaluation algorithms	Lookup time	Memory usage	Dimension scalability
Linear search [34]	$O(N)$	$O(N)$	unlimited
Grid-of-tries [34]	$O(W)$	$O(NW)$	2
Cross-producting [35]	$O(dW)$	$O(N^d)$	unlimited
Bit-parallelism [36]	$O(W \log N)$	$O(NW)$	2
Area-based Quad Tree [26]	$O(W)$	$O(NW)$	2
Fat-Inverted Segment [37]	$O((L + 1)W)$	$O(LN^{(1+1/L)})$	2
Segment tree [38]	$O(\log N)$	$O(N * \log N)$	2
RFC [39]	$O(d)$	$O(N^d)$	Unlimited
HiCuts [40]	$O(d)$	$O(N^d)$	Unlimited
Linear search on tuple [41]	$O(W^d)$	$O(N)$	Unlimited
Rectangle search [41]	$O(W)$	$O(NW)$	2
Binary search [34]	$O(\log^2 W)$	$O(N \log^2 W)$	2
Extended Grid-of-Tries [42]	$O(W)$	$O(NW)$	Unlimited
Leaf-pushed KD-tree	$O(d \log W)$	$O(Nd \log W)$	Unlimited

variable called *binth*. The *binth* controls the number of rules in a decomposed space. To provide a similar control mechanism, a modified form of the leaf-pushing KD-tree which was

TABLE V
COMPARISON OF THE SEARCH EFFICIENCY OF KD-TREE AND LEAF-PUSHING KD-TREE IN
TERMS OF THE NUMBER OF MEMORY ACCESSES (A: AVERAGE, W: WORST -CASE)

Dataset	# Rules	# packets	KD-tree				Leaf-pushed KD-tree			
			On-chip		Off-chip		On-chip		Off-chip	
			A_i	W_i	A_r	W_r	A_i	W_i	A_r	W_r
ACL	5K	13980	24	64	74	84	26.3	32	3.5	53
	10K	29205	27	64	164	240	26.8	33	3.9	4
	50K	48420	24	64	365	654	32.1	32	23	208
	100K	95340	26	64	594	742	27	33	124	214
IPC	5K	20610	25	64	182	224	25	31	11	34
	10K	24336	25	64	142	394	25	32	14	30
	50K	49047	28	64	524	642	28	33	43	245
	100K	94370	17	64	348	874	17	32	64	324
FW	5K	9201	7	64	589	784	7	30	89	98
	10K	13054	7	64	320	1247	7	31	20	125
	50K	49163	8	64	2312	5864	8	33	312	438
	100K	82473	9	64	3252	9865	9	30	352	569

maximum number of memory accesses in classifying a typical packet.

C. Evaluation

Table II compares the behavior of conventional KD-tree and leaf-pushing tree. While the number of nodes in the leaf-pushing tree does not show remarkable increase, the number of stored rules has significantly increased. The efficiency of the leaf-pushing tree strongly depends on the type of classifier as well as on the number of rules in the wild-card field because these rules tend to appear in many leaves. The FW rule set has a high rate of rule repetition.

Table III shows the memory required by the KD-tree and the leaf-pushing KD-tree. The size of the required on-chip memory (M_i) is calculated based on the width of a single node which includes the node type field, two child pointers, and one rule pointer. This width is then multiplied by the number of the nodes in the tree. This size is measured in KB, as opposed to the size of the off-chip memory which is measured in MB. As can be observed in the table, the increase in memory size is quite remarkable and the generated tree can be easily stored in the on-chip memory.

Table IV compares the complexity of the proposed algorithm with state-of-the-art algorithms. The total number of tuples in the classifier is Wd . The height of the KD-tree is $\log Wd$ or $d \log W$. The complexity of the structure is equal to the height of the balanced KD-tree, i.e. $O(d \log W)$. For the storage of N filters, the space complexity is $O(Nd \log W)$. Also, Table IV provides a comparison between the proposed algorithm and other classification algorithms. The proposed algorithm has an acceptable performance in terms of time and space complexity.

The average number of queries is related to the tree depth. The number of inputs to each rule set is shown in Table V.

The average number of rule comparisons is obtained by dividing the sum of comparisons for all inputs by the total number of inputs. The worst case of rule comparisons belongs to the input that causes the highest number of comparisons. Our evaluations show that the average number of access to the hash table in our algorithm is 1. The worst case of access

to the hash table is the maximum number of back-tracking as a result of the false positive of the Bloom filter. In this table, the number of accesses to the Bloom filter and the hash table is represented by A_i and W_i , respectively. The number of rule comparisons strongly depends on the type of sets and the features of the tree, particularly in the case of rules in which both prefix fields have the wild-card parameter. For example, the FW rule set has many such rules. As these rules are matched against the inputs after the BMR has been obtained from the leaf-pushing tree, the worst number of rule comparisons can be greater than the maximum number of rules in a leaf node. According to the results of our evaluations, the speedup achieved by the leaf-pushing KD-tree was 1 to 42 times as large as that achieved by the KD-tree. Fig. 7 compares the average number of accesses to the memory in each algorithm which refers to the number of rule comparisons. As can be seen in the figure, the proposed algorithm had a better performance in most of the sets in comparison with other algorithms. The reason is that in a Bloom filter with a remarkably low amount of error, access to the hash table is minimized. Moreover, since the numbers are sorted by their priority in the rule set, the number of rule comparisons is reduced as a result of decreased memory access. It can be seen in the figure that the number of memory accesses in the AQT algorithm has been reduced from 23 to 1.

In Fig. 8, the memory usage of the proposed modified structure is compared with that of HiCuts and AQT. Memory usage is directly related to the repetition of rules. The proposed modified structure can also be stored in an on-chip memory. Even if the on-chip memory is not sufficient, the significant reduction in the number of rule comparisons makes it possible to store the rule database in an off-chip memory without any concern about decrease in efficiency. As mentioned earlier, the number of stored rules has increased in the proposed method. The efficiency of the leaf-pushing tree strongly depends on the type of classifier as well as on the number of rules in the wild-card field because these rules tend to appear in many leaves. As the rate of rule repetition in FW and IPC rule sets is high, the memory usage of the proposed algorithm increases in these classifiers. As can be seen in the figure, the

memory usage of the algorithm is acceptable and there is a 77-percent reduction in comparison with AQT.

V. CONCLUSION

Software-defined intelligent vehicular networks require fast packet classification algorithms to provide several flow-based surveillance services to mobile applications on vehicular nodes. This requirement emerges when the scale of such networks grows exponentially and consequently results in a considerable delay in processing big streams of network packets to/from vehicular nodes. Using appropriate packet classification methods and enhancing their speed is a key solution to this problem.

In this paper, we first described the implementation of KD-tree which is an algorithm for packet classification and then discussed the structure of leaf-pushing tree. By using leaf-pushing, the prefix information in longer prefixes would significantly reduce the number of rules in a search path. The rules are kept only in leaf nodes. We showed that leaf-pushing technique can be efficiently used to separate the search process from the process of rule matching. To improve the performance of a previously generated tree, we used a Bloom filter and a hash table. The Bloom filter is used in our proposed method to search for a node that contains a rule that matches an incoming packet. The function of the hash table is to provide a pointer to the rule database when a node has been found to contain a matching rule. Finally, we also proposed a modified structure for our leaf-pushing KD-tree to enhance its performance and reduce the number of accesses to the off-chip memory.

We evaluated our method in terms of memory usage and memory access. Although the required memory increased only slightly, a significant improvement was observed in memory access. The obtained speedup is indicative of the efficiency of the proposed method. We compared the implementation results with other algorithms for geometric space decomposition such as AQT and HiCuts. The comparison proves that our modified structure is significantly more efficient in reducing the number of memory accesses. Our method could reduce this number from 23 to 1 and its memory usage was comparable to other algorithms.

To continue this research, parallel platforms like GPUs can be used for parallelization of the packet classification process. Given the larger number of computational cores in GPUs, it is predictable that the parallelization of the proposed algorithm would be expressively optimized on GPUs.

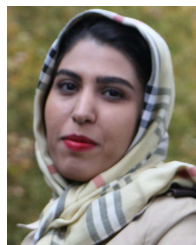
REFERENCES

- [1] M. Sredynski, G. Arnould, and D. Khadraoui, "The emerging applications of intelligent vehicular networks for traffic efficiency," in *Proc. 3rd ACM Int. Symp. Design Anal. Intell. Veh. Netw. Appl. (DIVANet)*, 2013, pp. 101–108.
- [2] M. B. Younes and A. Boukerche, "Safety and efficiency control protocol for highways using intelligent vehicular networks," *Comput. Netw.*, vol. 152, pp. 1–11, Apr. 2019.
- [3] J. Cheng, J. Cheng, M. Zhou, F. Liu, S. Gao, and C. Liu, "Routing in Internet of vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 5, pp. 2339–2352, Oct. 2015.
- [4] Z. Zhou, X. Chen, Y. Zhang, and S. Mumtaz, "Blockchain-empowered secure spectrum sharing for 5G heterogeneous networks," *IEEE Netw.*, vol. 34, no. 1, pp. 24–31, Jan. 2020.
- [5] W. Xu, H. Zhou, H. Wu, F. Lyu, N. Cheng, and X. Shen, "Intelligent link adaptation in 802.11 vehicular networks: Challenges and solutions," *IEEE Commun. Standards Mag.*, vol. 3, no. 1, pp. 12–18, Mar. 2019.
- [6] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," *Proc. IEEE*, vol. 108, no. 2, pp. 292–307, Feb. 2020.
- [7] X. Lin, J. Wu, S. Mumtaz, S. Garg, J. Li, and M. Guizani, "Blockchain-based on-demand computing resource trading in IoV-assisted smart city," *IEEE Trans. Emerg. Topics Comput.*, early access, Feb. 6, 2020, doi: 10.1109/TETC.2020.2971831.
- [8] G. Raja, A. Ganapathisubramanian, S. Anbalagan, S. B. M. Baskaran, K. Raja, and A. K. Bashir, "Intelligent reward-based data offloading in next-generation vehicular networks," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 3747–3758, May 2020.
- [9] L. Nkenyereye, L. Nkenyereye, S. M. R. Islam, Y.-H. Choi, M. Bilal, and J.-W. Jang, "Software-defined network-based vehicular networks: A position paper on their modeling and implementation," *Sensors*, vol. 19, no. 17, p. 3788, Aug. 2019.
- [10] F. A. Silva, A. Boukerche, T. R. M. B. Silva, E. Cerqueira, L. B. Ruiz, and A. A. F. Loureiro, "Information-driven software-defined vehicular networks: Adapting flexible architecture to various scenarios," *IEEE Veh. Technol. Mag.*, vol. 14, no. 1, pp. 98–107, Mar. 2019.
- [11] J. Bhatia, Y. Modi, S. Tanwar, and M. Bhavsar, "Software defined vehicular networks: A comprehensive review," *Int. J. Commun. Syst.*, vol. 32, no. 12, p. e4005, Aug. 2019.
- [12] J. C. Nobre *et al.*, "Vehicular software-defined networking and fog computing: Integration and design principles," *Ad Hoc Netw.*, vol. 82, pp. 172–181, Jan. 2019.
- [13] T. Ganegedara and V. K. Prasanna, "StrideBV: Single chip 400G+ packet classification," in *Proc. IEEE 13th Int. Conf. High Perform. Switching Routing (HPSR)*, Jun. 2012, pp. 1–6.
- [14] P. Gupta and N. McKeown, "Algorithms for packet classification," *IEEE Netw.*, vol. 15, no. 2, pp. 24–32, Mar./Apr. 2001.
- [15] C.-L. Hsieh and N. Weng, "Scalable many-field packet classification using multidimensional-cutting via selective bit-concatenation," in *Proc. 11th ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, May 2015, pp. 187–188.
- [16] W. Jiang and V. K. Prasanna, "Scalable packet classification on FPGA," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 20, no. 9, pp. 1668–1680, Sep. 2012.
- [17] Y. R. Qu, H. H. Zhang, S. Zhou, and V. K. Prasanna, "Optimizing many-field packet classification on FPGA, multi-core general purpose processor, and GPU," in *Proc. ACM/IEEE Symp. Archit. Netw. Commun. Syst. (ANCS)*, May 2015, pp. 87–98.
- [18] B. S. Tumari and W. LakshmiPriya, "FPGA implementation of binary-tree-based high speed packet classification system," *Int. J. Combined Res. Develop.*, vol. 2, no. 6, pp. 17–22, Jun. 2014.
- [19] K. Zheng, H. Che, Z. Wang, and B. Liu, "TCAM-based distributed parallel packet classification algorithm with range-matching solution," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Mar. 2005, pp. 293–303.
- [20] K. Zheng, H. Che, Z. Wang, B. Liu, and X. Zhang, "DPPC-RE: TCAM-based distributed parallel packet classification with range encoding," *IEEE Trans. Comput.*, vol. 55, no. 8, pp. 947–961, Aug. 2006.
- [21] Z. Cao, M. Kodialam, and T. V. Lakshman, "Traffic steering in software defined networks: Planning and online routing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, pp. 65–70, Feb. 2015.
- [22] K. G. Perez, X. Yang, S. Scott-Hayward, and S. Sezer, "A configurable packet classification architecture for software-defined networking," in *Proc. 27th IEEE Int. Syst.-Chip Conf. (SOCC)*, Sep. 2014, pp. 353–358.
- [23] S. Han, K. Jang, K. Park, and S. Moon, "PacketShader: A GPU-accelerated software router," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 195–206, 2011.
- [24] K. G. Perez, X. Yang, S. Scott-Hayward, and S. Sezer, "Optimized packet classification for software-defined networking," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 859–864.
- [25] H. Lim, Y. Choe, M. Shim, and J. Lee, "A quad-trie conditionally merged with a decision tree for packet classification," *IEEE Commun. Lett.*, vol. 18, no. 4, pp. 676–679, Apr. 2014.
- [26] H. Lim, M. Y. Kang, and C. Yim, "Two-dimensional packet classification algorithm using a quad-tree," *Comput. Commun.*, vol. 30, no. 6, pp. 1396–1405, Mar. 2007.
- [27] D. Pao and Z. Lu, "A multi-pipeline architecture for high-speed packet classification," *Comput. Commun.*, vol. 54, pp. 84–96, Dec. 2014.

- [28] S. Singh, F. Baboescu, G. Varghese, and J. Wang, "Packet classification using multidimensional cutting," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun.*, 2003, pp. 213–224.
- [29] V. Srinivasan and G. Varghese, "Fast address lookups using controlled prefix expansion," *ACM Trans. Comput. Syst.*, vol. 17, no. 1, pp. 1–40, Feb. 1999.
- [30] J. Lee, H. Byun, J. H. Mun, and H. Lim, "Utilizing 2-D leaf-pushing for packet classification," *Comput. Commun.*, vol. 103, pp. 116–129, May 2017.
- [31] H. Lim, N. Lee, G. Jin, J. Lee, Y. Choi, and C. Yim, "Boundary cutting for packet classification," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 443–456, Apr. 2014.
- [32] D. E. Taylor and J. S. Turner, "ClassBench: A packet classification benchmark," *IEEE/ACM Trans. Netw.*, vol. 15, no. 3, pp. 499–511, Jun. 2007.
- [33] D. E. Taylor and J. S. Turner, "ClassBench: A packet classification benchmark," in *Proc. 24th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Mar. 2005, pp. 2068–2079.
- [34] D. E. Taylor, "Survey and taxonomy of packet classification techniques," *ACM Comput. Surv.*, vol. 37, no. 3, pp. 238–275, Sep. 2005.
- [35] V. Srinivasan, G. Varghese, S. Suri, and M. Waldvogel, "Fast and scalable layer four switching," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, 1998, pp. 191–202.
- [36] T. V. Lakshman and D. Stiliadis, "High-speed policy-based packet forwarding using efficient multi-dimensional range matching," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 28, no. 4, pp. 203–214, Oct. 1998.
- [37] A. Feldman and S. Muthukrishnan, "Tradeoffs for packet classification," in *Proc. IEEE Conf. Comput. Commun., 19th Annu. Joint Conf. IEEE Comput. Commun. Soc. (INFOCOM)*, Mar. 2000, pp. 1193–1202.
- [38] C.-F. Su, "High-speed packet classification using segment tree," in *Proc. IEEE Global Telecommun. Conf. (Globecom)*, vol. 1, Nov./Dec. 2000, pp. 582–586.
- [39] P. Gupta and N. McKeown, "Packet classification on multiple fields," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, 1999, pp. 147–160.
- [40] P. Gupta and N. McKeown, "Packet classification using hierarchical intelligent cuttings," in *Proc. Hot Interconnects*, 1999, pp. 1–9.
- [41] V. Srinivasan, S. Suri, and G. Varghese, "Packet classification using tuple space search," in *Proc. Conf. Appl., Technol., Archit., Protocols Comput. Commun. (SIGCOMM)*, 1999, pp. 135–146.
- [42] H. Lu and S. Sahni, " $O(\log W)$ multidimensional packet classification," *IEEE/ACM Trans. Netw.*, vol. 15, no. 2, pp. 462–472, Apr. 2007.



Mahdi Abbasi received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from the Sharif University of Technology, Tehran, Iran, and the University of Isfahan, Isfahan, Iran, respectively. He is currently with the Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan, Iran. His research interests include computer architecture, signal and image processing, machine learning, the Internet of Things (IoT), and computer networks.



Hajar Rezaei received the M.Sc. degree in computer engineering (computer networks) from Bu-Ali Sina University, Hamedan, Iran, in 2019. Her research interests include computer networks, software defined networking, and the Internet of Things (IoT).



Varun G. Menon (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Sathyabama University, India, in 2017. He is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. His research interests include sensors, the IoT, fog computing, and underwater acoustic sensor networks. He is also a Distinguished Speaker of ACM.



Lianyong Qi (Member, IEEE) received the Ph.D. degree from the Department of Computer Science and Technology, Nanjing University, China, in 2011. He is currently an Associate Professor with the School of Information Science and Engineering, Qufu Normal University, China. He has already published more than 50 articles, including JSAC, TCC, TBD, FGCS, JCSS, CCPE, and ICWS. His research interests include services computing, big data, and the IoT.



Mohammad R. Khosravi is currently with the Department of Computer Engineering, Persian Gulf University, Iran. His main interests include statistical signal and image processing, medical bioinformatics, radar imaging and satellite remote sensing, computer communications, industrial wireless sensor networks, underwater acoustic communications, information science, and scientometrics.



A trust analysis scheme for vehicular networks within IoT-oriented Green City

Geetanjali Rathee^a  , Naveen Jaglan^b , Razi Iqbal^c , Sujesh P. Lal^d , Varun G. Menon^e 

Show more 

 Share  Cite

<https://doi.org/10.1016/j.eti.2020.101144> 

[Get rights and content](#) 

Abstract

A smart city refers to an intelligent environment obtained by deploying all available resources and recent technologies in a coordinated and smart manner. Intelligent Transportation System (ITS) is used for developing and maintaining a sustainable and environmental friendly transportation system leading to the evolution of the concept of a Green City. The expansion and development of green cities goes side by side with the development of smart cities. Internet-of-Vehicles (IoV) is an advanced version of transportation mechanisms to enhance the benefits mentioned above. Several authors have proposed various security transportation schemes, however, the response time and communication overhead still need to be considered to encourage business developments, and expanding the recycling's. In this paper, we have proposed an intelligent transportation mechanism that categorizes each device into various categories depending upon its communication behavior using decision based tree making system. Additionally, an Intrusion Detection System (IDS) method is proposed to further examine and analyze the continuous flow of data. The proposed mechanism is further validated through MATLAB simulator against number of metrics over a baseline method. The simulated results suggest that the proposed scheme leads to 89% efficiency in terms of better detection of legitimate nodes, altered records, records accuracy and response time during sensing of vehicles.

Introduction

The accelerated pace of urbanization in the recent decades has endangered the environment and economic sustainability by raising several social, technical and economic concerns. Therefore, for exploiting and optimizing the tangible and intangible assets, governments across the world have been taking an interest in adopting the concept of smart cities and their related infrastructure (Lombardi et al., 2012). A smart city refers to an intelligent environment obtained by deploying all available resources and technologies in a coordinated and smart manner with the means of developing urban centers (Neirotti et al., 2014, Mehmood et al., 2017). Traffic management, smart industry, smart grid, smart healthcare, public safety, secure e-voting, Intelligent Transportation Systems (ITS) and water management are the various developments related to smart cities that are being retro fired with Internet-of-Things (IoT) and smart sensors (Gubbi et al., 2013, Lee and Lee, 2015, Paul et al., 2016a, Paul and Jeyaraj, 2019). Another notion gaining prominence in the contemporary world is the concept of 'Green city' (Breuste, 2020). These cities are designed in a way to most efficiently handle traffic and over population and lessen their environment impact by reducing waste, expanding recycling, lowering emissions, increasing housing density and encouraging

the development of sustainable local businesses (Antrobus, 2011). The expansion and development of green cities goes side by side with the development of smart cities. The intelligent devices employed in smart cities are equally usable in green cities as they provide necessary tools to the authorities for handling the infrastructure requisite for green cities like recycling plants, emission control of factories in the city vicinity and etc. (Shabandri et al., 2020, Bănică et al., 2020). The IoT sensors and devices play a very crucial role in this regard. IoT sensors and recent technologies working mutually are steadily becoming more pervasive and accomplish users' desires more effectively and efficiently. Objects with internet protocol (IP) connectivity are connected to the internet to offer better usability in day-to-day actions. This interconnection amid devices creates a lot of data related to device status, energy usage and environmental behavior that can be aggregated, composed and then disseminated in an adopted, confined and secure manner. Also, as these devices are associated to the internet, they can be controlled at anytime and from everywhere. Therefore, the goal of smart cities is to make best public resource usage with improved services and quality of life by using advanced communication and information technology. Further, in order to ensure a reliable data exchange among devices, there is a need to use massive communication capabilities such as 5G and 6G technologies (Al-Turjman, 2019, Daniel et al., 2017).

Among the various applications of smart cities, we have considered smart vehicular networking with the aim of developing an Intelligent Transportation System (ITS) or a Green Transportation mechanism (Iqbal et al., 2019, Rathee et al., 2020, Paul et al., 2015). ITS can be used for a variety of purposes including efficient management of traffic, road congestion, road accidents, accurate mapping and positioning and monitoring the vehicular emissions to check on air pollution (Abbasi et al., 2020, Paul et al., 2016b). The daily life routine of human beings is being completely or partially replaced by automated or cognitive systems in various aspects. From home automation systems to smart industries and intelligent transportation, cognitive systems monitor or control each and every activity of environment without any human intervention. Now a days, VANET (Qu et al., 2020, Menon and Prathap, 2017) has become an advanced research area with the development of intelligent communication strategies.

A VANET is generally a combination of various stationary and mobile vehicles connected through several smart networks. The intelligent vehicular system is changing the view of transportation and communication where vehicles are increasingly communicating through various IoT sensors. Intelligent transportation mechanism is considered as the most emerging technology where number of devices are connected through internet to ensure an improved and secure communication mechanism. An Intelligent Transportation System (ITS) (Hussain et al., 2020, Daniel et al., 2016, Balasubramaniam et al., 2020) is defined as an intelligent way of ensuring an efficient and secure communication where devices may retrieve data through roadside sensors as depicted in Fig. 1. It depicts an example of ITS mechanism where vehicles communicate with each other through various smart sensors for accessing essential information related to road congestion, weather predictions, shorter routes etc. However, the utilization of IoT devices integrated with vehicular systems is still avoided by business organizations due to various safety concerns and associated costs of centralized firms and cloud servers (Nguyen et al., 2016, Erdinc et al., 2009).

In traditional vehicular techniques, communication without any smart devices raises concern regarding large number of road accidents and an efficient management of traffic. During traveling, the driver may sometimes fall asleep or may get stuck into a traffic jam in an accidental or natural disaster prone area among other reasons. Further, delay is considered to be a major concern where without any prior information of jams people may get stuck on roads for several hours. Though number of intelligent GPS based systems has been proposed by various researchers, delay due to lack of real-time information may lead to inefficient systems. Environmental issues related to vehicular emissions and pollution have been a major cause of concern in the recent years. This has led to large scale research for the development of cities called 'Green cities' that can efficiently manage pollution levels in a sophisticated manner by making use of methods like efficient management of traffic among other things. ITS can be a great contributing factor toward the development of these cities. Sensors embedded in vehicles and across the city may be used to check emissions from vehicles and pollution levels in various parts of the city. However, systems where a large number of smart devices work in a synchronous manner can always be potential targets for hackers which may affect the overall communication process. A number of intelligent vehicular techniques and security procedures have been proposed by several authors such as cryptographic, homomorphic, probabilistic etc., however, these existing solutions may further leads to increase in computational and communication complexity and cost inside the network. Therefore, it is further needed to propose a secure intelligent communication

mechanism in vehicular systems that is still in its early stages (Rajesh et al., 2019). In order to propose a secure and intelligent transportation mechanism for vehicular networks in green city, we have used a decision based tree system where upon transmission and receiving of information, smart devices are categorized into various entities such as malicious, trusted or prone to malicious devices. Further, the categorized devices are again analyzed continuously to examine their behavior in the communication process. All categories of nodes are examined after a specific interval of time to check their further behavior through IDS. However, the devices once identified as malicious will be permanently blocked by an intrusion detection system (IDS)(Antrobus, 2011, Shabandri et al., 2020, Bănică et al., 2020) to secure and speed up the communication mechanism in real time scenarios. The potential contribution of our paper is as follows.

1. Proposing a secure green transportation mechanism using device based tree system to analyze and categorize the vehicles into various categories.
2. Design of an IDS mechanism to further keep a check upon trusted and threat prone devices by examining their various communication parameters.
3. A proposed threshold value identity decided to compare and categorize the various devices during communication in green transportation system.

The remaining structure of the paper is organized as follows. Section2 discusses various security techniques in ITS with and without blockchain phenomenon. An intelligent vehicular mechanism using detection based tree and an IDS is elaborated in Section3. In addition, number of security measures such as response time, resource utilization, request processes and record accuracy are compared for an existing approach and the proposed phenomenon in Section4. Finally, Section5 concludes the paper and suggests a number of points that may be considered for future communication.

Section snippets

Related work

Number of researchers have proposed various security solutions in ITS environment for ensuring an efficient communication in various environments like smart cities and green cities. This section illustrates the number of security schemes proposed by various scientists. Now days, road accidents are increasing tremendously due to various reasons, it is very much needed to focus on this issue and propose several intelligent transportation mechanisms to secure or prevent from road attack. Numbers...

Proposed framework

In order to speed up the communication process, it is very essential to provide a secure procedure so that vehicles may interact effectively and share the information among each other safely. In addition, such efficient communication mechanisms are also required for various environments such as smart cities and green cities. Trust is considered as a very important factor to measure the security among nodes/vehicles. The computation of trust among communicating nodes not only provides an...

System state

In order to validate or identify the truly legitimate IoT devices, the proposed mechanism is simulated using MATLAB simulator. The numerical results are analyzed by examining the security environment against various network security metrics. Though, result measures in ITS application are considered to be crucial, however, the paper proposes a secure intelligent transportation system which not only ensures secure communication through legitimate devices but also provides a reliable and delay...

Conclusion

The development and expansion of green cities goes side by side with the development of smart cities. Green transportation mechanism can be used for a variety of purposes including efficient management of traffic, accurate mapping and checking vehicular emissions. In this paper, we have proposed a secure transportation mechanism based upon decision based tree model with an intrusion detection system and is analyzed against various security parameters. The proposed mechanism efficiently reduces...

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper...

References (40)

Al-TurjmanF.

[5g-enabled devices and smart-spaces in social-IoT: an overview](#)

Future Gener. Comput. Syst. (2019)

DanielA. *et al.*

[Big autonomous vehicular data classifications: Towards procuring intelligence in ITS](#)

Veh. Commun. (2017)

ErdincO. *et al.*

[A wavelet-fuzzy logic based energy management strategy for a fuel cell/battery/ultra-capacitor hybrid vehicular power system](#)

J. Power Sources (2009)

GaberT. *et al.*

[Trust-based secure clustering in WSN-based intelligent transportation systems](#)

Comput. Netw. (2018)

Gubbij. *et al.*

[Internet of things \(IoT\): A vision, architectural elements, and future directions](#)

Future Gener. Comput. Syst. (2013)

Leel. *et al.*

[The internet of things \(IoT\): Applications, investments, and challenges for enterprises](#)

Bus. Horizons (2015)

NeirottiP. *et al.*

[Current trends in smart city initiatives: Some stylised facts](#)

Cities (2014)

RatheeG. *et al.*

[A trust management scheme to secure mobile information centric networks](#)

Comput. Commun. (2020)

RatheeG. *et al.*


[A secure communicating things network framework for industrial IoT using blockchain technology](#)

Ad Hoc Netw. (2019)

AbbasiM. *et al.*

An efficient parallel genetic algorithm solution for vehicle routing problem in cloud implementation of the intelligent transportation systems

J. Cloud Comput. (2020)

 View more references

Cited by (10)

[A secure emotion aware intelligent system for Internet of healthcare](#)

2023, Alexandria Engineering Journal

[Show abstract](#) 

[AutoTrust: A privacy-enhanced trust-based intrusion detection approach for internet of smart things](#)

2022, Future Generation Computer Systems

Citation Excerpt :

...IoT has the potential to facilitate human beings with considerable services, which may completely transform the world. IoT is being merged with several areas, such as healthcare [3], agriculture [4], smart ecosystems [5], vehicular networks [6], smart grids [7], etc. CoT uses the same concept of connected IoT devices having access to the cloud [8]...

[Show abstract](#) 

[Agreement-Induced Data Verification Model for Securing Vehicular Communication in Intelligent Transportation Systems](#)

2023, IEEE Transactions on Intelligent Transportation Systems

[Inclusive green growth for sustainable development of cities in China: spatiotemporal differences and influencing factors](#)


2023, Environmental Science and Pollution Research

[Inclusive Green Growth for Sustainable Development of Cities in China: Spatiotemporal Differences and Influencing Factors](#)

2022, SSRN

[A Recent Survey on 6G Vehicular Technology, Applications and Challenges](#)

2021, 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO 2021

 View all citing articles on Scopus

Recommended articles (6)

Research article

[Triple band notched mushroom and uniplanar EBG structures based UWB MIMO/Diversity antenna with enhanced wide band isolation](#)

AEU - International Journal of Electronics and Communications, Volume 90, 2018, pp. 36-44

[Show abstract](#) 

Research article

Efficacy of incorporating PCM into the building envelope on the energy saving and AHU power usage in winter

Sustainable Energy Technologies and Assessments, Volume 43, 2021, Article 100969

[Show abstract](#) 

Research article

A trusted effective approach for forecasting the failure of data link and intrusion in wireless sensor networks

Theoretical Computer Science, Volume 941, 2023, pp. 1-13

[Show abstract](#) 

Research article

Energy simulation of residential house integrated with novel IoT windows and occupant behavior

Sustainable Cities and Society, Volume 78, 2022, Article 103594

[Show abstract](#) 

Research article

Optimizing the mobility management task in networks of four world capital cities

Journal of Network and Computer Applications, Volume 51, 2015, pp. 18-28

[Show abstract](#) 

Research article

Machine learning based trust management framework for vehicular networks

Vehicular Communications, Volume 25, 2020, Article 100256

[Show abstract](#) [View full text](#)

© 2020 Elsevier B.V. All rights reserved.



Copyright © 2023 Elsevier B.V. or its licensors or contributors.
ScienceDirect® is a registered trademark of Elsevier B.V.



All ▾

[Journals & Magazines](#) > [IEEE Transactions on Network Science and Engineering](#) > [Volume: 8 Issue: 4](#) ⓘ

I/Q Imbalance Aware Nonlinear Wireless-Powered Relaying of B5G Networks: Security and Reliability Analysis

Publisher: IEEE

[Cite This](#)[PDF](#)Xingwang Li  ; Mengyan Huang  ; Yuanwei Liu  ; Varun G Menon  ; Anand Paul  ; ... [All Authors](#)

53

Cites in
Papers

695

Full
Text Views

Abstract

Document
Sections

I. Introduction

II. System Model

III. Performance
AnalysisIV. Numerical
ResultsV. Conclusion and
Future Work[Show Full Outline](#)[Authors](#)[Figures](#)[References](#)[Citations](#)[Keywords](#)[Metrics](#)[Footnotes](#)

Abstract:

Physical layer security is known as a promising paradigm to ensure secure performance for the future beyond 5G (B5G) networks. In light of this fact, this paper elaborates on a tractable analysis framework to evaluate the reliability and the security of wireless-powered decode-and-forward (DF) multi-relay networks. More practical, the nonlinear energy harvesters, in-phase and quadrature-phase imbalance (IQI) and channel estimation errors (CEEs) are taken into account. To further enhance the secure performance, two relay selection strategies are presented: 1) suboptimal relay selection (SRS); 2) optimal relay selection (ORS). Specifically, exact analytical expressions for the outage probability (OP) and the intercept probability (IP) are derived in closed-form. For the IP, we consider that the eavesdropper can wiretap the signal from the source or the relay. In order to obtain more deep insights, we carry out the asymptotic analysis as well as the diversity orders for the OP in the high signal-to-noise ratio (SNR) regimes. Numerical results show that: 1) Although the mismatches of amplitude/phase of transmitter (TX)/receiver (RX) limit the OP performance, it can enhance IP performance; 2) Large number of relays yields better OP performance; 3) There are error floors for the OP due to the CEEs; 4) There is a trade-off for the OP and IP to obtain the balance between reliability and security.

Published in: [IEEE Transactions on Network Science and Engineering](#) (Volume: 8 , Issue: 4, 01 Oct.-Dec. 2021)

Page(s): 2995 - 3008

DOI: 10.1109/TNSE.2020.3020950

Date of Publication: 03 September 2020 Publisher: IEEE



ISSN Information:


Electronic ISSN: 2327-4697

CD: 2334-329X

[Home](#) > [Journal of Real-Time Image Processing](#) > [Article](#)

Special Issue Paper | [Published: 28 September 2020](#)

SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network

[Noman Khan](#), [Amin Ullah](#), [Ijaz Ul Haq](#), [Varun G. Menon](#) & [Sung Wook Baik](#) 

Journal of Real-Time Image Processing **18**, 1729–1743 (2021)

776 Accesses | **25** Citations | [Metrics](#)

Abstract

The advancements in computer vision-related technologies attract many researchers for surveillance applications, particularly involving the automated crowded scenes analysis such as crowd counting in a very congested scene. In crowd counting, the main goal is to count or estimate the number of people in a particular scene.

Understanding overcrowded scenes in real-time is important for instant responsive actions. However, it is a very difficult task due to some of the key challenges including clutter background, occlusion, variations in human pose and scale, and limited surveillance training data, that are inadequately

covered in the employed literature. To tackle these challenges, we introduce “SD-Net” an end-to-end CNN architecture, which produces real-time high quality density maps and effectively counts people in extremely overcrowded scenes. The proposed architecture consists of depthwise separable, standard, and dilated 2D convolutional layers. Depthwise separable and standard 2D convolutional layers are used to extract 2D features. Instead of using pooling layers, dilated 2D convolutional layers are employed that results in huge receptive fields and reduces the number of parameters. Our CNN architecture is evaluated using four publicly available crowd analysis datasets, demonstrating superiority over state-of-the-art in terms of accuracy and model size.

This is a preview of subscription content, [access via your institution.](#)

Access options

Buy article PDF

39,95 €

Price includes VAT (India)

Instant access to the full article PDF.

[Rent this article via DeepDyve.](#)

**Department of Computer Science and
Engineering, SCMS School of Engineering
and Technology, Ernakulam, 683576, India**

Varun G. Menon

Corresponding author

Correspondence to [Sung Wook Baik](#).

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Khan, N., Ullah, A., Haq, I.U. *et al.* SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network. *J Real-Time Image Proc* **18**, 1729–1743 (2021). <https://doi.org/10.1007/s11554-020-01020-8>

Received

10 May 2020

Accepted

13 September
2020

Published

28 September
2020

Issue Date

October 2021

DOI

<https://doi.org/10.1007/s11554-020-01020-8>

Service Offloading With Deep Q-Network for Digital Twinning-Empowered Internet of Vehicles in Edge Computing

Xiaolong Xu , Bowen Shen , Sheng Ding, Gautam Srivastava , Muhammad Bilal ,
 Mohammad R. Khosravi, Varun G Menon , Mian Ahmad Jan , and Maoli Wang 

Abstract—With the potential of implementing computing-intensive applications, edge computing is combined with digital twinning (DT)-empowered Internet of vehicles (IoV) to enhance intelligent transportation capabilities. By updating digital twins of vehicles and offloading services to edge computing devices (ECDs), the insufficiency in vehicles' computational resources can be complemented. However,

owing to the computational intensity of DT-empowered IoV, ECD would overload under excessive service requests, which deteriorates the quality of service (QoS). To address this problem, in this article, a multiuser offloading system is analyzed, where the QoS is reflected through the response time of services. Then, a service offloading (SOL) method with deep reinforcement learning, is proposed for DT-empowered IoV in edge computing. To obtain optimized offloading decisions, SOL leverages deep Q-network (DQN), which combines the value function approximation of deep learning and reinforcement learning. Eventually, experiments with comparative methods indicate that SOL is effective and adaptable in diverse environments.

Manuscript received September 6, 2020; revised October 28, 2020; accepted November 14, 2020. Date of publication November 24, 2020; date of current version October 27, 2021. This work was supported in part by the Financial and Science Technology Plan Project of Xinjiang Production and Construction Corps under Grant 2020DB005, in part by the NUIST Students' Platform for Innovation and Entrepreneurship Training Program under Grant 202010300024Z, and in part by the National Natural Science Foundation of China under Grant 61702277. Paper no. TII-20-4238. (Corresponding author: Maoli Wang.)

Xiaolong Xu is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China, is with the Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology (CICAET), Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China (e-mail: xlxu@ieee.org).

Bowen Shen is with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: bwshen@nuist.edu.cn).

Sheng Ding is with the Blockchain Laboratory of Agricultural Vegetables, Weifang University of Science and Technology, Weifang, 262700, China (e-mail: dingsheng@wust.edu.cn).

Gautam Srivastava is with the Department of Mathematics and Computer Science, Brandon University, Brandon, MB R7A 6A9, Canada, and also with the Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan (e-mail: srivastavag@brandou.ca).

Muhammad Bilal is with the Department of Computer and Electronics Systems Engineering, Hankuk University of Foreign Studies, Yongin-si 17035, Korea (e-mail: mbilal@kaist.ac.kr).

Mohammad R. Khosravi is with the Department of Computer Engineering, Persian Gulf University, Bushehr 75168, Iran, and also with the Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz 71557-13876, Iran (e-mail: mohammadkhosravi@acm.org).

Varun G Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala 683576, India (e-mail: varunmenon@scmsgroup.org).

Mian Ahmad Jan is with the Department of Computer Science, Abdul Wali Khan University, Mardan 23200, Pakistan (e-mail: mianjan@awkum.edu.pk).

Maoli Wang is with the School of Cyber Science and Engineering, Qufu Normal University, Qufu 273165, China (e-mail: wangml@qfnu.edu.cn).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TII.2020.3040180>.

Digital Object Identifier 10.1109/TII.2020.3040180

Index Terms—Deep reinforcement learning (DRL), digital twinning (DT), edge computing, Internet of vehicles (IoV), service offloading (SOL).

I. INTRODUCTION

THE INTERNET of Vehicles (IoV) is an evolution of vehicular ad hoc networks (VANETs), where vehicles are equipped with a variety of Internet of Things (IoT) equipments and envisioned as intelligent objects [1]. In the IoV, an intelligent vehicle is capable of vehicle to everything (V2X) communication. Specifically, an intelligent vehicle can share information with other vehicles through vehicle to vehicle (V2V) communications. Rather than observing the condition by a single car, V2V enables a broader view by sharing the traffic information observed by multiple vehicles, which can significantly reduce accidents caused by the blind spot [2]. Meanwhile, intelligent infrastructures like roadside units (RSUs) and smart traffic lights are deployed to analyze the vehicles in a specific region, then provide vehicles with external information through vehicle to infrastructure (V2I) communications [3]. Similarly, vehicle to pedestrian (V2P) communication enables vehicles and pedestrians to deliver commands and safety warnings [4]. With V2X communication in the IoV, intelligent vehicles have the potential to adjust the driving status in time and avoid the occurrence of traffic accidents and enhance the users driving experience.

Further, the digital twinning (DT) technology leverages machine learning and IoT technologies to create digital replicas of physical objects. The replica has its properties cloned from their original versions, and constantly update themselves with real-time data from sensors. Empowered by DT technology, a

virtual twin of vehicle in the IoV is generated and mapped to the physical vehicle with IoT technologies [5]. The DT-empowered IoV focuses on collecting the state information of the vehicle and surroundings through the smart sensor devices, and sharing the information with surrounding vehicles and infrastructures [6]. With the collected information, the digital twins are updated constantly to keep consistent with the physical vehicles. Then, through the technologies including augmented reality (AR) simulation and artificial intelligence (AI) predictive analytics, vehicles are provided with enhanced intelligence. Comparing with the traditional IoV, DT-empowered IoV can easily access the digital twins of vehicles instead of applying for and integrating numerous external data sources like the surveillance system and the remote sensing (RS) system. Under such circumstances, the data mining, simulation, and analytics of the IoV can be enhanced by DT.

As most of the collected data in the DT-empowered IoV are in the raw form (i.e., unprocessed images and videos), they cannot be directly used for control and services [7]. Thus, a powerful computing platform is required to refine the massive collected data, then feedback the extracted instructions to vehicles and passengers [8]. Usually, the processing of vehicular data requires technologies such as object detection and AR, which are computationally intensive operations [9]. To extend intelligent vehicles' capabilities, the cloud and edge computing solutions provide DT-empowered IoV with a platform as a service (PaaS) [10]. The data and service requests collected by vehicles are offloaded to the cloud data center through RSU. After data being processed at the cloud infrastructure, the refined data are fed back in the form of instructions or services [11]. Technically, the cloud data center is composed of centralized large-scale computer clusters with high performance. To reduce the cost of construction and facility maintenance, it is usually built in areas far away from end-users. Therefore, service offloading to the cloud will generate high latency during data transmission and is easy to cause bandwidth tension [12]. As a complementary paradigm of cloud computing, edge computing provides appropriate solutions in the DT-empowered IoV by offloading service requests to edge computing devices (ECDs), servers deployed close to vehicles and other end-users, for execution and data extraction [13].

Despite the advantages of fast transmission and sufficient bandwidth resources, edge computing has its own challenges. Considering the distributed manner of ECDs, the computing capacity of each independent ECD is smaller than the cloud data center. Thus, the resources in each ECD are supposed to be fully utilized to attain higher efficiency and quality of service (QoS) [14]. Further, the load balancing in ECD is an important issue, and mishandling of service offloading can cause load imbalance. Consequently, some devices in ECDs would underperform due to excessive service requests, and other would be underutilized. To enhance the performance of edge computing and provide reliable services to passengers, an effective service offloading method is needed in the DT-empowered IoV [15].

For the dynamic offloading control, deep reinforcement learning (DRL) is adopted to evaluate and choose decisions where the collective utilization is optimized [16]. Among the existing DRL

algorithms, the deep Q-network (DQN) has gained attention as a modification of Q-learning, which takes the advantage of temporal-difference learning from reinforcement learning (RL) and the function approximation from deep learning (DL) [17]. In this article, a dynamic service offloading method, named SOL, is proposed based on DQN in edge computing. Specifically, the contributions of this article are as follows.

- 1) Analyze the QoS level of DT-empowered IoV services in respect of response time in a multiuser offloading system.
- 2) Model the ECD as the agent and formalize the state, action, and reward in DRL to optimize the QoS level of the offloading system.
- 3) Apply DQN with experience replay and target network [17] to solve the problem of DT-empowered IoV service offloading in edge computing.
- 4) Conduct comparative experiments with a real-world IoV service dataset to evaluate the effectiveness and adaptability of SOL.

The rest of the article is organized as follows. In Section II, the related work is summarized. In Section III, the model of service offloading in edge computing is described. In Section IV, details of DRL and SOL are presented. Then, in Section V, comparison experiments are conducted. Finally, Section VI concludes this article.

II. RELATED WORK

So far, various applications in the DT-empowered IoV have been proposed to enhance the QoS, safety, and security of transportation [18]. However, the generated data of such applications are large in scale and has much redundancies, therefore not suitable for local computing and existing cloud computing paradigms [19]. Hu *et al.* [20] addressed the scale-sensitive problem of existing object detection, then modified the deep convolutional neural network for vehicle detection with a large variance of scales to guarantee the accuracy and safety in IoV. From another perspective, Liu *et al.* [21] exhibited the outstanding performance of edge computing on enhancing the security and QoS of autonomous vehicles, including extending computing capacity and reducing energy consumption.

The placement of ECDs has great impact on overall performance of edge computing. Zhao *et al.* [22] proposed a ranking-based near-optimal placement algorithm to minimize average access delay through SDN techniques in cloudlets placement. Wang *et al.* [23] studied the ES placement while considering load balancing as well as access delay and adopted mixed integer programming to find the optimal placement. After ECDs are located, task offloading can be taken into operation. He *et al.* [24] gave consideration to users' privacy and system cost in mobile edge computing, and proposed a novel task offloading scheme to enhance user experience. Zhou *et al.* [25] investigated the task offloading under information asymmetry and uncertainty in vehicular fog computing, and proposed a contract optimization to realize the effective server recruitment.

Owing to higher effectiveness of evolutionary algorithms (EAs), researchers widely adopted EAs as a tool for optimizing the offloading problems in edge computing. Guo *et al.* [26]

TABLE I
NOTATIONS AND DEFINITIONS

Notations	Definitions
N	The number of RSUs
M	The number of ECDs
K	The number of vehicles
R	The set of RSUs, $R = \{r_1, r_2, \dots, r_N\}$
E	The set of ECDs, $E = \{e_1, e_2, \dots, e_M\}$
V	The set of vehicles, $V = \{v_1, v_2, \dots, v_K\}$
$D(t)$	The data size of services at time t , $D(t) = \{d_1(t), d_2(t), \dots, d_K(t)\}$
C_e	The coverage of ECD
C_r	The coverage of RSU
$dist$	The distance between two network nodes
RT	The response time of services
S	The QoS level of services

comprehensively investigated the computation offloading as a mix integer nonlinear programming problem, and designed a computation algorithm based on the genetic algorithm and particle swarm optimization to minimize the energy consumption of the user equipment. However, EAs are usually iterative algorithms that find the global optimal solutions by updating the current solutions continuously. Thus, the dependency on global information and the considerable time complexity during the iteration of generations become significant drawbacks [27]. If EAs are adopted for the offloading of each service, the time overhead in controlling can be unaffordable for the practical implementation of edge computing-empowered IoV.

To obtain decentralized and time-efficient control in the IoV, DRL has been adopted in many aspects of the IoV. To achieve high QoS V2V communication, a decentralized resource allocation mechanism based on DRL is designed in [28]. Benefitting from the decentralized manner, DRL can significantly reduce the transmission overhead and the waiting time for global information. Apart from the efficiency, DRL also exhibits the advantage in adaptability. Liang *et al.* [29] adopted DRL to study the automatic determination of traffic signal duration based on the data collected from sensors. In their model, the actions are changes in the duration of a traffic light, and the reward is the difference in cumulative waiting time between two signal cycles. Meanwhile, Zhou *et al.* [30] proposed a DRL-based car-following model, which can make adjustments in driving behaviors under diverse traffic demands, to improve travel efficiency and safety at signalized intersections in real-time. Generally, DRL is promising in achieving distributed control in the dynamic environment of IoV.

III. SYSTEM MODEL AND PROBLEM DEFINITION

This section describes the system model and service offloading in edge computing. Table I presents the key notations and definitions used in this article.

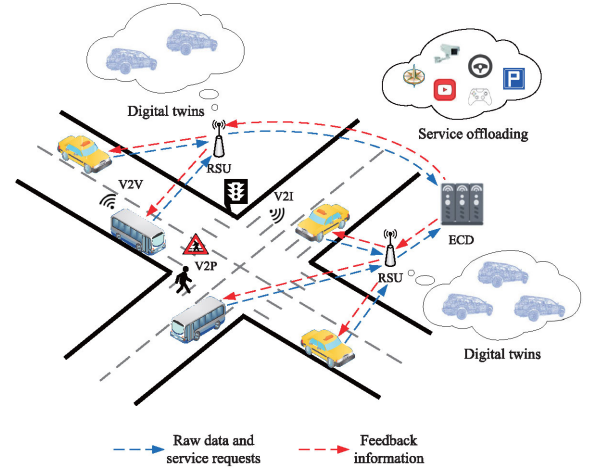


Fig. 1. Framework of service offloading in DT-empowered IoV with edge computing.

A. Framework of Service Offloading for DT-Empowered IoV in Edge Computing

In the proposed framework, vehicles are denoted by set $V = \{v_1, v_2, \dots, v_K\}$. For each vehicle, a digital twin of itself is generated with information of position, speed, vehicle gap, and dashcam videos collected by vehicular sensors and cameras. The raw data and service messages of vehicles can be sent to RSUs, denoted by set $R = \{r_1, r_2, \dots, r_N\}$. With the constant update, we can assume that the cloning is successful, and the functions of the digital twin keep pace with the entity's. Each vehicle can concurrently request one service at time t , and the data to be processed of each vehicular service is denoted by set $D(t) = \{d_1(t), d_2(t), \dots, d_K(t)\}$, while $d_i(t) = 0$ indicates that no service is requested by vehicle v_i . For RSUs are usually considered as communicating nodes and not capable of a large scale of computing tasks, ECDs are arranged to some certain districts to process the service requests based on digital twins of vehicles with massive data collected by RSUs. The ECDs are denoted by the set $E = \{e_1, e_2, \dots, e_M\}$. RSUs can communicate with each other as well as ECDs in their transmission range. Generally, the framework of task offloading in DT-empowered IoV with edge computing is shown in Fig. 1.

In the DT-empowered IoV, the coverage of each ECD is assumed to be the same and denoted by C_e , while for RSUs, the range is denoted by C_r . Then, every RSU, ECD, and vehicle can be, respectively, denoted by

$$r_i (\text{lat}_i, \text{lon}_i, C_r), \quad 1 \leq i \leq N \quad (1)$$

$$e_j (\text{lat}_j, \text{lon}_j, C_e), \quad 1 \leq j \leq M \quad (2)$$

$$v_k (\tilde{\text{lat}}_k(t), \tilde{\text{lon}}_k(t), d_k(t)), \quad 1 \leq k \leq K \quad (3)$$

where lat_n and lon_n represent the latitude and longitude of a network node, respectively, as the location of vehicle is dynamic with time, $(\tilde{\text{lat}}_k(t), \tilde{\text{lon}}_k(t), d_k(t))$ is used to represent the state of v_k at time t .

Based on the latitude and longitude, the distance between two nodes (i.e., RSU, ECD, or cloud access point) can be calculated by the Euclidean distance as

$$\text{dist}(\text{node}_i, \text{node}_j) = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{lon}_i - \text{lon}_j)^2}. \quad (4)$$

It is guaranteed that the data transmission between a vehicle and an RSU, as well as an RSU and an ECD, is a one-hop transmission. Specifically, each RSU is in the coverage of at least one ECD while each vehicle is in the coverage of at least one RSU as

$$\forall r_i \in R, \min_{e_j \in E} \text{dist}(r_i, e_j) \leq C_e \quad (5)$$

$$\forall v_k \in V, \min_{r_i \in R} \text{dist}(v_k(t), r_i) \leq C_r. \quad (6)$$

B. QoS Model of DT-Empowered IoV Services Offloading in Edge Computing

RSUs in the offloading system can independently choose their computing paradigm in each time period, namely, local computing or edge computing. The response time of a service request can be calculated as the sum of offloading time, execution time, and feedback time.

1) Local Computing Model: When vehicle v_k proposes a service request at time t , and locally executes it, the offloading indicator is $a_k(t) = 0$. In this case, local computing yields a response time of $RT_k^l(t)$, which only includes the execution time of the task by vehicular computing units. The execution time is determined by the processing capacity of resource units and the length of data to be executed. Considering that the processing requirements of vehicular services are usually different, a standard measurement is to divide the vehicular processor into multiple resource units with same local computing capacity of λ_l^{exec} , and u_a of these units are activated for the service. Then the local execution time is calculated as

$$RT_k^l(t) = RT_k^{\text{que}}(t) + \frac{f(d_k(t))}{u_a \cdot \lambda_l^{\text{exec}}} \quad (7)$$

where $RT_k^{\text{que}}(t)$ is the queuing time of the task, denoted by the difference between the execution starting time and requested time as $RT_k^{\text{que}}(t) = T_k^{\text{start}} - T_k^{\text{request}}$. Meanwhile, $f(d_k(t))$ represents the total computation of the service with the size $d_k(t)$ of raw data.

2) Offloading Computing Model: When the service of vehicle v_k is determined to be offloaded to ECD, the offloading indicator is $1 \leq a_k(t) \leq M$, which indicates that the offloading destination is the $a_k(t)$ th ECD in the offloading system. Accordingly, the response time $RT_k^o(t)$ is generated during three parts of offloading computing. First, the data and service request of vehicle are transmitted from v_k to the nearest RSU r_i , and r_i offloads the service to the destination ECD. During this phase, network latency occurs in the data transmission, calculated as

$$\begin{aligned} RT_k^{o,\text{tran}}(t) &= RT_v^{o,\text{tran}}(t) + RT_r^{o,\text{tran}}(t) \\ &= \frac{d_k(t)}{\lambda_v^{\text{tran}}} + \frac{d_k(t)}{\lambda_r^{\text{tran}}} \end{aligned} \quad (8)$$

where λ_v^{tran} is the data transmission rate between v_k and r_i while λ_r^{tran} are the data transmission rate between r_i and ECD. According to the Shannon–Hartley theorem, λ_v^{tran} and λ_r^{tran} is affected by the bandwidth B of the channel, signal power p_t , and the average power of the additive white Gaussian noise p_n . As the channel resources of an RSU are often utilized by several vehicles, the bandwidth utilized by each RSU is denoted by $\frac{B}{K_c}$ when K_c vehicles are utilizing the channel concurrently. Thus, λ_v^{tran} is calculated as

$$\lambda_v^{\text{tran}} = \frac{B_r}{K_c} \log_2 \left(1 + \frac{p_t}{p_n} \right) \quad (9)$$

analogously, the transmission rate λ_r^{tran} between the ECD and one of N_c RSUs is calculated as

$$\lambda_r^{\text{tran}} = \frac{B_e}{N_c} \log_2 \left(1 + \frac{p'_t}{p'_n} \right). \quad (10)$$

After the service and digital twin data of v_k being offloaded, the destination ECD will take time for execution. Analogous to (7), the execution time of ECD is calculated as

$$RT_k^{o,\text{exec}}(t) = RT_k^{\text{que}}(t) + \frac{f(d_k(t))}{u_a \cdot \lambda_o^{\text{exec}}} \quad (11)$$

where λ_o^{exec} represents the execution capacity of the ECD, usually considered as $\lambda_o^{\text{exec}} = n \cdot \lambda_l^{\text{exec}}$.

After the task is executed, the computing results are reported back to the RSU to update the digital twin and give instruction to the vehicle. Usually, the feedback data are condensed with a relatively small size of d'_k . Thus, the feedback time during feedback is considered negligible.

Based on (8) and (11), the total response time of the service proposed by v_k at time t by offloading computing is $RT_k^o(t) = RT_k^{o,\text{tran}}(t) + RT_k^{o,\text{exec}}(t)$.

3) QoS Measurement: To quantify and measure the QoS, the maximum tolerable response time RT_{th} is used as a standard to normalize the indicator of QoS. The QoS level of response time in local computing and offloading computing are calculated as

$$S_k^l(t) = 1 - \frac{RT_k^l(t)}{RT_{th}} \quad (12)$$

$$S_k^o(t) = 1 - \frac{RT_k^o(t)}{RT_{th}}. \quad (13)$$

C. Problem Definition

In the multiuser offloading system, the goal is to maximize the average QoS level of vehicular services through an optimal offloading strategies set $A(t) = \{a_1(t), a_2(t), \dots, a_K(t)\}$ at each time period t . Based on the models given above, the problem of service offloading in DT-empowered IoV is formulated as

$$\max_{A(t)} \sum_{k=1}^K \left[S_k^l(t) + \sum_{m=1}^M S_k^o(t) \Pr[a_k(t) = m] \right] / \sum_{k=1}^K \text{Sgn}(d_i(t)) \quad (14)$$

$$s.t. \quad \forall v_k \in V, a_k(t) \in [0, M] \quad (15)$$

$$\forall v_k \in V, S_k^o(t) \geq 0, S_k^l(t) \geq 0 \quad (16)$$

where $\Pr[a_k(t) = m]$ is the probability of $a_k(t) = m$, i.e., the value is 1 if $a_k(t) = m$, otherwise, 0. Meanwhile, $\text{Sgn}(d_i(t))$ is the sign of $d_i(t)$, i.e., $\text{Sgn}(d_i(t)) = 1$ indicates that $d_i(t)$ is positive, and when $d_i(t) = 0$, $\text{Sgn}(d_i(t)) = 0$. As an element of A , $a_k(t)$ represents the offloading destination, subject to constraint (15). When $a_k(t) = 0$, the service will be locally executed. Otherwise, it will be offloaded to the corresponding ECD for execution. Meanwhile, equation (16) indicates that the QoS is not negative, i.e., the service response time should be within the maximum tolerable time.

IV. SOL FOR DT-EMPOWERED IOV SERVICES OFFLOADING

In this section, SOL is designed for the service offloading in edge computing-enabled IoV. First, the framework of RL is introduced in service offloading. Then, the drawback of a primitive RL algorithm named Q-learning is analyzed, and a DRL algorithm named DQN is leveraged for SOL.

A. Framework of Reinforcement Learning in SOL

RL is one of the significant branches of machine learning alongside supervised learning and unsupervised learning. It refers to the process of achieving the highest cumulative rewards through the exploration of the environment and the exploitation of previous knowledge. During such a trial-and-error process, the agent in RL can obtain the perception of the environment and the decision-making strategy.

In the offloading system, the ECD is enabled the controlling of offloading decisions and viewed as the agent in RL. There are three key elements of an agent, namely, the state (s), the action (a), and the reward (R). Usually, the state is also considered the environment that the agent reacts to. In SOL, the state consists of two components, the available units of ECD, and the average QoS level of each vehicle in the offloading system calculated as (14). When the ECD receives a service request, it searches for an optimal action $a_k(t)$ available in its current state. Based on the action indicator $a_k(t)$, the ECD decides where to offload and execute the service request. After making offloading decision and execution, the QoS level of service is evaluated in terms of the vehicle's response time as $S_k(t)$, then fed back to ECD as the reward. In general, the goal of RL is to obtain the highest cumulative reward in a learning episode.

Among the RL algorithms, Q-learning has proved to be effective in model-free learning problems [31]. In Q-learning, the agent is given a Q-table which records the Q-value (i.e., quality) of each pair of state and action as $Q(s, a)$. For each step, the agent selects an action a_t at the state s_t which brings it the highest reward, then calculates and updates $Q(s_t, a_t)$ based on the action it chooses and the reward it gets as

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \delta_t \quad (17)$$

where α is the learning rate parameter that satisfies $0 \leq \alpha \leq 1$ and determines the extent to which the newly acquired knowledge overrides the old knowledge. Meanwhile, δ_t is the difference between the actual value of $Q(s_t, a_t)$ and the estimated

value of it through the Q-table, calculated as

$$\delta_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \quad (18)$$

where γ represents the discount factor of future reward, s_{t+1} is the next state after the agent performing a_t , and r_t is the instant reward experienced by the agent, also denoted as the QoS level of service. Notice that, if the response time exceeds the maximum tolerable time, the reward r_t is set as $r_t \leftarrow \min(r_t, 0)$ automatically as a punishment. Specifically, the discount factor satisfies $0 \leq \gamma \leq 1$, and the larger γ means that the agent has a clearer view toward the future while lower γ means that the agent is more focused on the instant reward. Usually, Q-learning starts with a lower discount factor and increases it toward its final value to accelerate learning.

As directly choosing the action with maximal Q-value encourages exploitation but lacks exploration, agents might fall into the local optimum. Thus, a certain degree of randomness is allowed by introducing the ϵ -greedy in strategy selection. Specifically, agents select the strategy with the highest Q-value with probability $\Pr[s_i(t) = s_{\text{best}}] = 1 - \epsilon$ to exploit knowledge, while with probability ϵ , they randomly select another action to explore for more available choice. Usually, ϵ decreases over time to encourage exploration during the early phase and limit the blindness and fluctuation of agents' decision-making in the later phase.

B. SOL With Deep Q-Network

The primitive reinforcement learning method has a significant disadvantage that it requires a Q-table to store the Q-values of all possible state-action pairs. However, the number of states is large or even infinite, the traverse and update of Q-table become time-consuming. Moreover, there exist many state-action pairs that are similar but not identical in a complex Q-table. Therefore, the traditional Q-learning method will become ineffective since the possibility of the agent to access a specific state-action pair is relatively small. To tackle the problem, a practical approach is to approximate the Q-values of different state-action pairs with deep neural network (DNN), which leads to the primary essence of DQN [17]. Intuitively, the differences between Q-learning and DQN in offloading decision-making are shown in Fig. 2.

Practically, the proposal of DQN successfully combined RL with DL while tackling the challenges in the inconsistency between them. Usually, DL assumes that the distribution of data samples is in an independent manner. However, the states and actions in RL are usually highly correlated, which is not consistent with the requirement of DL. To mitigate the correlation in data, a technique of experience replay is introduced. Technically, a structure of experience pool D , which stores the experience of each step as $e_t(s_t, a_t, r_t, s_{t+1})$, is adopted to enable experience replay in DQN. During the network training, a minibatch of the experience is randomly drawn from D for training, such that the distribution of data can be averaged, and the correlations can be alleviated.

Another feature of DQN is to generate a target Q value in a separate network (i.e., the target network Q^{tar}). Unlike the original network (i.e., the prediction network Q^{pre}) which updates

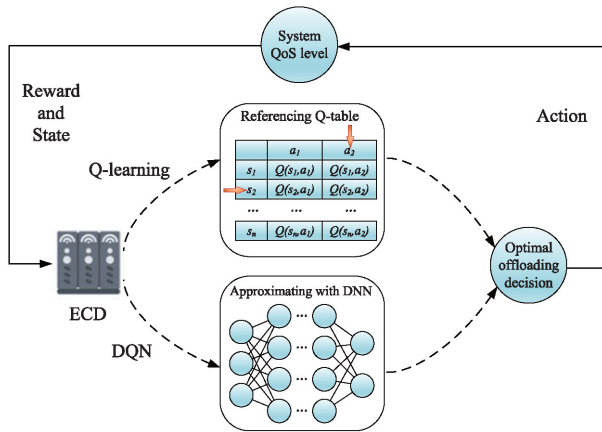


Fig. 2. Differences between offloading decision-making based on Q-learning and DQN.

the parameters θ in every iteration, θ^- in the target network are only periodically updated in every C iterations and stay fixed in other steps. Specifically, after C rounds of updates by the prediction network, the target network is updated by a copy of the prediction network. This feature adds a delay between the update of the network and the effect on the targets y_j and further stabilizes the performance of DQN.

With DNN, the Q-value of state-action pair (s_t, a) is estimated as $Q^{\text{pre}}(s_t, a; \theta) \approx Q(s_t, a)$, where the parameter θ is a vector of weights in the DNN. To evaluate the accuracy of the approximation and further train the network, the loss function is introduced as

$$L_i(\theta_i) = \mathbb{E} \left[(y_i - Q^{\text{pre}}(s, a; \theta_i))^2 \right] \quad (19)$$

where y_i represents the target Q-value generated by the target network of

$$y_i = r + \gamma \max_{a'} Q^{\text{tar}}(s_{t+1}, a', \theta_i^-). \quad (20)$$

By minimizing $L_i(\theta_i)$ through updating weight θ repeatedly, the network can be trained to be more accurate. Technically, minibatch stochastic gradient descent (MSGD) is applied to minimize the difference between the output of the target network and the prediction network. More precisely, the pseudo code of DQN is shown in Algorithm 1.

C. SOL Review

Generally, SOL is designed on the logical basis shown in Fig. 3. The basic idea of SOL is to enable the ECD to make optimal offloading decisions through RL. With the exploration of the unknown environment, the agent in RL can learn from the feedback reward. Meanwhile, the exploitation of experienced knowledge enables the agent to select optimal action at each state, jointly considering the instant reward and long-term reward. However, as the environment of the IoV service offloading system is dynamic and sophisticated, the space of states can be vast or infinite. If primitive RL algorithms like Q-learning are adopted, the update and search for optimal offloading decisions generate a significant overhead of storage and time. Moreover,

Algorithm 1: SOL With Deep Q-Network.

- 1: Initialize experience pool D with the size of N
- 2: Initialize Q^{pre} and Q^{tar} with same random weights θ
- 3: **for** episode = 1 to M **do**
- 4: **for** $t = 1$ to T **do**
- 5: Approximate Q-values of all actions at state s
- 6: Select the optimal offloading decision a_t based on ϵ -greedy policy
- 7: Perform service offloading or local computing according to a_t
- 8: Calculate the reward r_t and the next state s_{t+1}
- 9: Store experience $e_t(s_t, a_t, r_t, s_{t+1})$ in D
- 10: Perform MSGD to update the parameters θ of prediction network Q^{pre} through minimizing $L(\theta)$
- 11: Update the target network Q^{tar} every C steps
- 12: **end for**
- 13: **end for**

the similar but not identical states significantly increase the agent's exploration range and will lead to slow convergence of RL. To reduce the overhead in storage and time while fastening convergence, a DRL algorithm named DQN is adopted in SOL. Instead of referencing the Q-table to find the optimal decision, DQN introduced the function value approximation of DL to estimate the Q-value of state-action pairs. Also, with the features of experience replay and target network, DQN successfully alleviates the inconsistency between RL and DL, and can achieve satisfying performance in SOL.

V. EXPERIMENTAL EVALUATION

In this section, SOL is implemented and experiments are conducted based on the real-world IoV service requests. Then, comparative offloading strategies are introduced. Finally, the results of SOL and comparative offloading strategies under different circumstances are presented, and the effectiveness and adaptability of SOL are verified based on the experimental results.

A. Experiment Setup

Two real datasets of IoV service requests in Nanjing are applied in the experiment. One dataset contains details of 436 activated RSUs in Nanjing, including their latitude and longitude values. Based on the RSU locations, partitioning around medoids (PAM) clustering is adopted with the parameter $K=40$ to simulate the placement of ECDs and the assignment of RSUs. As shown in Fig. 4, on part of the brief road map of Nanjing, the RSUs and ECDs in one cell of the offloading system are marked with blue dots and red server icons, respectively. The 3 ECDs and 26 RSUs (including 3 colocated with each ECD) are analyzed in the experiments.

The other dataset contains vehicular service requests collected by RSUs in 30 consecutive days (from 00:00:00 Sep. 1st to 23:59:59 Sep. 29th). The total number of service requests is more than 160 million. From the second dataset, the service

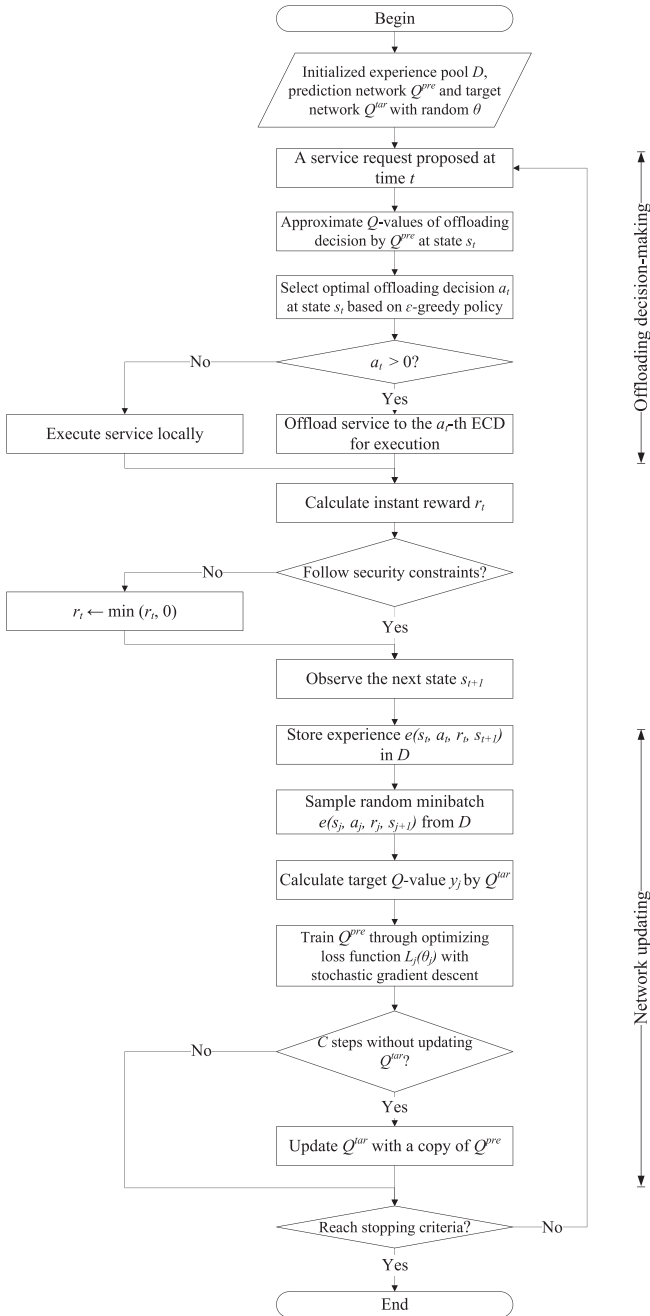


Fig. 3. Programming flowchart of SOL.

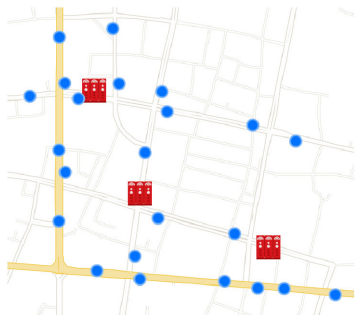


Fig. 4. Distribution of RSUs and ECDs in an offloading system.

TABLE II
CONTROLLED VARIABLE SETTINGS

Variable description	Controlled value
ECD execution capacity	$5 \times$ local execution capacity
Number of ECD	3
Number of service requests	5 per vehicle
Average size of raw data	50 MiB per request

requests in one cell of the offloading system are extracted for comparative analysis.

B. Comparative Offloading Strategies

1) *Entirely Local Computing*: Entirely local computing is a conventional paradigm which depends only on the vehicles' local execution capacity. Entirely local computing requires no additional controlling strategy, and is used as a baseline to evaluate the optimization capability of other offloading strategies.

2) *Nearest Neighbor Offloading Computing*: Contrary to entirely local computing, nearest neighbor offloading strategy enables all the service requests and raw data to be offloaded to the nearest ECD for execution. As the location of RSUs and ECDs are fixed, the nearest ECD of each RSU can be determined. When the computational resources of ECD are abundant, this strategy can achieve a high level of QoS without complicated controlling. However, as the local computing units are not utilized, and the distribution of workload is uneven, excessive offloaded services will increase the risk of ECD being overloaded and severely lower the QoS level of the offloading system.

3) *First Fit Offloading Computing*: First fit is an online algorithm where the service is offloaded to the nearest ECD that can accommodate it. When first fit algorithm begins, it searches for the closest ECD to the RSU which collected a service request. If the ECD has insufficient idle resource units for the service, it will be offloaded to the next closest ECD with sufficient resources. If no ECD is capable, the service will be executed by the computing devices of the vehicle which proposes the request.

C. Analysis on the Adaptability of Offloading Strategy

As the real condition of IoV services in cities are various, e.g., the number of vehicles and ECDs varies with the development of cities. Thus, the offloading strategy needs to be adaptive, so that it can be applied widely. To verify the adaptability of SOL, four sets of controlled experiments with diversity in services conditions are conducted, and the performance of SOL is evaluated.

The controlled value of variables in the comparative analysis are listed in Table II. In each set of experiment, there is one variable with its value fluctuating around the controlled value and the others remain unchanged.

1) *Analysis on the Variety of ECD Execution Capacity*: Experiments are conducted with different ECD's execution capacity, and the results are shown in Fig. 5. In this set of experiments, the ratio of ECD execution capacity to local execution capacity

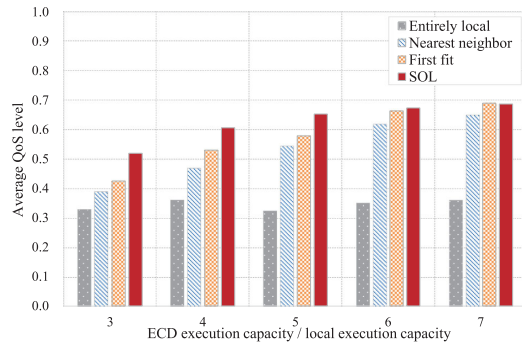


Fig. 5. Comparison of QoS level with variety in ECD capacity.

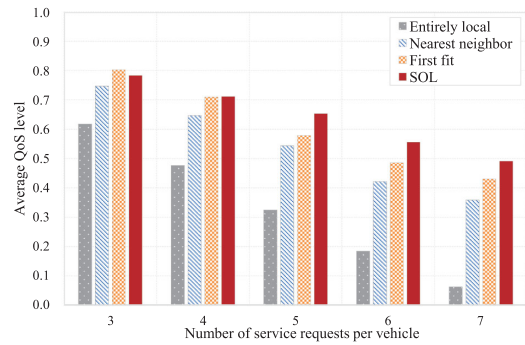


Fig. 7. Comparison of QoS level with variety in service number.

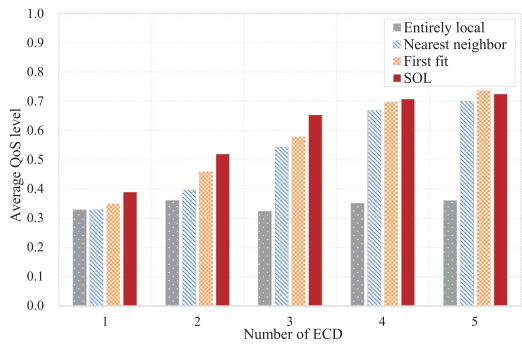


Fig. 6. Comparison of QoS level with variety in ECD number.

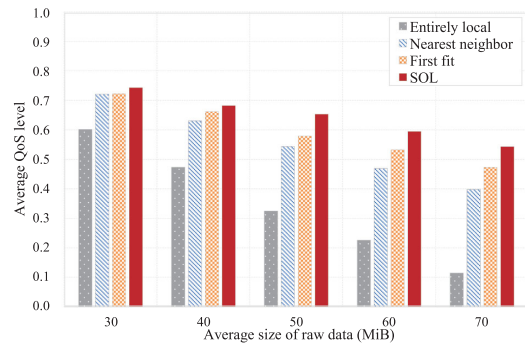


Fig. 8. Comparison of QoS level with variety in average size of raw data.

ranges from three to seven. As the results indicate, SOL outperforms entirely local computing, nearest neighbor offloading, and first fit offloading in response time. When the capacity of ECD is insufficient, the risk of ECD being overloaded is high if no effective offloading strategy is adopted. Thus, the QoS level of vehicular services by nearest neighbor offloading is severely reduced by long response time. In contrast, when the execution capacity of ECD is ample, the difference in response time between SOL and the other offloading strategies is small. As the ECDs can efficiently execute most of the services, offloading computing is usually the optimal choice.

2) Analysis on the Variety of ECD Number: When the number of ECDs in the offloading system are different, the QoS level of vehicular services are shown in Fig. 6. With other variables unchanged, the number of ECDs ranges from one to five in this set of experiments. The QoS level of SOL is generally the highest despite the little disadvantage over first fit when ECD number is five. When ECDs are sparsely deployed, the ECDs can be easily overloaded by the excessive service requests. Thus, SOL tends to assign the services to be executed locally and has a slight advantage over other strategies. In contrast, when ECDs are ample, the service requests and the workload of ECDs are more balanced with SOL or first fit offloading strategy, and overloading is unlikely to occur during offloading computing.

3) Analysis on the Variety of Service Number per Vehicle: Fig. 7 illustrates the impact on the QoS level by the number of services per vehicle. In this set of experiments, we assume each vehicle can propose multiple service requests at different time, and the number of proposed service requests per vehicle ranges

from three to seven, while other variables remain unchanged. The QoS level of response time by offloading method goes down as the number of services rises, while SOL keeps the decline smaller than first fit and nearest neighbor offloading. The advantage of SOL is that ECD selectively executes some of the services while others are executed locally, which reduces the latency in queuing. When the execution capacity of ECD goes beyond the service requests of vehicles, the QoS level of first fit and nearest neighbor offloading is close to the one of SOL and both outperform local computing. In addition, as the bandwidth of ECD is usually considered fixed, the intensive data transmission also has an impact on the offloading time when the communication is frequent.

4) Analysis on the Variety of Average Size of Raw Data: In Fig. 8, the QoS level with diversity in the average size of raw data is analyzed. Experiments are conducted with the average size of raw data ranging from 30 to 70 MiB, while the other variables remain unchanged. It is intuitive that the QoS level declines with the rise in the size of raw data. As the computing capacity of on-board devices is usually insufficient, the response time of local computing is intolerable. Simultaneously, the QoS level of nearest neighbor offloading, first fit offloading, and SOL also experience a drop. However, as the execution rate of ECD is much higher than on-board devices, the increase in response time by offloading methods is not significant. Instead, the time overhead generated in data transmission has an impact on the QoS level. Hopefully, 5G communication is promising in mitigating the data transmission time and further enhance the QoS level of service offloading by SOL.

VI. CONCLUSION

In this article, edge computing was adopted in the DT-empowered IoV to provide vehicular services with a high QoS level, and a service offloading method with deep reinforcement learning named SOL is proposed. First, a multiuser offloading system in DT-empowered IoV was modeled with consideration of response time. Then, DQN with experience replay and target network, which exerts the advantages of both RL and DL, was adopted in the offloading system to obtain optimal offloading strategy. The experiments were conducted with a real-world dataset of RSU locations and IoV service requests, and the results verified the effectiveness and adaptability of SOL.

To simplify the model, the IoV service offloading was modeled as a binary offloading process where the services are assumed atomic, i.e., services cannot be divided and executed on more than one devices. In future works, partial offloading can be taken into consideration where a service can be divided into several procedures and offloaded to different ECDs. In this case, computational resources can be better utilized. However, if partial offloading is adopted, the partibility, dependency, and priority in the procedures of services need to be thoroughly analyzed, and the offloading decisions are required a strict graph dependency constraint.

REFERENCES

- [1] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibañez, "Internet of Vehicles: Architecture, protocols, and security," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3701–3709, Oct. 2018.
- [2] K. Asano, N. Enami, T. Kamada, and C. Ohta, "Person reidentification for detection of pedestrians in blind spots through V2V communications," in *Proc. IEEE Intell. Transport Syst. Conf.*, 2019, pp. 764–770.
- [3] L. Chen and C. Englund, "Cooperative intersection management: A survey," *IEEE Trans. Intell. Transport Syst.*, vol. 17, no. 2, pp. 570–586, Feb. 2016.
- [4] X. Wang *et al.*, "Optimizing content dissemination for real-time traffic management in large-scale Internet of Vehicle systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1093–1105, Feb. 2019.
- [5] O. Veledar, V. Damjanovic-Behrendt, and G. Macher, "Digital twins for dependability improvement of autonomous driving," in *Systems, Software and Services Process Improvement*, Cham, Switzerland: Springer, 2019, pp. 415–426.
- [6] R. Minerva, G. M. Lee, and N. Crespi, "Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models," *Proc. IEEE*, vol. 108, no. 10, pp. 1785–1824, Oct. 2020.
- [7] X. Wang, L. T. Yang, L. Song, H. Wang, L. Ren, and J. Deen, "A tensor-based multi-attributes visual feature recognition method for industrial intelligence," *IEEE Trans. Ind. Informat.*, early access, doi: [10.1109/TII.2020.2999901](https://doi.org/10.1109/TII.2020.2999901).
- [8] M. Zhang, C. Chen, T. Wo, T. Xie, M. Z. A. Bhuiyan, and X. Lin, "SafeDrive: Online driving anomaly detection from large-scale vehicle data," *IEEE Trans. Ind. Informat.*, vol. 13, no. 4, pp. 2087–2096, Aug. 2017.
- [9] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, "Big data analytics in intelligent transportation systems: A survey," *IEEE Trans. Intell. Transport Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.
- [10] K. Djemame *et al.*, "PaaS-IaaS inter-layer adaptation in an energy-aware cloud environment," *IEEE Trans. Sustain. Comput.*, vol. 2, no. 2, pp. 127–139, Apr./Jun. 2017.
- [11] M. Abbasi, M. Rafiee, M. R. Khosravi, A. Jolfaei, V. G. Menon, and J. M. Koushyar, "An efficient parallel genetic algorithm solution for vehicle routing problem in cloud implementation of the intelligent transportation systems," *J. Cloud Comput.*, vol. 9, no. 1, 2020, Art. no. 6.
- [12] V. G. Menon, S. Jacob, S. Joseph, and A. O. Almagrabi, "SDN-powered humanoid with edge computing for assisting paralyzed patients," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5874–5881, Jul. 2020.
- [13] J. Ren, G. Yu, Y. He, and G. Y. Li, "Collaborative cloud and edge computing for latency minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 5031–5044, May 2019.
- [14] Q. He *et al.*, "A game-theoretical approach for user allocation in edge computing environment," *IEEE Trans. Parallel Distrib. Syst.*, vol. 31, no. 3, pp. 515–529, Mar. 2020.
- [15] J. Zhao, Q. Li, Y. Gong, and K. Zhang, "Computation offloading and resource allocation for cloud assisted mobile edge computing in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7944–7956, Aug. 2019.
- [16] Z. Ning *et al.*, "When deep reinforcement learning meets 5G vehicular networks: A distributed offloading framework for traffic big data," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1352–1361, Feb. 2020.
- [17] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [18] Z. Wang *et al.*, "A digital twin paradigm: Vehicle-to-cloud based advanced driver assistance systems," in *Proc. IEEE 91st Veh. Technol. Conf.*, 2020, pp. 1–6.
- [19] X. Wang, L. T. Yang, Y. Wang, L. Ren, and M. J. Deen, "ADTT: A highly-efficient distributed tensor-train decomposition method for iiot big data," *IEEE Trans. Ind. Informat.*, early access, doi: [10.1109/TII.2020.2967768](https://doi.org/10.1109/TII.2020.2967768).
- [20] X. Hu *et al.*, "Sinet: A scale-insensitive convolutional neural network for fast vehicle detection," *IEEE Trans. Intell. Transport Syst.*, vol. 20, no. 3, pp. 1010–1019, Mar. 2019.
- [21] S. Liu, L. Liu, J. Tang, B. Yu, Y. Wang, and W. Shi, "Edge computing for autonomous driving: Opportunities and challenges," *Proc. IEEE*, vol. 107, no. 8, pp. 1697–1716, Aug. 2019.
- [22] L. Zhao, W. Sun, Y. Shi, and J. Liu, "Optimal placement of cloudlets for access delay minimization in SDN-based Internet of Things networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1334–1344, Apr. 2018.
- [23] S. Wang, Y. Zhao, J. Xu, J. Yuan, and C.-H. Hsu, "Edge server placement in mobile edge computing," *J. Parallel Distrib. Comput.*, vol. 127, pp. 160–168, 2019.
- [24] X. He, R. Jin, and H. Dai, "Peace: Privacy-preserving and cost-efficient task offloading for mobile-edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 1814–1824, Mar. 2020.
- [25] Z. Zhou, H. Liao, X. Zhao, B. Ai, and M. Guizani, "Reliable task offloading for vehicular fog computing under information asymmetry and information uncertainty," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8322–8335, Sep. 2019.
- [26] F. Guo, H. Zhang, H. Ji, X. Li, and V. C. Leung, "An efficient computation offloading management scheme in the densely deployed small cell networks with mobile edge computing," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2651–2664, Dec. 2018.
- [27] P. A. Vikhar, "Evolutionary algorithms: A critical review and its future prospects," in *Proc. Int. Conf. Glob. Trends Signal Process., Inf. Comput. Commun.*, 2016, pp. 261–265.
- [28] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.
- [29] X. Liang, X. Du, G. Wang, and Z. Han, "A deep reinforcement learning network for traffic light cycle control," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1243–1253, Feb. 2019.
- [30] M. Zhou, Y. Yu, and X. Qu, "Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach," *IEEE Trans. Intell. Transport Syst.*, vol. 21, no. 1, pp. 433–443, Jan. 2020.
- [31] N. Kumar, S. N. Swain, and C. Siva Ram Murthy, "A novel distributed Q-learning based resource reservation framework for facilitating D2D content access requests in LTE-A networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 15, no. 2, pp. 718–731, Jun. 2018.



Xiaolong Xu received the Ph.D. degree in computer science and technology from Nanjing University, Nanjing, China, in 2016.

He was a Research Scholar with Michigan State University, East Lansing, MI, USA, from 2017 to 2018. He is currently a Professor with the School of Computer and Software, Nanjing University of Information Science and Technology, Nanjing, China. He has authored or coauthored more than 80 peer-review articles in international journals and conferences. His research interests include edge computing, the Internet of Things (IoT), cloud computing, and big data.

Prof. Xu is a fellow of EAI (European Alliance for Innovation). He was the recipient of the Best Paper Award from the IEEE Cloud and Big Data (CBD) 2016, IEEE CPCSCOM 2020, and International Conference on Security and Privacy in Digital Economy (SPDE) 2020.

Authorized licensed use limited to: Scms School Of Engineering And Technology. Downloaded on July 27, 2023 at 09:39:31 UTC from IEEE Xplore. Restrictions apply.



Bowen Shen is currently working toward the B.S. degree in computer science and technology with the School of Computer and Software at Nanjing University of Information Science and Technology, Nanjing, China.

His research interests include edge computing and IoT.



Sheng Ding received the graduate degree in computer science and technology, from East China University of Technology, Fuzhou, China, in 2003, and the master's degree in computer application technology from Ocean University of China, Qingdao, China, in 2012.

Since graduation, he has been working in the forefront of education, and has accumulated rich teaching experience. He has authored or coauthored many high-quality articles in various academic journals.



Gautam Srivastava (Senior Member, IEEE) received the Ph.D. degrees from the University of Victoria, Victoria, BC, Canada, in 2012.

In 2014, he joined a tenure-track position with Brandon University, Canada, and was promoted to the rank Associate Professor, in 2018. He has authored or coauthored more than 50 papers in high-impact conferences and high-status journals, including IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS (TWC), IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING

(TNSE), IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS (TCSS), IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS (TIA), and IEEE COMMUNICATIONS LETTERS. His research interests include social network and big data.

Prof. Srivastava was the recipient of the Best Oral Presenter Award in FSDM 2017. He is the Associate Editor for several international journals, such as IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE ACCESS, etc. He has served as the leading Guest Editor FOR IEEE TRANSACTIONS ON FUZZY SYSTEMS.



Muhammad Bilal (Senior Member, IEEE) received the Ph.D. degree in information and communication network engineering from the School of Electronics and Telecommunications Research Institute (ETRI), Korea University of Science and Technology, Daejeon, South Korea, in 2017.

From 2017 to 2018, he was with Korea University, where he was a Postdoctoral Research Fellow with the Smart Quantum Communication Center. Since 2018, he has been an Assistant

Professor with the Division of Computer and Electronic Systems Engineering, Hankuk University of Foreign Studies, Yongin, South Korea. His research interests include design and analysis of network protocols, network architecture, network security, the IoT, named data networking, blockchain, cryptology, and future Internet.

Dr. Bilal serves as an Editor for the IEEE FUTURE DIRECTIONS ETHICS AND POLICY IN TECHNOLOGY NEWSLETTER and the IEEE INTERNET POLICY NEWSLETTER.



Mohammad R. Khosravi received the B.Sc. degree from Shiraz Unuversity, Iran, in 2013, the M.Sc. degree fom Persian Gulf University, Iran, in 2015, and the Ph.D. degree from the Shiraz University of Technology, Iran, 2020, all in electrical engineering.

He is currently with the Department of Computer Engineering, Persian Gulf University, Bushehr, Iran, and has been with Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz, Iran. His main

interests include statistical signal and image processing, medical bioinformatics, radar imaging and satellite remote sensing, computer communications, industrial wireless sensor networks, underwater acoustic communications, information science, and scientometrics.



Varun G Menon (Senior Member, IEEE) He received the Ph.D. in computer science and engineering from Satyabhama University, India, in 2017.

He is currently an Associate Professor and Head of the Department of Computer Science and Engineering at the SCMS School of Engineering and Technology, Ernakulam, Kerala, India. His research interests include Internet of Things, 5G communications, fog computing and networking, underwater acoustic sensor net-

works, hijacked and predatory journals, ad hoc networks, opportunistic routing, and wireless sensor networks.

Dr. Menon was the recipient of the Top Peer Reviewer Award by Publons, in 2018 and 2019. He is a Distinguished Speaker of ACM. He is an Associate Editor for *Physical Communications* and *IET Quantum Communications*, Technical Editor for *Computer Communications*, and also an Editorial Board Member for IEEE FUTURE DIRECTIONS: TECHNOLOGY POLICY AND ETHICS. He has served over 20 conferences in leadership capacities including program Co-Chair, Track Chair, Session Chair, and Technical Program Committee member.



Mian Ahmad Jan received the Ph.D. degree in computer systems from the University of Technology Sydney (UTS), Sydney, Australia, in 2016.

He is currently an Assistant Professor with the department of computer science, Abdul Wali Khan University Mardan, Pakistan. His research interests include energy-efficient and secured communication in wireless sensor networks and internet of things, and has recently been actively involved in machine learning, big data analytics, smart cities infrastructure, and vehicular ad hoc networks.

Dr. Jan was the recipient of International Research Scholarship (IRS), UTS and Commonwealth Scientific Industrial Research Organization (CSIRO) scholarships. He was the recipient of the best researcher awarded for the year 2014 at the University of Technology Sydney Australia. He had been the recipient of various prestigious scholarships during his Ph.D. studies.



Mao-Li Wang received the B.S. degree in automation from Qufu Normal University, Jining, China, in 2004, and the M.S. and the Ph.D degrees in control theory and control engineering from Harbin Engineer University, China, in 2008.


He is a Professor with the School of Cyber Science and Engineering, Qufu Normal University. His research interests include Internet of Things, blockchain, and artificial intelligence.

RESEARCH

Open Access



A secure data deduplication system for integrated cloud-edge networks

Shynu P. G.¹, Nadesh R. K.¹, Varun G. Menon², Venu P.³, Mahdi Abbasi^{4*}  and Mohammad R. Khosravi^{5,6}

Abstract

Data redundancy is a significant issue that wastes plenty of storage space in the cloud-fog storage integrated environments. Most of the current techniques, which mainly center around the static scenes, for example, the backup and archive systems, are not appropriate because of the dynamic nature of data in the cloud or integrated cloud environments. This problem can be effectively reduced and successfully managed by data deduplication techniques, eliminating duplicate data in cloud storage systems. Implementation of data deduplication (DD) over encrypted data is always a significant challenge in an integrated cloud-fog storage and computing environment to optimize the storage efficiently in a highly secured manner. This paper develops a new method using Convergent and Modified Elliptic Curve Cryptography (MECC) algorithms over the cloud and fog environment to construct secure deduplication systems. The proposed method focuses on the two most important goals of such systems. On one side, the redundancy of data needs to be reduced to its minimum, and on the other hand, a robust encryption approach must be developed to ensure the security of the data. The proposed technique is well suited for operations such as uploading new files by a user to the fog or cloud storage. The file is first encrypted using the Convergent Encryption (CE) technique and then re-encrypted using the Modified Elliptic Curve Cryptography (MECC) algorithm. The proposed method can recognize data redundancy at the block level, reducing the redundancy of data more effectively. Testing results show that the proposed approach can outperform a few state-of-the-art methods of computational efficiency and security levels.

Keywords: Convergent encryption (CE), Modified elliptic curve cryptography (MECC), Edge computing, Integrated cloud and fog networks, Hash tree. Secure hash algorithm (SHA)

Introduction

The data gathered through different sources and the Emergence of the Internet of Things in all aspects of applications increases data volume from petabytes to yottabytes, necessitating cloud computing paradigm and fog networks to process and store the data. Cloud computing (CC) produces a network-centered environment vision to users which provides access to the internet, to a collective pool of programmable grids, servers, software, storage, and amenities that could be quickly freed, with less supervision and communication to the cloud service provider. Data processing in all ways is carried out

remotely in the cloud server with the help of internet connectivity. Fog computing provides the local infrastructure to process the application locally and then connects to the cloud. The fog environment reduces delay when compared to the application connected to the cloud for processing. The application developed to process and store the data needs end-to-end security, communication protocols, and resources to access information stored in the cloud and fog environments. Smart applications are built with the help of sensors and actuators, and the data is stored in the cloud environment; and edge computing facilities are also used along with the local infrastructure, termed as fog, to process the data without delay. Internet of Things does not end up with an information system but tries to build a cyber-physical system [1]. Edge computing provisions the

* Correspondence: Abbasi@basu.ac.ir

⁴Department of Computer Engineering, Engineering Faculty, Bu-Ali Sina University, Hamedan 65178-38695, Iran
Full list of author information is available at the end of the article

feature of mobility for the user to process and store data on the move. Mobile edge computing provides seamless integrity among multiple applications, vendors, mobile subscribers, and enterprises [2].

Sending data to the cloud was the bulbous trend in the past decades, which is now changing to fog, edge, and cloudlet due to delay-sensitive and context-aware services. To address these challenges, the centralized cloud computing paradigm is moving to distributed edge cloud computing and this makes computing transparent [3]. Fog computing is an attractive solution to the distributed edge cloud computing for any type of applications and benefits in low-latency, mobility, and geo-distributed services distinguished from the cloud with several access control schemes [4–6]. When fog computing is considered the one-step solution for reducing computation tasks' latency, some schemes are described for offloading the task focusing on reducing the latency, energy efficiency, and reliability [7]. Admission control, computational resource allocation and power control are some of the critical parameters considered before offloading the intensive task from the cloud. The performance can be further improved only by the efficient resource allocation methods available for cloud and fog environment, thus increasing the reliability and transparency in application processing [8]. Resource allocation in a period allows the moving user to offload the task to the nearest cloudlet and extend the services from the fog environment. This type of offloading reduces delays in computational tasks with more significant mobility features [9].

Various application services impulse the possibility of edge computing by offering cloud capabilities at the network edge closer to mobile devices. Edge computing is an encouraging paradigm to decide several vital challenges in the Internet of Things in all domains, such as delay, low bandwidth, energy issues, latency in transmission, and data security and privacy [10, 11]. A comprehensive study of information security and privacy requirements, tasks, and tools in edge computing, cloud computing, fog computing and the cryptography-based technologies for solving security and privacy issues are analyzed before incorporating the cloud and fog networks [12, 13]. Hybrid encryption techniques using AES in CBC Mode and HMAC-SHA-1 (Hash-based Message Authentication Code) with lightweight procedures improve the robust encryption at user-level security in a cloud computing environment [14]. There are many more technical developments, but they exhibit other issues that have to be resolved, encompassing processing and storing data, securing sensitive information, and protecting user privacy [15].

Data deduplication (DD) stands as a universal data redundancy removal technology. DD primarily classifies

identical data, stores one copy of data, and substitutes other similar copies with undirected references instead of keeping full copies [16]. DD involves three major processes: a) *chunking*, b) *hashing*, and c) *comparing* hashes to recognize redundancy. The chunking process breaks a file into many smaller files termed as chunks. The chunk level deduplication method ameliorates the storage of unique chunks by contrasting it with all incoming chunks for duplicate recognition [17]. Once the data is being uploaded to the cloud, the owner could not assure the security of the data in remote storage systems. Performing encryption is necessary to make data secure; simultaneously, performing deduplication is imperative for attaining optimized storage. Hence, encryption and deduplication should be done simultaneously for ensuring optimized and secured storage [18]. DD could be employed within a file, across files, across applications, or across clients over a particular period of time. It is utilized in archiving and backing up the file systems, databases with low change rate, Network Attached Storage, VMware environment, Local Area Network, and Storage Area Network. By adopting them, the key utilized for encryption and also decryption is itself attained from the data and would resist further attacks [19].

This paper proposes a secure data deduplication system using convergent and MECC algorithms over the integrated cloud-fog-edge environment. The convergent encryption appears to be the right choice for the implementation of deduplication with encryption in the cloud storage domain. But there is the possibility of *dictionary attacks* concerning this scheme as the encryption key is formed using the plaintext. An intruder who has gained access to the storage can compare the ciphertexts produced after the encryption of distinguished plaintext values from a dictionary where the ciphertexts are being stored. Moreover, even if encryption keys are encrypted with the users' private keys and stored somewhere else, the cloud provider, who has no access to the encryption key but has access to the encrypted chunks (blocks), can efficiently perform offline dictionary attacks and determine the expected data. Hence, to solve the above problem, the encrypted and deduplicated data using the CE are once again encrypted by the proposed modified elliptic curve cryptography (MECC) technique. The combined CE and MECC technique ensure efficient deduplication and secured encryption of cloud-fog-edge storage with less computational overhead compared to existing data deduplication techniques.

The significant contributions of this paper can be summarized as follows.

- A new method of constructing a secure deduplication (DD) system using Convergent and Modified Elliptic Curve Cryptography (MECC)

algorithms over the cloud and fog/edge environment.

- Performance evaluation of the proposed technique, based on its computational efficiency and level of security is done.
- We validated the proposed deduplication technique's ability to recognize data redundancy in the level of blocks, which can reduce the redundancy of data more effectively to minimize the storage space in the cloud environment.

The draft structure of the manuscript is arranged as follows: section 2 surveys the associated works, section 3 provides the proposed methodology, section 4 explores the tentative outcome and section 5 contains the conclusion and scope for future work.

Related works

The secure deduplication system abandons the duplicate copies of data, and it also proffers security to the data. Convergent Encryption (CE) is utilized to encrypt or decrypt the data to the file level with a convergent key that is generated as the file content itself [20]. After encrypting such files, the cloud user just holds the encryption key and outsources the ciphertext (CT) to the CS to save storage space. By updating the CT saved in the central cloud and user-level public keys without knowing the private keys, consistent privacy is rendered [21]. Kwon et al [22]. proffered a secure deduplication framework with user revocations. The system comprises of '3' phases, namely: upload, revocation, and download. The proffered framework is executed via a privilege-centric re-encryption methodology over convergent-encryption. Liet al [23]. recommended Dekey, a construction wherein users need not handle any keys but rather securely disseminate the convergent key shares across multi servers. Dekey upholds the block-level and file-level deduplication. File-level deduplication eradicated the storage of any redundant files, and block-level deduplication separated the files into a smaller variable or fixed-sized blocks and wiped out the storage of any redundant blocks. Security analysis delineated that Dekey was secure. The Dekey centric Ramp secret centered framework elucidated that Dekey incurred limited overhead in factual environments.

Kwon et al. [24] recommended a deduplication framework at the server-side for the encrypted data. It permits the Cloud Server to control the access to the outsourced data even while the ownership altered dynamically with the secured ownership group key distributions and exploited randomized CEs. It can avert data leakage to the revoked users though they formerly owned the data and even to the cloud storage. The system assures data integrity like the tag inconsistency attack. The efficiency

estimation results corroborated that the scheme was almost as effectual as the former framework, while the extra overhead in computations was insignificant. Yuan et al [25]. developed a primitive termed and a fully randomized framework (R-MLE2). It comprises of '2' schemes: i) static and ii) dynamic, where the latter one permitted tree adjustment by elevating specific computation cost. The primary trick of the framework was to utilize the interactive protocol centered on dynamic or static decision trees. The security and performance analyses evinced that the frameworks were Path-PRV-CDA2 secured which attained multiple orders of magnitude and high-level performance for the data equality tests, more than the R-MLE2 framework, when the count of data items was comparatively large. Han et al. [26] proffered a multi-bit secret channel in the cloud storage service, and also recommended a framework that attained good security and high-level data transmission rate. In the recommended algorithm, the data upload was simplified via multi-bit file depiction. It eradicated the need to upload "0" to diminish the number of uploaded files, thereby made it hard for the attacker to spot the covert channel and also effectually ameliorated the security of cloud user data upload. Tawalbeh et al. [27] reconsidered the security and privacy for cloud and fog environments with the case study of health care systems using fog simulator and enhanced the performance and trust among the end-users. Similarity and emergence centered indexing for high-performance deduplication of data was introduced by Zhanget al [28]. which provides quick responses to fingerprint queries. Houet al [29]. suggested to check the truthfulness of cloud data beneath the condition that the remote server stores only a single copy of the same file from different users.

Deduplication has confirmed to achieve great space and cost investment, and a higher number of distributed storage suppliers are currently embracing it. Deduplication can weaken capacity needs by up to 90–95% for corroboration [30]. As more users outsource their data to remote server storage, the latest data breach occurrences make end-to-end encryption increasingly desirable. Enhanced Secure Threshold Data Deduplication Pattern for remote storage helps to maintain end-to-end encryption [31]. A flexible admission control tool called Proxy re-encryption (PRE) has been recently hosted. PRE is an effective tool for creating cryptographically imposed admission control systems [32]. These schemes show competence in computational cost and ciphertext size.

A confidentiality-preserving deduplication technique for remote storage in public cloud services is discussed in [33]. The authors have proposed a secure file deduplication mechanism on the encrypted file, supporting public reliability and auditing in the deduplication of the cloud

storage system. A chaotic fuzzy transformation method is projected to provision protected fuzzy keyword indexing, storage, and query for fog systems that aid in raising the privacy and confidentiality of the end-user data and also by saving the resources of the mobile user devices [34]. A comprehensive study on various security problems associated with outsourced data on the cloud and their existing solutions is described using access control models for the cloud computing environment [35].

A framework to mine structures statically and dynamically from malware that imitates the performance of its code, such as the Windows Application Programming Interface (API) classifies malware with high accuracy and low false alarm rates [36]. The public-key-based schemes obviate the security vulnerability inherent to symmetric-key-based μ TESLA-like schemes. But their signature verification is time-consuming [37].

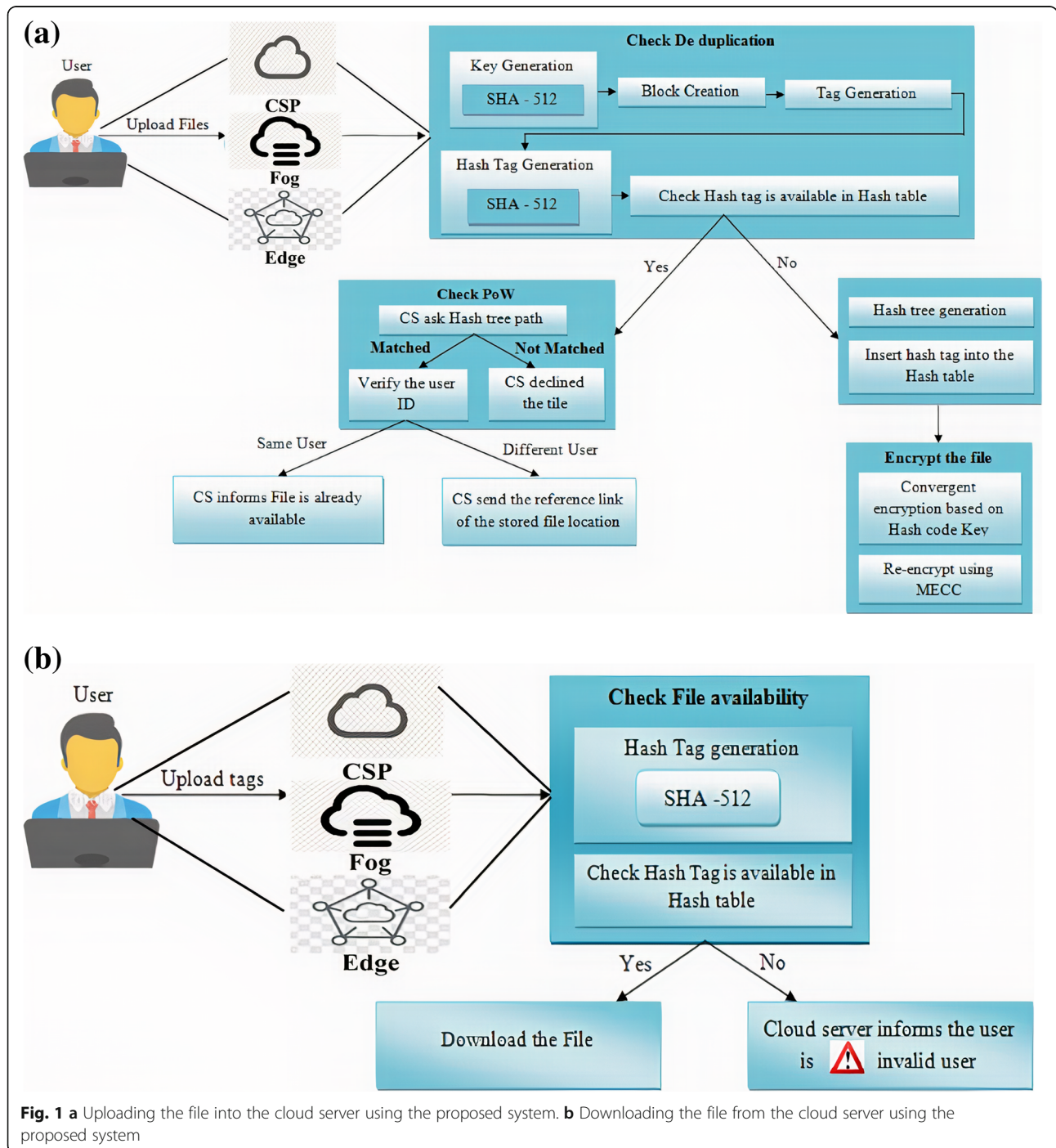


Fig. 1 a Uploading the file into the cloud server using the proposed system. **b** Downloading the file from the cloud server using the proposed system

Proposed secure deduplication approach

Cloud Service Provider provides many resources to users as a service, for instance, vastly available storage space. Managing the ever-elevating volumes of data in the cloud is a noteworthy task. The DD technique makes data management more scalable in CC. But security is the main problem in Data Deduplication. To overcome this problem, this paper proposes a secure data deduplication system using convergent and MECC algorithms over the integrated cloud-fog environment.

The proposed methodology is analyzed in four ways, i.e., a) when a new user tries to upload a new file, b) when the same user tries to upload the same file c) when different users try to upload the same file to the cloud server and d) when the users try to download the file. The proposed methodology could be expounded in detail using the block diagram evinced in Fig. 1a and b.

When a new user tries to upload a new file

Initially, the new user browses a file and uploads it to the CS. Then, the CS generates the hash code (HC) key for the appropriate file using SHA 512 algorithm. The input file is then appended with padding and fixed 128bit length field. The enlarged message is partitioned as blocks. A 64-bit word is derived as of the current message block utilizing 8 constants based on the square root of the first 8 prime numbers. In the subsequent level, a 512-bit buffer is updated. SHA-512 operation can be comprehended using

Fig. 2, and then, the original input file is split into blocks. Next, tag values are assigned for each block and hash code (HC) is created for each tag value of a particular block utilizing the same SHA 512 algorithm. Cloud server (CS) verifies whether the hashtag is available in the HT. If it is unavailable, then the hash tree is generated for Proofs of Ownership (PoW) grounded on the hashtag value. Next, the file is encrypted using a convergent encryption (CE) method. The CE takes the HC key of the file as input. Next, the convergent based encrypted file is again encrypted utilizing the MECC algorithm.

The encryption process is done for securely uploading the data to the CS. The CE and MECC based encryption of the particular file is expounded as follows.

Convergent encryption (CE)

In data deduplication, CE improves data confidentiality. The convergence key (CK) is denoted as the generated hash code (HC) value of the file. Using this CK, all blocks of data copy are encrypted. To detect the duplicate file in the CSP, a tag will be derived for each data block. If two data files are the same, the same tags will be provided. Before storing the data file to the CSP, its tag would be forwarded to the CSP for detecting duplicate data files. At last, this encrypted data block, along with its tag, is saved in the CSP. The phases in Convergent encryption (CE) are described as follows.

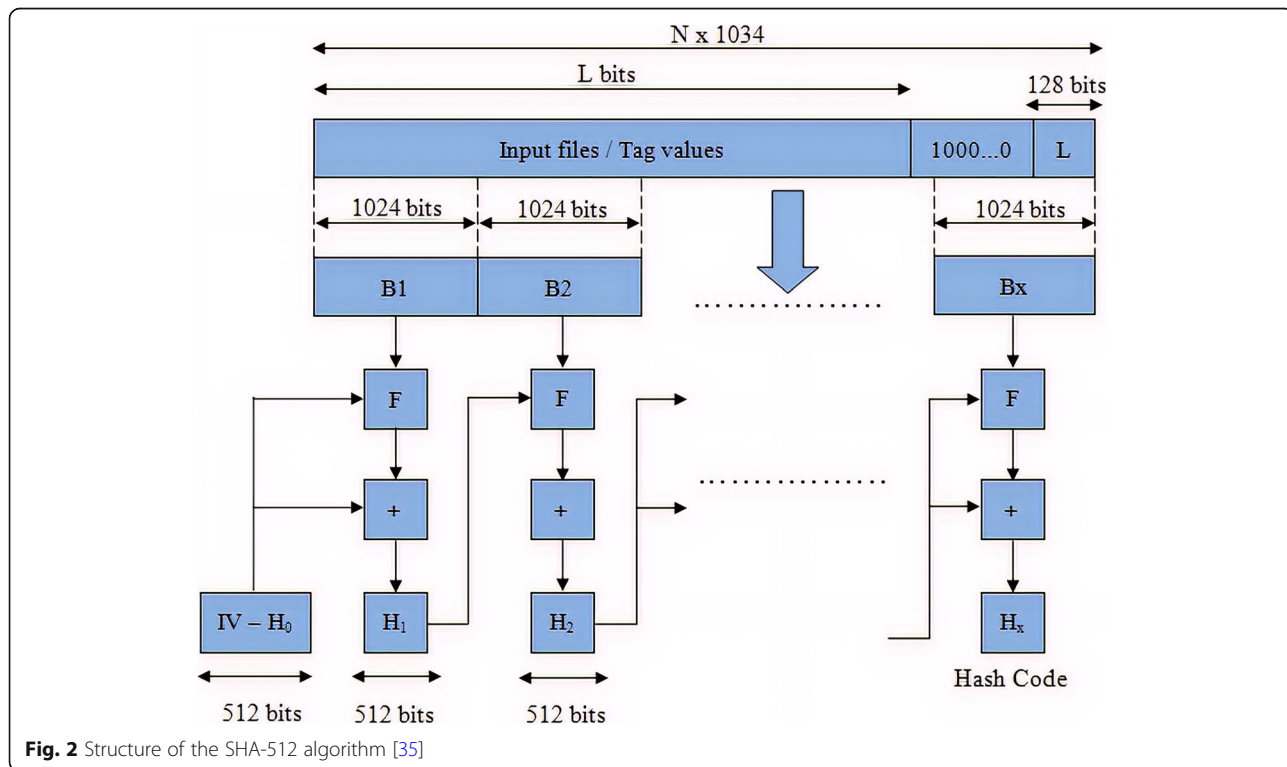


Fig. 2 Structure of the SHA-512 algorithm [35]

Convergent encryption

During encryption, the original file (f) and κ_h are given as input for the encryption algorithm and E_y is the encryption function. Finally, this encryption algorithm gives ciphertext (C_t) as output.

$$C_t = E_y(f, \kappa_h) \tag{1}$$

Convergent decryption

During decryption, the encrypted file f or C_t is inputted to the decryption algorithm. Finally, this decryption algorithm outputs f and C_t .

$$f = D_y(C_t, \kappa_h) \tag{2}$$

The CE algorithm gives better performance when compared with other existing methods, but it has a limitation as the CE is not secure since it may be affected by the dictionary attack. To avoid this, the proposed methodology once again encrypts the above convergent encrypted file utilizing the Modified ECC algorithm, which is discussed as follows.

Elliptic Curve Cryptography (ECC) algorithm is centered on a curve with specific base points and a prime number function. This function is utilized as a maximum limit. ECC is a kind of algorithm that is used in the implementation of public-key cryptography. The mathematical model of the ECC with g and e as integers is given below.

$$w^2 = v^3 + gv + e, 4g^3 + 27e^2 \neq 0 \tag{3}$$

In a cryptographic procedure, the potency of the encryption technique depends mainly on the mechanism that is deployed for the key generation. In the recommended system, three types of keys have to be generated. The main step is to generate the public key (α_k) from the server and encrypting it. In the next step, a private key (β_k) is produced on the server-side, and the message is decrypted. The last step is to generate a secret key (o_k) from α_k , β_k , and point on the curve (ρ_c). Using the succeeding equation, the α_k is generated,

$$\alpha_k = \beta_k * \rho_c \tag{4}$$

The eq. (5) elucidates o_k generation,

$$o_k = \alpha_k * \beta_k * \rho_c \tag{5}$$

After o_k generation, the file is encrypted. This encrypted file contains two CTs, and mathematically, they are depicted as,

$$C_1 = ((K=1, 2, \dots, (n-1)) * \rho_c) + o_k \tag{6}$$

$$C_2 = (f + ((K=1, 2, \dots, (n-1)) * \alpha_k)) + o_k \tag{7}$$

Here, C_1 and C_2 represents the two CTs, K is the random number generated in $(1, \dots, (n-1))$ interval. During encryption, o_k is added to the CTs. During decryption, o_k is subtracted with the two CTs, and the original file f is given by,

$$f = (((C_2 - \beta_k) * C_1) - o_k) \tag{8}$$

When the same user tries to upload the same file

When the same user tries to upload the same file again, the CS calculates the hash value with the CK by utilizing the SHA 512 algorithm. Next, for every single input file, the binary depiction of the file is split into fixed-sized blocks. The size of the data block finds the level of granularity of deduplication. As the data block size decreases, the level of deduplication increases. Meanwhile, it might bring complex metadata management. The proposed approach considered the file block-sizes of 5 MB, 10 MB, 15 MB, 20 MB, and 25 MB. Then, the tag key is created for each of the divided blocks. Next, the hash value is computed for all the tag keys utilizing the same SHA-512 algorithm.

In the uploading phase, the CS checks the hashtag (HT) for a particular input file. If the hashtag value of the input file is in that HT, then the CS queries the path of the hash tree to the users. If a user sends the correct path, then the CS verifies the user id. If the id is the same, then the CS does not store the file again. Generally, the hash tree path has the succeeding format,

$$P(H_t) = \{RLL, RLR, etc.\} \tag{9}$$

Where $P(H_t)$ denotes the path of the hash tree, RLL represents the ‘‘Root, Left, Left’’, RLR , denotes the ‘‘Root, Left, Right.’’ The leaf node is not added to the hashed tree path. The same user trying to upload the same file is mathematically denoted as,

$$S_u \xrightarrow{\text{Uploads}} S_f \left(CS \xrightarrow{\text{Informs}} A \right) \tag{10}$$

Where, S_u denotes the same user, S_f represents the same file, and CS means the cloud server, which informs the file is previously available (A).

When different users try to upload the same file to the cloud server

When different users try to upload the same files to the CS, the file is split into several blocks, and a tag is created for checking the duplicate data copies in CSP. Then, each tag is converted into HC, and it is called a hashtag value. The CS checks the HT for the input file grounded on the hashtag value. If the hashtag value is available in the HT,

then the CS asks the path of the hash tree of the input file. If a user sends the correct path, then the CS verifies the user id. If the id is different, then the CS sends the reference link of the particular stored file's location to the user. Different users trying to upload the same file to the CS is expressed as,

$$D_u \xrightarrow{\text{Uploads}} S_f \left(CS \xrightarrow{\text{asks}} P(H_t) \right) \quad (11)$$

$$P(H_t)_{\text{matched}} \rightarrow \left(CS \xrightarrow{\text{Send}} R_t(f) \right) \quad (12)$$

$$P(H_t)_{\text{Not matched}} \rightarrow \left(CS \xrightarrow{\text{Informed}} I_u \right) \quad (13)$$

Where D_u denotes the different users. R_t is the reference link and I_u denotes an invalid user.

When users try to download the file

Here, the user sends the tag value of the specified file. Then, the CS generates the hash value utilizing the SHA512 algorithm. The CS now checks the hashtag value, whether it is in the HT. If the value is available, then the CS lets the user download the file, else the CS considers them as an invalid user. It is mathematically denoted as,

$$H(T)_{\text{matched}} \rightarrow D_u(\downarrow) \quad (14)$$

$$H(T)_{\text{Notmatched}} \rightarrow I_u \quad (15)$$

Where $H(T)$ denotes the hashtag value. Pseudocode for the proposed secure deduplication system is evinced below,

Algorithm 1: Uploading file into the Cloud Server

Input: Original file

Output: Upload the file into the CS

begin

initialize key k , tag T , hashtag $H(T)$,

blocks (B_1, B_2, \dots, B_n) , and hash tree H_t

for n -number of files **do**

{

generate k using SHA-512

divide f into blocks (B_1, B_2, \dots, B_n)

generate tag T

generate $H(T)$ using SHA-512

if $(H(T) == Hash_table)$ **then**

check PoW

else

generate H_t and insert $H(T)$ into the-hash table and encrypt the file f

store the file in Cloud Server (CS)

end if

}

end for

end

Algorithm 2: Downloading file from the Cloud Server

Input: Tag value

Output: Download the original file

begin

initialize tag T , hashtag $H(T)$, original-file f

for all tags **do**

{

generate $H(T)$ using SHA-512

if $(H(T) == Hashtable)$ **then**

Cloud Server allows the user to download-the file $f(\downarrow)$

else

Cloud Server informs as invalid user

end if

}

end for

end

Result and discussions

The implemented deduplication methodology is deployed in the JAVA programming environment with the following system configuration. The system performance is analyzed, centered on different file sizes varying from like 5 MB to 25 MB with an increase of 5 MB after each iteration. In this section, the performance scrutiny is done on the proposed system. First, the performance revealed by the proposed MECC security algorithm is contrasted to the existing security algorithms, say, Diffie-Hellman (DH), ECC and Rivest Shamir Adelman (RSA) in respect of encryption time (Et), decryption time (Dt), key generation time, and security analysis.

Performance analysis of proposed encryption technique

Encryption time

E_t is considered as the time that an encryption algorithm consumes to generate encrypted data as of the inputted data. Encryption time is computed as the difference between the encryption ending time and encryption starting time. It is evaluated as,

Table 1 Performance Comparison of Proposed MECC and Existing Techniques in terms of Encryption Time

File Size in MB	Encryption Time (sec)			
	Proposed MECC	DH	Existing ECC	Existing RSA
5	6.21	10.22	14.47	12.44
10	10.04	19.64	22.56	21.76
15	15.54	28.32	28.00	30.35
20	21.0	36.33	36.65	35.61
25	27.55	46.29	44.32	47.28

Table 2 Performance Comparison of Proposed MECC and Existing Techniques in terms of Decryption Time

File Size in MB	Decryption Time (sec)			
	Proposed MECC	DH	Existing ECC	Existing RSA
5	5.28	12.21	13.50	11.51
10	10.22	18.11	22.66	20.53
15	16.74	26.43	29.43	28.54
20	20.36	34.56	38.64	36.87
25	25.44	45.44	45.81	48.43

$$E_t = E_e - E_s \tag{16}$$

Where E_t is the encryption time, E_e is the encryption ending time and E_s is the encryption starting time.

Decryption time

D_t is defined as the difference between the decryption ending time and decryption starting time. It is evaluated as,

$$D_t = D_e - D_s \tag{17}$$

where D_t is the decryption time, D_e is the decryption ending time and D_s is the decryption starting time.

Security

Security is highly essential for cloud storage. The security level is computed by dividing the hacked data with the number of the original text. The security level of the system is expressed as,

$$S = H_d / O_d \tag{18}$$

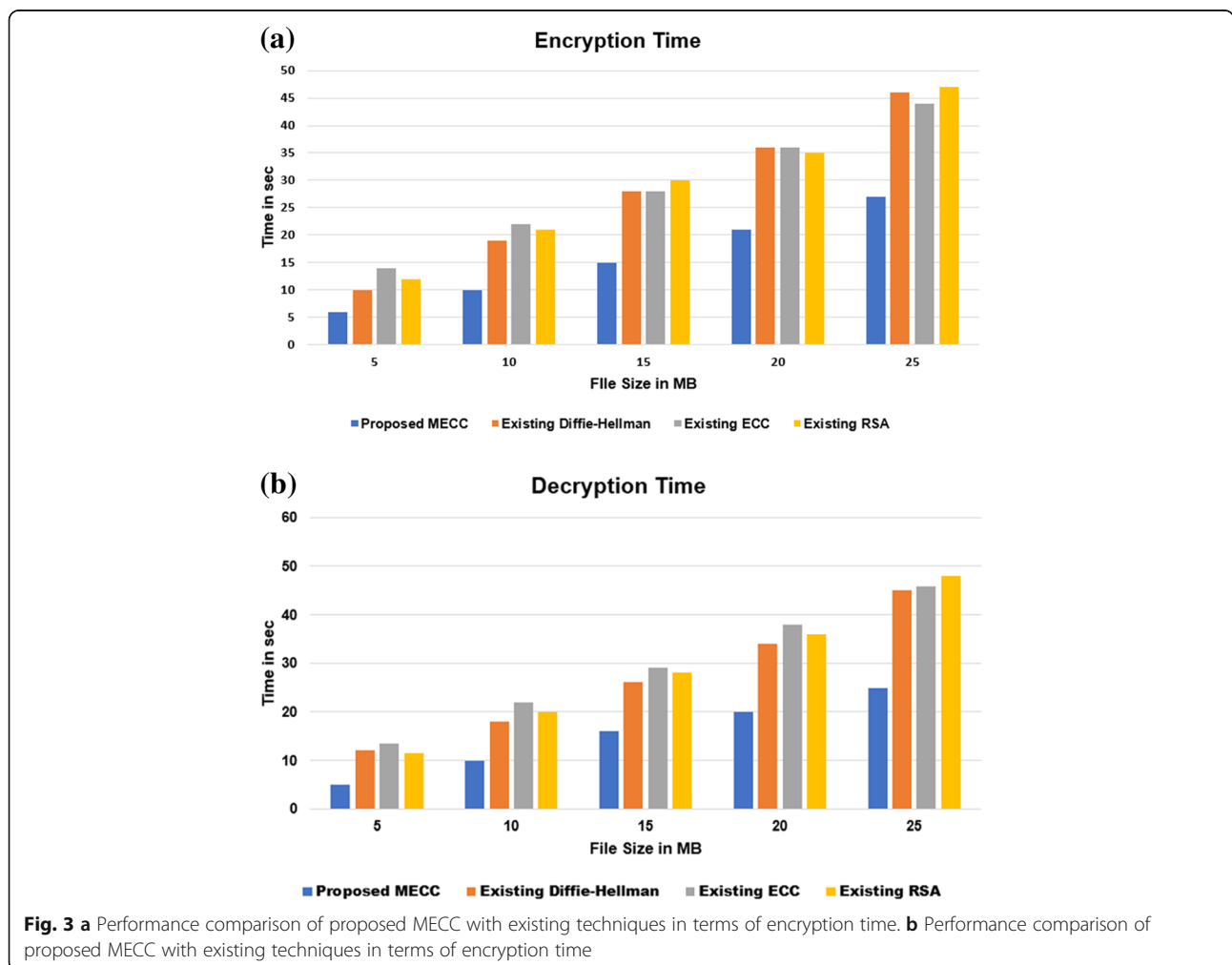


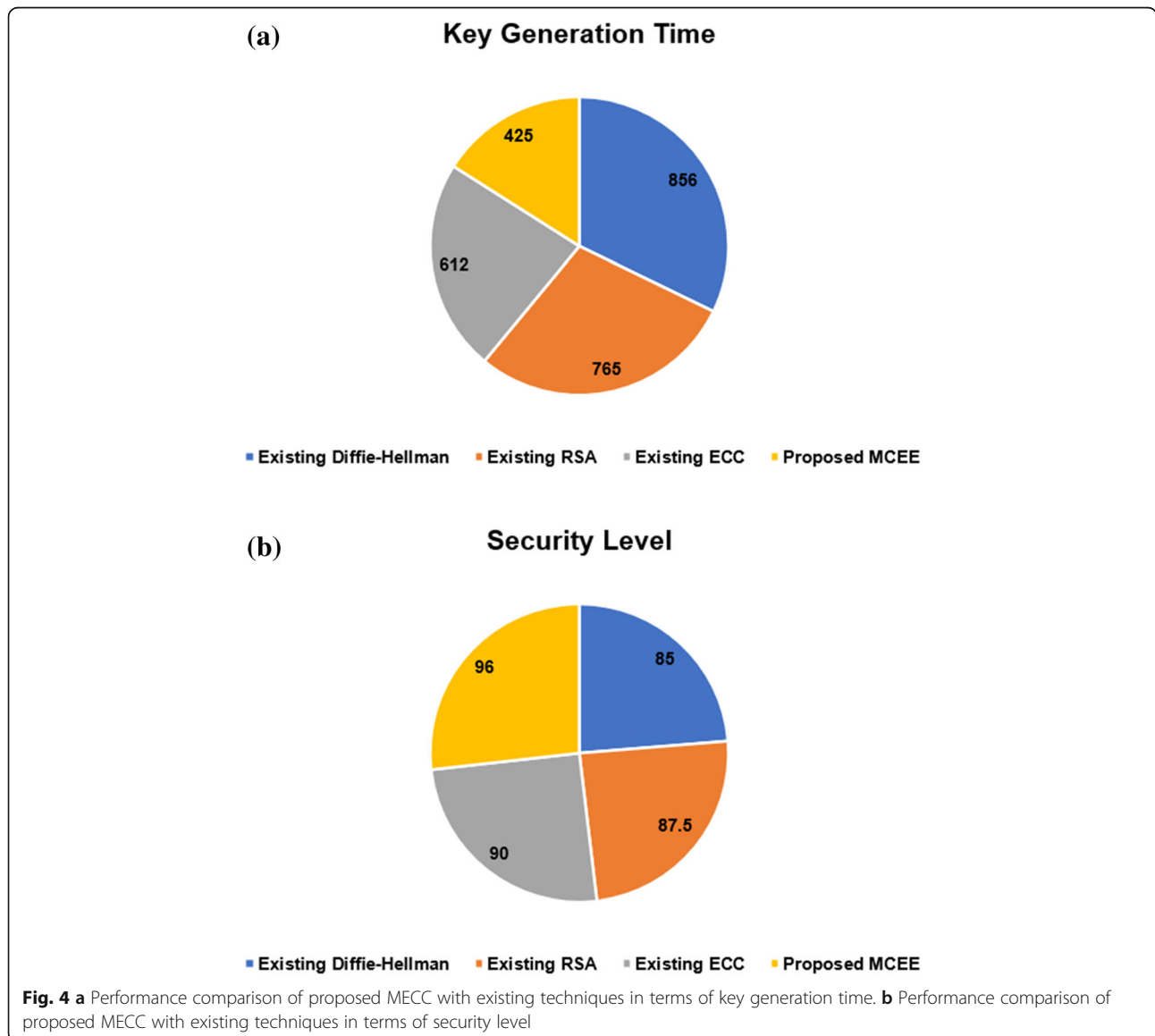
Table 3 Performance Comparison of Proposed MECC and Existing Techniques in terms of Key Generation Time and Security Level

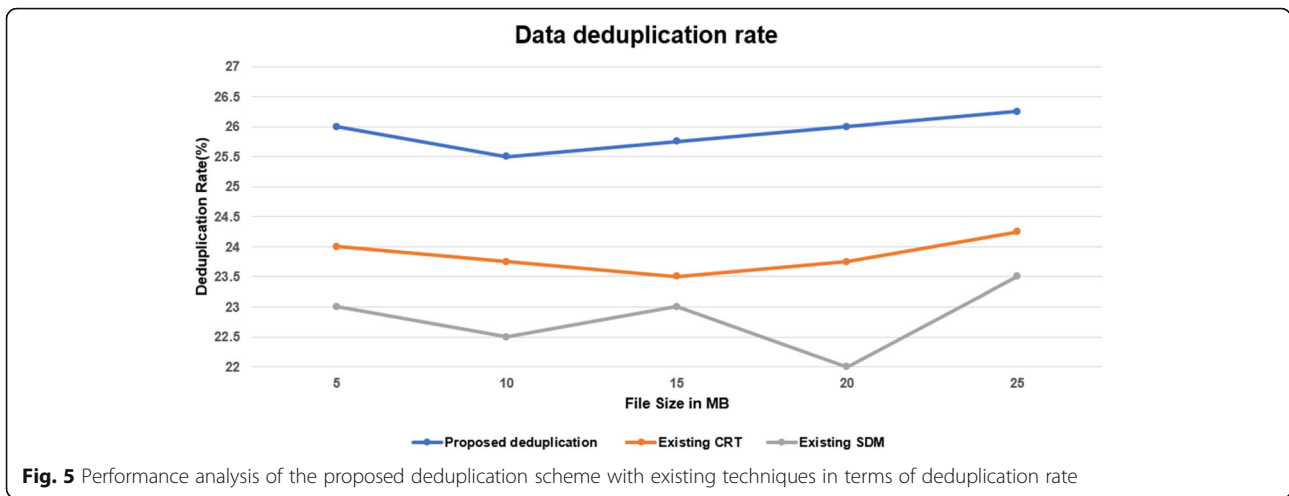
Sl. No	Encryption Algorithms	Key Generation Time (ms)	Security (%)
1	Proposed MECC	425.21	96
2	ECC	612.32	90
3	RSA	765.54	87.5
4	DH	856.33	85

where S denotes the security level, H_d is a hacked data, and O_d denotes the number of original data.

Tables 1 and 2 elucidate the performance comparison of proposed MECC with the prevailing DH, RSA, and ECC techniques concerning E_t and D_t . The comparison is performed, centered on the uploaded file size. The E_t and D_t are denoted in seconds (s). The proposed MECC takes 6 s to encrypt the 5mb file, whereas the existing

DH, ECC, and RSA take 10, 14, and 12 s to encrypt the same 5mb file, which is high when contrasted to the proposed MECC. Similarly, for the remaining file sizes (10 to 25 MB), the proposed method takes less time to encrypt the data. For the same 5 Mb file, the proposed MECC takes 5 s to decrypt the data, but the prevailing DH, RSA, and ECC takes 12, 13.5, and 11.5 s to decrypt the data. So, it is inferred that the suggested MECC





algorithm takes less E_t and D_t when contrasted to the remaining techniques. Tables 1 and 2 are graphically plotted and are displayed in Fig. 3a and b.

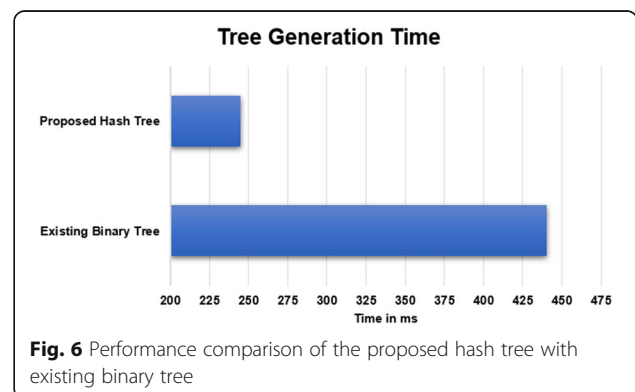
Fig. 3a and b analyze the performance proffered by the proposed MECC approach with other existing techniques. The E_t and D_t time varies centered on the file sizes. Here, the file sizes range from 5mb to 25mb. For 10mb file size, the E_t and D_t of the MECC are 10s, but the existing DH, RSA, and ECC take 19 s, 21 s, and 22 s for encryption and 18 s, 20s, and 22 s for decryption. Similarly, for all file sizes, the proposed MECC takes lesser E_t and D_t . So, it is deduced that the MECC attains the best performance when contrasted to others.

Table 3 compared the performance rendered by the proposed MECC technique with the prevailing methods concerning the key generation time and security level. The proposed MECC takes 425 ms to generate a key, whereas the existing ECC, RSA, and DH methods take 612 ms, 765 ms, and 856 ms to generate a key. Here, the existing Diffie-Hellman (DH) method takes more time for key generation. But the proposed MECC takes less time to generate a key when contrasted to other techniques. Furthermore, the security level of the proposed and existing methods is compared with existing techniques. The proposed MECC gives the highest security value (96%), but the existing ECC, RSA, and DH methods give 90%, 87.5%, and 85% of security. So, it is inferred that the proposed MECC proffers high performance for both key generation and security. Table 2 is graphically illustrated as displayed in Fig. 4a for the Key Generation Time and Fig. 4b for the Security Level.

Performance analysis of proposed deduplication technique

The proposed deduplication scheme is contrasted to existing techniques such as the Chinese Remainder Theorem (CRT) centered secret sharing and Smart Deduplication for Mobile (SDM) in respect of deduplication rate, which is evinced in Fig. 5. Here, the proposed deduplication scheme is contrasted to other techniques concerning the deduplication rate and tree generation time.

Figure 5 contrasts the performance of the proposed deduplication with the prevailing methods, say CRT and SDM. The deduplication rate varies centered on the file size. For the 5mb file, the deduplication rate of the proposed method is 26%, whereas the existing SDM and CRT give 23% and 24%, which are relatively low when contrasted to the proposed method. For the 25mb file, the existing SDM and CRT give 23.5% and 24.2% of the



deduplication rate, but the deduplication rate of the proposed method is 26.2%. Similarly, for other file sizes such as 10mb, 15mb, and 20mb, the proposed deduplication scheme gives superior results contrasted to CRT and SDM. The performance comparison of the proposed hash tree used in deduplication with the existing binary tree in respect of tree generation time is evinced in Fig. 6.

Figure 6 contrasts the performance shown by the proposed hash tree with the existing binary tree regarding tree generation time. The proposed hash tree takes 245 ms to generate a tree, whereas the existing binary tree takes 440 ms to generate a tree, which is high when contrasted to the proposed hash tree generation approach. So it is deduced that the proposed hash tree approach shows high-level performance compared to binary tree generation methodology.

Conclusion

Deduplication is the utmost notable Data compression methodology. Many existing methods introduced different deduplication methods, but they provided low security. This paper proposed a secure deduplication system using convergent and MECC algorithms over the cloud-fog environment. The proposed method is analyzed in four ways: a) when new users try to upload the new file, b) when the same user tries to upload the same file, c) when different users try to upload the same file, and d) when different users try to download the file. The performance of the recommended system was analyzed by using various file sizes ranging from 5 MB to 25 MB, with an incremental of 5 MB each in every iteration. The performance analysis corroborated that the recommended system has 96% security, which is a promising result and higher than the other existing encryption methods.

The assessment result elucidates that the recommended system is extremely secure and effective for data deduplication for an integrated cloud environment. This proposed model may be extended in the future for any kind of Internet of Things (IoT) applications that use dynamic resources management at the edge environment. It can also be used in building cyber-physical systems by studying the different use cases having different payload with the variant data formats. The proposed technique would certainly be a promising model for increasing the security and optimizing the computation time and storage in an integrated environment such as IoT or cyber-physical systems.

Abbreviations

DD: Data deduplication; MECC: Modified Elliptic Curve Cryptography; PoW: Proofs of Ownership; CE: Convergent Encryption; SHA: Secure Hash Algorithm; CS: Cloud Server; CC: Cloud Computing; HMAC: Hash-based Message Authentication Code; CT: Ciphertext; PRE: Proxy re-encryption;

ECC: Elliptic Curve Cryptography; CRT: Chinese Remainder Theorem (CRT); SDM: Smart Deduplication for Mobile; DH: Diffie Hellman Algorithm

Acknowledgements

Authors thank editor and reviewers for their time and consideration.

Authors' contributions

All authors have participated in the design of the proposed method and practical implementation. SPG and MRK have coded the method. SPG, NRK, VGM, VP, MA and MRK have completed the first draft of this paper. All authors have read and approved the manuscript.

Authors' information

Not applicable.

Funding

Not applicable.

Availability of data and materials

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India. ²Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala 683576, India. ³Department of Mechanical Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala 683576, India. ⁴Department of Computer Engineering, Engineering Faculty, Bu-Ali Sina University, Hamedan 65178-38695, Iran. ⁵Department of Computer Engineering, Persian Gulf University, Bushehr 75169-13817, Iran. ⁶Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz 71555-313, Iran.

Received: 23 March 2020 Accepted: 5 November 2020

Published online: 19 November 2020

References

- Lohstroh M, Kim H, Eidson JC et al (2019) On enabling Technologies for the Internet of important things. *IEEE Access* 7:27244–27256. <https://doi.org/10.1109/ACCESS.2019.2901509>
- Abbas N, Zhang Y, Taherkordi A, Skeie T (2018) Mobile edge computing: a survey. *IEEE Internet Things J* 5:450–465. <https://doi.org/10.1109/JIOT.2017.2750180>
- Ren J, Zhang D, He S et al (2019) A survey on end-edge-cloud orchestrated network computing paradigms: transparent computing, mobile edge computing, fog computing, and cloudlet. *ACM Comput Surv* 52. <https://doi.org/10.1145/3362031>
- Zhang P, Liu JK, Richard Yu F et al (2018) A survey on access control in fog computing. *IEEE Commun Mag* 56:144–149. <https://doi.org/10.1109/MCOM.2018.1700333>
- Menon VG, Jacob S, Joseph S, Almagrabi AO (2019) SDN powered humanoid with edge computing for assisting paralyzed patients. *IEEE Internet Things J*:1. <https://doi.org/10.1109/jiot.2019.2963288>
- Menon VG, Prathap J (2017) Vehicular fog computing. *Int J Veh Telemat Infotain Syst* 1:15–23. <https://doi.org/10.4018/jvvtis.2017070102>
- Liu J, Zhang Q (2018) Offloading schemes in Mobile edge computing for ultra-reliable low latency communications. *IEEE Access* 6:12825–12837. <https://doi.org/10.1109/ACCESS.2018.2800032>
- Li S, Zhang N, Lin S et al (2018) Joint admission control and resource allocation in edge computing for internet of things. *IEEE Netw* 32:72–79. <https://doi.org/10.1109/MNET.2018.1700163>
- Nadesh RK, Aramudhan M (2018) TRAM-based VM handover with dynamic scheduling for improved QoS of cloud environment. *Int J Internet Technol Secur Trans*:8. <https://doi.org/10.1504/IJITST.2018.093340>
- Ning Z, Kong X, Xia F et al (2019) Green and sustainable cloud of things: enabling collaborative edge computing. *IEEE Commun Mag* 57:72–78. <https://doi.org/10.1109/MCOM.2018.1700895>

11. Rajesh S, Paul V, Menon VG, Khosravi MR (2019) A secure and efficient lightweight symmetric encryption scheme for transfer of text files between embedded IoT devices, pp 1–21
12. Zhang J, Chen B, Zhao Y et al (2018) Data security and privacy-preserving in edge computing paradigm: survey and open issues. *IEEE Access* 6:18209–18237. <https://doi.org/10.1109/ACCESS.2018.2820162>
13. Nadesh RK, Srinivasa Perumal R, Shynu PG, Sharma G (2018) Enhancing security for end users in cloud computing environment using hybrid encryption technique. *Int J Eng Technol* 7
14. Abbasi M, Rafiee M, Khosravi MR et al (2020) An efficient parallel genetic algorithm solution for vehicle routing problem in cloud implementation of the intelligent transportation systems. *J Cloud Comput* 9. <https://doi.org/10.1186/s13677-020-0157-4>
15. Subramanian N, Jeyaraj A (2018) Recent security challenges in cloud computing. *Comput Electr Eng* 71:28–42. <https://doi.org/10.1016/j.compeleceng.2018.06.006>
16. Jiang S, Jiang T, Wang L (2017) Secure and efficient cloud data Deduplication with ownership management. *IEEE Trans Serv Comput* 12: 532–543. <https://doi.org/10.1109/TSC.2017.2771280>
17. Yoon MK (2019) A constant-time chunking algorithm for packet-level deduplication. *ICT Express* 5:131–135. <https://doi.org/10.1016/j.icte.2018.05.005>
18. Wang L, Wang B, Song W et al (2019) Offline privacy preserving proxy re-encryption in mobile cloud computing. *Inf Sci (Ny)* 71:38–43. <https://doi.org/10.1016/j.jksuci.2019.05.007>
19. Wang L, Wang B, Song W, Zhang Z (2019) A key-sharing based secure deduplication scheme in cloud storage. *Inf Sci (Ny)* 504:48–60. <https://doi.org/10.1016/j.ins.2019.07.058>
20. Kwon H, Hahn C, Kim D, Hur J (2017) Secure deduplication for multimedia data with user revocation in cloud storage. *Multimed Tools Appl* 76:5889–5903. <https://doi.org/10.1007/s11042-015-2595-4>
21. Akhila K, Ganesh A, Sunitha C (2016) A study on Deduplication techniques over encrypted data. *Procedia Comput Sci* 87:38–43. <https://doi.org/10.1016/j.procs.2016.05.123>
22. Kwon H, Hahn C, Kang K, Hur J (2019) Secure deduplication with reliable and revocable key management in fog computing. *Peer-to-Peer Netw Appl* 12:850–864. <https://doi.org/10.1007/s12083-018-0682-9>
23. Li J, Chen X, Li M et al (2014) Secure deduplication with efficient and reliable convergent key management. *IEEE Trans Parallel Distrib Syst* 25: 1615–1625. <https://doi.org/10.1109/TPDS.2013.284>
24. Koo D, Hur J (2018) Privacy-preserving deduplication of encrypted data with dynamic ownership management in fog computing. *Futur Gener Comput Syst* 78:739–752. <https://doi.org/10.1016/j.future.2017.01.024>
25. Liu J, Wang J, Tao X, Shen J (2017) Secure similarity-based cloud data deduplication in ubiquitous city. *Pervasive Mob Comput* 41:231–242. <https://doi.org/10.1016/j.pmcj.2017.03.010>
26. Li S, Xu C, Zhang Y (2019) CSED: client-side encrypted deduplication scheme based on proofs of ownership for cloud storage. *J Inf Secur Appl* 46:250–258. <https://doi.org/10.1016/j.jisa.2019.03.015>
27. Tawalbeh LA, Saldamli G (2019) Reconsidering big data security and privacy in cloud and mobile cloud systems. *J King Saud Univ - Comput Inf Sci*. <https://doi.org/10.1016/j.jksuci.2019.05.007>
28. Zhang P, Huang P, He X et al (2017) Resemblance and mержence based indexing for high performance data deduplication. *J Syst Softw* 128:11–24. <https://doi.org/10.1016/j.jss.2017.02.039>
29. Hou H, Yu J, Hao R (2019) Cloud storage auditing with deduplication supporting different security levels according to data popularity. *J Netw Comput Appl* 134:26–39. <https://doi.org/10.1016/j.jnca.2019.02.015>
30. Khanaa V, Kumaravel A, Rama A (2019) Data deduplication on encrypted big data in cloud. *Int J Eng Adv Technol* 8:644–648. <https://doi.org/10.35940/ijeat.F1188.08865219>
31. Stanek J, Kencl L (2018) Enhanced secure Thresholded data Deduplication scheme for cloud storage. *IEEE Trans Dependable Secur Comput* 15:694–707. <https://doi.org/10.1109/TDSC.2016.2603501>
32. Zeng P, Choo KKR (2018) A new kind of conditional proxy re-encryption for secure cloud storage. *IEEE Access* 6:70017–70024. <https://doi.org/10.1109/ACCESS.2018.2879479>
33. Wu J, Li Y, Wang T, Ding Y (2019) CPDA: a confidentiality-preserving Deduplication cloud storage with public cloud auditing. *IEEE Access* 7: 160482–160497. <https://doi.org/10.1109/ACCESS.2019.2950750>
34. Awad A, Matthews A, Qiao Y, Lee B (2018) Chaotic searchable encryption for Mobile cloud storage. *IEEE Trans Cloud Comput* 6:440–452. <https://doi.org/10.1109/TCC.2015.2511747>
35. Shynu P G, John Singh K (2016) A comprehensive survey and analysis on access control schemes in cloud environment. *Cybern Inf Technol* 16:19–38. <https://doi.org/10.1515/cait-2016-0002>
36. Alazab M (2015) Profiling and classifying the behavior of malicious codes. *J Syst Softw* 100:91–102. <https://doi.org/10.1016/j.jss.2014.10.031>
37. Benzaid C, Lounis K, Al-Nemrat A et al (2016) Fast authentication in wireless sensor networks. *Futur Gener Comput Syst* 55:362–375. <https://doi.org/10.1016/j.future.2014.07.006>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)

Efficient Flow Processing in 5G-Envisioned SDN-Based Internet of Vehicles Using GPUs

Mahdi Abbasi¹, Ali Najafi, Milad Rafiee², Mohammad R. Khosravi³,
Varun G. Menon⁴, *Senior Member, IEEE*, and Ghulam Muhammad⁵, *Senior Member, IEEE*

Abstract—In the 5G-envisioned Internet of vehicles (IoV), a significant volume of data is exchanged through networks between intelligent transport systems (ITS) and clouds or fogs. With the introduction of Software-Defined Networking (SDN), the problems mentioned above are resolved by high-speed flow-based processing of data in network systems. To classify flows of packets in the SDN network, high throughput packet classification systems are needed. Although software packet classifiers are cheaper and more flexible than hardware classifiers, they could only deliver limited performance. A key idea to resolve this problem is parallelizing packet classification on graphical processing units (GPUs). In this paper, we study parallel forms of Tuple Space Search and Pruned Tuple Space Search algorithms for the flow classification suitable for GPUs using CUDA (Compute Unified Device Architecture). The key idea behind the offered methodology is to transfer the stream of packets from host memory to the global memory of the CUDA device, then assigning each of them to a classifier thread. To evaluate the proposed method, the GPU-based versions of the algorithms were implemented on two different CUDA devices, and two different CPU-based implementations of the algorithms were used as references. Experimental results showed that GPU computing enhances the performance of Pruned Tuple Space Search remarkably more than Tuple Space Search. Moreover, results evinced the computational efficiency of the proposed method for parallelizing packet classification algorithms.

Index Terms—5G, flow processing, GPU, Internet of Vehicles, intelligent transport systems, SDN.

I. INTRODUCTION

OUR intelligent vehicular world is connected by the Internet of vehicles (IoV). With the advent of the fifth-generation (5G) wireless networks, the significant development of network bandwidth and growth of hardware

technologies has increased the speed of communications considerably [2]–[4]. That is, the communications speed is reached to terabits per second. SDN enhances the performance by accelerating network systems to process the packets with the rate of vehicular network communications [3], [5], [6]. For this purpose, a new technology, packet classification, is used in the architecture of modern network systems, which lets them be flow-aware. In such systems, after classifying the received packets into flows, the corresponding actions are performed on each flow. A variety of modern network systems, including high-speed core routers, network firewalls, intelligent intrusion detection systems, and high-level network management systems, run packet classification as their fundamental process. In SDN-based IoV, as shown by Fig. 1, the flows of data produced by billions of IoV sensing devices are transferred to SDN controllers via high-speed switches and routers [7]. The intermediate SDN switches should apply flow-based actions on the received streams and process them at the speed of vehicular network links. Therefore, to accelerate the flow classification as the fundamental process of these systems, GPUs with their rich computational resources are highly interested.

To classify an incoming packet, first, its header information is compared against the filters of the classifier according to specified fields to find a match. In this comparison, more than one filter may match the packet, or no match may be found. In the former case, the priority of filters is used to select the best-matched filter. The action of this filter is applied to the packet. In the latter case, a default action is applied by the network processor on the unclassified packet. The standard fields extracted from the packet header to be used in packet classification include Service Type, Destination IP, Source IP, Destination Port, and Source Port. The classifier may access to other fields of the packet header like Source MAC and Destination MAC addresses [8].

Packet classification algorithms are generally implementable in hardware and software. Heavy computations of packet classification are performed on parallel processors and mainly, graphics processing units (GPUs) [9]. GPUs, these new emerging commodities, have increased the computational speed of modern systems compared to multi-core processors. By the introduction of useful modules like CUDA (Compute Unified Device Architecture) [10] and the OpenCL (Open Computing Language) [11], many pieces of research have focused on using GPUs to solve computation-intensive problems in various domains of science.

Manuscript received March 22, 2020; revised August 11, 2020 and October 4, 2020; accepted October 30, 2020. Date of publication December 7, 2020; date of current version August 9, 2021. The Associate Editor for this article was S. Garg. (*Corresponding author: Mahdi Abbasi.*)

Mahdi Abbasi, Ali Najafi, and Milad Rafiee are with the Department of Computer Engineering, Engineering Faculty, Bu-Ali Sina University, Hamedan 65178-38695, Iran (e-mail: abbasi@basu.ac.ir; a.najafi92@basu.ac.ir; m.rafee@alumni.basu.ac.ir).

Mohammad R. Khosravi is with the Department of Computer Engineering, Persian Gulf University, Bushehr 75168, Iran, and also with the Telecommunications Group, Shiraz University of Technology, Shiraz 71557-13876, Iran (e-mail: m.r.khosravi.taut@gmail.com; mohammadkhosravi@acm.org).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Karukutty 683576, India (e-mail: varunmenon@ieee.org).

Ghulam Muhammad is with the Department of Computer Engineering, College of Computer and Information Sciences, King Saud University (KSU), Riyadh 11451, Saudi Arabia (e-mail: ghulam@ksu.edu.sa).

Digital Object Identifier 10.1109/TITS.2020.3038250

1558-0016 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See <https://www.ieee.org/publications/rights/index.html> for more information.

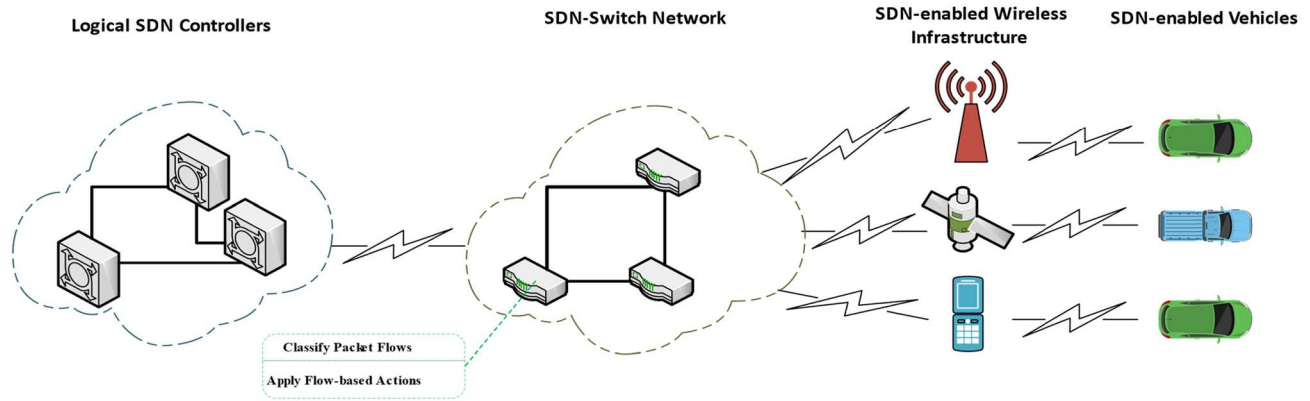


Fig. 1. SDN-Enabled internet of vehicles.

In this paper, we investigate GPUs to accelerate software-based packet classification algorithms. Two typical packet classification algorithms are selected for their simplicity and highly parallelizable structures. The current work has the following contributions:

- Different scenarios have been offered for the parallelization of two important algorithms of packet classification. The difference between the proposed plans is how the GPU uses different memory configurations. Due to the limited memory capacity and the data structure size of packet classification algorithms, a suitable method is proposed for optimal use of the GPU.
- The complexity of any parallelization scenario is highly dependent on using the memory hierarchy of GPU and the technique of exploiting the maximum concurrency among threads. Hence, we present an asymptotic analytic model to estimate the computational complexity of the proposed parallelization scenarios. This model helps us to compare the efficiency of different scenarios before running their corresponding kernels.

The rest of this paper is structured as follows. In the next section, related works on parallel packet classification via GPU are reviewed. In section 3, structures of TSS and TPS algorithms are evaluated for parallelization. In section 4, after establishing the CUDA programming model, the proposed method to parallelize TSS and TPS is explained. Experiments and their results are offered and analyzed in section 5. The conclusions and future works are presented in the final section.

II. RELATED WORK

Different classification algorithms, implementable in hardware or software, fall into one of the four categories, including *linear search*, *decision tree-based*, *decomposition-based*, and *tuple space-based* [8].

In all of the algorithms mentioned above, the classification process is launched per packet. That is, to classify each input packet, the classification algorithm is executed again. This key observation motivated limited researches to investigate the parallel forms of packet classification algorithms.

Nottingham and Irwin [11] provided valuable articles to pioneer the concept of parallel packet classification on GPU using CUDA and OpenCL platforms. However, their work does not include any implementation of algorithms and related performance comparisons. Han *et al.* [12] showed that by utilizing GPU co-processors, one might increase the classification throughput. Their proposed GPU-based IP routing algorithm, named PacketShader, could improve the performance compared with CPU-based IP routing methods. Hung *et al.* [13] presented parallelized versions of BPF and BitMap algorithms on GPUs. They also, utilized different memory architectures for GPU and compared numerical results revealing related computation performance. Kang and Deng [14] parallelized linear search and DBS algorithms and implemented them on GPU. Comparing the computation time of sequential versions of algorithms with that of GPU-based versions, they noticed that the performance of linear search on GPU is higher than DBS on GPU. Zhou *et al.* [9], [15] implemented the Bit-Vector algorithm on GPU. Their results showed that the performance of the algorithm, utilizing K computation threads on GPU, is enhanced to $\log_2 k$ times. Recently, Varvello *et al.* [16] used GPU computations to accelerate the Bloom filter algorithm. More recently, Zheng *et al.* [17] has presented an analogous study to enhance the performance of the HiCuts algorithm via parallelizing it on GPU.

Considering the researches mentioned above, some packet classification algorithms have not yet been considered for implementation on GPU. Moreover, none of the aforementioned studies have offered a general methodology to parallelize every packet classification algorithm. In the following, after reviewing the structure of the TSS and TPS, a general methodology is presented that simply makes the GPU-based implementation of these packet classification algorithms feasible.

III. BACKGROUND

A. Tuple Space Search

Srinivasan *et al.* [18] introduced the TSS technique for packet clarification. The key idea behind this method is to make the n the scope of a search on multiple fields of packet

TABLE I
TUPLE SPACE

Filter	(Source Prefix, Destination Prefix)	Tuple
R1	(00*,00*)	(2,2)
R2	(0*,01*)	(1,2)
R3	(1*,0*)	(1,1)

TABLE II
FILTER-SET

Filter	Src IP	Dst IP	Src Port	Dst Port	Protocol
R0	0001*	00001*	0,65535	25,25	6
R1	010*	0010*	53,53	443,443	4
R2	010*	001*	53,53	1024,65535	17
R3	110*	00*	53,53	443,443	4
R4	1111*	1*	53,53	25,25	4
R5	0101*	0*	0,65535	2788,2788	17
R6	0*	1101*	53,53	5632,5632	6
R7	*	10*	53,53	25,25	6
R8	1*	*	0,65535	2788,2788	17

header narrower by dividing the filters to mutually exclusive subsets according to definable “tuples”. Each tuple is a list of n values; each of them is the length of a field in a filter. For example, in filter-set on five prefix fields, the tuple [3, 5, 7, 0, 12] shows that the size of the first prefix field is 3 bits, the size of the second prefix is 5 bits, the size of the next prefix is 7, the size of the fourth one is 0 (or a wildcard field), and the size of the prefix of final field is 12.

TSS algorithm reduces the number of distinct tuples as compared with the number of filters in the original filter-set. Therefore, the filters of a filter-set are partitioned into the distinct tuple groups. Then, the algorithm performs lookups across all the identical tuples to find the most suitable filter that is highly matched with the packet. Finally, among the multiple reported filters, the highest priority one is reported as the best-matched filter.

In order to visualize the idea of tuples in the TSS algorithm, a sample filter-set is presented in Table I. The prefix of the source IP address and destination IP address of three filters are presented in Table I. The tuple of each filter is formed by putting the lengths of those prefixes together according to a predefined order. For example, since the size of the source and destination IP prefixes are 1 and 2, respectively, the tuple of R2 is (1,2).

In order to show the classification of packets with TSS, a sample filter-set containing nine filters is presented in Table II. Each filter of this table has five fields. Two binary trees are constructed using source and destination IP prefixes. For this purpose, according to the presence of 0 or 1 in the successive bits of the prefix, a branch is added to the left or right of the under-construction node of the tree.

Fig. 2 shows the binary trees corresponding to the source address field and the destination address field of the sample filter-set. Black nodes represent filters. Alongside these nodes, several relevant nodes are also provided.

The algorithm works for input (11010, 00001, 53, 443, 4) with a set of filters in Table II as follows. Input (11010, 00001, 53, 443, 4) contains five fields of IP headers. These fields are shown from left to right in Table II, respectively.

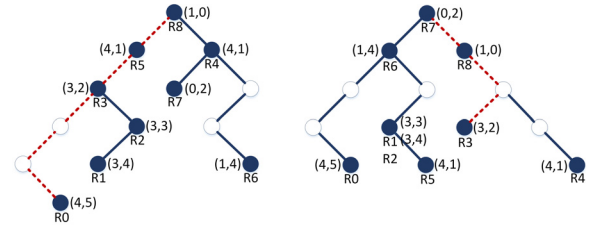


Fig. 2. The binary tree of source and destination addresses.

TABLE III
SELECTED FILTERS OF THE FILTER-SET

Filter	Src. IP	Dst. IP	Src. Port	Dst. Port	Protocol
R0	0001*	00001*	0,65535	25,25	6
R3	110*	00*	53,53	443,443	4
R4	1111*	1*	53,53	25,25	4
R5	0101*	0*	0,65535	2788,2788	17
R7	*	10*	53,53	25,25	6
R8	1*	*	0,65535	2788,2788	17

TABLE IV
SELECTED TUPLES FROM PRUNED TUPLE SPACE

Filter	Src.IP IP	Dst. IP	Src. Port	Dst. Port	Protocol
R3	110*	00*	53,53	443,443	4
R8	1*	*	0,65535	2788,2788	17

In this example, the search is performed by traversing the constructed binary trees using corresponding values extracted from the packet header. The dashed line in Fig.2 shows the search path. The filters corresponding to the tuples found in the navigation path are extracted.

Table III shows the extracted filters for this example. As can be seen, the number of these filters is a small fraction of the filter-set. As a result, a substantial depletion of the irrelevant filters in the large-size filter-sets is expectable.

Finally, to find the best-matched filter, a linear search is conducted on the small subset of filters that were resulted from the previous step. For the above example, the best adaptive filter is the R4 filter.

B. Pruned Tuple Space Search (TPS)

The original tuple space search algorithm performs a linear search on all tuples obtained from the lookups over two binary trees. But the pruned tuple space search algorithm first shares some of the results obtained by traverse in the source address and destination address tree. Then, an exhaustive search is performed on the small subset of filters that are found from the intersection of tuples. Table IV shows the result of applying the TPS algorithm to the example of Table III.

Comparison of Tables III and IV shows that the speed of the TPS algorithm is higher than that of the TSS algorithm. This is well illustrated in the evaluations carried out. But the speed of this algorithm is still far from the ideal speed. One of the best ways to accelerate the packet classification algorithms is to execute them in parallel on graphics processing units. In the last few years, there has been some work on parallelizing

packet sorting algorithms to the graphics processing unit, which will be reviewed later.

It can be easily deduced from the above discussion that the TSS and TPS algorithms for packet classification benefit significantly from the parallelizable search. In the following section, parallel implementation of the TSS and TPS packet classification algorithms is explained.

C. CUDA Programming Model

Generally, the GPUs are processing systems in which many scalar processors execute threads of code in parallel. Modern GPUs include a number of stream processors (SMs) that manage some scalar processors (SPs). To utilize the computational power of these processors, standard platforms like CUDA are provided by Nvidia [10], [19].

CUDA programs classically include two related modules, a commonly sequential module that is executable on the CPU of the system, or the host, and a heavier-but-parallel module called the kernel that runs on the SPs. The former controls the transfer of the input data for the latter from system memory to the GPU memory hierarchy. Also, it copies the computation results from the memory hierarchy of the device to the system memory. The programmer should concisely allocate the required kernel memory and minimize the kernel communication overhead.

Kernel invokes a predefined set of threads on SPs that compute results for a slice of input data. To handle these threads, several blocks of thread are defined, each of which contains 512 concurrent threads. The blocks of threads are typically organized in a two/three-dimensional grid. By using the unique index of each thread, the data elements which should be processed by that thread are specified. Each block of threads is run by a single multiprocessor, which orchestrates the execution of the kernel on the corresponding SPs and coordinates the access of threads to input data elements and also storing the computation results through shared memory.

CUDA devices have different memory modules with varying delays of access. The performance of the kernel is positively related to the memory modules used in the kernel code.

Fig. 3 illustrates the proposed plan to keep filter-sets and packets in the CUD device. Given the size of the filter-set and packet headers, we prefer to store the filter-set and packets in the shared memory and the global memory of the device, respectively. The CUDA application programming interface (API) coordinates the memory access on the device, and provides a set of functions to communicate the data between the host memory and allocated device memory. Note that, among different data-transfer models, the streaming model is the best. It is a pipeline of data transmission. In this model, sequences of operations are scheduled to be performed progressively on the CUDA device [10], [20]. This model is used for data transfer in the proposed parallelization methodology.

IV. PARALLEL IMPLEMENTATION OF TSS AND TPS

This paper presents three different methods for parallelizing the search using a GPU. The first method uses global memory, and the two other methods use shared memory. Our GPU

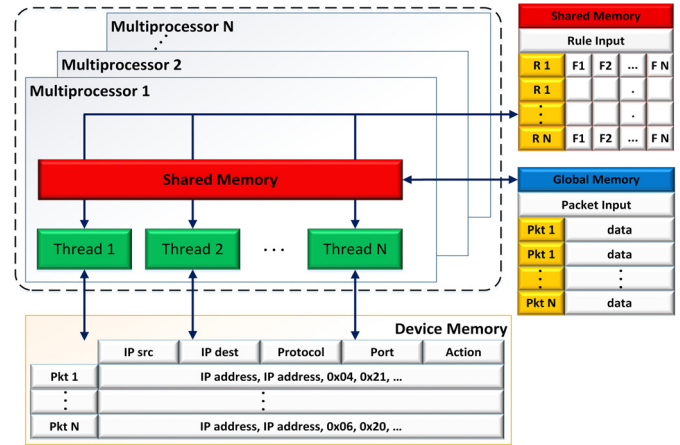


Fig. 3. The memory storage for filter-sets and packet data on CUDA device.

version of TSS and TPS are vastly parallel programs, in which fine-grained parallelism is realized by exploiting the maximum number of concurrent threads to classify packets. Note that, since the precise number of references to the memory during the runtime is not precisely determinable, the exact number of floating-point computations per packet could not be estimated. But as a lower bound, the balanced TSS/TPS algorithms should perform at least one floating-point operation for each packet classification.

Algorithm 1 represents the main steps of implementing TSS, and TPS algorithms using CUDA. Initially, the filter-set and packets are loaded from two separate files into the host memory. Two binary tries, one for storing source address prefixes of filters and the other for storing destination address prefixes of filters are constructed in host memory. Then, breadth-first traversals of these two trees are stored in host memory. Next, after allocating the required memory, these data, which reveal tree structures, filters of filter-set, and packets are transferred from the system memory to the global memory of GPU. In addition to these items, corresponding to each packet a specific space, denoted by a result, is allocated in the global memory of GPU to store its flow number.

Algorithm 1: The Parallel Form of TSS and TPS Algorithms

- 1: $Global\ Memory \leftarrow Host\ Memory (Tree, structure$
 $Packets, Filters)$
- 2: $p \leftarrow ReadPacket(threadIdx)$
- 3: $BMR \leftarrow Classification(p, tuples)$
- 4: $Result[threadIdx] \leftarrow BMR$
- 5: $HostMemory \leftarrow Result$

As shown in the second step of the pseudo-code of Algorithm 1, the proposed form for parallel kernel is very fine grain. That is, each thread picks a packet from the pool of packet in the global memory and classifies it according to the third step of the process. The output of the calcification is the identification number of the best-matched filter. As shown in Algorithm 1, in the fourth step of the pseudo-code, this result is stored in a pre-allocated memory space corresponding to

each classified packet. After classifying all packets, the result is transferred from device storage to host storage. Note that after transferring the result to hot memory, the algorithm frees the allocated space in device memory.

A. Scenario 1: Using Global Memory

In the proposed mechanism for parallelizing the packet classification operation first, the source tree and destination tree corresponding to the filters are constructed by the CPU. Then the remaining data structure including two trees, the filter-set, the packet header, and the result array is transferred from the system memory to the pre-specified memory module on GPU. The reason for keeping this data in global memory is that they are reachable to all threads.

It should be noted that the Result array stores the best matching filter of each packet within a corresponding index of it. These operations are illustrated in lines 1 and 2 of Algorithm 2. Each thread classifies the maximum number of packets equal to $\frac{\text{the number of headers}}{3072}$. The number of threads used is 3072. Finally, each thread stores the result of the classifying each packet in the position corresponding to the index of that packet in the output array. This operation is illustrated in lines 3 through 8 of Algorithm 2. The classification process is performed in parallel with all threads. When all threads are done, the output array is transferred to the host memory. This operation is illustrated in lines 9 and 10 of Algorithm 2.

Algorithm 2: Global Memory Scenario

Input : *rulesR, header H*
Output: *rule indexes I for each header h ∈ H*
 1 *array results[number of headers];*
 2 *Global Memory ← Host Memory(Source Tree structure, Destination Tree structure, H, R, results);*
 3 *tid ← threadIdx;*
 4 **while** *tid < number of headers do*
 5 *p ← Readheader(tid);*
 6 *BMR ← Classification(p, tree);*
 7 *Result[tid] ← BMR;*
 8 *tid ← tid + threadIdx;*
 9 *__syncthreads();*
 10 *HostMemory ← Result*
 11 **end While**

B. Using Shared Memory

In this study, three different scenarios of shared memory-based parallel packet classification are presented. These three methods have commonalities, which will be discussed below.

The access rate and the capacity of memory modules of GPU are reversely related. That is, by increasing the access-rate, the capacity of the memory module is decreased. In the classification of the packet on GPU, two essential parts of the dataset, including filters and headers, should be stored on the most suitable modules of the memory hierarchy of GPU. Each thread accesses the header of the packet once, but it

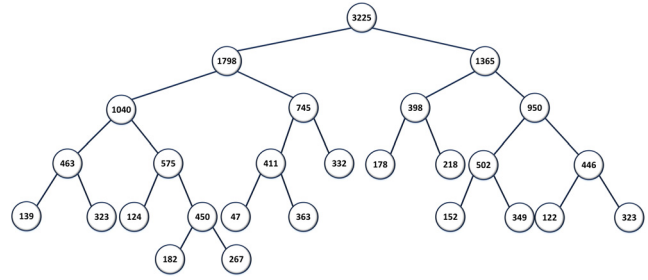


Fig. 4. Creating ACL2_1K destination tree with a 6-Block for each processing kernel.

requires access to the filters set multiple times. Therefore, keeping the filter-set in a fast memory results in a considerable decrease in the classification time of packets. On the other hand, the filter-sets in network processors of core routers include millions of entries. Regarding those two major concerns, it seems that the shared memory of GPUs with 48KB of storage is the best module for keeping the filter-set. This memory module is so advantageous since all threads in a CUDA block can simultaneously access to the content of this memory.

Each SM uses shared memory. This memory space is split between blocks in the GPU. The storage space is so small that if all this memory is allocated to a block, the source and destination trees created corresponding to the filters, will not be included in the associated memory. The solution used in this article is to break the tree from different levels. To do this, it is first estimated the amount of data that should be communicated between the CPU and the GPU. This is based on the amount of space required for tree storage and storage space. Then the blocking policy is specified. Based on the policy selected, the amount of memory allocated to each block is specified. The number of nodes in a block can be specified and used as a threshold. Then the breadth-first traversal is done on the tree and each node encountered along the path of the traversal is taken as the root of a separate tree. Then, the algorithm gets the number of allocated nodes in this new binary-tree. If the number of allocated nodes is less than the given threshold, a cut is done on the current node, and the corresponding subtree is separated from the main tree. This is done as long as the breadth-first traverse continues. It should be noted that this traverse does not include beneath cut trees. These operations are illustrated in lines 1-2 of the algorithm in Algorithm 3. Fig.4 demonstrates how to cut the ACL2_1K (Access Control Lists type 2 with 1K Rules) destination tree. The cutting process is explained below.

By estimating the number of memory transfers between the memory of the system and device, and choosing the maximum number of blocks, six blocks are defined in the processing kernel. By choosing a 6-block policy for each processor, each block reaches 8kB of shared memory, holding $512 = 8kB / 16B$ nodes. Note that the amount of space needed to store every node of trees is 16 bytes. The numbers in each node represent the sum of all nodes to the left and right of that node plus the number of filters in that node. By cutting on the target tree of ACL2_1K, the leaf nodes in Fig. 4 are each considered as the root of a separate tree. As can be seen, the number of nodes in all newly created trees is less than 512.

Algorithm 3: Shared Memory Scenarios

Input : $rulesR, headers H$
Output: $rule indexes I$ for each header $h \in H$

```

1 Source tree break into Subtrees;
2 Destination tree break into Subtrees;
3 array  $O[|rules|][|headers|]$ ;
4 for  $i \leftarrow 0$  to  $|headers| - 1$  do
5    $O[foundruleoftree][i] += 1$ 
6 end for
7 Global Memory  $\leftarrow$  Host Memory(subtrees,  $H, O$ );
8 for  $i \leftarrow 1$  to  $\text{ceilf}(|subtrees|/|blocks|)$  do
9    $tid \leftarrow threadIdx$ ;
10   $\_shared\_Tree$  tree[maximum size of the subtrees];
11  If (subtree(blockIdx.x*i)  $\neq$  Null) then
12    while  $tid <$  size of subtree(blockIdx.x*i) do
13      tree[tid] = subtree(blockIdx.x*i);
14       $tid \leftarrow tid + threadIdx$ 
15    end While
16     $\_syncthread$ s()
17  end if
18  for  $j \leftarrow 0$  to  $\text{ceilf}(|headers|/dimBlock)$  do
19     $index \leftarrow threadIdx.x + j * dimBlock$ 
20    If ( $index <$   $|headers|$ ) then
21       $O[found rule of tree][index] += 1$ 
22    end if
23  end for
24 end for
25 array results[|headers|];
26 Global Memory  $\leftarrow$  Host Memory( $R, H, Results$ );
27 Shared Memory  $\leftarrow$  Global Memory( $R$ );
28  $tid \leftarrow threadIdx$ ;
29 while  $tid <$   $|headers|$  do
30    $p \leftarrow Readheader(tid)$ ;
31    $BMR \leftarrow Classification(p)$ ;
32    $Result[tid] \leftarrow BMR$ ;
33    $tid \leftarrow tid + threadIdx$ ;
34 end While
35 Host Memory  $\leftarrow$  Result

```

With this type of cutting, the nodes on the top of the tree do not include any newly created trees. To solve this problem, a straightforward classification operation is done for each incoming packet. For this purpose, a two-dimensional array is created first. The number of packets and the size of the filter-set specifies the size of two corresponding dimensions of this array. All cells of this array are preset, first. The corresponding cells of this array are incremented in each step for the filters of the matched tuples in the traversal path. This operation is illustrated in lines 3 through 6 of Algorithm 3.

Then the new trees, the headers of the packets, along with this two-dimensional array, are transferred to the machine's global memory (line 6-7 of Algorithm 3). The next line of Algorithm 3 is executed when the number of trees created is likely to exceed the number of available blocks. In this case, the shared memory of the blocks must be emptied and reused for the remaining trees. The trees are transferred to their own

block shared memory. Each thread gets one or more nodes from the tree corresponding to its block and copies it to its block memory.

The number of nodes that each thread copies in its shared memory is obtained by $\frac{\text{the size of subtree}}{\text{dimBlock}}$ relation. Here it is necessary to make a complete copy of the tree in order to proceed to the next steps. This operation is illustrated in lines 9 through 17 of Algorithm 3.

Each thread picks $\frac{\text{the number of headers}}{\text{dimBlock}}$ of packets from the global memory and performs search operations on the tree in its block. The algorithm records the filters in the tuples encountered in the traversal path in the 2D array in the global-memory module. When all the trees are transferred to the shared-memory module and the initial search operation is performed on all packets on the trees, the two-dimensional array created in the block is transferred to the host memory. This operation is illustrated in lines 18 to 23 of the algorithm in Algorithm 3.

Then a new kernel function is created to perform the second phase of the search, which is an exhaustive search. Then the structure of the packet-headers, the filter-set, the two-dimensional array completed in the previous step, along with a new vector, and the number of test packets is copied to the pre-assigned memory of GPU to store the classification results. This operation is illustrated in lines 25 and 26 of Algorithm 3.

The linear search in the TSS is done on the filters with the corresponding value for the packet in the two-dimensional array being one or two. In the pruned tuple space algorithm, this value should be two. The reason for this difference is that the linear search in the TSS is run on the filters in the tuples that exist in the path of the source or destination tree traversal, or both, while in the pruned tuple space algorithm, the linear search on the filters which exist the ways of traversing source and destination tree. Finally, when all threads are processed, the Result array is transferred to the global storage of GPU. This transfer is performed by the functions of lines 28 to 35 of Algorithm 3. The differences between the scenarios implemented in the shared memory are presented below. Line 27 will also be executed if the linear search used in both of these is executed on the shared memory.

C. Scenario 2

This method tries to use the maximum number of blocks to decrease the number of sweeps between the CPU and GPU. Here, the key idea is to break down the tree more. This will reduce the workload of the threads in the block. Linear search is also performed on global memory in this scenario.

D. Scenario 3

The difference of this scenario with the previous one is to perform an exhaustive search on the shared memory of the block. This scenario copies the filters to the shared memory to access each thread earlier.

TABLE V
ANALYSIS PARAMETERS [1]

Parameter	Description
Q	Number of cores per core group (SMs)
L	Time for a slow global memory access
P	Total number of processors (cores)
T_1	Total number of operations in the program (work)
M	Number of global memory transactions
τ	Number of threads per core
n	Number of filters in classifier
a	Number of packets
B_a	Number of active thread blocks on each SM
B_r	Number of requested thread blocks for parallel algorithm
n_T	Number of threads on each block

V. ANALYZE USING THE CALIBRATED ASYMPTOTIC FRAMEWORK

Analytically estimating the complexity of parallel kernels is of great importance in designing resource-efficient algorithms. Recently several methods have been offered that estimate the efficiency of parallelized algorithms on many-core machines like GPUs [21]–[24].

Recently, a comprehensive method is presented for analyzing the complexity of intricate parallel algorithms on many-core computing systems like GPUs [1]. In this method, to estimate the efficiency of parallel kernels, different parameters, including the running time of the sequential algorithm, the number of SPs, the data communication time, and the number of concurrently running threads on each SP are considered. Table V, introduces the parameters that are used in this framework [1].

In our analysis model, the total time of running the parallel kernel on the graphic processor according to the proposed scenarios is obtained by equation (1):

$$Time_{total} \propto \left[\frac{a}{n_T} \right] \times \max \left(T_1, \frac{M \times L}{\tau} \right) \times \left[\frac{Q B_r}{P B_a} \right] \times \frac{1}{P} \quad (1)$$

In the above equation, T_1 and $\frac{ML}{\tau}$ represent the computational burden of the sequential deployment of the algorithm and the number of simultaneous accesses to the storage, correspondingly. The former depends on the nature of the sequential algorithm and the latter depends on the kernel policy that dictates how to deposit the data structure of the algorithm on the hierarchy of memory modules of the GPU. Given that P/Q represents the number of SMs, if $B_r > B_a \times (\frac{P}{Q})$, to completely execute the kernel, it is required to re-fill the shared memory of defined blocks $\left[\frac{Q B_r}{P B_a} \right]$ times. The $(M \times L)$ term in equation (1), represents the memory complexity of the kernel.

A. Parameters Required for the Proposed Parallel Model

According to equation (1), the parameters of our analysis in different scenarios have different values regarding the size of the filter-set, the number of the SMs and SPs, and the type of used memory modules. Table VI shows the value of these parameters for the proposed parallel model.

TABLE VI
VALUE OF REQUIRED PARAMETERS

Q	L	P	T_1	n	B_a	B_r	n_T	τ
192	100	384	$O(W^2) + O(N + N)$	1024	6	24	256	8

TABLE VII
THE MEMORY COMPLEXITY

Scenario	Memory Complexity
Scenario 1	$O(W^2) * L + O(N + N) * L$
Scenario 2	$O(N) * L + O(W^2) + O(N + N) * L$
Scenario 3	$O(N) * L + O(W^2) + O(N + N)$

The parameters B_r and B_a show the number of reconstructed subtrees and the number of exploited functional blocks in each SM.

The value of parameter T_1 is computed to the extent of $O(W^2) + O(N + N)$ given the searching process of the algorithm and the way of comparing the fields of each filter of filter-set with the values extracted from their corresponding fields of the packet header. Parameter τ is computed according to the following equation and reflects the computational load for each SP.

$$\tau = \frac{n_T \times \frac{P}{Q} \times B_a}{P} \quad (2)$$

B. The Complexity of the Proposed Scenarios

The complexity of transferring the data from the global memory to the shared memory is $O(n)$. Also, the total complexity of searching the tree to find all matching filters is $O(w^2) + O(n + n)$. Note that the type and combination of memory modules used for storing the essential data determine the worst-case memory complexity of each scenario. The memory complexity of all scenarios is calculated and shown in Table VII. The linear search in the second and third scenarios is performed in both global and shared memories. Hence, the memory complexity of the third scenario is the lowest as compared to others. Based on the parameters of the parallel model, Table VIII shows the complexity of the proposed kernels on a CUDA device with specifications represented in Table IX.

Fig. 5 shows the complexity diagram of the various scenarios presented for parallel packet classification using the TSS algorithm. The obtained complexity is calculated according to the parameter values of Table VII and the scenarios of Table IX. Given the complexity results, it is predicted that the third, first, and second scenarios will yield the best results, respectively.

VI. IMPLEMENTATION AND EVALUATION

The first part of this section is focused on the technical details of the implementation of the parallel kernels and the ClassBench tool [16]. Finally, the results of the implementation of both algorithms on different filter-sets are compared.

TABLE VIII
THE COMPUTATIONAL COMPLEXITY OF THE PRESENTED PARALLEL SCENARIOS ON THE FILTER-SET UNDER EVALUATION
BASED ON THE CALIBRATED ASYMPTOTIC FRAMEWORK

Scenario	Complexity
<i>Scenario 1</i>	$Time_{total} \propto \left\lceil \frac{a}{256} \right\rceil \times \max \left(O(W^2) + O(N + N), \frac{O(W^2) * L + O(N + N) * L}{8} \right) \times \left\lceil \frac{QB_r}{PB_a} \right\rceil \times \frac{1}{P}$
<i>Scenario 2</i>	$Time_{total} \propto \left\lceil \frac{a}{256} \right\rceil \times \max \left(O(W^2) + O(N + N), \frac{O(N) * L + O(W^2) + O(N + N) * L}{8} \right) \times \left\lceil \frac{QB_r}{PB_a} \right\rceil \times \frac{1}{P}$
<i>Scenario 3</i>	$Time_{total} \propto \left\lceil \frac{a}{256} \right\rceil \times \max \left(O(W^2) + O(N + N), \frac{O(N) * L + O(W^2) + O(N + N)}{8} \right) \times \left\lceil \frac{QB_r}{PB_a} \right\rceil \times \frac{1}{P}$

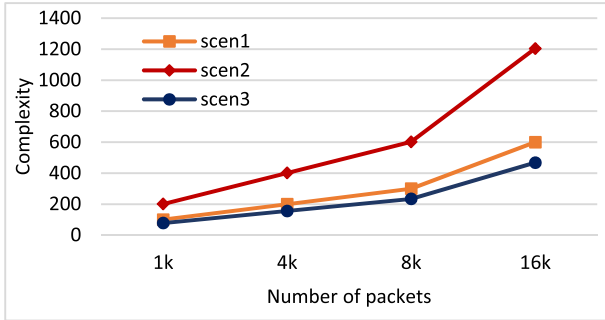


Fig. 5. The complexity of different parallel classification scenarios using TSS.

A. Implementation

The required synthetic data is produced using ClassBench. It is used to generate required filter-sets and packets based on input parameter files [25], [26]. The presence of this tool resolves the need for collecting real packet profiles, and real filter-sets of Firewalls (FW), IP Chains (IPC), and Access Control Lists (ACL). A different number of filters, related to ACL2 and FW2, with $N = 1k$ filters, was generated using ClassBench and used in all of our experiments. Corresponding to each of the filter-sets above, different profiles, including 1k to 64k synthetic packet headers were generated using ClassBench and used in all experiments. The parallelized version of TSS and TPS algorithms were developed by C programming language. The CUDA 9 platform was used to implement the GPU-based programs. The provided scenarios were tested on a CUDA device with specifications represented in Table IX.

B. Evaluation

In this section, the proposed scenarios are examined from different aspects of performance, such as classification time and throughput. The classification throughput shows the number of packets that are classified per unit time. The proposed scenarios were run the algorithms on the filter-sets ten times, and then the average of the aforementioned parameters was calculated. All times quoted are in a millisecond.

Fig. 6 shows the packet classification times of 1k to 16k of packets using the parallel TSS algorithm and the parallel TPS algorithm. Charts a and b show the results of the ACL filter-set, and the c and d charts show the result of the FW filter-set.

According to the results of the two filter-sets, the lowest classification time belongs to the third scenario. Because in

TABLE IX
SPECIFICATION OF THE SYSTEM

Device	Specification
VGA	Nvidia GeForce GT 645M (Kepler) / 2GB DDR3
CUDA cores	384
Graphics clock	708 MHz
Memory bandwidth	28.80 GB/s
Shared memory	48 KB
CPU	Intel®Core™i7-3630QM
RAM	32 GB
Operation System	Windows 7 Ultimate, 64-bit (SP1)

this scenario, a binary search is performed to match all of the packet fields in shared memory. The first scenario, where the tree is wholly held in global memory, is faster than the second scenario where the tree is formed from small sub-trees in shared memory. The results obtained from the implementation confirm the predictions made in the previous section proposed by scenarios.

After traversing two binary trees according to the corresponding fields of the packet, all of the visited tuples and all of the commonly visited tuples during the traversal of the respectively TSS, and TPS algorithms are inspected to find the best-matched filter. Since the number of inspected tuples in TPS is much lower than that of TSS, the packet classification time by the former would be considerably lower than the latter. The plots of classification time of the ACL and the FW datasets in Fig. 6 shows that in all scenarios, the TPS is faster than the TSS algorithm. For example, in the second scenario, for classifying 16K of packets, the speedup of TPS and TSS to the sequential version of the algorithms is 9.28 and 6.76, respectively. Fig. 7 shows the throughput of the proposed scenarios of the parallelization of the TSS and the TPS algorithms. Bar chart (a) shows the Throughput of ACL filter-set and bar chart (b) shows that of the FW filter-set per kilo packet per second. The Throughput of the TPS algorithm is more than TSS in all scenarios and filter-sets.

The main reason for the higher throughput of the TPS algorithms as compared to the TSS algorithm is that as compared with the latter, the former inspects a lower number of tuples for finding the best-matched filter at the final step of classifying packets. Among the scenarios presented, the third, first, and second scenarios have the highest amount of throughput, respectively. The highest pass rates in the ACL and FW filter-sets are 1028.28 and 926.46 KPPS, respectively, which are achieved in the third TPS scenario. This result introduces the third scenario as the best one for the real-time classification of the vehicular network packet.

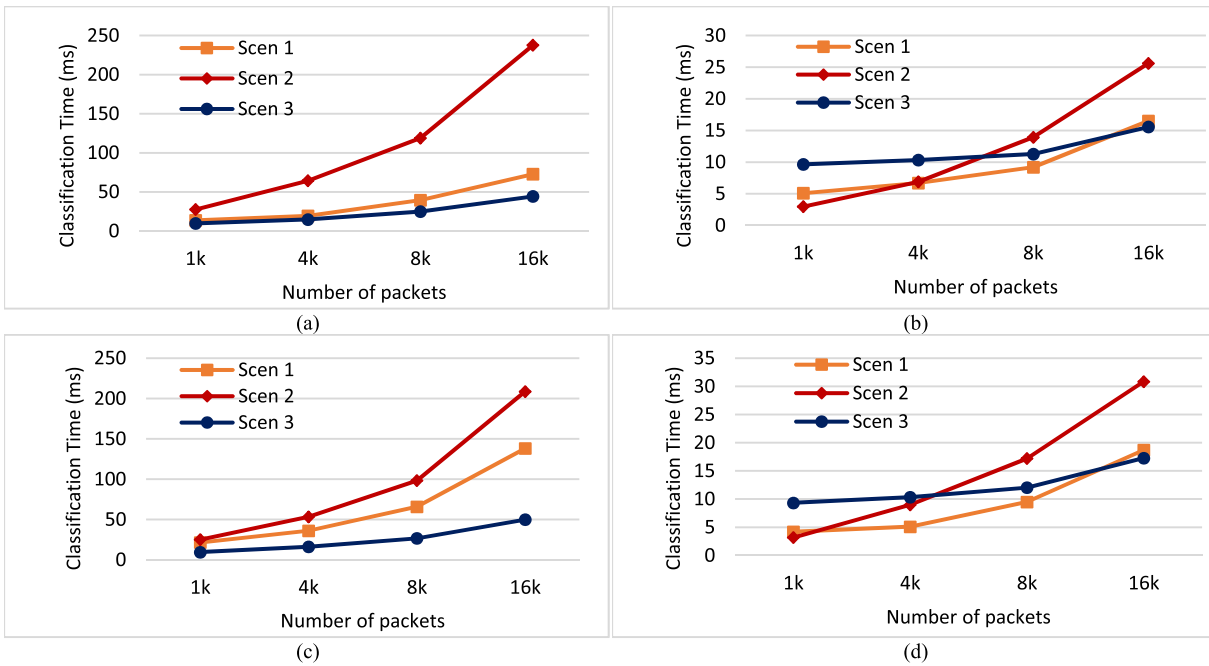


Fig. 6. Packet classification time in different scenarios and filter-sets. (a) TSS-ACL. (b) TPS-ACL. (c) TSS-FW. (d) TPS-FW.

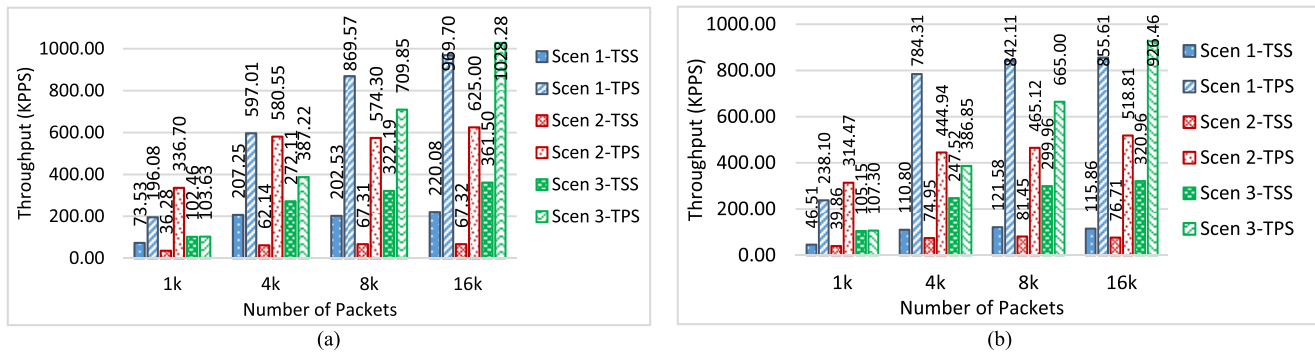


Fig. 7. Throughput of different packet classification scenarios. (a) Throughput of ACL filter-set. (b) Throughput of FW filter-set.

VII. CONCLUSION

Flow classification is considered as the key technique for accelerating the different functions of modern network systems in SDN-based IoT. The critical challenge in this regard is deploying an engine that can classify the incoming packets with speed near to the wire speed. The engine should also optimize memory usage.

Existing solutions have failed to make a fair tradeoff between the time and amount of memory consumed. This paper focuses on the TSS and TPS packet classification algorithms, which are considered suitable algorithms for parallel implementation. These algorithms work well to reduce the amount of memory consumed, but they suffer from low speed. This study has attempted to resolve this problem by parallel implementation of these algorithms. These two algorithms are implemented in the GPU in three different scenarios. To predict the classification speed of the proposed scenarios, their complexity is calculated. The computational complexity is obtained by the calibrated asymptotic model. The main parameter in evaluating the performance of parallel algorithms on the graphics processing unit is the packet classification time.

Analyzing the evaluation results show that the different modes of memory utilization in GPU have a significant effect on the speed of packet classification by TSS and TPS algorithms. Also, comparing the results of the formal complexity analysis and the corresponding experimental results confirms the efficiency of the hybrid scenario in parallelization of the TSS and TPS algorithms.

The present study would be extended by parallelizing the parallelizable algorithms on a cluster of GPUs. Without any doubt, the design of efficient GPU cluster-based parallel classifier engines requires a comprehensive analytical platform that can precisely estimate the effect of specific parameters on the complexity of kernels.

REFERENCES

- [1] M. Abbasi and M. Rafiee, "A calibrated asymptotic framework for analyzing packet classification algorithms on GPUs," *J. Supercomput.*, vol. 75, no. 10, pp. 6574–6611, Oct. 2019.
- [2] E. Benalia, S. Bitam, and A. Mellouk, "Data dissemination for Internet of vehicle based on 5G communications: A survey," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 5, p. e3881, May 2020.

- [3] S. Garg, K. Kaur, G. Kaddoum, S. H. Ahmed, and D. N. K. Jayakody, "SDN-based secure and privacy-preserving scheme for vehicular networks: A 5G perspective," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8421–8434, Sep. 2019.
- [4] E. Khaledian, N. Movahedinia, and M. Khayyambashi, "Spectral and energy efficiency in multi hop OFDMA based networks," in *Proc. 7th Int. Symp. Telecommun. (IST)*, Sep. 2014, pp. 123–128.
- [5] L. Huo, D. Jiang, and H. Qi, "A security traffic measurement approach in SDN-based Internet of Things," in *Proc. Int. Conf. Simulation Tools Techn.*, 2019, pp. 146–156.
- [6] K. Kaur, S. Garg, G. Kaddoum, E. Bou-Harb, and K.-K.-R. Choo, "A big data-enabled consolidated framework for energy efficient software defined data centers in IoT setups," *IEEE Trans. Ind. Informat.*, vol. 16, no. 4, pp. 2687–2697, Apr. 2020.
- [7] J. Wee, J.-G. Choi, and W. Pak, "Wildcard fields-based partitioning for fast and scalable packet classification in vehicle-to-everything," *Sensors*, vol. 19, no. 11, p. 2563, Jun. 2019.
- [8] D. E. Taylor, "Survey and taxonomy of packet classification techniques," *ACM Comput. Surv.*, vol. 37, no. 3, pp. 238–275, Sep. 2005.
- [9] S. Zhou, S. G. Singapura, and V. K. Prasanna, "High-performance packet classification on GPU," in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, Waltham, MA, USA, 2014, pp. 1–6, doi: 10.1109/HPEC.2014.7041005.
- [10] P. Du, R. Weber, P. Luszczek, S. Tomov, G. Peterson, and J. Dongarra, "From CUDA to OpenCL: Towards a performance-portable solution for multi-platform GPU programming," *Parallel Comput.*, vol. 38, no. 8, pp. 391–407, Aug. 2012.
- [11] A. Nottingham and B. Irwin, "GPU packet classification using OpenCL: A consideration of viable classification methods," in *Proc. Annu. Res. Conf. South Afr. Inst. Comput. Scientists Inf. Technologists (SAICSIT)*, 2009, pp. 160–169.
- [12] S. Han, K. Jang, K. Park, and S. Moon, "PacketShader: A GPU-accelerated software router," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 40, no. 4, pp. 195–206, 2010.
- [13] C.-L. Hung, H.-H. Wang, S.-W. Guo, Y.-L. Lin, and K.-C. Li, "Efficient GPGPU-based parallel packet classification," in *Proc. IEEE 10th Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Nov. 2011, pp. 1367–1374.
- [14] K. Kang and Y. S. Deng, "Scalable packet classification via GPU metaprogramming," in *Proc. Design, Autom. Test Eur.*, Mar. 2011, pp. 1–4.
- [15] S. Zhou, Y. R. Qu, and V. K. Prasanna, "Multi-core implementation of decomposition-based packet classification algorithms," *J. Supercomput.*, vol. 69, pp. 34–42, Jun. 2014.
- [16] M. Varvello, R. Laufer, F. Zhang, and T. V. Lakshman, "Multi-layer packet classification with graphics processing units," presented at the 10th ACM Int. Conf. Emerg. Netw. Exp. Technol., Sydney, NSW, Australia, 2014.
- [17] J. Zheng, D. Zhang, Y. Li, and G. Li, "Accelerate packet classification using GPU: A case study on HiCuts," in *Computer Science and Its Applications*, vol. 330, J. J. Park, I. Stojmenovic, H. Y. Jeong, and G. Yi, Eds. Berlin, Germany: Springer, 2015, pp. 231–238.
- [18] V. Srinivasan, S. Suri, and G. Varghese, "Packet classification using tuple space search," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 4, pp. 135–146, Oct. 1999.
- [19] NVIDIA. (Mar. 2015). *NVIDIA CUDA Compute Unified Device Architecture Programming Guide, Version 6.5*. [Online]. Available: http://docs.nvidia.com/cuda/pdf/CUDA_C_Programming_Guide.pdf
- [20] A. R. Brodtkorb, T. R. Hagen, and M. L. Sætra, "Graphics processing unit (GPU) programming strategies and trends in GPU computing," *J. Parallel Distrib. Comput.*, vol. 73, no. 1, pp. 4–13, Jan. 2013.
- [21] M. Amaris, D. Cordeiro, A. Goldman, and R. Y. D. Camargo, "A simple BSP-based model to predict execution time in GPU applications," in *Proc. IEEE 22nd Int. Conf. High Perform. Comput. (HiPC)*, Dec. 2015, pp. 285–294.
- [22] Y. Deng, X. Jiao, S. Mu, K. Kang, and Y. Zhu, "NPGPU: Network processing on graphics processing units," in *Theoretical and Mathematical Foundations of Computer Science*. Berlin, Germany: 2011, pp. 313–321.
- [23] K. Nakano, "Simple memory machine models for GPUs," in *Proc. IEEE 26th Int. Parallel Distrib. Process. Symp. Workshops PhD Forum*, May 2012, pp. 794–803.
- [24] K. Nakano, "The hierarchical memory machine model for GPUs," in *Proc. IEEE Int. Symp. Parallel Distrib. Process., Workshops Phd Forum*, May 2013, pp. 591–600.
- [25] D. E. Taylor and J. S. Turner, "Classbench: A packet classification benchmark," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Societies*, Mar. 2005, pp. 2068–2079.
- [26] M. Abbasi, R. Tahouri, and M. Rafiee, "Enhancing the performance of the aggregated bit vector algorithm in network packet classification using GPU," *PeerJ Comput. Sci.*, vol. 5, p. e185, Apr. 2019.



Mahdi Abbasi is currently an Associate Professor with the Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran. He is also the Director of the Computer Architecture Research Laboratory, Bu-Ali Sina University. His areas of research interests include computer architecture, network embedded systems, multi-objective optimization and control, distributed computing, the IoT, and signal processing.



Ali Najafi received the M.S. degree in information technology from Bu-Ali Sina University, Hamedan, Iran, in 2016. His research interests include computer networks, parallel and distributed computing, and GPU programming.



Milad Rafiee is currently a Lecturer and a Research Assistant with the Department of Computer Engineering, Bu-Ali Sina University, Hamedan, Iran. His research interests include computer networks, the IoT, software-defined networking, parallel and distributed computing, and GPU programming.



Mohammad R. Khosravi is currently with the Department of Computer Engineering, Persian Gulf University, Iran. His main interests include signal and image processing, computer networks, and distributed computing.



Varun G. Menon (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering, Kerala, India. His main interests include security and privacy, computer networks, and cyber-physical systems.



Ghulam Muhammad (Senior Member, IEEE) is currently a Professor with the Department of Computer Engineering, King Saud University (KSU), Riyadh, Saudi Arabia. His research interests include multimedia data processing, healthcare systems, and machine learning.

Received December 3, 2020, accepted December 18, 2020, date of publication December 24, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047136

Secrecy Outage Probability of Relay Selection Based Cooperative NOMA for IoT Networks

HUI LI¹, YAPING CHEN¹, MINGFU ZHU², JIANGFENG SUN³, (Member, IEEE),
DINH-THUAN DO⁴, VARUN G. MENON⁵, (Senior Member, IEEE),
AND SHYNU P. G.⁶, (Member, IEEE)

¹School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo 454003, China

²Huawei-Chuaitian 5G Edge Computing Laboratory, Hebei 458000, China

³College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003, China

⁴Department of Computer Science and Information Engineering, College of Information and Electrical Engineering, Asia University, Taichung 41354, Taiwan

⁵Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India

⁶School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India

Corresponding author: Jiangfeng Sun (sunjiangfeng@bupt.edu.cn)

This work was supported in part by the Basic and Advanced Technology Research Project of Henan Province under grant 152300410103, in part by Science and Technology Research Project of Henan Province under grant 202102310299, and in part by Opening Project of Henan Engineering Laboratory of Photoelectric Sensor and Intelligent Measurement and Control of Henan Polytechnic University under grant HELPSIMC-2020-002.

ABSTRACT As an important partner of fifth generation (5G) communication, the internet of things (IoT) is widely used in many fields with its characteristics of massive terminals, intelligent processing, and remote control. In this paper, we analyze security performance for the cooperative non-orthogonal multiple access (NOMA) networks for IoT, where the multi-relay Wyner model with direct link between the base station and the eavesdropper is considered. In particular, secrecy outage probability (SOP) for two kinds of relay selection (RS) schemes (i.e., single-phase RS (SRS) and two-phase RS (TRS)) is developed in the form of closed solution. As a benchmark for comparison, the SOP for random RS (RRS) is also obtained. To gain more meaningful insights, approximate derivations of SOP under the high signal-to-noise ratio (SNR) region are provided. Results of statistical simulation confirm the theoretical analysis and testify that: i) Compared with RRS scheme, SRS and TRS may improve secure performance because of obtaining smaller SOPs; ii) There exists secrecy performance floor for the SOP in strong SNR regime, which is dominated by NOMA protocol; iii) The security performance can be enhanced by augmenting the quantity of relays for SRS and TRS strategies. The purpose of this work is to provide theoretical basis for the analysis and design of anti-eavesdropping for NOMA systems in IoT.

INDEX TERMS Non-orthogonal multiple access, physical layer security, secrecy outage probability, single-phase relay selection, two-phase relay selection.

I. INTRODUCTION

Recently, the rapid development of IoT makes all walks of life get convenient and fast services. However, due to the importance of ownership and privacy protection, the IoT system must provide corresponding security mechanisms. The classical method to deal with the security problem is complex encryption and decryption scheme [1]. Quantum computing can crack complex keys. Moreover, the terminals of IoT are often limited in size and power, and do not have strong computing power. These contradictions lead that the classical method is not so effective in many scenarios [2]. So an alternative mechanism, i.e., physical layer security (PLS) exhibits

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenyu Zhou¹.

more advances. The method of PLS was initially discussed by Wyner from the standpoint of information theory [3]. PLS is a new approach to enhance network security by utilizing the characteristics of channels, which has caught quantity of attention due to the randomness of fading channels rather than encryption technology [4]–[7]. In [8], the authors studied the secrecy behaviours for underlay cooperative relaying networks. Recently, NOMA is deemed to have a bright prospect in 5G networks on account that it can improve the band-efficient and spectral efficiency [9]–[13]. Serving multiple users working at the same frequency band with different power-split is the core thought of NOMA [14]. Do *et al.* put forward a model which can serve cellular networks better in NOMA [15]. The authors of [16] discussed a large-scale network with an antenna and multiple antennas in NOMA

systems, and derived the SOP. Lei *et al.* researched a security NOMA system including two different forms of eavesdropping [17]. The ambient backscatter NOMA systems was studied in terms of the secure performance [18]. Jiang *et al.* analyzed the secure performance for uplink NOMA including multiple eavesdroppers in [19]. Therefore, exploring PLS in NOMA systems has also aroused the interests of many researchers.

Cooperative communication is a specially efficient method by furnishing greater diversity and expanding network coverage [20]. At present, two fields are mainly included in cooperative communication for NOMA's research. On the one hand, the use of NOMA in cooperative networks was discussed in [21]–[24]. On the other hand, cooperative NOMA was first put forward by Ding *et al.* in [25] and researched in [26]–[29]. Choi studied the transmission rates on the cooperative system in [21]. The authors studied the interruption probability (IP) and systematic capacity of NOMA using decoded and forward (DF) in relaying systems [22]. In [23], the SOP was investigated based full-duplex (FD) in cooperative communication using optimizing power allocation jointly. Men *et al.* discussed the outage character using amplify and forward (AF) protocol on Nakagami- m distribution for NOMA in [24]. The core idea of cooperative NOMA is that nearby NOMA users are treated as DF relaying to transfer the messages for far NOMA users. The secrecy behaviors for both AF and DF relaying strategies were investigated in cooperative NOMA system [26]. Simultaneous wireless information and power transmission (SWIPT) was adopted by nearby NOMA users which were counted as DF relaying [27]. The work researched the security transmission and proposed an optimal power distribution scheme with maximum secrecy sum rate, where the precondition was that the users' quality of service (QoS) met the conditions [28]. The authors of [29] employed FD and artificial noise (AN) methods in two-way relaying networks based on NOMA. The mathematical expressions for the ergodic secrecy rate were discussed under containing and excluding eavesdroppers.

As a popular transmission scheme, relay selection (RS) has the advantages of low complexity due to taking full advantage of spatial variety and high spectrum-efficient [30], [31]. Considering that there might be some differences between two users in the QoS requirements, two-stage single-relay-selection and dual-relay-selection strategies were put forward, respectively [32]. Ding *et al.* derived closed-form expressions for the precise and asymptotic outage probability (OP) by employing single-stage RS and two-stage RS strategies in cooperative NOMA. And the two NOMA users were classified as nearby and far users by their QoS, rather than their channel conditions [33]. Accurate analytical formulae for the OP and IP were analyzed by using two relay selection strategies (i.e., optimal RS and suboptimal RS) in wireless communication networks (WCNs) [34]. Under three wiretapping cases including one eavesdropper, non-colluding and colluding eavesdroppers, the secrecy outage behaviors of the TRS strategy based on the system over Nakagami- m

fading channels were investigated in [35]. Zhang *et al.* analyzed the SOP with optimal relay selection, suboptimal relay selection and multiple relays uniting schemes. In addition, the confidentiality of a cognitive DF relaying network over Nakagami- m fading channels with independent but not necessarily identical distributed was also surveyed in [36].

Although these previous contributions provided a firm foundation for understanding collaborative NOMA and RS technologies, it still needs further developments and applications. It should be pointed out that RS schemes can meet the requirements of actual IoT situation. In this paper, we investigate the SRS and TRS methods which can achieve the minimum SOP. As far as we know, there is no research on the security performance of SRS and TRS schemes in cooperative NOMA networks considering direct link between base station and eavesdropper. To this end, we explore SOP using RS schemes for basing on half-duplex (HD) NOMA networks over independently Rayleigh distribution. More specifically, the rate of data transmission for the far user is assured to choose a relay as its auxiliary equipment to forward the messages in the SRS strategy. Under the premise of guaranteeing the data transmission rate of far user, the maximum data rate of the service is provided for nearby user to select the relay opportunistically in the TRS scheme. The key contributions of this paper are summarized as follows:

- This paper describes system model of cooperative NOMA for IoT and focuses on two kinds of relay selection strategies (i.e., SRS and TRS schemes). Moreover, the direct link between the eavesdropper and the base station is considered. The eavesdropper uses selective combination (SC) technique to process the received signals from two slots.
- The theoretical SOP is derived by employing the SRS strategy over Rayleigh fading channel. In addition, the SOP for RRS scheme is also analyzed as a contrast. The results show that SRS strategy obtains the lower SOP. To better understand secure outage performance, the asymptotic behaviors of SOP are analyzed with RRS and SRS schemes in cooperative NOMA.
- We also derive the formulas of SOP for TRS scheme in cooperative NOMA based on HD. What's more, experimental results prove that TRS scheme can obtain the superior SOP. To get more insights, the approximate SOP of TRS scheme under high SNR regime is analyzed in cooperative NOMA. The results also verify that the security performance can be enhanced distinctly by augmenting the quantity of relays.

The specific arrangement of each section is as follows. In Section II, the network system of HD NOMA's RS schemes is established. Section III deduces new analytic formulae of SOP for the RRS, SRS and TRS schemes. In Section IV, the asymptotical SOPs in high SNR regime are derived. Section V presents numerical results and systematic performance. The conclusions are shown in Section VI in the paper.

Notations: The $\mathcal{CN}(\mu, \sigma^2)$ denotes the complex Gaussian distribution with expectation μ and standard variance σ .

The $\Pr(\cdot)$ and $\mathbb{E}(\cdot)$ are the probability and expectation operation. $f_X(\cdot)$ and $F_X(\cdot)$ are the probability density function (PDF) and the cumulative distribution function (CDF), respectively.

II. SYSTEM MODEL

As illustrated in Fig.1, a typical dual-hop NOMA relaying system for IoT includes a base station (*BS*), K half-duplex relays, a couple of legitimate users (e.g., the nearby user D_1 and far user D_2) and one eavesdropper (*Eve*). It should be noted that the direct links from *BS* to D_j ($j = 1, 2$) are not considered, but the direct link between *BS* and *Eve* exists. So, the *Eve* processes the received signals by using SC arithmetic. Adopting a multi-access scheme, multiple users can be easily partitioned into many groups in this cooperative model, each of these groups implements the NOMA protocol [37]. In the network model, all relaying nodes are equipped with receiving and transmitting antennas, but *BS* and users have only one antenna. The *BS* tries to communicate the users via relays, but there exists eavesdropping between transmissions, and the information leakage exists in the transmission between two slots. Each relay is assumed to use DF protocol. All wireless channels are affected by additive white Gaussian noise (AWGN) and modeled as independent non-selective Rayleigh distribution. The distance from X to Y is represented as d_{XY} , α denotes exponent for the path loss, h_{XY} denotes the channel coefficient from X to Y , $XY \in \{SR_i, R_iD_1, R_iD_2, SE, R_iE\}$ and $h_{XY} \sim \mathcal{CN}(0, 1)$. The PDF and CDF for $|h_{XY}|^2$ have an exponential distribution as

$$f_{|h_{XY}|^2}(x) = \frac{1}{g_{XY}} \exp\left(-\frac{x}{g_{XY}}\right), \quad (1)$$

and

$$F_{|h_{XY}|^2}(x) = 1 - \exp\left(-\frac{x}{g_{XY}}\right), \quad (2)$$

respectively, where g_{XY} is the mean channel power gain [38]. The two legitimate users are segmented into nearby and far users on the basis of their QoS. More specifically, with the assistance of relay chosen, the QoS requirements of legal users can be effectively provided for the IoT scenario. Therefore, we assume that D_1 can serve opportunely with low target data rates, D_2 needs to be served quickly.

During the first stage, the *BS* transmits composite messages $\sqrt{a_1 P_s} x_1 + \sqrt{a_2 P_s} x_2$ to the assistances on the basis of NOMA theory, and normalization method of x_1 and x_2 signal is adopted respectively, i.e. $\mathbb{E}(|x_1|^2) = \mathbb{E}(|x_2|^2) = 1$, P_s and P_r denote the transmitted power from the *BS* and R_i . a_1 and a_2 are the corresponding power allocation coefficients. In fact, in order to provide better fairness and QoS requirements among users [39], we hypothesize that $a_2 > a_1$ and $a_1 + a_2 = 1$. Hence the received messages at the i th relay R_i can be expressed as

$$y_{SR_i} = \frac{h_{SR_i}}{\sqrt{d_{SR}^\alpha}} \left(\sqrt{a_1 P_s} x_1 + \sqrt{a_2 P_s} x_2 \right) + n_{SR_i}, \quad (3)$$

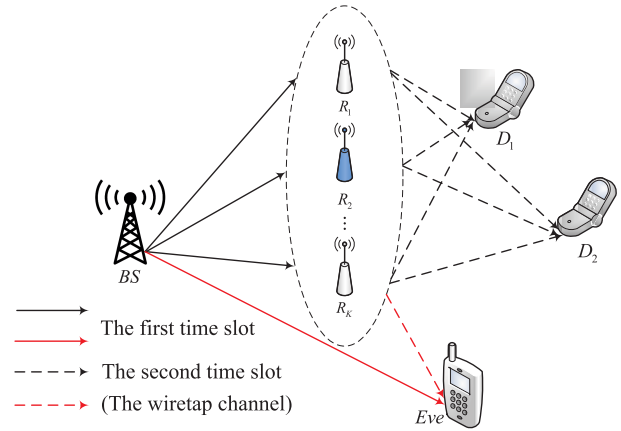


FIGURE 1. System model for IoT.

where n_{SR_i} is written as the superimposed Gaussian noise at relay i .

In order to reduce the interference for decoding signal x_1 of D_1 at R_i , the successive interference cancellation (SIC) method is employed to detect the information x_2 of D_2 firstly with the high power allocation coefficient. Therefore, the received signal-to-interference-plus-noise ratios (SINRs) to decode x_1 and x_2 at R_i are shown by

$$\gamma_{R_i, D_2} = \frac{a_2 \rho_s |h_{SR_i}|^2}{a_1 \rho_s |h_{SR_i}|^2 + d_{SR}^\alpha}, \quad (4)$$

and

$$\gamma_{R_i, D_1} = \frac{a_1 \rho_s |h_{SR_i}|^2}{d_{SR}^\alpha}, \quad (5)$$

where $\rho_x = \frac{P_x}{N_0}$, $x \in (s, r)$ is the transmit SNR, and N_0 is the mean power of the AWGN in this system.

In the same way, the message received at *Eve* can be expressed as

$$y_{SE} = \frac{h_{SE}}{\sqrt{d_{SE}^\alpha}} \left(\sqrt{a_1 P_s} x_1 + \sqrt{a_2 P_s} x_2 \right) + n_{SE}. \quad (6)$$

where n_{D_j} , n_X denote the Gaussian noise at users D_j and X ($X = SE, RE$).

We analyse the SINR for wiretapper to decode x_1 and x_2 . Considering a direct link between *BS* and *Eve* in this paper. Therefore, in this time slot, the instantaneous SINRs at *Eve* that eavesdrops the messages from legal users D_j are written by

$$\gamma_{SE_1} = \frac{a_1 \rho_s |h_{SE}|^2}{d_{SE}^\alpha}, \quad (7)$$

and

$$\gamma_{SE_2} = \frac{a_2 \rho_s |h_{SE}|^2}{a_1 \rho_s |h_{SE}|^2 + d_{SE}^\alpha}. \quad (8)$$

During the second stage, it is assumed that relay R_i can decode received messages and transmit signals to the target nodes, the following situations are met in this phase,

i) $\log\left(\frac{1+\gamma_{R_i,D_1}}{1+\gamma_{E_1}}\right) \geq R_{D_1}$, ii) $\log\left(\frac{1+\gamma_{R_i,D_2}}{1+\gamma_{E_2}}\right) \geq R_{D_2}$, where γ_{E_j} is the SNR at *Eve* and will be further analyzed later, R_{D_j} denotes the target data transmission rate. Selective relay R_i forwards the signals to the user, so the signals received in D_j can be represented as

$$y_{D_j} = \frac{h_j}{\sqrt{d_{RD_j}^\alpha}} \left(\sqrt{a_1 P_r} x_1 + \sqrt{a_2 P_r} x_2 \right) + n_{D_j}. \quad (9)$$

The signals received in *Eve* in this phase can be represented as

$$y_{RE} = \frac{h_{RE}}{\sqrt{d_{RE}^\alpha}} \left(\sqrt{a_1 P_r} x_1 + \sqrt{a_2 P_r} x_2 \right) + n_{RE}. \quad (10)$$

It is assumed that perfect SIC can be used in D_2 to detect messages from D_1 with a higher transmitting power. Therefore, D_2 detects the SINR of x_1 given by the following formula,

$$\gamma_{D_1,D_2} = \frac{a_2 \rho_r |h_1|^2}{a_1 \rho_r |h_1|^2 + d_{RD_1}^\alpha}. \quad (11)$$

Then, the received SINR at D_1 is given by

$$\gamma_{D_1} = \frac{a_1 \rho_r |h_1|^2}{d_{RD_1}^\alpha}. \quad (12)$$

Meanwhile, D_2 decodes messages x_2 by regarding x_1 as interference, and the SINR can be shown as

$$\gamma_{D_2} = \frac{a_2 \rho_r |h_2|^2}{a_1 \rho_r |h_2|^2 + d_{RD_2}^\alpha}. \quad (13)$$

For the second time slot, the instantaneous SINRs at *Eve* to wiretap the messages are expressed as

$$\gamma_{R_i E_1} = \frac{a_1 \rho_r |h_{R_i E}|^2}{d_{RE}^\alpha} \quad (14)$$

$$\gamma_{R_i E_2} = \frac{a_2 \rho_r |h_{R_i E}|^2}{a_1 \rho_r |h_{R_i E}|^2 + d_{RE}^\alpha} \quad (15)$$

III. SOP ANALYSIS

In this part, the SOPs of the cooperative NOMA system using three kinds of relay selection schemes are studied.

To get the SOP for every user, channel statistics for the users and *Eve* are analyzed primarily. Combined with (5) and (12), the CDF of SINR from BS to D_1 can be written as

$$\begin{aligned} F_{\gamma_1}(x) &= \Pr(\min(\gamma_{R_i,D_1}, \gamma_{D_1}) < x) \\ &= 1 - \Pr(\gamma_{R_i,D_1} > x) \Pr(\gamma_{D_1} > x) \\ &= 1 - \Pr\left(\frac{a_1 \rho_s |h_{SR_i}|^2}{d_{SR}^\alpha} > x\right) \Pr\left(\frac{a_1 \rho_r |h_1|^2}{d_{RD_1}^\alpha} > x\right) \\ &= 1 - e^{-\frac{Ax}{a_1}}, \end{aligned} \quad (16)$$

where $\gamma_1 = \min\{\gamma_{R_i,D_1}, \gamma_{D_1}\}$, $A = \frac{d_{SR}^\alpha}{\rho_s g_{SR_i}} + \frac{d_{RD_1}^\alpha}{\rho_r g_1}$.

In similar, the CDF of SINR from BS to D_2 is given as

$$F_{\gamma_2}(x) = \begin{cases} 1 - e^{-\frac{Bx}{(a_2 - a_1)x}} & x \leq \frac{a_2}{a_1} \\ 1 & x > \frac{a_2}{a_1}, \end{cases} \quad (17)$$

where $\gamma_2 = \min\{\gamma_{R_i,D_2}, \gamma_{D_1,D_2}, \gamma_{D_2}\}$, and $B = \frac{d_{SR}^\alpha}{\rho_s g_{SR_i}} + \frac{d_{RD_1}^\alpha}{\rho_r g_1} + \frac{d_{RD_2}^\alpha}{\rho_r g_2}$.

For the signals received in *Eve* ($BS \rightarrow E, R_i \rightarrow E$), the SC algorithm is employed. Then, according to (5), (7) and (14), the CDF of γ_{E_1} is expressed as

$$\begin{aligned} F_{\gamma_{E_1}}(x) &= \Pr(\max(\gamma_{SE_1}, \min(\gamma_{R_i,D_1}, \gamma_{R_i E_1}) < x)) \\ &= \Pr(\gamma_{SE_1} < x) (1 - \Pr(\min(\gamma_{R_i,D_1}, \gamma_{R_i E_1}) > x)) \\ &= \Pr(\gamma_{SE_1} < x) (1 - \Pr(\gamma_{R_i,D_1} > x) \Pr(\gamma_{R_i E_1} > x)) \\ &= \left(1 - e^{-\frac{d_{SE}^\alpha x}{a_1 g_{SE}}}\right) \left(1 - e^{-\frac{x}{a_1} \left(\frac{d_{SR}^\alpha}{g_{SR_i}} + \frac{d_{RE}^\alpha}{g_{R_i E}}\right)}\right) \\ &= \left(1 - e^{-\frac{Ex}{a_1}}\right) \left(1 - e^{-\frac{Cx}{a_1}}\right), \end{aligned} \quad (18)$$

where $C = \frac{d_{SR}^\alpha}{\rho_s g_{SR_i}} + \frac{d_{RE}^\alpha}{\rho_r g_{R_i E}}$, and $E = \frac{d_{SE}^\alpha}{\rho_s g_{SE}}$.

The PDF of γ_{E_1} can be obtained as

$$f_{\gamma_{E_1}}(x) = \frac{E}{a_1} e^{-\frac{Ex}{a_1}} + \frac{C}{a_1} e^{-\frac{Cx}{a_1}} - \frac{D}{a_1} e^{-\frac{Dx}{a_1}}, \quad (19)$$

where $D = \frac{d_{SR}^\alpha}{\rho_s g_{SR_i}} + \frac{d_{RE}^\alpha}{\rho_r g_{R_i E}} + \frac{d_{SE}^\alpha}{\rho_s g_{SE}}$.

Referring to the derivation of γ_{E_1} , the PDF of γ_{E_2} can be derived as

$$\begin{aligned} f_{\gamma_{E_2}}(x) &= \frac{E a_2}{(a_2 - a_1 x)^2} e^{-\frac{Ex}{a_2 - a_1 x}} \\ &\quad + \frac{C a_2}{(a_2 - a_1 x)^2} e^{-\frac{Cx}{a_2 - a_1 x}} \\ &\quad - \frac{D a_2}{(a_2 - a_1 x)^2} e^{-\frac{Dx}{a_2 - a_1 x}}. \end{aligned} \quad (20)$$

A. SOP FOR RRS

The SOP is a very important benchmark to evaluate systematic secure performance, we can formulate it as [40]

$$P_{out} = \Pr\left(\lceil C_{D_j} - C_{E_j} \rceil^+ < R_{th}\right), \quad (21)$$

where $\lceil X \rceil^+ = \max\{X, 0\}$, R_{th} is the threshold of rate.

The SOP for RRS can be rewritten as

$$\begin{aligned} SOP_{RRS} &= \Pr\left(\lceil C_{D_1} - C_{E_1} \rceil^+ < R_{th_1} \text{ or} \right. \\ &\quad \left. \lceil C_{D_2} - C_{E_2} \rceil^+ < R_{th_2}\right) \\ &= 1 - \Pr\left(\lceil C_{D_1} - C_{E_1} \rceil^+ > R_{th_1}, \right. \\ &\quad \left. \lceil C_{D_2} - C_{E_2} \rceil^+ > R_{th_2}\right) \\ &= 1 - \Pr\left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} > \varepsilon_1, \frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right), \end{aligned} \quad (22)$$

where $\varepsilon_j = 2^{R_{th_j}}$ with R_{th_j} being the target rates of D_j .

Note that the variables γ_j, γ_{E_j} in (22) are related, acquiring an accurate expression of SOP is difficult. Therefore, the upper bound of SOP is given using the basic probability theory, (22) can be rewritten as

$$\begin{aligned}
 &SOP_{RRS} \\
 &\leq \min \left\{ 1, 2 - \Pr \left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} > \varepsilon_1 \right) \right. \\
 &\quad \left. - \Pr \left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2 \right) \right\} \\
 &= \min \left\{ 1, \underbrace{\Pr \left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} < \varepsilon_1 \right)}_{p_1^{out}} + \underbrace{\Pr \left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} < \varepsilon_2 \right)}_{p_2^{out}} \right\}. \tag{23}
 \end{aligned}$$

The term p_j^{out} in (23) represents the SOP for RRS at D_j and can be calculated as

$$\begin{aligned}
 p_j^{out} &= \Pr (\gamma_j < \varepsilon_j (1 + \gamma_{E_j}) - 1) \\
 &= \int_0^\infty f_{\gamma_{E_j}}(x) F_{\gamma_j}(\varepsilon_j (1 + x) - 1) dx. \tag{24}
 \end{aligned}$$

Then, on the basis of (16), (19) and (24), p_1^{out} can be obtained as (25), shown at the bottom of the page.

Take full advantage of (24), the SOP for RRS at D_2 can be written as

$$\begin{aligned}
 p_2^{out} &= \int_0^\mu f_{\gamma_{E_2}}(x) F_{\gamma_2}(\varepsilon_2(1+x) - 1) dx \\
 &\quad + \int_\mu^\infty f_{\gamma_{E_2}}(x) dx, \tag{26}
 \end{aligned}$$

where $\mu = \frac{1}{a_1 \varepsilon_2} - 1$.

With the combination of (17), (20), (26) and the Gaussian-Chebyshev quadrature method, the SOP for RRS at D_2 is given by (27), which is shown at the bottom of the page, where $\phi_t = \cos \left(\frac{2t-1}{2N} \pi \right), t \in \{l, m, n\}$, and

$$\begin{aligned}
 \varphi_1(x) &= \frac{1}{(a_2 - a_1 x)^2} e^{-\frac{Ex}{(a_2 - a_1 x)}} e^{-\frac{B(\varepsilon_2 + \varepsilon_2 x - 1)}{(a_2 - a_1(\varepsilon_2(1+x) - 1))}}, \\
 \varphi_2(x) &= \frac{1}{(a_2 - a_1 x)^2} e^{-\frac{Cx}{(a_2 - a_1 x)}} e^{-\frac{B(\varepsilon_2 + \varepsilon_2 x - 1)}{(a_2 - a_1(\varepsilon_2(1+x) - 1))}}, \\
 \varphi_3(x) &= \frac{1}{(a_2 - a_1 x)^2} e^{-\frac{Dx}{(a_2 - a_1 x)}} e^{-\frac{B(\varepsilon_2 + \varepsilon_2 x - 1)}{(a_2 - a_1(\varepsilon_2(1+x) - 1))}}.
 \end{aligned}$$

Combining (23), (25) and (27), the SOP for RRS is shown by (28), shown at the bottom of the page.

B. SOP FOR SRS

In this part, we consider the SRS scheme for HD-based cooperative NOMA. BS can randomly select a relay as its auxiliary to transpond the messages. Maximizing the minimum data transmission rate D_2 is the main idea of SRS method. What's more, the range of data rate for D_2 is dominant by three different data rates: i) the transmission rate for the relay R_i to decode messages x_2 , ii) the transmission rate for D_1 to decode messages x_2 . iii) the transmission rate for D_2 to decode messages x_2 . In relaying networks, the SRS scheme activates a relay, which can be expressed as

$$\begin{aligned}
 i_{SRS}^* &= \arg \max_i \left\{ \min \left\{ \log(1 + \gamma_{R_i, D_2}), \right. \right. \\
 &\quad \left. \left. \log(1 + \gamma_{D_1, D_2}), \log(1 + \gamma_{D_2}) \right\}, i \in S_R^1 \right\}, \tag{29}
 \end{aligned}$$

where S_R^1 reveals the amount of relays in the network. Pay attention that the HD-based SRS scheme inherits the advantage of guaranteeing the data rate of D_2 ,

$$\begin{aligned}
 p_1^{out} &= \int_0^\infty f_{\gamma_{E_1}}(x) F_{\gamma_1}(\varepsilon_1(1+x) - 1) dx \\
 &= 1 - \int_0^\infty \left(\frac{E}{a_1} e^{-\frac{Ex + A(\varepsilon_1(1+x) - 1)}{a_1}} + \frac{C}{a_1} e^{-\frac{Cx + A(\varepsilon_1(1+x) - 1)}{a_1}} - \frac{D}{a_1} e^{-\frac{Dx + A(\varepsilon_1(1+x) - 1)}{a_1}} \right) dx \\
 &= 1 - \left(\frac{E}{E + A\varepsilon_1} + \frac{C}{C + A\varepsilon_1} - \frac{D}{D + A\varepsilon_1} \right) e^{-\frac{A(\varepsilon_1 - 1)}{a_1}}. \tag{25}
 \end{aligned}$$

$$\begin{aligned}
 p_2^{out} &= 1 - \int_0^\mu f_{\gamma_{E_2}}(x) e^{-\frac{B(\varepsilon_2 + \varepsilon_2 x - 1)}{(a_2 - a_1(\varepsilon_2(1+x) - 1))}} dx \approx 1 - \frac{E a_2 \mu \pi}{2N} \sum_{l=0}^N \sqrt{1 - \phi_l^2} \varphi_1 \left(\frac{\mu \phi_l + \mu}{2} \right) \\
 &\quad - \frac{C a_2 \mu \pi}{2N} \sum_{m=0}^N \sqrt{1 - \phi_m^2} \varphi_2 \left(\frac{\mu \phi_m + \mu}{2} \right) + \frac{D a_2 \mu \pi}{2N} \sum_{n=0}^N \sqrt{1 - \phi_n^2} \varphi_3 \left(\frac{\mu \phi_n + \mu}{2} \right). \tag{27}
 \end{aligned}$$

$$\begin{aligned}
 SOP_{RRS} &= \min \left\{ 1, 2 - \frac{E}{E + A\varepsilon_1} e^{-\frac{A(\varepsilon_1 - 1)}{a_1}} - \frac{C}{C + A\varepsilon_1} e^{-\frac{A(\varepsilon_1 - 1)}{a_1}} + \frac{D}{D + A\varepsilon_1} e^{-\frac{A(\varepsilon_1 - 1)}{a_1}} - \frac{a_2 \mu \pi}{2N} \times \right. \\
 &\quad \left. \left(E \sum_{l=0}^N \sqrt{1 - \phi_l^2} \varphi_1 \left(\frac{\mu \phi_l + \mu}{2} \right) + C \sum_{m=0}^N \sqrt{1 - \phi_m^2} \varphi_2 \left(\frac{\mu \phi_m + \mu}{2} \right) - D \sum_{n=0}^N \sqrt{1 - \phi_n^2} \varphi_3 \left(\frac{\mu \phi_n + \mu}{2} \right) \right) \right\}. \tag{28}
 \end{aligned}$$

where applications for lower target data rate can be implemented.

In accordance with the above investigations, Ξ_1 denotes that either the relay i_{TRS}^* or any of the legal users is unable to decode x_2 safely. So, the SOP based on SRS scheme with HD can be obtained as follows,

$$\begin{aligned} SOP_{SRS} &= Pr(\Xi_1) = Pr(|S_R^1| = 0) \\ &= \prod_{i=1}^K \left(1 - Pr\left(\frac{1 + \min(\gamma_{R_i, D_2}, \gamma_{D_1, D_2}, \gamma_{D_2})}{1 + \gamma_{E_2}} > \varepsilon_2\right) \right) \\ &= \prod_{i=1}^K \left(1 - Pr\left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right) \right), \end{aligned} \quad (30)$$

where $|S_R^1|$ denotes the size of S_R^1 .

Substituting (27) into (30), the SOP for SRS scheme can be obtained, that is shown by (31) at the bottom of the page.

C. SOP FOR TRS

For HD-based cooperative NOMA, TRS consists of two main periods. In the first period, the objective data rate of D_2 is met. In the second period, we expect to make the data transmission rate of D_1 as high as possible under the condition that the data transmission rate of D_2 is satisfied. Therefore, the first period activates the relays that meet the following conditions,

$$S_R^2 = \left\{ \log(1 + \gamma_{R_i, D_2}) \geq R_{D_2}, \log(1 + \gamma_{D_1, D_2}) \geq R_{D_2}, \right. \\ \left. \log(1 + \gamma_{R_i, D_2}) \geq R_{D_2}, 1 \leq i \leq K \right\}, \quad (32)$$

where S_R^2 denotes these relays satisfying the objective data rate of D_2 in the first stage.

For all relays from S_R^2 , the second period chooses a relay to transmit messages and maximizes the data rate of D_1 , the selected relay is

$$i_{TRS}^* = \arg \max_i \left\{ \min \left\{ \log(1 + \gamma_{R_i, D_1}), \right. \right. \\ \left. \left. \log(1 + \gamma_{D_1}) \right\}, i \in S_R^2 \right\}. \quad (33)$$

$$\begin{aligned} SOP_{SRS} = Pr(\Xi_1) \approx & \left(1 - \frac{a_2 \mu \pi}{2N} \left(E \sum_{l=0}^N \sqrt{1 - \phi_l^2} \varphi_1 \left(\frac{\mu \phi_l + \mu}{2} \right) \right. \right. \\ & \left. \left. - C \sum_{m=0}^N \sqrt{1 - \phi_m^2} \varphi_2 \left(\frac{\mu \phi_m + \mu}{2} \right) + D \sum_{n=0}^N \sqrt{1 - \phi_n^2} \varphi_3 \left(\frac{\mu \phi_n + \mu}{2} \right) \right) \right)^K. \end{aligned} \quad (31)$$

$$T_1 = \min \left\{ 1, \left[\frac{\left(\frac{E}{E+A\varepsilon_1} - \frac{C}{C+A\varepsilon_1} + \frac{D}{D+A\varepsilon_1} \right) e^{-\frac{A(\varepsilon_1-1)}{a_1}}}{\frac{a_2 \mu \pi}{2N} \left(E \sum_{l=0}^N \sqrt{1 - \phi_l^2} \varphi_1 \left(\frac{\mu \phi_l + \mu}{2} \right) + C \sum_{m=0}^N \sqrt{1 - \phi_m^2} \varphi_2 \left(\frac{\mu \phi_m + \mu}{2} \right) - D \sum_{n=0}^N \sqrt{1 - \phi_n^2} \varphi_3 \left(\frac{\mu \phi_n + \mu}{2} \right) \right)} \right]^i \right\}. \quad (37)$$

As can be seen from the above explanations, excepting for guaranteeing the data rate of D_2 , the TRS scheme based on HD can support D_1 to perform some background tasks.

It is worth noting that the total SOP events can be classified as

$$SOP_{TRS} = Pr(\Xi_1) + Pr(\Xi_2), \quad (34)$$

where Ξ_2 means that the relaying i_{TRS}^* , D_1 and D_2 can successfully decode x_2 , while the i_{TRS}^* and D_1 cannot successfully decode x_1 . Considering the analysis of the second period, $Pr(\Xi_2)$ is expressed as

$$Pr(\Xi_2) = \sum_{i=1}^K \underbrace{Pr\left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} < \varepsilon_1 \mid |S_R^2| = i\right)}_{T_1} \underbrace{Pr\left(|S_R^2| = i\right)}_{T_2}, \quad (35)$$

where $|S_R^2|$ represents the value of S_R^2 .

Because of the mathematical intractability in (22), T_1 can be given as

$$\begin{aligned} T_1 &= 1 - Pr\left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} > \varepsilon_1 \mid |S_R^2| = i\right) \\ &= \left[1 - \frac{Pr\left(\frac{1 + \gamma_1}{1 + \gamma_{E_1}} > \varepsilon_1, \frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right)}{Pr\left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right)} \right]^i. \end{aligned} \quad (36)$$

So, the term T_1 can be rewritten as (37) by substituting (25) and (27) into (36), it is shown at the bottom of the page.

Moreover, there exist i relays in S_R^2 , so the corresponding probability T_2 is calculated by

$$\begin{aligned} T_2 &= \binom{K}{i} \left(Pr\left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right) \right)^i \\ &\quad \times \left(1 - Pr\left(\frac{1 + \gamma_2}{1 + \gamma_{E_2}} > \varepsilon_2\right) \right)^{K-i} \\ &= \binom{K}{i} (1 - p_2^{out})^i (p_2^{out})^{K-i}. \end{aligned} \quad (38)$$

Combining (31), (35), (37) and (38) and employing some arithmetical operations, the SOP for TRS scheme can be

expressed as

$$SOP_{TRS} = \sum_{i=0}^K \binom{K}{i} (1 - p_2^{out})^i (p_2^{out})^{K-i} \times \min \left(1, \left[\frac{p_1^{out}}{1 - p_2^{out}} \right]^i \right). \quad (39)$$

IV. ASYMPTOTIC SOP ANALYSIS

To gain deeper insights, the asymptotical SOPs of cooperative NOMA over Rayleigh fading channel are analyzed under these RS schemes. As $\rho \rightarrow \infty$ ($\rho_s = \rho_r$), specifically, the SOP of cooperative NOMA systems under each RS scheme depends on far user D_2 when $\gamma_2 \rightarrow \infty$. The asymptotical SOP for cooperative NOMA is shown as

$$ASOP_{RRS} \approx \Pr \left([C_{D_2} - C_{E_2}]^+ < R_{th_2} \right). \quad (40)$$

Substituting (27) into (40), the asymptotic SOP for RRS scheme when $\rho \rightarrow \infty$ can be obtained by

$$ASOP_{RRS} = 1 - \frac{a_2 \mu \pi}{N} \left\{ \sum_{l=0}^N \frac{2E \sqrt{1 - \phi_l^2}}{(2a_2 - a_1 \mu (\phi_l + 1))^2} + \sum_{m=0}^N \frac{2C \sqrt{1 - \phi_m^2}}{(2a_2 - a_1 \mu (\phi_m + 1))^2} - \sum_{n=0}^N \frac{2D \sqrt{1 - \phi_n^2}}{(2a_2 - a_1 \mu (\phi_n + 1))^2} \right\}. \quad (41)$$

From the asymptotic expression of SOP, it can be seen that there exists secure performance floor in cooperative NOMA system, which depends on the NOMA protocol. The main cause for this situation is that the realizable data rate of far user D_2 is restricted by the power distribution coefficient, a_2/a_1 . However, there is no such restriction to realize the data rate in *Eve*.

Based on (31), we observe that $\Pr(\Xi_1)$ tends to a constant. Therefore, the asymptotic SOP under SRS scheme is given by

$$ASOP_{SRS} = \left\{ 1 - \frac{a_2 \mu \pi}{N} \left\{ \sum_{l=0}^N \frac{2E \sqrt{1 - \phi_l^2}}{(2a_2 - a_1 \mu (\phi_l + 1))^2} + \sum_{m=0}^N \frac{2C \sqrt{1 - \phi_m^2}}{(2a_2 - a_1 \mu (\phi_m + 1))^2} - \sum_{n=0}^N \frac{2D \sqrt{1 - \phi_n^2}}{(2a_2 - a_1 \mu (\phi_n + 1))^2} \right\} \right\}^K. \quad (42)$$

Furthermore, taking into account the second stage, we have

$$\Pr(\Xi_2) = 0, \rho \rightarrow \infty. \quad (43)$$

According to the above analysis, the asymptotic SOP under TRS scheme is similar to SRS scheme. That is to say, $ASOP_{TRS} = ASOP_{SRS}$.

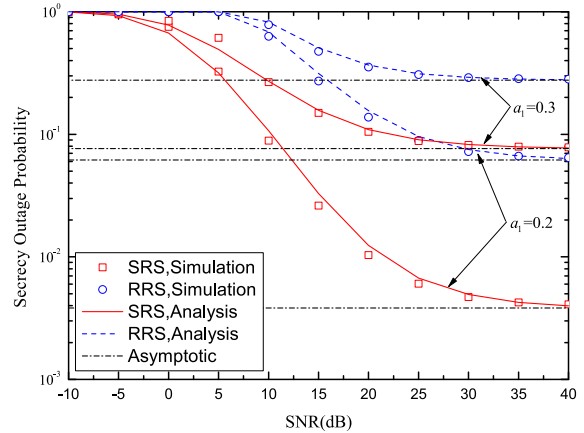


FIGURE 2. SOP versus the transmit SNR for RRS and SRS schemes with $K = 2, \alpha = 3, R_{th_1} = 2$ and $R_{th_2} = 0.5$.

By comparing asymptotic SOP under RRS scheme with SRS and TRS schemes, we find that the SRS and TRS schemes prominently improve the secrecy outage performance, and the interesting discovery is that increasing the amount of relays can further enhance the security performance.

V. NUMERICAL RESULTS

In this section, theoretical and practical simulation results are provided. The abbreviation for the bit-per-channel-use is BPCU. Combined with complexity and exactitude, we set up tradeoff parameter: $N = 30$.

Fig. 2 is drawn to describe the SOP of cooperative NOMA under RRS and SRS schemes for different power distribution coefficients with $K = 2, \alpha = 3, d_{SR} = 0.5, d_{RD_1} = 0.3, d_{RD_2} = 0.5, d_{SE} = 0.8, d_{RE} = 0.6, R_{th_1} = 2$ and $R_{th_2} = 0.5$, where $a_2 > a_1$ and $a_2 = 1 - a_1$. The blue circles and dash curves indicate the accurate SOP of RRS scheme for HD-based NOMA. The red squares and solid curves are the SRS strategy for cooperative NOMA, as can be seen from the accurate result obtained in (31). The curves of theoretical SOP coincide with the statistical simulation results. No matter what SNR situation is, the performance of SRS strategy is preferable to RRS strategy. Moreover, the SOP of the HD-based SRS and RRS schemes under $a_1 = 0.2$ and $a_2 = 0.8$ outperforms the SRS and RRS schemes under $a_1 = 0.3$ and $a_2 = 0.7$, respectively. Another phenomenon can be clearly obtained that HD-based NOMA RRS scheme under $a_1 = 0.2$ and $a_2 = 0.8$ is superior to SRS scheme under $a_1 = 0.3$ and $a_2 = 0.7$ in high SNR range. The reason for this situation is that the power distribution coefficient has a great influence in HD-based RS strategies. In addition, the simulation results also show that the security requirements of the nearby user D_1 have no effect on the security performance layer, which also proves the conclusion of the approximate SOP analyzed in the previous discussion.

Fig. 3 depicts the SOP of cooperative NOMA under RRS and TRS schemes for different power

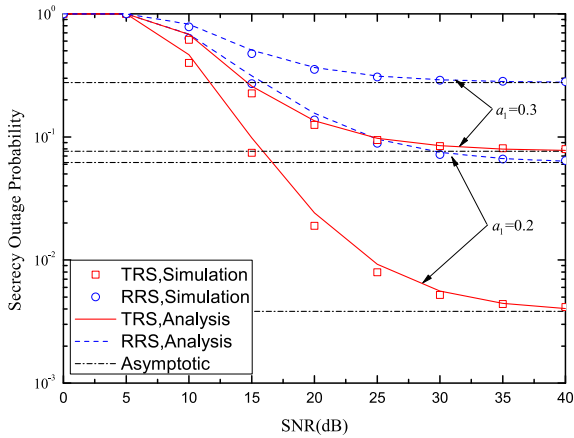


FIGURE 3. SOP versus the transmit SNR for RRS and TRS schemes with $K = 2$, $\alpha = 3$, $R_{th1} = 2$ and $R_{th2} = 0.5$.

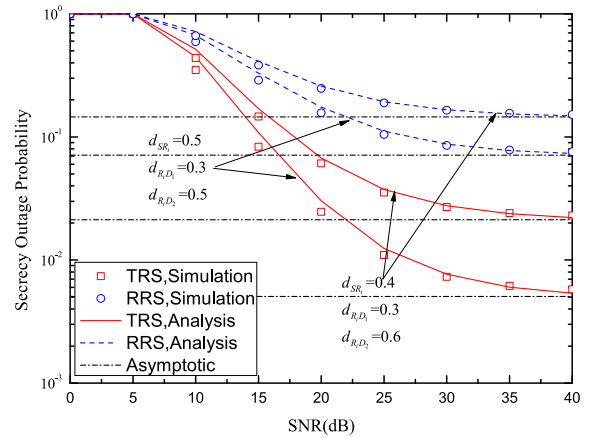


FIGURE 5. SOP versus the transmit SNR for RRS and TRS schemes for the different distances with $R_{th1} = 1$ and $R_{th2} = 0.3$.

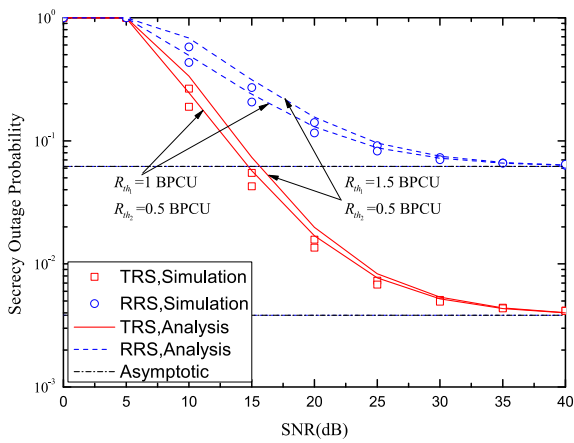


FIGURE 4. SOP versus the transmit SNR for RRS and TRS schemes for the different target rates with $a_1 = 0.2$ and $K = 2$.

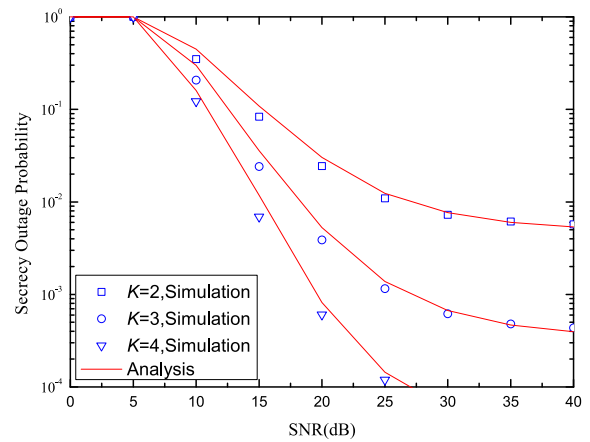


FIGURE 6. SOP versus the transmit SNR for TRS schemes with $K = 2, 3, 4$, $R_{th1} = 1$ and $R_{th2} = 0.3$.

distribution coefficients. The red lines represent TRS scheme for cooperative NOMA, and are consistent with the results obtained in (39). According to the analysis results, the TRS strategy can strengthen security performance. Similarly, the SOP of TRS and RRS schemes under $a_1 = 0.2$ and $a_2 = 0.8$ outperforms the TRS and RRS schemes under $a_1 = 0.3$ and $a_2 = 0.7$, respectively. Moreover, when $SNR < 25$ dB, the security performance of RRS scheme with $a_1 = 0.2$ is inferior to the TRS scheme with $a_1 = 0.3$. When $SNR > 25$ dB, the security performance of RRS scheme with $a_1 = 0.2$ begins to improve and surpasses the security performance of TRS scheme with $a_1 = 0.3$. Therefore, power distribution coefficient has a great influence on the security performance. It is also worth noting that the SOP is saturated in high SNR, and the target confidentiality rate of legal user D_2 can determine the lower performance. The primary cause for this phenomenon is that the NOMA protocol limits the available data rate for weak user D_2 .

In Fig. 4, we compare the SOP using RRS and TRS strategies with different target transmission rates. An interesting observation is that transforming the NOMA user's target rate can affect the security outage behaviors for HD-based RRS

and TRS schemes. As the target rate value reduces, the two kinds of schemes provide better outage performance, but the advantage fades away in high SNR range. Even if an effective RS scheme is implemented, there also exists secrecy performance floor. This is because the application of these two schemes does not eliminate the limitations (e.g., a_2/a_1) imposed by the NOMA protocol.

In Fig. 5, the SOP of cooperative NOMA under RRS and TRS schemes for different distances with $K = 2$, $a_1 = 0.2$, $a_2 = 0.8$, $\alpha = 3$, $d_{SE} = 0.6$, $d_{RE} = 0.4$, $R_{th1} = 1$ and $R_{th2} = 0.3$. This paper normalizes the distances for d_{SR} and d_{RD2} , where $d_{RD1} < d_{RD2}$ and $d_{RD1} + d_{RD2} = 1$, because D_1 is the nearby user, whereas D_2 is the far user. It is observed that the security performance of TRS scheme is superior to RRS scheme when changing the distance. Compared with RRS scheme, there are more obvious variations for TRS scheme with different distances on security performance. Therefore, the distance from BS to R_i and from R_i to D_j has a significant impact on the secure outage performance for HD-based systems. Similarly, the SOP of cooperative NOMA can be influenced by d_{SE} and d_{RE} .

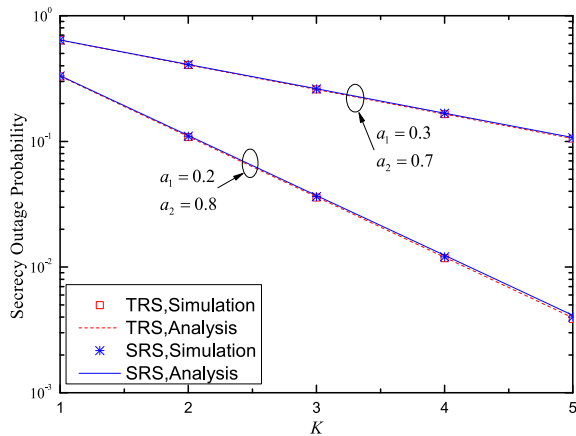


FIGURE 7. SOP versus K for SRS and TRS schemes with $R_{th1} = 1$ and $R_{th2} = 0.5$.

Fig. 6 paints the SOP employing TRS scheme when the number of relays is $K = 2, 3, 4$. The results show that the quantity of relays in this model has great effect on the performance of HD-based TRS schemes. When the number of relays increases, the RS schemes can achieve the lower outage probability. The reason is that the number of relays is positively correlated with diversity gain, thus it can improve the reliability of the cooperative networks.

Fig. 7 shows the SOP for both SRS and TRS schemes with respect to the number of relays K in high SNR region. It is observed that the analytic curves are precisely consistent with the simulated results. It can be concluded that the SOP using RS schemes reduces as the quantity of relays K increases. The security performance is improved due to the application of the efficient RS schemes that take advantage of the diversity of relaying networks. Moreover, from the analysis in section IV and expressions (31), (39), the SOP of both SRS and TRS schemes is coincident on account of $p_1^{out} = 0$ in strong SNR. Another conclusion is that the SOP of the RS schemes becomes smaller when the increasing of a_2/a_1 distinctly.

VI. CONCLUSION

This paper has studied the security performance for cooperative HD-based NOMA IoT systems over Rayleigh-distributed under the influence of different relay selection methods. The closed-form formulae of SOP for two users are derived. Further analysis shows that the SRS/TRS scheme can achieve the best secure performance, and RRS strategy may increase the SOP compared with SRS/TRS strategy. The security performance can be enhanced by augmenting the quantity of relays. Whereas, it is pointed out that due to the adoption of NOMA system, each RS scheme exists secrecy performance floor that cannot be deleted by RS schemes and power allocation strategy.

REFERENCES

[1] X. Chen, L. Guo, X. Li, C. Dong, J. Lin, and P. T. Mathiopoulos, "Secrecy rate optimization for cooperative cognitive radio networks aided by a wireless energy harvesting jammer," *IEEE Access*, vol. 6, pp. 34127–34134, 2018.

[2] Y. Cao, N. Zhao, G. Pan, Y. Chen, L. Fan, M. Jin, and M.-S. Alouini, "Secrecy analysis for cooperative NOMA networks with multi-antenna full-duplex relay," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5574–5587, Aug. 2019.

[3] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.

[4] F. Jameel, S. Wyne, G. Kaddoum, and T. Q. Duong, "A comprehensive survey on cooperative relaying and jamming strategies for physical layer security," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2734–2771, 3rd Quart., 2019.

[5] J.-H. Lee, "Full-duplex relay for enhancing physical layer security in multi-hop relaying systems," *IEEE Commun. Lett.*, vol. 19, no. 4, pp. 525–528, Apr. 2015.

[6] A. Mukherjee, "Physical-layer security in the Internet of Things: Sensing and communication confidentiality under resource constraints," *Proc. IEEE*, vol. 103, no. 10, pp. 1747–1761, Oct. 2015.

[7] L. Qing, H. Guanyao, and F. Xiaomei, "Physical layer security in multi-hop AF relay network based on compressed sensing," *IEEE Commun. Lett.*, vol. 22, no. 9, pp. 1882–1885, Sep. 2018.

[8] A. Pandey, S. Yadav, T. Do, and R. Kharel, "Secrecy performance of cooperative cognitive AF relaying networks with direct links over mixed Rayleigh and double-Rayleigh fading channels," *IEEE Trans. Veh. Technol.*, early access, Oct. 29, 2020, doi: 10.1109/TVT.2020.3034729.

[9] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE 24th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2013, pp. 611–615.

[10] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[11] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[12] X. Li, Q. Wang, M. Liu, J. Li, H. Peng, J. Piran, and L. Li, "Cooperative wireless-powered NOMA relaying for B5G IoT networks with hardware impairments and channel estimation errors," *IEEE Internet Things J.*, early access, Oct. 9, 2020, doi: 10.1109/JIOT.2020.3029754.

[13] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[14] T. Nakamura, A. Benjebbour, Y. Kishiyama, S. Suyama, and T. Imai, "5G radio access: Requirements, concept and experimental trials," *IEICE Trans. Commun.*, vol. E98.B, no. 8, pp. 1397–1406, 2015.

[15] D. Do, T. Nguyen, K. M. Rabie, X. Li, and B. M. Lee, "Throughput Analysis of Multipair Two-Way Repeating Networks With NOMA and Imperfect CSI," *IEEE Access*, vol. 8, pp. 128942–128953, 2020.

[16] Y. Liu, Z. Qin, M. El-kashlan, Y. Gao, and L. Hanzo, "Enhancing the physical layer security of non-orthogonal multiple access in large-scale networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1656–1672, Mar. 2017.

[17] H. Lei, J. Zhang, K.-H. Park, P. Xu, Z. Zhang, G. Pan, and M.-S. Alouini, "Secrecy outage of Max-Min TAS scheme in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 6981–6990, Aug. 2018.

[18] X. Li, M. Zhao, Y. Liu, L. Li, Z. Ding, and A. Nallanathan, "Secrecy analysis of ambient backscatter NOMA systems under IQ imbalance," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12286–12290, Oct. 2020.

[19] K. Jiang, T. Jing, Y. Huo, F. Zhang, and Z. Li, "SIC-based secrecy performance in uplink noma multi-eavesdropper wiretap channels," *IEEE Access*, vol. 6, pp. 19664–19680, 2018.

[20] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, Dec. 2004.

[21] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Lett.*, vol. 18, no. 2, pp. 313–316, Feb. 2014.

[22] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037–2040, Nov. 2015.

[23] Q. Li, P. Ren, and D. Xu, "Security enhancement and QoS provisioning for NOMA-based cooperative D2D networks," *IEEE Access*, vol. 7, pp. 129387–129401, 2019.

- [24] J. Men, J. Ge, and C. Zhang, "Performance analysis for downlink relaying aided non-orthogonal multiple access networks with imperfect CSI over Nakagami- m fading," *IEEE Access*, vol. 5, pp. 998–1004, 2017.
- [25] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.
- [26] J. Chen, L. Yang, and M.-S. Alouini, "Physical layer security for cooperative NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4645–4649, May 2018.
- [27] N. Dahi and N. Hamdi, "Relaying in non-orthogonal multiple access systems with simultaneous wireless information and power transfer," in *Proc. 14th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2018, pp. 164–168.
- [28] Y. Zhang, H. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.
- [29] B. Zheng, M. Wen, C.-X. Wang, X. Wang, F. Chen, J. Tang, and F. Ji, "Secure NOMA based two-way relay networks using artificial noise and full duplex," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1426–1440, Jul. 2018.
- [30] Y. Zou, X. Wang, and W. Shen, "Optimal relay selection for physical-layer security in cooperative wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 10, pp. 2099–2111, Oct. 2013.
- [31] H. Lei, H. Zhang, I. S. Ansari, Z. Ren, G. Pan, K. A. Qaraqe, and M.-S. Alouini, "On secrecy outage of relay selection in underlay cognitive radio networks over Nakagami- m fading channels," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 4, pp. 614–627, Dec. 2017.
- [32] J. Zhao, Z. Ding, P. Fan, Z. Yang, and G. K. Karagiannidis, "Dual relay selection for cooperative NOMA with distributed space time coding," *IEEE Access*, vol. 6, p. 20440–20450, 2018.
- [33] Z. Ding, H. Dai, and H. Vincent Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.
- [34] X. Li, H. Mengyan, Y. Liu, V. G. Menon, A. Paul, and Z. Ding, "IQ imbalance aware nonlinear wireless-powered relaying of B5G networks: Security and reliability analysis," *IEEE Trans. Netw. Sci. Eng.*, early access, Sep. 3, 2020, doi: [10.1109/TNSE.2020.3020950](https://doi.org/10.1109/TNSE.2020.3020950).
- [35] X. Yue, Y. Liu, S. Kang, A. Nallanathan, and Z. Ding, "Spatially random relay selection for Full/Half-duplex cooperative NOMA networks," *IEEE Trans. Commun.*, vol. 66, no. 8, pp. 3294–3308, Aug. 2018.
- [36] H. Zhang, H. Lei, I. S. Ansari, G. Pan, and K. A. Qaraqe, "Security performance analysis of DF cooperative relay networks over Nakagami- m fading channels," *KSII Trans. Internet Inf. Syst.*, vol. 11, no. 5, pp. 2416–2432, May 2017.
- [37] Y. Liu, Z. Qin, M. El-kashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.
- [38] J. G. Proakis, *Digital Communications*, 4th ed. Boston, MA, USA: McGraw-Hill, 2001.
- [39] G. Liu, Z. Wang, J. Hu, Z. Ding, and P. Fan, "Cooperative NOMA Broadcasting/Multicasting for low-latency and high-reliability 5G cellular V2X communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7828–7838, Oct. 2019.
- [40] B. Li, Y. Zou, J. Zhou, F. Wang, W. Cao, and Y.-D. Yao, "Secrecy outage probability analysis of friendly jammer selection aided multiuser scheduling for wireless networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3482–3495, May 2019.



HUI LI received the B.Sc. degree in communication engineering from the School of Information Engineering, in 1999, and the M.Sc. degree in communication and information system and the Ph.D. degree in information and communication engineering from the Nanjing University of Science and Technology, in 2004 and 2008, respectively. He was a Visiting Scholar with Charles Darwin University, Australia, in 2013, and North Carolina A & T State University, in 2014. He is currently a Professor with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo China. His research interests include wireless communications and intelligent signal processing.



YAPING CHEN received the B.Sc. degree in electronic information science and technology from the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo China, in 2019, where she is currently pursuing the M.Sc. degree in communication and information systems. Her research interests include physical layer security (PLS) and cooperative communication.



MINGFU ZHU received the B.Sc. degree from Tianjin University, in 2000, the M.Sc. degree from East China Normal University, in 2004, and the Ph.D. degree from the University of California, Los Angeles (UCLA), in 2007.

From 2011 to 2013, he was working as an Executive Director with Mayyard Photoelectric Technology Company Ltd., Ningbo, China. Since 2013, he has founded and served as the Chairman of Hebei National Optoelectronics Technology Company Ltd., Hebei, China. He is currently the Chairman of Henan Chuangzhi Technology Company Ltd., and a General Manager of Henan Chuitian Technology Company Ltd., Hebei. His research interests include chip packaging and intelligent light development and manufacturing. With innovative ideas, intelligent lights are used to build IOL, integrate IOL into IoT, and upgrade to 5G IoT. By building scientific research platforms and manufacturing bases, 5G industry ecosystem is built to interact with upstream and downstream enterprises. He received several awards and achievements, which include the excellent builder for the socialist cause with Chinese characteristics, special government allowance under the State Council and leading talents in Science and Technology Innovation of National Ten Thousand Talents Plan and so on. He has served as a member of Henan CPPCC and a Vice President of the Henan Euro-American Alumni Association. He is also a President of the Henan Alumni Association of Tianjin University and the Director of the Henan Mechanical Engineering Society.



JIANGFENG SUN (Member, IEEE) received the M.S. degree in communication and information system from Zhengzhou University, in 2009. He is currently pursuing the Ph.D. degree with the Beijing University of Posts and Telecommunications. He is currently a Lecturer with the School of College of Computer Science and Technology, Henan Polytechnic University. He has several papers published in journal and conferences. His current research interests include physical layer

security, cooperative communications, and performance analysis of fading channels.



DINH-THUAN DO received the B.S., M.Eng., and Ph.D. degrees in communications engineering from Vietnam National University (VNU-HCMC), in 2003, 2007, and 2013, respectively. From 2003 to 2009, he was a Senior Engineer with the VinaPhone Mobile Network. From 2009 to 2010, he was a Visiting Ph.D. Student with the Communications Engineering Institute, National Tsinghua University, Taiwan. His name and his achievements will be reported in special

book entitled *Young talents in Vietnam 2015-2020*. His research interests include signal processing in wireless communications networks, NOMA, full-duplex transmission, and energy harvesting. His publications include over 80 SCIE/SCI-indexed journal articles, over 45 SCOPUS-indexed journal articles, and over 50 international conference papers. He is sole author in one textbook and one book chapter. He was a recipient of the Golden Globe Award from Vietnam Ministry of Science and Technology, in 2015, (Top 10 most excellent scientist nationwide). He is currently serving as an Editor of *Computer Communications* (Elsevier), an Associate Editor of the *EURASIP Journal on Wireless Communications and Networking* (Springer), and an Editor of *KSII Transactions on Internet and Information Systems*.



VARUN G. MENON (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. His research interests include the Internet of Things, fog computing and networking, underwater acoustic sensor networks, cyber psychology, hijacked journals, ad-hoc networks, and wireless sensor networks. He is a Distinguished Speaker of ACM Distinguished

Speaker. He is also a Guest Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE SENSORS JOURNAL, the *IEEE Internet of Things Magazine*, and the *Journal of Supercomputing*. He is an Associate Editor of *IET Quantum Communications*. He is also an Editorial Board Member of the IEEE Future Directions: Technology Policy and Ethics.



SHYNU P. G. (Member, IEEE) received the Ph.D. degree in computer science from the Vellore Institute of Technology (VIT), Vellore, India, and the master's degree in engineering in computer science and engineering from the College of Engineering, Anna University, Chennai, India. He is currently working as an Associate Professor with the School of Information Technology and Engineering, VIT. He has published over 30 research papers in refereed international conferences and


journals. His research interests include deep learning, cloud security and privacy, ad-hoc networks, and big data.

• • •

[Home](#) > [Journal of Real-Time Image Processing](#) > Article

Special Issue Paper | [Published: 24 June 2020](#)

Dual-mode power reduction technique for real-time image and video processing board

[Sunil Jacob](#), [Varun G. Menon](#) , [Saira Joseph](#) & [Paramjit Sehdev](#)

Journal of Real-Time Image Processing **17**, 1991–2004 (2020)

183 Accesses | **2** Citations | [Metrics](#)

Abstract

In real-time image and video processing boards, power, speed, and area are the most often used measures for determining the performance of motion imagery applications. Due to technological advancement, power consumption has gained major attention in real-time image processing ability compared to speed. The increase in on-chip temperature due to larger power consumption has resulted in reduced operating life of chip and battery-driven devices. In this work, a new logic family has been introduced i.e., dual-mode logic (DML), which provides flexibility between the optimization of energy and delay (E-D

optimization). This gate can be switched between two modes of operation that is a static mode (CMOS-like mode), which provides low power consumption and dynamic mode, which provides high speed. Recently, power leakage has become a dominant problem due to continuous data transfer among a large number of connected devices. Thus, to reduce power leakage, a self-controllable voltage level (SVL) power reduction technique is used along with DML logic. In the SVL technique, a maximum dc voltage is provided to the active load circuit on-demand or decrease the dc supplied to the load circuit in the standby mode. Integrating DML with the SVL technique reduces power consumption as well as leakage power. A 4-bit RCA, 8-bit RCA, and 16-bit RCA are used for verifying the proposed method and comparison of performance parameters is done with a conventional circuit. Complete circuit implementation and simulation are carried out in TANNER EDA version 13 tools with operating voltage of 1 V. The proposed system is further applied to real-time image, and we obtain the finest resolution level with minimum power consumption.

This is a preview of subscription content, [access via your institution](#).

Access options

Buy article PDF

23. Müller, H., Unay, D.: Retrieval from and understanding of large-scale multi-modal medical datasets: A review. *IEEE Transactions on Multimedia* **19**(9), 2093–2104 (2017)

Acknowledgement

Authors would like to thank Dr. Mohammad Khosravi, Department of Computer Engineering, Persian Gulf University, Bushehr, Iran for his valuable inputs that has substantially helped in revising and improving the research paper.

Author information

Authors and Affiliations

Center for Robotics, SCMS School of Engineering and Technology, Cochin, 683576, India

Sunil Jacob

Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Cochin, 683576, India

Varun G. Menon

Department of Electronics and Communication Engineering, SCMS School of Engineering and Technology, Cochin, 683576, India

Saira Joseph

Department of Mathematics and Computer Science, Coppin State University, Baltimore,

MD, 21216, USA

Paramjit Sehdev

Corresponding author

Correspondence to [Varun G. Menon](#).

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Jacob, S., Menon, V.G., Joseph, S. *et al.* Dual-mode power reduction technique for real-time image and video processing board. *J Real-Time Image Proc* **17**, 1991–2004 (2020). <https://doi.org/10.1007/s11554-020-00992-x>

Received

10 October 2019

Accepted

10 June 2020

Published

24 June 2020

Issue Date

December 2020

DOI

<https://doi.org/10.1007/s11554-020-00992-x>

Keywords

Dual-mode logic (DML)

An efficient and adaptable multimedia system for converting PAL to VGA in real-time video processing

Authors:  Deepak Kumar Jain,  Sunil Jacob,  Jafar Aizubi,  Yarun Menon [Authors Info & Claims](#)

Journal of Real-Time Image Processing, Volume 17, Issue 6 • Dec 2020 • pp 2113–2125 • <https://doi.org/10.1007/s11554-019-00869-4>

Published: 01 December 2020 [Publication History](#)

6 0






Abstract

Abstract

Real-time video processing has found its range of applications from defense to consumer electronics for surveillance, video conferencing, etc. With the advent of Field Programmable Gate Arrays (FPGAs), flexible real-time video processing systems which can meet hard real-time constraints are easily realized with short development time. Most of the existing solutions have high utilization of system resources and are not quite flexible with many applications. Here we propose a hardware–software co-design for an FPGA-based real-time video processing system to convert video in standard Phase Alternating Line (PAL) 576i format to standard video of Video Graphics Array (VGA)/ Super Video Graphics Array (SVGA) format with little utilization of resources. Switching between multiple video streams, character/text overlaying, and skin color detection are also incorporated with the system. The system is also adaptable for rugged applications. VHDL Hardware Description Language (VHDL) codes for the architecture were synthesized using Altera Quartus II and targeted for Altera Stratix I FPGA. Results achieved confirm that the proposed system performs efficient conversion with very less resource utilization compared to the existing solutions. Since the proposed system is also flexible, many other applications can be incorporated in the future.

References

- Kehtamavaz N, Gamaelia M: Real-time image and video processing: from research to reality. *Synth Lectur Image Video Multimed Process* (2006) 2(1), 1 - 108. [10.2200/S00021ED1V01Y200604NM005](https://doi.org/10.2200/S00021ED1V01Y200604NM005)  
- Wenge, Z., Huiming, H.: FPGA-based video image processing system research. In: 2010 3rd International Conference on Computer Science and Information Technology, Chengdu, China, vol. 4, pp. 680–682. IEEE (2010) 
- Gao, X., Wei, X., Liu, Y.: An FPGA implementation of multi-channel video processing and 4 K real-time display system. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Shanghai, China, pp. 1–6. IEEE (2017) 

Show All References

Cited By

[View all](#) 

Fan C. (2023). Evaluation of machine learning in recognizing images of reinforced concrete damage. *Multimedia Tools and Applications*. **82**:19. (30221-30248). Online publication date: 1-Aug-2023.

<https://doi.org/10.1007/s11042-023-14911-2>

Zhang N, Shan Z, Yao Q and Jain D. (2022). Multimedia Based Medical Big Data Analysis of Leg Swing Strike Effect in Wushu Sanda Sports. *Wireless Communications & Mobile Computing*. **2022**. Online publication date: 1-Jan-2022.

<https://doi.org/10.1155/2022/7181370>

Zhang P, Hou J and Khan R. (2022). Physical Education Teaching Strategy under Internet of Things Data Computing Intelligence Analysis. *Computational Intelligence and Neuroscience*. **2022**. Online publication date: 1-Jan-2022.

<https://doi.org/10.1155/2022/5296497>

Show All Cited By

Index Terms

An efficient and adaptable multimedia system for converting PAL to VGA in real-time video processing

Computer systems organization

Computing methodologies

Embedded and cyber-physical systems

Artificial intelligence

Real-time systems

Hardware



Association for Computing Machinery

Browse

About

Proceedings Books SIGs Conferences People

Search ACM Digital Library



Periodical Home Latest Issue Archive Authors Affiliations Award Winners

Index terms have been assigned to the content through auto-classification.

Numerical Study on Cyclic Loading Effects on the Undrained Response of Silty Sand



M. Akhila , P. C. Jithesh, K. Rangaswamy, and N. Sankar

Abstract The soil liquefaction is a major earthquake disaster which causes tremendous damages to all infrastructure facilities. The examples during past earthquakes have shown the evidence of liquefaction-induced ground failures in fine-grained soils. Until recently, liquefaction-related studies concentrated on clean sands believing that only sands are susceptible to liquefaction. However, the earthquakes like 1976 Tangshan earthquake, the 1989 Loma Prieta earthquake, the 1999 Kocaeli earthquake, the 2010 Chile earthquake, and the 2011 Christchurch earthquake, etc., showed that sand with fines could also liquefy. The present study deals with the numerical simulations on cycling loading effects on the undrained response of silty sand. The material parameters for the numerical model are found after conducting basic experimental tests. The response of silt sand under the undrained condition of cyclic triaxial loading is analyzed using the hypoplastic constitutive model. The influence of soil parameters, i.e., void ratio/relative density and consolidation pressure level on undrained response of soil is examined from model simulations.

Keywords Hypoplastic model · Silty sand · Undrained response

1 Introduction

The hypoplastic constitutive model has been utilized for all geotechnical applications of field studies since 1985. The original version of the hypoplastic model is coined by Kolymbas (1985) and is improved in further versions. The present improved version of the hypoplastic model is more advanced over the elasto-plastic models. P.-A. von Wolffersdorff (1996) has described the mathematical formulations involved in

M. Akhila

Department of Civil Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala, India
e-mail: akhila144@gmail.com

P. C. Jithesh (✉) · K. Rangaswamy · N. Sankar

Department of Civil Engineering, NIT Calicut, Kozhikode 673601, Kerala, India
e-mail: pcjithesh08@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

M. Latha Gali and R. R. P. (eds.), *Geotechnical Characterization and Modelling*, *Lecture Notes in Civil Engineering* 85, https://doi.org/10.1007/978-981-15-6086-6_86

1067

1 **PM10 source identification using the trajectory based potential source** 2 **apportionment (TraPSA) toolkit at Kochi, India**

3 Afifa K. Shanavas¹, Chuanlong Zhou², **Ratish Menon**¹, Philip K. Hopke^{2,3*}

4

5 ¹Dept. of Civil Engineering, **SCMS School of Engineering and Technology, Karukutty,**
6 Ernakulam, Kerala, India - 683582.

7 ²Center for Air Resources Engineering & Science, Clarkson University, Potsdam,
8 NY 13699 USA.

9 ³Department of Public Health Sciences, University of Rochester School of Medicine and
10 Dentistry, Rochester, NY 13642 USA

11

12 **Abstract**

13 A recently developed open source tool kit named the Trajectory based Potential Source
14 Apportionment (TraPSA) was used to identify the sources of respirable particulates at Kochi,
15 India. Using 24-hour average particulate matter data from samples collected at five regulatory
16 monitoring stations over the five-year period from January 2011 to October 2016, local and
17 regional scale analyses were made. Concentration field analysis was performed using back
18 trajectories generated by Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT)
19 model with inputs from atmospheric reanalysis data. Conditional bivariate probability function
20 analyses were made using local meteorology data to identify the influence of local sources. Most
21 of the stations indicated the contribution from local traffic activities during low wind conditions
22 and from a nearby industrial area especially during high speed winds. Back trajectory analysis
23 identified potential source areas in Kerala as well as in nearby state of Tamil Nadu as contributing
24 to the air quality at Kochi. Arabian sea on the western side was also observed to be a potential
25 source area for Kochi. The study demonstrated the utility of TrapSA as a tool for deriving
26 information about the potential source areas affected particulate matter mass concentrations.

27

28 **Keywords:** PM₁₀, Air pollution, TraPSA, Back trajectory receptor model, HYSPLIT, Kochi

*Corresponding Author Email: phopke@clarkson.edu

Cite this article

Ranga Swamy K, Akhila M and Sankar N
Effects of fines content and plasticity on liquefaction resistance of sands.
Proceedings of the Institution of Civil Engineers – Geotechnical Engineering,
<https://doi.org/10.1680/jgeen.19.00270>

Research Article

Paper 1900270
Received 17/11/2019;
Accepted 24/06/2020

Keywords: dynamics/geotechnical
engineering/seismic engineering

ICE Publishing: All rights reserved

Effects of fines content and plasticity on liquefaction resistance of sands

K. Ranga Swamy PhD

Associate Professor, Department of Civil Engineering, NIT Calicut, Calicut, Kerala, India

M. Akhila PhD

Assistant Professor, SCMS School of Engineering and Technology, Ernakulam, Kerala, India (corresponding author: akhila144@gmail.com)
(Orcid:0000-0002-0514-0841)

N. Sankar PhD

Professor, Department of Civil Engineering, NIT Calicut, Calicut, Kerala, India

A detailed experimental programme was conducted to evaluate the liquefaction susceptibility of non-plastic and low-plasticity soils subjected to cyclic loading under undrained triaxial loading conditions. The study mainly focused on examining the influence of the amount of fines and plasticity indices on the liquefaction resistance of sands. After mixing silt and clay fractions into fine sand, 16 soil combinations were prepared. The silty sands contained up to 40% non-plastic fines and the low-plasticity soils contained a clay fraction of 5–40%. Each cylindrical soil specimen was constituted to a medium relative density and saturated specimens were subjected to a confinement pressure of 100 kPa. The consolidated specimens were then subjected to various levels of cyclic stress amplitudes using a sinusoidal wave load form at a frequency of 1 Hz. The results showed that both the non-plastic and low-plasticity clay soils were less resistant to liquefaction than the fine sand. The soils belonging to categories SM and SC (silty sand and clayey sand, respectively, as per Indian standard soil classifications) were susceptible to liquefaction if the fines passing through a 75 μm sieve were $\leq 40\%$, the liquid limit was $\leq 40\%$, the plasticity index was < 15 and the saturated water content was about 0.86 times the LL.

Notation

C_c	coefficient of curvature
C_u	uniformity coefficient
D_{50}	mean size
e_c	consolidation void ratio
e_o	initial void ratio
e_{\max}	maximum void ratio
e_{\min}	minimum void ratio
e_{sk}	sand skeleton void ratio
N_L	number of cycles to liquefaction
G	specific gravity
w	water content
Δu	change in pore water pressure
ε_A	axial strain
σ_3	effective applied consolidation pressure

1. Introduction

Liquefaction-induced soil failures in earthquakes have been known to occur in several soil deposits in a loose to medium-density state containing non-plastic to low-plasticity fines. In general, those types of soils cause a rapid build-up of excess pore pressures during seismic excitations. It is difficult to dissipate the excess pore pressures within a short duration of an earthquake event due to the presence of small voids in fine-grained soils. Therefore, foundation soils containing non-plastic and low-plasticity clay fractions underneath buildings and geotechnical structures are susceptible to liquefaction during an earthquake or other dynamic event. Liquefaction causes severe damage to building structures, the soil and

soil-retaining structures, in the form of settlements, lateral spreading, tilting, ground damage and so on. The prevention of such disastrous damage is thus required and current research has focused on understanding the mechanisms and finding suitable mitigation measures.

A review of the literature on the undrained response and liquefaction susceptibility of non-plastic and plastic soil mixtures reveals that unique conclusions have not been found. The liquefaction resistance of sand may vary with the fines content (FC) and the plasticity of the fines. The effects of the non-plastic FC on the liquefaction strength of sands have been extensively investigated, with contradictory findings on the basis of comparisons of liquefaction resistance such as the (global or total) void ratio, skeleton void ratio, relative density and so on. Based on in situ tests, Seed *et al.* (1985) reported that the presence of fines induces an increase in liquefaction resistance. However, with an increase in the FC, some laboratory investigations found an increase in the liquefaction resistance (Amini and Qi, 2000; Chang *et al.*, 1982; Dezfulian, 1984; Fei, 1991; Vaid, 1994) while reversal behaviour was observed in other studies (Finn *et al.*, 1994; Kuerbis *et al.*, 1988; Lade and Yamamuro, 1997; Zlatovic and Ishihara, 1997). The liquefaction resistance of sands has been reported to decrease up to a certain FC but then increases with a further increase in FC (Koester, 1994; Polito and Martin, 2001; Troncoso, 1990). According to Shen *et al.* (1977) and Kuerbis *et al.* (1988), the sand skeleton void ratio is the best parameter to evaluate the liquefaction resistance of non-plastic soils. Based on a review of various studies, Carraro *et al.*

Numerical Study on the Undrained Response of Silty Sands Under Static Triaxial Loading



M. Akhila, K. Rangaswamy and N. Sankar

Abstract The silty soils are more susceptible to liquefaction, even under static loading, than the coarse sands. Pore pressure developed during dynamic events may not dissipate easily due to the presence of more number of small voids. Hence, the rate of pore pressure build-up under static/dynamic loading conditions is much faster in silty sands, which lead to a reduction in the soil strength. This phenomenon may be assessed in terms of either contraction or dilation behaviour under triaxial loading. Therefore, it is necessary to analyse the undrained response of silty sands under triaxial loading so that the damages occurring during future dynamic events may be predicted. The present study involves both the experimental and numerical simulations on various silty sands, which contain 0, 10, 20, 30 and 40% silt fines. Initially, experimental static triaxial testing was performed to determine the undrained response of silty sands moulded to cylindrical specimens at medium relative density. The saturated samples are isotropically consolidated at 100 kPa pressure before shearing. Further, numerical simulations were performed on silty sands by inputting the material parameters into the hypoplastic model. This model requires eight material constants as input including critical friction angle, hardness coefficients, limited void ratios, peak state and stiffness coefficients. These constants were determined for each silty sand combination after conducting basic laboratory tests according to the formulations build in the hypoplastic model program. The experimental trends were compared with numerical model simulations under triaxial testing. The effect of the initial state of soil and the amount of silt fines on the undrained response of fine sands is discussed in detail. The liquefaction susceptibility of silty sand is described based on steady state line concept. The results indicate that the silt sands behave as highly contractive, i.e. more liquefiable when compared with sands.

M. Akhila (✉)

Department of Civil Engineering, SCMS School of Engineering and Technology,
Ernakulam, Kerala, India
e-mail: akhila144@gmail.com

K. Rangaswamy · N. Sankar

Department of Civil Engineering, NIT Calicut, Calicut, India

© Springer Nature Singapore Pte Ltd. 2020

A. Prashant et al. (eds.), *Advances in Computer Methods and Geomechanics*, Lecture Notes in Civil Engineering 56,
https://doi.org/10.1007/978-981-15-0890-5_17

9th World Engineering Education Forum 2019, WEEF 2019

Service Learning in Engineering Education: A Study of Student-Participatory Survey for Urban Canal Rejuvenation in Kochi, India

Sunny George¹, Ratish Menon¹, Pramod Thevanoor¹ and John Tharakan^{2,*}

¹SCMS Water Institute, SSET Campus, Karukutty, Kerala, India

^{2,*}College of Engineering and Architecture, Howard University, Washington DC 20059, USA

Abstract

It is widely accepted that learning through doing, or service learning (SL) and engaging students in community centred project based learning (PBL) is transformative in terms of enhancing student learning and employability, effectively improving both technical and soft skills that are sought after by employers, while at the same time growing and developing an informed and educated citizenry. Participatory learning here is a pedagogical approach in which students involve themselves in a community-based project, which has proven to be more effective than direct lecture based transfer and absorption of knowledge. In this paper, we present a case study from Kochi city in Kerala, India, where undergraduate (UG) engineering students from the environmental engineering (EE) program at SCMS School of Engineering and Technology (SSET) participated voluntarily in the comprehensive survey of a 10.87 km canal running through busy, dense and heavily populated urban area of Kochi City. This Thevara-Perandoor (T-P) canal was a heavily used commercial artery for the city. Unfortunately, the T-P canal is now totally degraded, primarily due to unregulated solid waste dumping and untreated sewage inflows at numerous locations along its course throughout the urban space. The UG students of the CE program at SSET voluntarily came forward to do the study on behalf of Kochi Municipal Corporation (KMC). This partnership, between an academic program and a community based entity, such as a municipal corporation or any other community based entity, establishes a model for integrating meaningful service learning into engineering education. The partnership provided an immense opportunity for the students to implement whatever they had learned in the classroom and doing so by working for the benefit of the community in which they themselves were resident. This paper describes the practices that are being followed in this service learning exercise. The paper also focuses on the impediments as well as the opportunities that exist for both widening and deepening the knowledge domain of the students, while working on the mentioned urban canal survey. The value and impact of the model described through the examined case study is especially important, given that the notion of service learning as a pedagogical approach is gaining momentum in the Indian engineering education sector, and when programs such as *Unnat Bharath Abhiyan* which focuses on and mandates utilizing student voluntary work for rural development are being implemented.

© 2020 The Authors. Published by Elsevier B.V.

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>)

Peer-review under responsibility of the scientific committee of the 9th World Engineering Education Forum 2019.

Keywords: Service Learning; Urban Canal Rejuvenation; Engineering Education; Kochi; Thevara-Perandoor Canal;

A Novel Spectrum Sharing Scheme using Dynamic Long Short-Term Memory with CP-OFDMA in 5G Networks

Sunil Jacob, Member, IEEE, Varun G Menon, Senior Member, IEEE, Saira Joseph, Member, IEEE, Vinoj P G, Alireza Jolfaei, Senior Member, IEEE, Jibin Lukose and Gunasekaran Raja, Senior Member, IEEE

Abstract—With the rapid increase in communication technologies, shortage of spectrum will be a major issue faced in the coming years. Cognitive radio is a promising solution to this problem and works on the principle of sharing between cellular subscribers and ad-hoc Device to Device (D2D) users. Existing 5G spectrum sharing techniques work as per a fixed rule and are pre-established. Also, recent game theoretic approaches for spectrum sharing uses unrealistic assumptions with very less practical implications. Here, a novel spectrum sharing technique is proposed using 5G enabled bidirectional cognitive deep learning nodes (BCDLN) along with dynamic spectrum sharing long short-term memory (DSLSTM). Joint spectrum allocation and management is carried out with wireless cyclic prefix orthogonal frequency division multiple access (CP-OFDMA). The BCDLN self-learning nodes with decision making capability route information to several destinations at a constant spectrum sharing target, and cooperate via DSLSTM. BCDLN based on time balanced and unbalanced channel knowledge is also examined. With the proposed framework, expressions are derived for the spectrum allocated to multiple sources to obtain their spectrum targets as a variant of the participation node spectrum sharing ratio (PNSSR). The impression of noise when all nodes broadcast with equal spectrum allocation is also investigated.

Index Terms—5G, Artificial Intelligence, Cognitive Nodes, CP-OFDMA, Deep learning, PNSSR, Spectrum sharing

I. INTRODUCTION

THE enormous demand for localized services has forecasted twofold increase in mobile data traffic compared to the existing fixed IP traffic [1], while the cellular networks fails to deliver this growing demand due to restricted bandwidth and shortage of spectrum [2-5]. Device-to-device (D2D) communication has been proposed and integrated in the next generation mobile network as a possible solution to meet this

rising mobile traffic demand. D2D communication allows user to transmit and share cellular data among close proximity users without the involvement of the base stations [6]. Two approaches are employed in D2D to enhance the quality of service: the direct approach and collaborative approach. The cooperative D2D scheme utilizes the cellular user to relay information between base station and end-user to speed up the communication. Decode-and-forward and amplify-and-forward are the prominent relay schemes employed to achieve cooperation among nodes [7]. One of the important factors in multi-hop communication is relay selection. Although various relay selection methods are used to achieve fast data transmission, many of them suffer from limitations in transmission delay, security.

D2D communication happens via the route discovery protocol with increased power savings as base stations are not involved. Once ad-hoc D2D users determine their route, they can share the spectrum with cellular users, i.e. they can operate in the same frequency spectrum as licensed cellular radio network. The existing 5G spectrum sharing has a fixed rule and it is pre-established [8-9]. To overcome the issues, 5G enabled bidirectional cognitive deep learning nodes (BCDLN) along with dynamic spectrum sharing long short-term memory (DSLSTM) is proposed. BCDLN are self-learning proactive and predictive node with decision making capability that creates dynamically adaptable clusters. The joint spectrum allocation and management for 5G access wireless CP-OFDMA communication system with numerous source and multiple destination BCDLN based on time variant and invariant channel knowledge is examined. Each of the BCDLN in the network, sends data to different receivers at a fixed spectrum sharing target and cooperates via DSLSTM with different frequency slots.

Sunil Jacob is with Center for Robotics, SCMS School of Engineering and Technology, Karukutty 683576 India. Email: suniljacob@scmsgroup.org

Varun G Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Karukutty 683576 India. Email: varunmenon@scms.org (Corresponding Author)

Saira Joseph is with the Department of Electronics and Communication Engineering, SCMS School of Engineering and Technology, Karukutty 683576 India. Email: saira_joseph@scmsgroup.org

Vinoj P G is with the Department of Electronics and Communication Engineering, SCMS School of Engineering and Technology, Karukutty 683576 India. Email: vinojp@scmsgroup.org

Alireza Jolfaei is with the Department of Computing, Macquarie University, Australia. Email: alireza.jolfaei@mq.edu.au

Jibin Lukose is with Center for Robotics, SCMS School of Engineering and Technology, Karukutty 683576 India. Email: jbin2nd@gmail.com

Gunasekaran Raja is with the Department of Computer Technology, Anna University, India. Email: dr_gunasekaran@ieee.org

The hidden role Patriarchy in Malayalam Cinema: An analysis of the movie 'sufiyum sujathayum'

Febini M Joseph¹, Cefy Joice J²

¹Assistant Professor of English, SCMS School of Engineering and Technology, Karukutty, Ernakulam, Kerala, India

²Student, Kerala Law Academy, Trivandrum, India

Received: 09 Nov 2020; Received in revised form: 23 Nov 2020; Accepted: 26 Nov 2020; Available online: 14 Dec 2020

©2020 The Author(s). Published by Infogain Publication. This is an open access article under the CC BY license

(<https://creativecommons.org/licenses/by/4.0/>).

Abstract— *Movies are the most popular medium that represents the popular taste and culture. Malayalam cinema has undergone several radical shifts throughout the past years, But still there are movies to satisfy or reinstate the traditional gender roles and patriarchy. The movie Sufiyum Sujathayum is about how society transforms our way of thinking based on the strict adherence to patriarchy. More than love people consider social acceptance as the most important priority. The paper is an analysis of the movie by using theories such as male gaze and feminism.*

Keywords— *Gender Stereotypes, Male Gaze, Patriarchy.*

Gender has been seen as a principle set up by theorists like Simone de Beauvoir. Rather than the result of sexual differences, it is represented as the consequence of social customs and practices which incorporates support from movies to practice of regular discussions as described in the book History of Sexuality by Michel Foucault. Foucault in this manner sums up that sex is the set impacts delivered on bodies, practices' and social relations by the sending of 'complex political advancements'.

Foucault's talk on the advances of sex is reformulated by Teresa De Lauretis, whose Technologies of Gender expresses that sexual orientation is a portrayal of connection having a place with a class, a gathering or a classification. The portrayal of gender starts from the principal material that contacts a child's body. The cliché garments in blue or pink breaks the perfect world and drives the youngster into a framework of portrayals and images. To cite Teresa De Lauretis, that sexual orientation isn't sex, a condition of nature however the portrayal of every person as far as a specific social connection which pre exists the individual and is predicated on the theoretical and inflexible resistance of two organic genders, which establish 'the sex-sex framework'.

Through the sexual orientation people start to fortify certain 'proper' practices, young ladies get prepared in craft and music while young men takeover the play areas. Barbie dolls and kitchen sets enhance the rooms of girls when the young boys play with automatic rifles and autos. The ongoing inclination of selecting young ladies to karate classes is furthermore just to build up their 'safeguard component' which accentuation that 'you are fragile and could be a weakling!' Our famous legends additionally strengthen the indistinguishable thought. The chivalric male warriors wandering around to abstain from squandering the moaning females, or the sovereign appeal coming to spare heaps of the alluring Rapunzel from the hands of the witch underlines the indistinguishable factor. The temperate, sensitive, delightful, and crying women are consistently princesses where on the grounds that the forceful females are consistently witches. The case of sexual orientation generalizing in legends is that the depiction of guys as globe-trotters and pioneers and females as aides or supporters.

The visual media particularly, film is one among the various innovations of sexual orientation. Laura Mulvey in her exposition, Visual Arts and Narrative Cinema clarifies the effect of visual expressions as a decent social innovation in deciding one's belief systems.

The enchantment of movies emerged from the gifted and fulfilling control of visual joys. The suggestive is being coded to the language of the prevailing male centric request in conventional standard movies and this is regularly the earlier element of achievement in each entertainment world. Movies as an assortment grasp both elitist and mainstream ideas of craftsmanship and work intimately with abstract style. The verbal and visual works of art don't appear to be only equal however intuitive and associated. A film will be considered as a social ancient rarity, which speaks to the way of life and convention to which it has a place. When it enters the social texture of a general public, it progressively impacts the way of life additionally. The entertainers likewise assume a significant job inside the methods for articulation in film.

"The visual medium offers tremendous decisions which the composed account may not. There's a more prominent opportunity inside the decision of point of view; the organizations are various camera eye, storyteller, lights, utilization of room, the language, visual correspondence, face comparatively in light of the fact that the hushes. There's likewise the vital projection of generalizations." The sex generalizations are made by these verbal and visual media's which assembles the social ideas and philosophies of the moving toward ages too. The idea of perfect spouse, perfect mother and so on are remoulded in movies. The perfect ladies in Kerala are thought as "Malayali Manka". She is considered as a kind of a goddess figure and she or he complies with each and every standards in society. She is considered on the grounds that the encapsulation of gentility. In her we will see the blooming of female temperance.

The high social improvements in Kerala lists has offered ascend to the 'fantasy of Malayali ladies 'as getting a charge out of a superior status than their partners somewhere else inside the nation, particularly the high female education inside the state. This legend has been enlarged and supported by proof that matrilineal kinds of connection designs were common in specific networks in Kerala. The elevated level of female proficiency and work, 33% reservation of seats in nearby administration bodies, high sex proportion and low fruitfulness rates alongside high female physical wellbeing accomplished a specific measure of social and political strengthening inside the property right.

Even feminine images in visual media are intended to fulfil the male looks. Intentionally or accidentally ladies emulates the vivid screen to satisfy the other gender. Malayalam film neglects to speak to the encounters of ladies from alternate point of view. At the point when a female situated film is delivered the star

esteem is a low in light of the fact that a lady assumes the principle job. Malayalam film reflects Malayali tastes, wants and dreams; one would then be compelled to surrender that so as to comprehend the contemporary public activity of karalla we ought to likewise view the delicate pornography motion pictures which once made the Kerala entertainment world drifting. Similar watchers of Adoor and Chandran films likewise delighted in Shakeela motion pictures.

The current movies or the so called movies which represent nuances in the way of presentation still gives picture of woman who are always under the control of men in the family. Obeying orders and living according to the unwritten norms are the fate of so called Malayali Mankas (A term used to represent ideal woman in Kerala) represented in movies. It is considered as usual and acceptable to everyone. But knowingly or unknowingly it provides a wrong message to the audience. The symbolic representations also denote the struggle taken by a woman when she transcends her limits. Sufiyum Sujathayum is a 2020 Malayalam movie directed and written by Naranipuzha Shanavas and produced by Vijay Babu under the banner of Friday film house. Sujatha is mute daughter of Mallikarjunan and Kamala. Sujatha was a talented dancer and an energetic girl in her village. One day she meets Sufi on her bus journey who is a disciple who returns to meet his master Ustad. Soon after their meeting both of them falls in love and they decided to elope accidentally her parents caught her and married off to Rajeev who lives in Dubai. After ten years Sufi returns to the village to meet Ustad but he was no more alive. Sufi gives out a prayer call (bank) Sufi passes away during the prayer Sujatha's husband Rajeev decides to bring her back to her village to attend Sufi's burial. Rajeev pays a visit to Sufi but Sujatha was not allowed to see him according to their beliefs woman were not allowed inside. At that evening Rajeev's passport seemed missing and they searched everywhere and he got reminded of the incident that the passport may fell into Sufi's grave and. Rajeev and his father in law decided to unearth Sufi's grave with the help of their tenant. They could not find his passport in the grave at the same time Sujatha arrives there with his passport and she throws that Misbahha (A chain with Green Beads used for prayer by Muslims).As given by Sufi gave her as her Mehar and she wanted to give him back the misbhaha that his mother gave him she placed it on his grave when her husband opened it.

The heroine is dumb and her thoughts are expressed through written words and gestures. She is lovable and everyone gives her freedom until she falls in love with a man from another religion. Her father tries to

stop her, but she plans to elope with her lover. At the moment, like several other movie scenes father tries to persuade her by sentiments. She was not able to protest and that is another symbolic way that represented the tragedy of several women. She never gets a chance to unleash her thoughts through spoken form.

The system of marriage is praised and the value is reinstated in the movie. Even though she is not mentally ready to live with her husband, she leads a troublesome life for ten long years. And the husband is always keeps jealous over her past relationship and hates her lover. When Sufi died, he ardently tries to make her realize that her love is gone forever. And when they travel together in climax scene, she holds his hands with love. And in the tomb of Sufi she throws away his ornament that she kept for all these years. It's a symbolic representation of grabbing herself from an unseen bond of love and longing.

Sujatha's grandmother was a person who was more modern in thoughts and deeds. She always respected her granddaughter's ideas and thoughts. When the groom's family came to see Sujatha, she said to them,

Avalude lokam molila...aa lokam avrude onnu kanatte
(Her world is above, let them see it too)

And when she talks with her grandmother, they discuss about a plant and her grandmother told that

"Dead bodies are buried in that place and we (woman) can't enter there. But I have gone there

These simple dialogues convey the progressive thoughts from a woman who lived a traditional life. But she deviates from the one way path of tradition. The death of grandmother is a symbolic one because it is the disappearance of a ray of hope and dreams for Sujatha.

Even though Sujatha enjoys freedom on all aspects, when she confronts with her lover or family, she sacrifices her true desires for the sake of family. Her supportive father changes completely when she is in love with a man from another religion. The conventional behaviour patterns and patriarchal ideologies are hidden while her decision making power is offended.

The movie reinstates the patriarchal ideologies that are deep rooted in Kerala. The feminine and pleasing appearance of the heroine also demands obeying and sacrificing role. At the concluding part, like a typical woman in India, she starts living in accordance with her husband. In the beginning also she awakes from a dream as if something gets dragged from their body. The various elements that are introduced contribute to reinstate the tastes of Malayali audience.

REFERENCES

- [1] Butler, Judith. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge, 1999. Print.
- [2] Beauvoir, Simone de. *The Second Sex*. New York: Vintage Books 1989, c1952. Print.
- [3] Pillai, Meena T. *Women in Malayalam Cinema: Naturalising Gender Hierarchies*. New Delhi: Orient Black Swan, 2010.
- [4] Woolf, Virginia. *A Room of One's Own*. New York: Harcourt, Brace and Company, 1929.
- [5] Foucault, Michel. *The History of Sexuality*:New York :Pantheon Books, 1978.
- [6] Bretl, Daniel J., and Joanne Cantor. "The portrayal of men and women in U.S. television commercials: A recent content analysis and trends over 15 years." *Sex Roles* 18.9-10 (1988): 595-609. Print.
- [7] Yue, Ming-Bao. "Gender and Cinema: Speaking through Images of Women." *Asian Cinema*, vol. 22, no. 1, 2012, pp. 192-207. Print
- [8] Knight, Julia. "Cinema of Women." University of Illinois Press, 2017, doi:10.5406/illinois/9780252039683.003.0017.
- [9] Devasundaram, Ashvin Immanuel. "Indian Cinema Beyond Bollywood." 2018, doi:10.4324/9781351254267.
- [10] Austin, Guy. "Representing Gender." *Algerian National Cinema*, 2019, doi:10.7765/9781526141170.000009.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Synthesis and characterization of Co-5Cr-RHA hybrid composite using Powder metallurgy

U. Arunachalam^a, G.R. Raghav^{b,*}, S. Dhanesh^c

^a Department of Mechanical Engineering, University College of Engineering, Nagercoil 629004, Tamilnadu, India

^b Department of Mechanical Engineering, SCMS School of Engineering and Technology, Vidyayanagar Karukutty, Ernakulam 683576, India

^c Department of Mechanical Engineering, SNS College of Engineering, Coimbatore, Tamilnadu 641107, India

ARTICLE INFO

Article history:

Received 5 March 2021

Received in revised form 7 April 2021

Accepted 10 April 2021

Available online xxxxx

Keywords:

Powder metallurgy

Wear

Corrosion

RHA

ABSTRACT

Cobalt-Chromium alloys are in high demand as a material for prosthetics and dental implants. Powder metallurgy was used to create Co-5Cr-RHA (Rice Husk Ash) hybrid composites in this research. RHA is made by heating rice husk in a furnace to 700 degrees Celsius. The surface morphology of the Co-5Cr-RHA hybrid composites is analysed using a scanning electron microscope. Due to the RHA reinforcement, the Micro hardness of the Co-5Cr-10RHA hybrid composite increased by 8% as compared to other samples. The density of the hybrid composites has decreased as a result of the addition of RNA. The compressive strength of the Co-5Cr-10RHA (130 MPa) hybrid composites has increased by 4%. The addition of RNA reinforcement has a positive effect on tribological behaviour, according to tribological studies. Because of the oxides in the RHA, wear loss and COF have decreased significantly. The after-wear SEM analysis confirms that abrasive wear is the primary wear mechanism. The corrosion behaviour of the Co-5Cr-RHA hybrid composites was investigated using the electrochemical workstation in the presence of a 3 percent NaCl electrolytic solution. Of all specimens, Co-5Cr-10RHA hybrid composites have a stronger E_{corr} value of -0.812 V.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

1. Introduction

The need for materials with exceptional properties in the field of bio implants, such as dental and orthopaedic implants, has resulted in substantial research and development activities for Cobalt matrix composites. As opposed to ceramic matrix composites, the advantage of using metals as matrix materials is their superior mechanical and wear resistance [1–3]. Ceramic matrix composites, on the other hand, are known for their high temperature stability and corrosion resistance. The reinforcement of natural ceramic particles such as fly-ash in metal matrix composites enhanced various properties of the composites [4]. One of the most important factors that define the mechanical properties of composite materials is the consistent spreading of reinforcements. Many composite manufacturing techniques, such as welding, coating processes such as HVOF, and physical vapour deposition, make it difficult to achieve uniform reinforcement dispersion. The P/M

(powder metallurgy) method can easily achieve a uniform amalgamation of matrix and reinforcements [2,5–9].

Because of its remarkable mechanical properties (young's modulus = 210 GPa, hardness = 1040 MPa, density = 8.90 g/cm³), cobalt is being considered for bio implants. They also have outstanding temperature control. The above properties make them ideal for biomedical alloys. Some materials suitable for biomedical alloys, such as Nickel and Titanium, are allergic to humans. As a result, cobalt is being researched as a possible substitute for the above materials in dental prosthetics and other bio implant applications [5,6,10].

Fuzeng Ren et al. evaluated the various tribo-corrosion properties of nano cobalt developed through a P/M (powder metallurgy) process. The final results show that nano Cobalt's mechanical properties have greatly improved [11]. The nano Cobalt's corrosion resistance has decreased, but it still has strong wear resistance. CoCrMo hybrid composites intended for bio implants were investigated by H. Stevenson et al. The wear tests were performed in Human Synovial Fluid and Bovine Calf Serum [12]. Yanjin Lu et al used the laser melted method to create CoCrW alloy for dental

* Corresponding author.

E-mail address: raghavmechklnc@gmail.com (G.R. Raghav).



Nature and environmentalism: Post-colonial eco critical rereading of selected Nigerian poems

Divya MS

Assistant Professor, Department of English, SCMS School of Engineering and Technology, Ernakulam, Kerala, India

Abstract

This paper is an attempt to discuss about ecology and environmentalism in the selected poems of Nigerian poets Wole Soyinka, Tanure Ojaide and Niyi Osundare in a Post-colonial Eco critical review. In literature, ecocriticism is a mode of aesthetics that deals with the nature of relation between literature and the natural environment. Its adherents investigate human attitudes towards the world as reflected in writing about nature. It is a diverse genre known by many names, including green cultural studies, eco poetics and literary analysis of the environmental. The study seeks to explore selected poems in Nigerian literature from an Eco critical perspective. The relationship between man, the environment and nature is documented in literature. Eco-critical insights are studied in the poetry of Wole Soyinka, Tanure Ojaide and Niyi Osundare. Literature resides where creation exists, and where nature exists, life exists. Literature is an imperative tool for having a historical understanding of the relationship between man and also for determining the way man treats nature in future. In the 1990's, ecocriticism gained significant prominence in the Western academia as a domain of literary research. This does not, however, indicate that the literature of earlier periods ignored ecologically conscious concerns. Similarly, ecological scepticism seeks to explain how nature is expressed in literature and how the meaning of nature and the relationship between man and nature have changed over time as they are perceived in literature. In recent decades, the natural environment has progressively become threatened by man's activities. The chosen poems are full of varied environmental details. The poets responded to their plight in distinctive perspectives through their poetry. Extreme ecological issues such as global warming, increased pollution levels, recurrent coastal flooding, tsunami and cyclones, earthquakes and floods have culminated from the incessant cutting of trees for human use and deforestation, the use of weapons and arms, radioactive elements in nuclear power plants, industrial pollution and many more. Not only has this disruption to nature caused a catastrophic change in the atmospheric conditions around the world, but the ozone layer, our earth's defensive shield, has also been destructive. And now there is a growing and crucial need to conserve our environment and make our earth a better place to live. In Nigerian Literature, the study provides a more detailed introduction to the Eco theory from its beginnings to the present. It will also address the relationship between nature and culture, the gradual progression of ecocriticism, and its related concepts.

Keywords: ecocriticism, eco psychology, eco poetics, ecological issues

Introduction

Ecocriticism is literature is an analytical method that examines the importance of the relationship between literature and the natural environment. With several names, green cultural studies, eco poetics and environmental literary criticism, it is a diverse genre. Ecocriticism began to gain prominence in Western academia in the 1990s as a sphere of literary research. Ecological criticism seeks to analyse how nature is presented in literature and how, as seen in literature, both the interpretation of nature and the relationship between man and nature have grown over time. British colonial rulers formed a chain of command in many British colonies, such as Anglo-Egyptian Sudan and Nigeria, in which colonial officials ruled over indigenous African leaders, who then governed the majority of the African indigenous population. Colonialism in Africa is primarily responsible for the continent's lack of cultural, social, and political development. The so-called empirical scrutiny of agricultural practices imposed in northern Nigerian communities by successive British colonial era authorities is an example of a European influenced paradigm pursued by African elites. Irrigation, forest

management, and extensive use of chemical fertilizers were emphasized by the colonial scientific scrutiny system. The system provided very little benefit for the region from economic development and disrupted the traditional farming practices that for centuries had sustained the local population. Researchers and academic investigators have largely overlooked the effects of postcolonial Nigeria's economic growth. The colonization process resulted in the realignment of power, with European trading companies imposed by the colonial authority replacing the hitherto domestic Nigerian authority centers such as Opobo's Ja Ja, Oguta's Kalabari and Ibadan's Ijebu.

"Ecocriticism speaks for the earth by rendering an account of the indebtedness of culture to nature while acknowledging the role of language in shaping the view of the world"

(Campbell 5)

Thus Ecocriticism begins from the conviction that the arts of creativity and the research there of will make a major contribution to the understanding of environmental issues and the various types of eco-degradation affecting planet Earth today. Global warming, which triggers rapid climate change

as a result of unequal human interactions with nature, is a real concern that marked the end of the twentieth and early twenty-first centuries. Ecological issues are caused by climate change and have become an important concern for interdisciplinary / multidisciplinary studies. Under the concept of ecocriticism, multiple literature disciplines have embraced this style of work, centred on ecological issues. The ambivalent relationships between man and nature are old and either require or need to overcome and master human romantic devotion to nature. In the foreseeable future, climate change has arisen from these anthropocentric relationships. The reality of climate change is threatening every corner of the world. Yet he believes that lethal silence is a big impediment to resolving and mitigating climate change problems. Wangari Maathai is unveiling the true global warming issues that would have dramatic consequences on Africa. At the global stage, the query is answered as:

“Africa is the continent that will hit hardest by the climate change. Unpredictable rains and floods, prolonged droughts, subsequent crop failures and rapid desertification, among other signs of global warming, have in fact already begun to change the face of Africa.”

(as cited by Toulmin, 2008, p. 1).

In environmental concerns and philosophies, there are several expressions that share similar denominators in the objective of environmental conservation. For Graham Huggan and Helen Tiffan:

“Postcolonial ecocriticism and Ecocriticism are hedged about with seemingly insurmountable problems. The two fields are notoriously difficult to define not least by their own practitioners.... Thus, internal divisions...e.g. the commitment to social and environmental justice or differences... and large scale distinctions based on the attractive view that postcolonial studies and eco/environmental studies offer mutual correctives to each other turn out to... be perilous” (3). Postcolonial ecocriticism, on the other hand, is a plurality of ecocriticism that discusses: “concerns with conquest, colonisation, racism, sexism along with its investments in theories of indigeneity and diaspora and the relations between native and invader, societies and cultures” (Huggan and Tiffan 6) to explicate Eco critical modes of feminist ecocriticism, romantic ecocriticism and postcolonial Ecocriticism “need to be understood as particular ways of reading” (Huggan and Tiffan 13). Regardless of the numerous discourses on ecocriticism and postcolonial ecocriticism, this research indicates that postcolonial ecocriticism cannot be evaluated without delving into environmental problems, and ecocriticism or eco-environmental studies cannot be discussed without discussing postcolonial concerns alongside imperialism, a metaphor that examines ideologies of supremacy and socio-history.

It is in this regard that am going to analyse some ecological problems in Wole Soyinka's poem “*Dedication for Moremi 1963*” with a post-colonial perspective. The concept of the poem is about the natural order of things, and also about bringing a child into the world. It begins with the consummation of the child, and then the birth of the child into the universe of this child, a miracle created by love. It's almost like a prayer to the Earth, and a dedication to the child. It

speaks of our human life as a whole, and also of our journey back to earth. He makes use of many poetic devices in the poem, including metaphors and a lot of imagery. The line in which he says, “your tongue arch / to scorpion tail.” is one instance that stands out as a good metaphor. A pretty metaphor compares a crying baby's tongue at birth to a scorpion tail when it flicks in terror when feeling threatened. It gives us this impression of the child being born with a venomous tongue, which later brings trouble to the parents- as well as presenting this picture of a baby's squirming tongue as it clears its lungs and wails in fear of being so unexpectedly brought into this world. There are plenty of imagery examples, including the moment where he says, “Earth's honeyed milk, wine of the only rib / Now roll your tongue into honey until your cheeks are / Swarming honeycombs — your world needs sweetening kids. Through this, we get this image of taste and touch and sight all in one, the very thought makes my mouth water. The poem is full of deep inner meanings that invoke a radiant feeling, make us wonder what it means, see these peculiar literal images that attack our senses, and give us the emotions that the poet wants us to experience. The tone of this poem is joy and wonder at the birth of a child, and all those involved can feel the spiritual journey. He relates this miracle of life to the earth, as a woman bears a child, and her fruits are brought forth by it. The sound is gentle and ties us to the earth, as if every part of this birth was nature, just like every part of any animal or plant birth. In many of his words, like baobab, roots, rain, plumb her deep for life, season, fruits, and embrace, he creates the earthy and joyful sound. They all give us the feeling of a warm earth coming together to bring this happy occasion to life. In the midst of the independence of Nigeria, Soyinka recalls the many events that took place throughout his life, such as the birth of his daughter and the opening of the first National Park in Nigeria. Soyinka writes through many frames that the poem can be read through, one being a nourishing tone for his daughter, as well as one that protects the earth and its resources. The earth can be seen as a symbol of the daughter and the daughter can be seen as a symbol of the earth. Poet gives an insight to his daughter regarding the endless parallels and metaphors about the world, and how it functions. He says, “my child- your tongue arch to scorpion tail, spit straight and return to danger's threats yet coo with the brown pigeon, tendril dew between your lips.” This is the example of Soyinka asking his daughter to be as sharp and dangerous as a scorpion but also to be caring, gentle and kind as a pigeon. He clearly shows the paternal qualities he imparts to his daughter in a manner similar to the way he tells the people of Nigeria to protect their new park. He wraps up the poem with the idea that we too must let the world depend on us in the same way we rely so heavily on the sun. We have to give earth back in the way it gives us. Soyinka evokes the past not as a dead past, but as a living one whose positive or negative results catch the present and influence the future, not historical but archetypal any more. Either to condemn those suicidal attitudes or to laud the current resistant wilderness, he evokes pastoral imagery, recalls the less anthropocentric past as a less troubled model, and projects a green future as a common dream. As the only way to face fundamental and sustainable growth, Soyinka urges readers

and listeners to take on the soil. As an expression of inextricable human ties with it, this communion with one's land at every level includes mind-set, commitment, love, and respect for oneself and all of its inhabitants. As a result of technical and scientific developments, the African holistic world view that imperialists saw as "savage" has become the global solution to the danger that climate change presents today. It's not too religious to ask "who was wild and who was civilized" if the "savage" incriminated African world view has since become a "worldwide genius" response to the climate change problem.

Tanure Ojaide is a significant Literary voice of Nigerian post-war poetry, distinguished by his recourse to the orator of his birthplace. Ojaide takes oratory as a locus of an esthetic that is conscious of rural people's arts and politics, particularly in the face of a viperous, modernity-driven establishment. The focus of his poetry on orality implies its rootedness in nature. But the point that nature in Ojaide's poetry is not merely evoked as an esthetic technique, an embellishment of what many have regarded in his poetry as an overwhelming political theme, is much more crucial to this paper. Nature is also addressed as home (the natural world, biodiversity, flora and fauna), now a forgotten home in the face of modernity and global petrodollar capitalism. In the sense of postcolonial ecocriticism, I try to point out from a reading of his poetry that the nature (environment) of the Niger Delta area from which the poet comes from is a victim of exploitation and injustice caused by large-scale oil extraction in the region, just like the people living in it; and it is no longer the pristine home it used to be. Tanure Ojaide's fifteenth poetry book, "The Tale of the Harmattan" (2007), offers poetry readers and those familiar with his work a critical insight into the Niger Delta region's bleak socio-political and economic circumstances. The plurality of the poet's concerns are oil extraction and its negative environmental and human family effects. The poems differ in style and form; however, what makes the collection a publication of substance is the poet's ability to discuss contemporary problems with a spectator's eyes, and the sincerity of an empathically inspired one. This compilation illustrates the degradation of the biodiversity and climate of the Niger Delta as a result of the extraction of oil and the marginalization of the ethnic minority in whose territories the oil is mined. In one poetry collection divided into three parts with a glossary that familiarizes the reader with the landscape, politics, Urhobo mythology, and various historical and mythical figures of Nigeria, the prolific Nigerian scholar-poet Tanure Ojaide uses bold rhetoric and a variety of techniques to claim the person of the poet as an eyewitness to historical events, especially the destruction of the destruction of the Niger Delta's ecosystem and environment as a result of oil exploitation and the marginalization of the ethnic minority people in whose land oil is exploited. He shows concern for the underprivileged and oppressed in society, whose fight for equality, fairness and justice he supports, in the course of this poetic story. Conscious of the postcolonial situation in Nigeria, his native nation, he condemns the rampant corruption that drains the country's enormous wealth. Affirming humanity, he condemns the perpetrators of genocide, as in the Darfur region of Sudan, in the strongest

possible words. The fact that what happens in Nigeria's troubled oil-rich yet poor Niger Delta region affects the worldwide price of oil demonstrates the degree of local and global connectivity, what is now described as 'glocal.' The Harmattan Tale (2007) argues that his research on the indigenous peoples (especially women) of Nigeria's Niger Delta offers an important way to revise our understanding of postcolonial theory in order to step beyond the outdated notion of colonial nations to colonialist power as sitting in multinational corporations that transcend national origin. My research combines elements from environmental, political, and socio-cultural images to analyze how Ojaide's work exposes the relationship between environmental problems and government collusion with multinational corporations, while calling for a vision of environmental justice to be accomplished by the movement of the Delta people. Ojaide's definition of historic environmental destruction and devastating oil contamination caused by multinational oil firms in the Niger Delta region is part of an interdisciplinary and multi-theoretical view of neo-colonial literature. The dialogic development of a variety of discourses is part of his complex literary style; his work involves feminist discourse and eco-critical interpretation of environmental issues, as well as post-colonial discourse that has become a defining feature of contemporary African literature. Ojaide's earlier-generation poetry and establishes him in post-colonial African poetry as a significant voice. The poems in *The Harmattan Tale* share Ojaide's love for exploring ancient African folklore with readers. In these poems, Ojaide's concerns owe much of their connection to his sensibilities and affinities towards his homeland. He does not surrender his creative inclinations or call for a Marxist agenda for political sloganeering or writing poetry, as one can admit, unaware of the genius of his imaginary complexity.

The fourth collection of poetry "The Eye of the Earth" by Niyi Ariyoosu Osundare (1986) ^[10], Nigerian ecology is celebrated in this work and focus is given to the common man where it portrays one of the fiercest indictments of the people and alien destructive powers of modern economic culture. *The Eye of the Earth* (1986) by Osundare is divided into three sections: back to earth, eye-ful glances of rain songs and home call with eighteen poems. This study investigates ecological implications in such poems as "forest echoes", "The Rocks Rose to meet me", "harvest call", "Let the earth's pain Be Soothed", "First rain", "Rain-coming", "Rain drum", "farmer-born", "They too Are the Earth", "Ours to Plough, Not to Plunder" and "Our Earth Will Not Die". *The Eye of the Earth* poetry is divided into poems of varying lengths that lament the harm to the Nigerian climate for economic reasons and technological development. The poet's memories and impressions are captured by a series of confessional and lyrical poetry. The environmental views of Osundare are drawn precisely from the Yoruba world view of traditional values taken from African culture. He claims that nature promotes a coherent equilibrium between microscopic species, insects, plants and humans and calls for the protection of the environment in Nigeria from the destruction of modern civilizations. It takes a pictorial account of man-and-earth violence. In other words, in the quest for better leadership by

alternative order, *Eye of the Earth* (1986) is dedicated to reclaiming the earth that has been forced to prostrate by capitalist processes. The poetry of Osundare is based on a vigorous, sustained concern for one of the oldest producers in the world: the peasants, those who till the land, and their quasi-mythical links to the earth. His goal is to immerse the realities and multiple lineaments of Africa's underdevelopment and poet laments on the ecological collapse and future which threatens the Nigerian landscape showing the increasing level of environmental degradation by the world's mining industries. The poet's concern for the pathetic condition of the Nigerian environment and the propensity of the Nigerian ruling class to safeguard and exploit land, power and income resources at the cost of ecological balance and the well-being of the oppressed people is self-evident in this volume of poetry. The poet is concerned with both fact and the relationship between the individual and his environment. Therefore, it is not surprising that the whole volume is dedicated to poems about man engaging with nature's physical aspects. Really, the opening poem 'Forest Echoes' is a harbinger of what's to come. The poet saunters into the Ubo Abusoro forest in the poem, from where he allows his sea of memory to flood unimpeded. The first thing that strikes the poet when he enters the forest is the destruction by timber traders of the land and the trees referred to as *agbegilodo* in the poem. From this position, the poet laments the fact that, as a consequence of exploitation, these economic trees were reduced to mere stumps. There is the palm-wine tree which is described as conqueror of rainless seasons/mother of nuts and kernels/bearer of wine and life. In 'Forest Echoes,' Osundare portrays man, the ground, animals, plants (actually all of nature) interacting and celebrating at this period of universal productivity in one festive mood. It's set in the past but it's meant to reinforce our current understanding. The second poem in the collection '*The Rocks Rose to Meet Me*' is an encounter with the rocks – another aspect of physical nature. Before the rock of Olosunta, the poet is standing and waiting like Christopher Okigbo at heavens gate. And the Olosunta rock began to address the poet in the following words:

“You have been long, very long, and far
Unwearying wayfarer,
Your feet wear the mud of distant waters
Your hems gather the bur
Of farthest forests;
I can see the west most sun
In the mirror of your wandering eyes”
(Osundare the Eye of the Earth, 13).

In these lines, Osundare is doing some kind of homecoming. He is a renegade and is now trying to establish vital links with the past. As he put it:

‘The Rocks Rose to Meet Me’ is a homecoming of a Kind, a journey back (and forth) into a receding past Which still has a right to live. The rocks celebrated in This section... occupy a central place in the cosmic Consciousness of Ikere people; they are worshipped and frequently appeased with rare gifts, thunderous

Drumming and dancing
(Osundare the Eye of The Earth ‘Preface’ xiii).

The truth is that Osundare honors the rocks of Olosunta in Ikere cosmology, since they are both an aspect of physical existence and have a supernatural dimension. It is mother earth and natural laws require that the resources of nature should be used to advance society. Osundare also revolves around the cosmology of Ikere individuals in 'Harvest Call'. The rocks that rose in the previous poem to meet the poet are also named guardians of the spirit of harvest in Ikere's worldview. Thus, in this portion of the collection, all the poems speak of crops, harvest and bounty. The assumption is that the earth is a source of development and growth. Fertile and generous, it is. It will create food and resources for the good of mankind. In fact, the earth means abundance and abundance. The Earth is seen as the centre of wealth and life. Yet the rain acts as an agent or regulator between man and Earth. In his poetry, Osundare explores and praises these two facets of nature through introspection and nostalgia. Osundare also makes the suggestion in his celebration of the theme of nature that the dispossession of the world by some powers in society is capable and can actually threaten the full life of man as a human being.

References

1. Abdu, Saleh. *The Peoples Republic: Reading the Poetry of Niyi Osundare*. Kano: Benchmark Publishers, 2003.
2. Abrams, M.H. *A Glossary of Literary Terms*. Canada: Wadson Cengage Learning, 2009.
3. Ascroft B, Gareth G, Helen T. *Postcolonial Studies: The key Concepts*. New York: Routledge, 2007.
4. Byron, Lord George ‘Childe Harold’s Pilgrimage’ Ed. Frank Kermonde and John Hollander. *The Oxford Anthology of English Literature*. Vol. 2. London: Oxford University Press, 1973, 2.
5. Barret, Lindsay. “The Niger Delta Conundrum” *New African* 483, 2009. Print. Betty Roszak and Theodore Roszak, „Deep Form in Art and Nature”. *The Green Studies Reader: From Romanticism to Ecocriticism*. Laurence Coupe (ed) New York: Routledge, 2000.
6. Bodunde, Charles. “Niyi Osundare and the Materialist Vision: A Study of the Eye of the Earth.” *Ufahamu Journal of the African Activist*, 1997; 5:81.
7. Charles E. “The Possibilities of Hope: Africa in Niyi Osundare’s Poetry”. *Lagos Papers in English* 2, 2007, 62-63
8. Chiwenzu. *Towards the Decolonisation of African Literature Vol.1*. Enugu: Fourth Dimension Publishers, 1980.
9. Edward. Said. *Culture and Imperialism*. London: Chatto and Windus, 1993.
10. Osundare, Niyi. *The Eye of the Earth*. Ibadan: Heinemann, 1986.
11. Ruecket, William. “Literature and Ecology: An Experiment in Ecocriticism” *The Ecocriticism Reader: Landmarks in Literary Ecology*. Glotfelty Cheryl and Harold Fromm. Athens: University of Georgia P, 1996.
12. Russell S. Sanders. “Speaking a Word for Nature”. *The*

- Ecocriticism Reader: Landmarks in Literary Ecology. Cheryll Glotfelty and Harold Fromm (ed). Athens: University of Georgia P, 1986.
13. Walunywa, Joseph. Postcolonial African Theory and Practice: Wole Soyinka. PhD Dissertation. Syracuse: Syracuse University, 1997.

Full Length Research Paper

City scale water audit of a pilgrimage town in South India

Merin Mathew^{1*}, Sunny George², Ratish Menon¹ and John Tharakan³

¹SCMS School of Engineering and Technology, Karukutty, Kerala 683 576, India.

²SCMS Water Institute, SSET Campus, Karukutty, Kerala 683 576, India.

³College of Engineering and Architecture, Howard University, Washington DC 20059, USA.

Received 9 December, 2020; Accepted 9 February, 2021

The water need of a religious pilgrimage town in South India would typically be much larger than a regular town where religious pilgrims and ritual activities do not add to the water burden of the municipality. To understand this added water burden, a city scale audit was carried out to estimate the water supply, demand and deficit at Guruvayoor, which is a pilgrimage town in South India. Guruvayoor is popularly known for the Sri Krishna Temple which is visited daily by an average of 10,000 devotees. For the entire municipality, 11,117 open wells, including 144 public wells within the municipal area. The study revealed that increased dependency on ground water sources without proper implementation of rainwater harvesting (RWH) facilities demonstrated a potential threat for the water security of the town. Increased water distribution by water tanker trucks, mostly operated by the unorganized private sector, imported 2.5 MLD of water from the outer bounds of the city to meet the commercial and institutional demand. The results of this investigation showed that urban water security will likely be subject to such external water suppliers, suggesting the need for further research to understand the implications of such a distributed water supply panel on urban water security.

Key words: Urban water security, Pilgrimage town, Guruvayoor, water audit, water demand, South India.

INTRODUCTION

Water is one of the essential resources for the existence of life. The requirement of water has essentially increased over a period of time, especially due to explosive population growth, urbanization, and industrialization. In the last century, water use has been growing at more than twice the rate of population increase. It is predicted

that the water withdrawals will increase by 50% in developing countries and by 18% in developed countries, and as many as 1.8 billion people will be living in countries or regions with absolute water scarcity, with as much as two-thirds of the world's population potentially under water-stress conditions (UN-Water, 2015; UNDP,

*Corresponding author. E-mail: merinmathew@scmsgroup.org.

Liquefaction resistance improvement of silty sands using cyclic preloading

Akhila M^{1,3}, Rangaswamy K² and Sankar N²

¹Department of Civil Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala, India

²Department of Civil Engineering, NIT Calicut, Kerala, India

³E-mail: akhilam@scmsgroup.org

Abstract. Liquefaction induced damages are plenty and cause various levels of destruction to civil engineering infrastructure. It is possible to prevent liquefaction-induced hazards by understanding the mechanism and adopting some improvement techniques or design the structure to resist the soil liquefaction. In the present study, the influence of cyclic preloading on the liquefaction resistance of sand-silt mixtures is analyzed by conducting undrained cyclic triaxial tests on the cylindrical samples reconstituted at medium dense conditions ($D_r = 50\%$). All samples were tested at an effective confining pressure of 100 kPa by varying the cyclic stress ratios (CSR) in the range of 0.127 to 0.178 using a sinusoidal waveform of frequency 1 Hz. The results are presented in the forms of the pore pressure build-up, axial strain variation and liquefaction resistance curves. Test results indicate that the liquefaction resistance of silty sands is increased substantially with the application of preload under drained conditions.

1. Introduction

Liquefaction induced damages are plenty and cause various levels of destruction to civil engineering infrastructure. It is possible to prevent liquefaction-induced hazards by understanding the mechanism and adopting some improvement techniques or design the structure to resist the soil liquefaction. The first possibility is to avoid the construction on liquefiable soil deposits as far as possible. However, it is mandatory to utilize the available land for the various infrastructure developments due to scarcity in the availability of land even it does not satisfy the required properties. Hence, the second option is to make the structure resistant to liquefaction by adopting deep foundations. Nevertheless, the deep pile foundations may not prevent liquefaction damages in all cases. Piles are causing to deflect in liquefaction susceptibility zones. Hence, the third option is liquefaction mitigation which involves improving the strength, density, and drainage characteristics of the soil. The selection of the most appropriate ground improvement method for a particular application could depend on many factors including the type of soil, level, and magnitude of improvement to be attained, required depth and extent of the area to be covered. This paper presents an experimental study regarding the applicability of preloading for the improvement of liquefaction resistance.

2. Literature review

Preloading of the soils occurs naturally (for eg., erosion, the flow of groundwater, etc) or artificially (purposeful preloading to improve the soil properties, demolition of structures, etc). A few researchers have analyzed the liquefaction resistance of preloaded soils. The details are given in Table 1.





A review on the use of ferrocement with stainless steel mesh as a rehabilitation technique

Juby Mariam Boban, Anjana Susan John*

Department of Civil Engineering, SCMS School of Engineering and Technology, Kerala, India

ARTICLE INFO

Article history:

Received 29 September 2020

Received in revised form 6 December 2020

Accepted 10 December 2020

Available online 3 February 2021

Keywords:

Rectangular columns

Preload

Stainless steel

Ferrocement confinement

Rehabilitation

Ultimate load

ABSTRACT

One of the major issue faced by the construction industry is the degradation of structures due to different loads acting on the structure. So retrofitting and rehabilitation has become quite inevitable and it can help in regaining the original strength of the structure. Use of ferrocement is an effective method and it is used in developed countries as it is considerably cheap and materials of construction are easily available. Ferrocement is a system of construction using reinforced mortar or plaster applied over an armature of metal mesh, woven expanded-metal or metal-fibers and closely spaced thin steel rods such as rebar. The skill required is of low level and it has superior strength properties as compared to conventional reinforced concrete. The main drawback of ferrocement is corrosion. Thus to avoid corrosion stainless steel jacketing is employed for rehabilitation within the study that opens the scope for a new jacketing methodology.

© 2020 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Second International Conference on Recent Advances in Materials and Manufacturing 2020.

1. Introduction

Concrete is the most popular construction material which is made of cement, aggregate and water. Water is acting as the bonding agent between the component. On adding water, the concrete is in a plastic state and acquires strength with time. Portland cement is the ordinarily used type of cement for production of concrete. Concrete is used in the construction of the major structural elements like foundations, columns, beams, slabs and other load bearing components. The use of traditional construction materials such as steel and concrete showed signs of deterioration due to prolonged action of loads which results in degradation of overall strength of the structure which makes it futile. This degradation is a result of poor construction techniques, flaws in designing process or may be due to poor updating of the methods specified in design codes. Proper maintenance is a partial solution. So is a necessity of an effective rehabilitation technique which will improve the life expectancy of the structure. Earlier studies focused on steel meshes which is prone to corrosion. My study focuses on a non corrosive technology for rehabilitation. The scope of stainless steel as a jacketing method is not studied formerly.

In most of the developed countries, the development trade has almost reached saturation. So there is an increasing demand to ameliorate and strengthen the existing structure instead of demolishing. The damages are mainly due to the environment degradation, design inadequacies, poor construction practices, irregular maintenance, requirement of revision of codes in practice, increase in the loads and seismic conditions etc. Rehabilitation is one of the practical solution for such structural collapse and it can be done effectively by strengthening the load bearing components or by strengthening the vital components of the building which results in the failure of the building. Therefore, rehabilitation and upgrading of degraded structure has become one among the foremost vital challenges in development industry. In several cases, the whole demolition of the existing structure is not an economical answer as it becomes an exaggerated money burden. So upgrading or repairing the structure is an effective practical approach. Column is the major compression load bearing component member and the failure of which results in the failure of the whole building. During earthquakes, columns are likely to undergo brittle failure. So the ductility of columns has to be improved to prevent the inelastic deformation occurred during earthquakes. Whereas repair and rehabilitation using ferrocement enhance the strength and ductility of the column. Proper selection of the strengthening material is inevitable to enhance the properties of the column.

* Corresponding author.

E-mail address: anjanajohn@scmsgroup.org (A. Susan John).



Surface modification of tungsten fillers for application in polymer matrix composites

E. Jenson Joseph^{a,*}, V.R. Akshayraj^b, K. Panneerselvam^c

^a Faculty of Mechanical Engineering, SCMS School of Engineering & Technology, Ernakulam, India

^b Production and Industrial Engineering, SCMS School of Engineering & Technology, India

^c Faculty of Production Engineering, National Institute of Technology, Tiruchirappalli, India

ARTICLE INFO

Article history:

Available online 10 February 2021

Keywords:

Surface modification

Silane coupling agent

Tungsten particle

Fourier Transform Infrared Spectroscopy

Thermogravimetric analysis

ABSTRACT

In this research, a new class of treated metal fillers that can be used as reinforcements in polymer matrix composites have been developed. Surface modification of the tungsten metal particles is carried out using a suitable silane coupling agent. These composites are a modern type of alternative material to conventionally filled polymers. The peculiar properties of tungsten such as the highest melting point, highest tensile strength, and radiation resistance find application especially in the field of radiation shielding. Initially, the tungsten metal powder of 2 μm is treated with suitable silane i.e. 3-Glycidyloxypropyl Tri Methoxy Silane (GPTMS) for improving the wettability of the tungsten metal fillers. Fourier Transform Infrared Spectroscopy (FTIR) and Thermo Gravimetric Analysis (TGA) was carried out to test GPTMS grafting on particles of tungsten. FTIR confirms the grafting of the silane coupling agent on tungsten particles. It also shows the reaction between these agents. TGA reveals the uniform coating of the silane coupling agent on the tungsten particles.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the 2nd International Conference on Materials, Manufacturing, and Machining for Industry 4.0.

1. Introduction

Surface modification is the scientific technique of depositing a thin layer of silane on the surface of the filler material to improve the wettability of the fillers in polymer matrix composites. Improvement in wettability increases the adhesion between the filler material and the polymer matrix in which the fillers are introduced. Surface modification is a rapidly growing sector inside the fields of nanotechnology and production. Surface modification is required to stabilize the particles and to prevent aggregation of particles. Materials may exhibit desired properties but are inappropriate due to their morphology on the surface, ionic conditions, and phobia. The modification aims to create consistency to avoid compatibility issues between two phases, thereby increasing the availability and usability of the properties of materials in their application. Polymers in combination with metal fillers offer cost-effective, high-strength, and lightweight composite materials. Metal particle-reinforced polymer composites constitute a new class of alternative material to traditionally filled polymers and

have some remarkable exceptional properties. The main problem in metals fillers is the agglomeration of metal particles due to the high force of Van der Waals's force existing between them. Agglomerated particles in polymer matrix composites ultimately decrease the mechanical and tensile strength of the composites. In our previous work [1] untreated tungsten particles are introduced into the polymer matrix and achieved an improvement of 10% in mechanical strength. If the metal fillers are treated with a suitable compatibilizer the mechanical strength of the composites could be improved further. The filler and matrix material needed to be in strong adhesion to attain high strength. Therefore, the metal particles are subject to surface alteration to achieve stronger adhesion with the matrix medium. Chemical modification of the filler by the use of coupling agents and subsequent casting by the use of high shear forces produced by homogenizers is a common technique for processing polymer composites. Surface modification can be used to provide improved compatibility of nanoparticles towards dispersing media to avoid convergence of nanoparticles and to make chemically reactive nanoparticles. Coupling agents Silane are important ligands for oxide nanoparticles to act. They are a bifunctional group with features of trialkoxy group and organic head group.

* Corresponding author.

E-mail address: jenson@scmsgroup.org (E. Jenson Joseph).

For the surface treatment of metal particles, the authors have used several techniques with varieties of coupling agent. Xavier et al. [2], GPTMS WO₃ nanoparticles with GPTMS and introduced it into epoxy resin to boost the functional group interactivity of nanoparticles and epoxy resin present in GPTMS on WO₃. The risk of corrosion is significantly reduced in the developed composites. An outstanding barrier property is displayed, also increased mechanical properties were reported due to improved adhesion. Chang et al. [3], modified nano ZnO particles with 3-Aminopropyltriethoxysilane using mechanical stirring and heating. The treated fillers are then introduced into ultra-high molecular weight polyethylene polymers. The developed composites exhibited improved wear resistance properties. Yu et al. [4], surface modified alumina particles with γ -aminopropyl triethoxysilane using chemical processing technique. The treated filters are introduced into the epoxy resin matrix. The developed composites displayed better thermal properties and flexural properties. Rallini et al. [5], treated boron carbide particles with triethylenetetramine using mechanical stirring. The treated particles are then introduced into the epoxy resin matrix. The developed composites exhibit excellent thermal and fire-retardant properties. Tjong et al. [6], treated ZnO particles with Maleated styrene-ethylene butylene-styrene block copolymer and introduced it into polyethylene composites. The resulting composites developed improved electrical properties.

In the case of non-metal particles authors have done many works related to surface property alteration. Owing to the low wettability of non-active filler metal on these materials, the joining of graphite materials is problematic in particular. Chen et al. [7], the magnetron sputtering deposition of Cr film on graphite to alter the graphite surface succeeded in overcoming this problem. Lamastra et al. [8], researched diatomite fillers that can chemically bind to elastomeric molecules during vulcanization, chemically adjusted at 85 °C in H₂O: NaOH: H₂O solution. A technique that does not require a toxic solvent was then used to silanize the modified diatomite with bis(triethoxysilylpropyl) disulfide. Strong interfacial adhesion and fine dispersion were given by the resultant composite. According to Zafar et al. [9], Hydroxyapatite layer between bone and implants made of calcium phosphate (CaP) promotes good contact, thus promoting osseointegration for bonding and enhancing the durability of dental implants, which has been achieved by electrospinning. It was observed that the amount of work performed with tungsten metal particles was inadequate. Hence in this work tungsten metal particles were treated with a suitable coupling agent to modify their surface properties.

In this research, tungsten metal powder is treated with a GPTMS silane coupling agent. The treating method has been discussed. The treated metal fillers are subjected to FTIR analysis and TGA for testing the surface modification properties.

2. Materials and methods

2.1. Materials

Tungsten metal powder is chosen for surface modification in this analysis was supplied by Sigma Aldrich, Bangalore, India. Tungsten has the highest melting point (3422 °C, 6192 °F), lowest vapor pressure (at temperatures above 1650 °C, 3000 °F), and the highest tensile resistance of all metals in pure form. Tungsten has the lowest thermal expansion coefficient on any pure metal. Acetone was used in the initial stage as a cleaning agent before the GPTMS was applied and the final stages before heating. Acetone is an effective cleaning agent in the center of metal particles and can wash away dirt and impurities. GPTMS is a bifunctional silane agent with three methoxy groups on one side, and an epoxy

ring on the other was supplied by Sigma Aldrich, Bangalore. GPTMS is extremely water-resistant and can be used as a connecting agent between the silica surface and the polymeric matrix.

2.2. Surface modification of tungsten particles

Initially, tungsten metal particles of 10 g were placed inside a vacuum chamber with pressure 10^{-3} mbar and a temperature of 140 °C for 1 h. The particles are then cleansed in 75 ml of acetone with 300 rpm magnetic stirrer for 1 h at 25 °C and 60 min of sonication was done. Then the dispersion was supplemented with 5 g of GPTMS and stirred for 24 h using a mechanical stirrer. In the final stage, the acetone was used after centrifuging to wash the excess residue. Then the resulting material was allowed to dry at 60 °C in a vacuum oven for 48 h. FTIR and TGA were carried out on the treated particles to test GPTMS grafting on the tungsten particles.

2.3. Testing and characterization

2.3.1. Fourier Transform Infrared Spectroscopy (FTIR)

In FTIR, the infrared radiations are passed through the treated metal particles. Certain radiations are absorbed by the particles and certain radiations are transmitted through the particles. The resulting spectrum obtained represents the fingerprint of the molecules present in the treated particles. FTIR spectroscopy uses KBr pellet to conduct FTIR spectroscopy. About thirty-five scans were obtained in the spectrum in the 400–4000 cm^{-1} range, with 4 cm^{-1} resolution.

2.3.2. Thermogravimetric analysis (TGA)

The thermal stability of the surface treated and untreated tungsten particles were evaluated by the Thermo Gravimetric Analyser NETZSCH model STA (Germany) 449F3. The percentage reduction in the weight of the sample was found as a function of temperature as per the standard ASTM E1131. Treated and treated metal particle samples of 10 mg were loaded into an aluminum crucible and heated at a rate of 10 °C / min from 25 °C to 600 °C. The resulting thermograms of reduction in weight of the sample as a function of temperature are plotted as a graph.

3. Results and discussion

3.1. Analysis of FTIR spectra

The infrared spectroscopy of GPTMS treated tungsten particles was monitored to verify the presence of GPTMS on tungsten particles. Fig. 1 shows the infrared spectroscopy analysis of tungsten particles treated with GPTMS. The presence of sharp peaks is found at 2940 cm^{-1} , 2840 cm^{-1} and 860 cm^{-1} indicating the reaction bands. Surface modification of tungsten particles by silane is confirmed with the bands present here. Bands present at 2940 and 2840 cm^{-1} are the indication of the presence of the alkyl group that belongs to the silane-modified tungsten particles. The presence of a band at 860 cm^{-1} is the result of the reaction that occurred between the methoxy group of silanes and tungsten particles [10]. Hence it confirms the interaction between the methoxy group of silane and tungsten particles. Thus, it confirms that the tungsten particles are effectively surface treated with a silane coupling agent.

3.2. TGA analysis

The effective concentration of silane can be determined by TGA. Fig. 2 shows the TGA thermograms of untreated and treated tung-

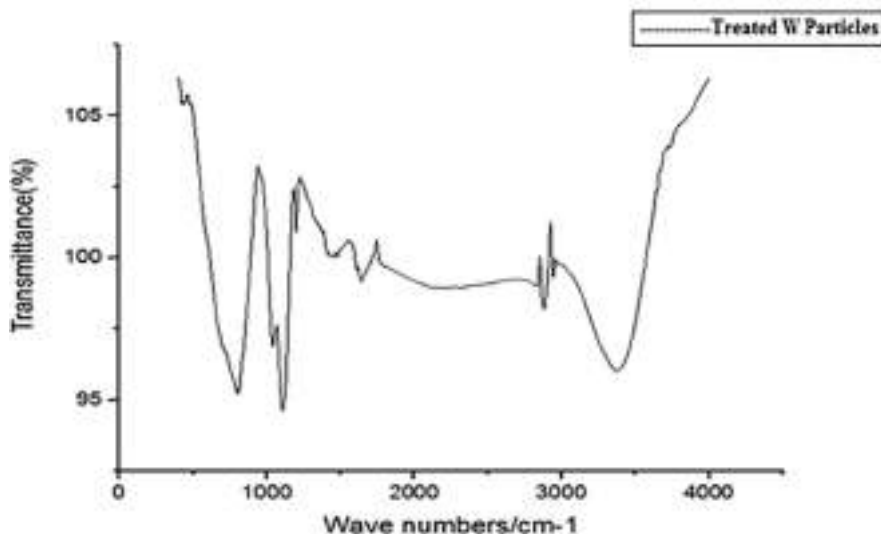


Fig. 1. FTIR Spectra of GPTMS treated W particles.

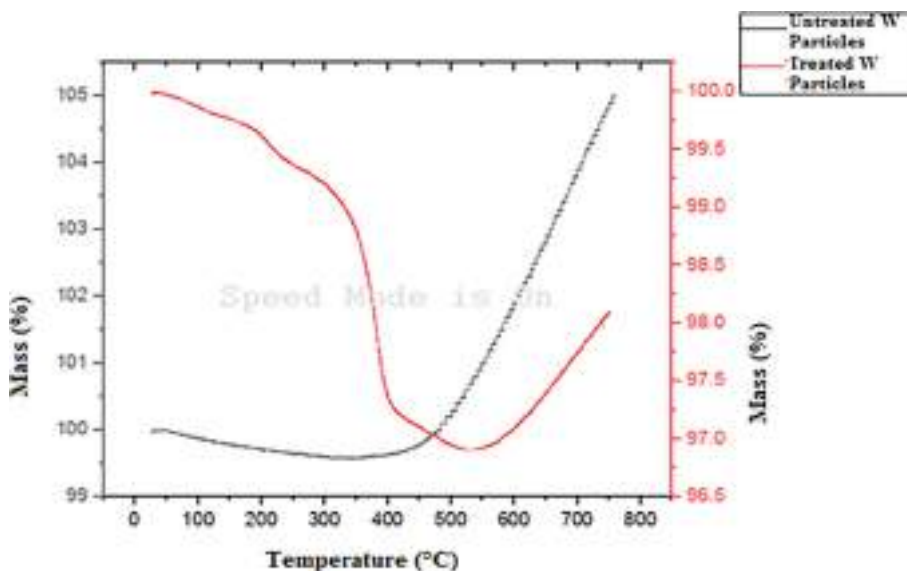


Fig. 2. TGA thermograms of untreated and treated W particles.

sten particles. TGA thermograms of treated tungsten particles exhibit three different weight loss regions. Initial weight loss primarily happens because of the evaporation of moisture content in the particles and this exists at a range of 50 °C and 200 °C. The second region occurs in a range of 200 °C to 400 °C. here the weight loss is very rapid, and it is due to the decomposition of a silane coupling agent that is treated around the tungsten particles. The final region is at 400 °C to 500 °C which shows a very minimal drop in weight percentage and it is due to the removal of burnt gases of the volatile components. The untreated tungsten particles do not show a decrease in weight and it is because tungsten is an extremely high melting point metal and it will decompose at very high temperature and there is no decomposition happening here, there is no decrease in mass.

4. Summary and conclusion

Surface modification of tungsten particles is successfully done by treating them with GPTMS. GPTMS being a silane coupling agent, surface modification of metal fillers with GPTMS improves the wettability of the filler materials. Surface modified tungsten metal particles using silane coupling are investigated by FTIR and TGA. FTIR shows the chemical reaction between the tungsten particles and GPTMS. Thus proving the successful surface modification of tungsten particles with GPTMS. TGA indicates the decomposition of the silane coupling agent and also ensures the presence of remaining tungsten particles. Using a silane coupling agent, surface modification of metal particles is used to provide more outstanding particle compatibility with dispersing media to avoid agglomeration of the particles and to impart chemical reactivity to the particles.

CRediT authorship contribution statement

E. Jenson Joseph: Conceptualization, Methodology, Data curation, Writing - original draft. **V.R. Akshayraj:** Visualization, Investigation, Writing - review & editing. **K. Panneerselvam:** Validation, Supervision, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank the HOD and lab assistants of the Polymer Science department of CUSAT, Kalamssery, Kerala. Dr. Ajith James, Assistant Professor, Department of Chemistry St. Berchmans College, Changanassery, Kerala, for his valuable suggestions and support.

References

- [1] E. Jenson Joseph, K. Panneerselvam, Investigation on the influence of tungsten particulate in mechanical and thermal properties of HD50MA180 high-density polyethylene composites, *Mater. Res. Express* 7 (4) (2020) 045306.
- [2] J. Xavier, Effect of surface modified WO₃ nanoparticle on the epoxy coatings for the adhesive and anticorrosion properties of mild steel, *J. Appl. Polym. Sci.* 137 (5) (2019) 48323.
- [3] B.P. Chang, H.M. Akil, R.B.M. Nasir, Comparative study of micro-and nano-ZnO reinforced UHMWPE composites under dry sliding wear, *Wear* 297 (1–2) (2013) 1120–1127.
- [4] Z.Q. Yu, S.L. You, Z.G. Yang, H. Baier, Effect of surface functional modification of nano-alumina particles on thermal and mechanical properties of epoxy nanocomposites, *Adv. Compos. Mater* 20 (5) (2011) 487–502.
- [5] M. Rallini, M. Natali, J.M. Kenny, L. Torre, Effect of boron carbide nanoparticles on the fire reaction and fire resistance of carbon fiber/epoxy composites, *Polymer* 54 (19) (2013) 5154–5165.
- [6] S.C. Tjong, G.D. Liang, Electrical properties of low-density polyethylene/ZnO nanocomposites, *Mater. Chem. Phys.* 100 (1) (2006) 1–5.
- [7] Z. Chen, H. Bian, S. Hu, X. Song, C. Niu, X. Duan, J. Cao, J. Feng, Surface modification on wetting and vacuum brazing behavior of graphite using AgCu filler metal, *Surf. Coat. Technol.* 348 (2018) 104–110.
- [8] F. Lamastra, S. Mori, V. Cherubini, M. Scarselli, F. Nanni, A new green methodology for surface modification of diatomite filler in elastomers, *Mater. Chem. Phys.* 194 (2017) 253–260.
- [9] M.S. Zafar et al., Bioactive surface coatings for enhancing osseointegration of dental implants, *Biomed. Therapeut. Clin. Applications Bioactive Glasses* (2019) 313–329.
- [10] J.R. Xavier, Effect of surface modified WO₃ nanoparticle on the epoxy coatings for the adhesive and anticorrosion properties of mild steel, *J. Appl. Polym. Sci.* 137 (5) (2020) 48323.

PAPER • OPEN ACCESS

Mechanical and tribological performance of Al-Fe-SiC-Zr hybrid composites produced through powder metallurgy process

To cite this article: G R Raghav *et al* 2021 *Mater. Res. Express* **8** 016533

View the [article online](#) for updates and enhancements.

You may also like

- [Effect of heat treatment on the wear behavior of zircon reinforced aluminium matrix composites](#)
Sandeep Sharma, Suresh Kumar, Tarun Nanda et al.
- [Microstructure characterization and biocompatibility behaviour of TiNbZr alloy fabricated by powder metallurgy](#)
Mehmet Kaya, Fahrettin Yakuphanolu, Ebru Elibol et al.
- [The Effect of Normal Force on Tribocorrosion Behaviour of Ti-10Zr Alloy and Porous TiO₂-ZrO₂ Thin Film Electrochemical Formed](#)
E Dnil and L Benea

Materials Research Express



PAPER

Mechanical and tribological performance of Al-Fe-SiC-Zr hybrid composites produced through powder metallurgy process

OPEN ACCESS

RECEIVED

16 October 2020

REVISED

12 January 2021

ACCEPTED FOR PUBLICATION

13 January 2021


PUBLISHED

22 January 2021

Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](#).

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



G R Raghav¹ , Sheeja Janardhanan², E Sajith², Vidya Chandran² and V Sruthi³

¹ Department of Mechanical Engineering, KLN College of Engineering, Pottapalayam, Sivagangai Dt. Tamil Nadu 630612, India

² Department of Mechanical Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala, India

³ Department of Basic Science and Humanities, SCMS School of Engineering and Technology, Ernakulam, Kerala, India

E-mail: raghavmechklnc@gmail.com

Keywords: powder metallurgy, wear, microhardness, FE-SEM, XRD

Abstract

In this work a ternary Al-Fe-SiC metal matrix composites were reinforced using Zr particles through powder metallurgy process. The Al matrix and the reinforcements were mixed in high energy ball mill at a speed of 250 rpm over a period of 5 h so as to develop a homogenously dispersed composite material. The composite powders are then pressed at 500 MPa using hydraulic press. The compressed composite green compacts are then sintered at 500 °C for 2 h and allowed to cool under furnace atmosphere. The densities, micro hardness and compressive strength of Al-Fe-SiC-Zr composites were investigated and reported. The composite materials were characterized using SEM, EDS and XRD. The density of Al-10Fe-10SiC-10Zr hybrid composites was found to be around 3.44 g cm⁻³. The Zr particles have influenced the micro hardness of the composite materials. The micro hardness of the Al-10Fe-10SiC-10Zr hybrid composites was found to be better compared to Al-10Fe and Al-10Fe-10SiC hybrid composites. The compressive strength of the Al-10Fe-10SiC-10Zr hybrid composites was around 205 MPa which is 44% higher than the Al-10Fe composite material. The porosity of the hybrid composites has reduced when compared to that of Al-10Fe and Al-10Fe-10SiC hybrid composites. The wear studies reveal that Al-10Fe-10SiC-10Zr bear out better wear resistance. The predominant wear mechanism was identified as adhesive wear followed by plastic deformation. This improved wear resistance was due to the formation of oxides layers such Al₂O₃, Fe₂O₃ and also due to the presence of AlFe₃ and Al₃Zr₄ intermetallics.

1. Introduction

The utilization of hybrid composite materials as a replacement of conventional materials has increased drastically in many areas such as aerospace industries, automobile industries and also in various industrial applications where better mechanical, wear and corrosion characteristics are needed [1–7]. Therefore the main objectives of the development of hybrid composites are to develop materials with low density and better strength along with superior wear and corrosion resistance [8–10].

In the development of composite materials it is important to select the matrix materials, reinforcements, percentage of reinforcement and finally the method and production parameters as per the requirements. Now a day due to the economic considerations the industries are opting for low cost materials, in order to overcome high production cost. The most widely used matrix used reinforcements are Al₂O₃, TiO₂, SiC and graphite [11–19]. T Sathish Kumar *et al* investigated the wear behavior of AA6082 alloy reinforced with Y₂O₃ and graphite particles. The studies revealed that the hybrid composites have micro hardness which is 40% higher than that of base alloy [20]. T Sathish kumar *et al* also studied the effect of heat treatment on tribological properties of Al-7Si-ZrSiO₄ hybrid composites manufactured using stir casting processes. The results revealed that the wear resistance of the hybrid composites is much superior to that of base alloy [21].

The other important factor to be considered is the wettability of Al matrix when fabricated using powder metallurgy process. The ceramic reinforcements such as Al₂O₃ and TiO₂ does not easily wetted as a result of

surface oxides on Al matrix. In order to improve the wettability of the Al based composites other reinforcements such as Fe, SiC, and Zr are added [22–27]. The addition of these reinforcements increases the mechanical properties as well as tribological and corrosion resistance properties of the composite material.

Another major area of concern is the uniform dispersion of reinforcements with the matrix materials. Even though there are various methods for fabricating Al based composites such as stir casting method, the major disadvantage was the lack of homogenous dispersion of the reinforcements as the result of agglomeration and cluster formation. The powder metallurgy is one among those methods by which uniform dispersion of reinforcements can be achieved. Moreover the powder metallurgy has been proven to be one of the cost efficient and most reliable methods for fabrication of high melting point materials [5, 6, 8, 26, 27]. There are many literatures based on light weight reinforcements so as to improve the mechanical properties of Al based composite materials; however there are very few studies based on high density hybrid reinforcements so as to improve the mechanical wear and corrosion characteristics of Al based composite materials.

Novelty of this work is to study the effect of Zr reinforcement on the Al-10Fe-10SiC hybrid composites. It obvious, that the addition of ceramic particles Such as SiC will improve the mechanical properties and wear resistant properties of the composites. But there will be some negative effects in terms of increase in porosity and thereby making the composites more brittle in nature compared to the base material. The addition of Zr as reinforcement might improve the ductile nature of the composites by reducing the porosity since Zr particles have a density of 6.49 g cm^{-3} . Further the Zr particles exhibit good mechanical hardness and better wear resistant properties even at high temperatures.

In this work various proportions of Al-Fe-SiC-Zr hybrid nanocomposites were produced using powder metallurgy process. The hybrid composites are then fabricated into 8 mm cylindrical pellets using high speed steel die. The compacted green pellets are then sintered using muffle furnace. The sintered composite pellets are subjected to mechanical characterizations such as density, microhardness and compressive strength. The wear resistance properties were studied using pin on disc apparatus. Thus the main objective of this work is to develop hybrid nanocomposite materials with superior mechanical and tribological properties that can be utilized in automobile, aerospace and other industrial applications.

2. Materials and method

2.1. Materials

The pure aluminum was used as the base material and the Fe, SiC and Zr are used as reinforcements in weight percentage. All the materials used in this research work are of research grade and of purity level 99.5% respectively. The figure 1 shows the Scanning Electron Microscope images of Pure Al, Fe, SiC and Zr. The micrographs were taken in Secondary electron mode operated at 10 kV. The morphology of pure Al resembles a flake like structure with an average particle size of $50 \mu\text{m}$. The Fe powders were elliptical in nature with a particle size of $20 \mu\text{m}$. The SiC and Zr powders were crystalline in nature and their particle size was found to be around $5 \mu\text{m}$ and $3 \mu\text{m}$ respectively.

2.2. Production of hybrid composite materials

The table 1 shows the various proportions of Al-10Fe-10SiC-Zr hybrid nanocomposites. The selected proportions of matrix and reinforcements are then fed into a high energy ball mill consisting of tungsten carbide balls. The ball milling process was carried out for 5 h at a speed of 250 rpm under the presence of toluene as a process control agent so as to obtain homogenous and reaction free hybrid composite materials. The homogeneously mixed composite powders are then compacted using uniaxial hydraulic press at 500 Mpa so as to develop an 8 mm cylindrical green pellet. The green pellets are then sintered at a temperature of $500 \text{ }^\circ\text{C}$ for 2 h and cooled under furnace atmosphere.

2.3. Microhardness and density

The microhardness of the Al-10Fe-10SiC-Zr hybrid composites was carried out using Vickers hardness equipment at a uniform load of 1 kg. The dwell time for the entire process was maintained at 20 s. The results of the experiments represent an average of 10 measurements and the standard deviation values were reported. The density of the composite specimens after sintering process was measured using Archimedes principle and the relative density and porosity of the composite materials were calculated by the relations.

$$\text{Relative Density} = 1 - \text{Porosity} \quad (1)$$

$$\text{Porosity} = \frac{\text{Theoretical Density} - \text{Actual Density}}{\text{Theoretical Density}} \times 100 \quad (2)$$

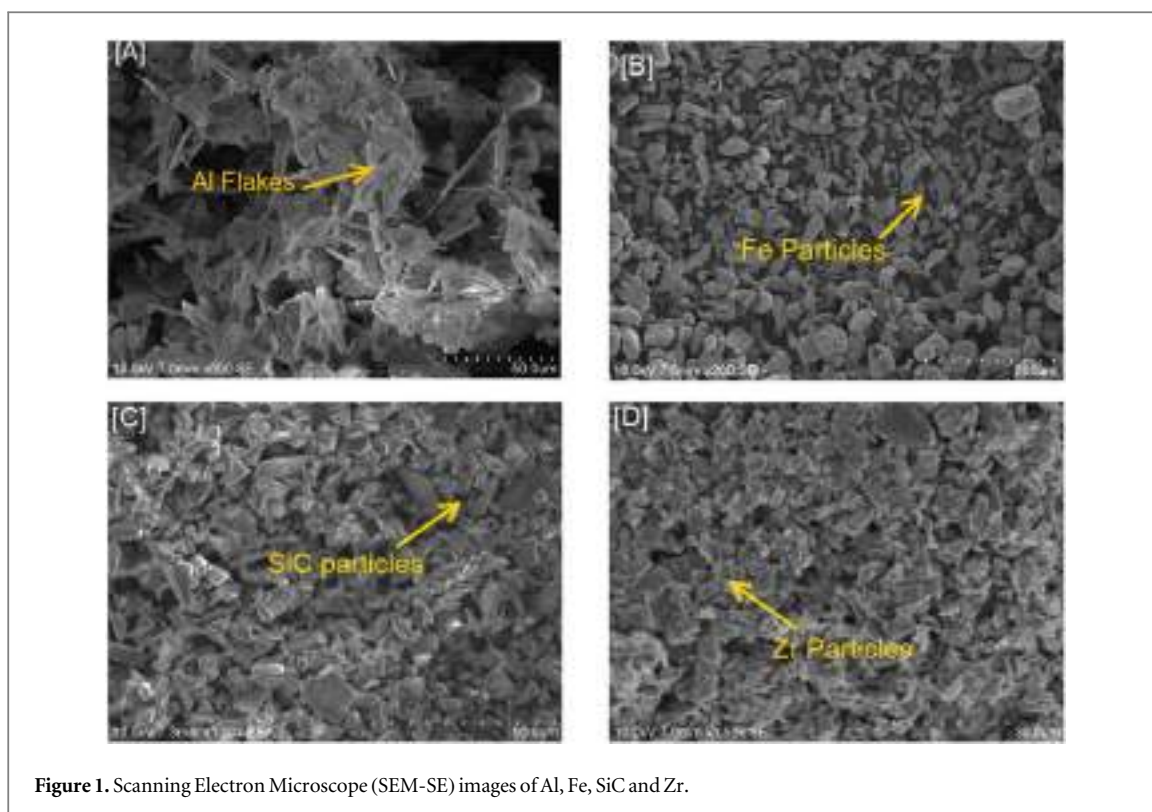


Figure 1. Scanning Electron Microscope (SEM-SE) images of Al, Fe, SiC and Zr.

2.4. Compressive strength

The universal testing machine UTM was utilized to study the compressive strength of Al-10Fe-10SiC-Zr composite materials. The 8 mm diameter composite pellets are compressed at a uniform and gradual speed rate of 5 mm min^{-1} .

2.5. Microstructural characterization

The Scanning Electron Microscope (SEM) was used to explore the microstructures of the Al-10Fe-10SiC-Zr composite materials. The topographical characterization was carried out using Atomic Force Microscope (AFM). The XRD analysis was used to explore the chemical compositions present in the hybrid composites. The EDS analysis was used to confirm the presence of various elements in the hybrid composites.

2.6. Wear analysis

The Al-10Fe-10SiC-Zr hybrid composite pellets of 8 mm diameter and 30 mm long were used as test specimen. The DUCOM make pin on disc apparatus was used to perform wear test as per the ASTM-G99 standard. The wear analysis was carried out for various conditions say applied load, sliding distance and sliding speed. The tests were performed for five different trials for each specimen and the average values are tabulated. The composite wear specimens were weighed before and after the experiments using electronic weighing scale [28, 29].

3. Results and discussion

3.1. Characterization

The figure 2 shows the high resolution Fe-SEM of Al-10Fe-10SiC-Zr hybrid composites of varying Zr content at the magnification of $10,000\times$ at an operating voltage of 10 kV. From the figure it can be understood that the reinforcements are uniformly dispersed into the Al matrix as the result of 5 h milling time. It can be observed that the average particle size of Al powder was also reduced considerably due to ball milling process. The figure 3 represents the EDS mapping of Al-10Fe-10SiC-10Zr hybrid composite powders. From the spectra it can be confirmed that the composite materials has the presence of Al, Fe, SiC and Zr content. Moreover there is also formation of AlFe_3 , Al_3Zr_4 intermetallics and AlFe_3C compound and ZrO_2 which can be inferred from the EDS mapping. The AFM image of Al-10Fe-10SiC-10Zr hybrid composite is shown in figure 4. From the image it can be understood that the there is uniform dispersion of reinforcements with the Al matrix and also it can be noted that there is formation of surface oxides due to the ball milling process. The x-ray diffraction spectra of Al-10Fe-10SiC-5Zr and Al-10Fe-10SiC-10Zr hybrid composites are shown in figure 5. The XRD analysis were carried out

Table 1. Density and Microhardness of Al-10Fe-10SiC- Zr hybrid composites.

S.no	Composition	Composition notation	Actual density (g/cm ³)	Theoretical density (g/cm ³)	Relative density (%)	Porosity (%)	Micro hardness (HV)
1	Al-10Fe	C1	2.98	3.22	92.55	7.45	101
2	Al-10Fe-10SiC	C2	3.05	3.27	93.27	6.73	118
3	Al-10Fe-10SiC-2.5Zr	C3	3.14	3.36	93.45	6.55	120
4	Al-10Fe-10SiC-5Zr	C4	3.23	3.46	93.36	6.64	132
5	Al-10Fe-10SiC-10Zr	C5	3.44	3.65	94.25	5.75	135



Figure 2. High Resolution Field Emission Scanning Electron Microscope (FE-SEM) image of Al-10Fe- 10SiC- 10Zr hybrid composites.

using Xpert-3 diffractometer (45 kV, 30 mA) with Cu anode ($\lambda = 0.15406$ nm). The XRD spectra exhibit the characteristic peaks of Al, Fe, SiC and Zr which confirms the uniform dispersion of reinforcements in Al matrix. The peaks at $2\theta = 39.5^\circ, 44.25^\circ, 65^\circ, 77.25^\circ$ and 82.3° confirm the presence of Al in composite materials as per the JCPDS No: 34-0529, 06-0696. The characteristic low intensity 2θ peaks at 37.5° and 82.3° corresponds to SiC which authenticates its presence in the composite materials (JCPDS No: 42-1172). The peaks at $2\theta = 44.25^\circ, 65^\circ$ and 82.3° also prove the presence of Fe particles in the composite materials (JCPDS No: 45-1203). The 2θ peaks at $77.25^\circ, 14.5^\circ, 35.87^\circ, 60.14^\circ$ are the characteristics peaks of Zr (JCPDS No: 41-0814). The formation of AlFe_3 ,

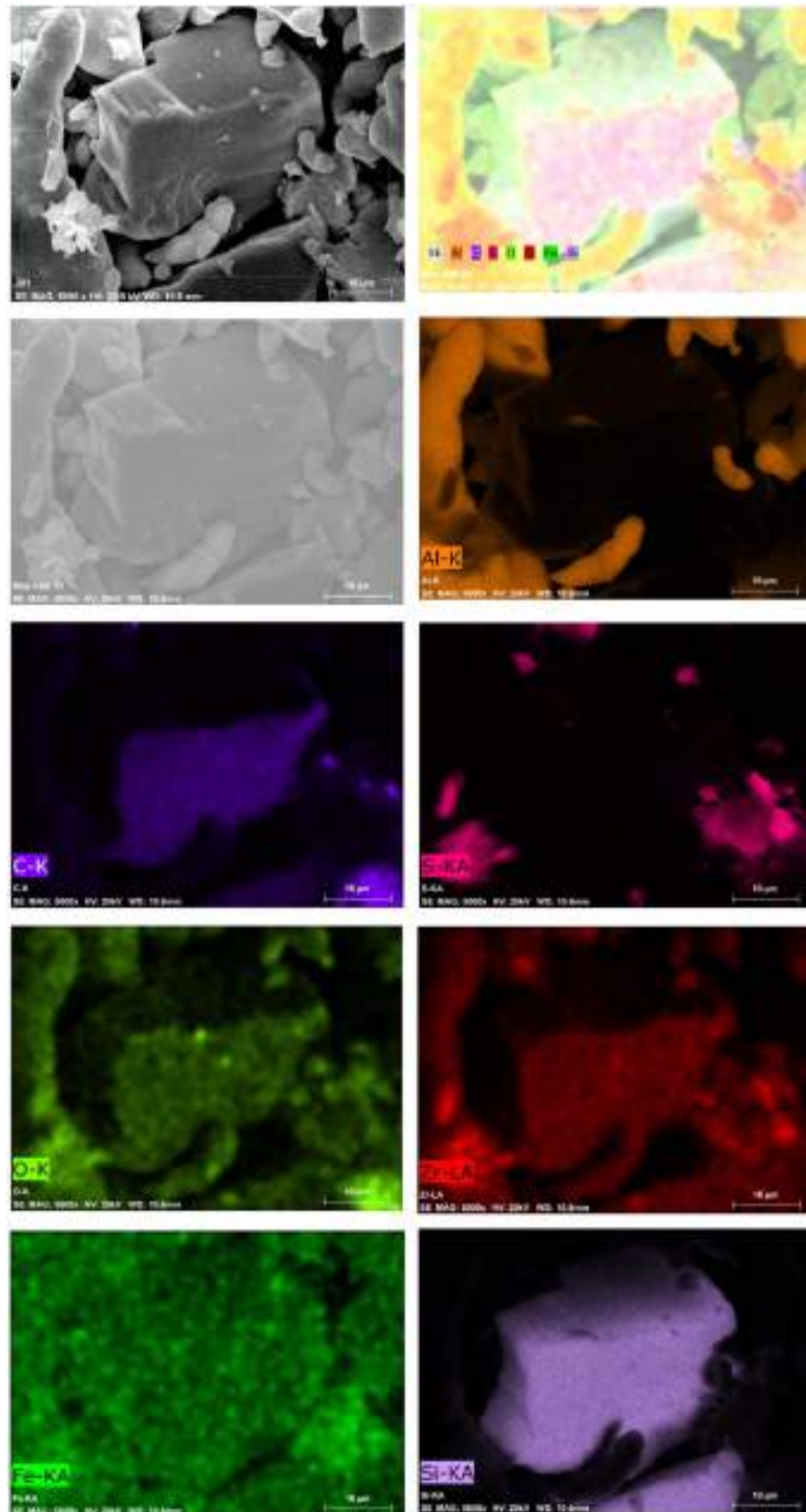


Figure 3. Field Emission Scanning Electron Microscope (FE-SEM) mapping of Al-10Fe-10SiC-10Zr hybrid composites.

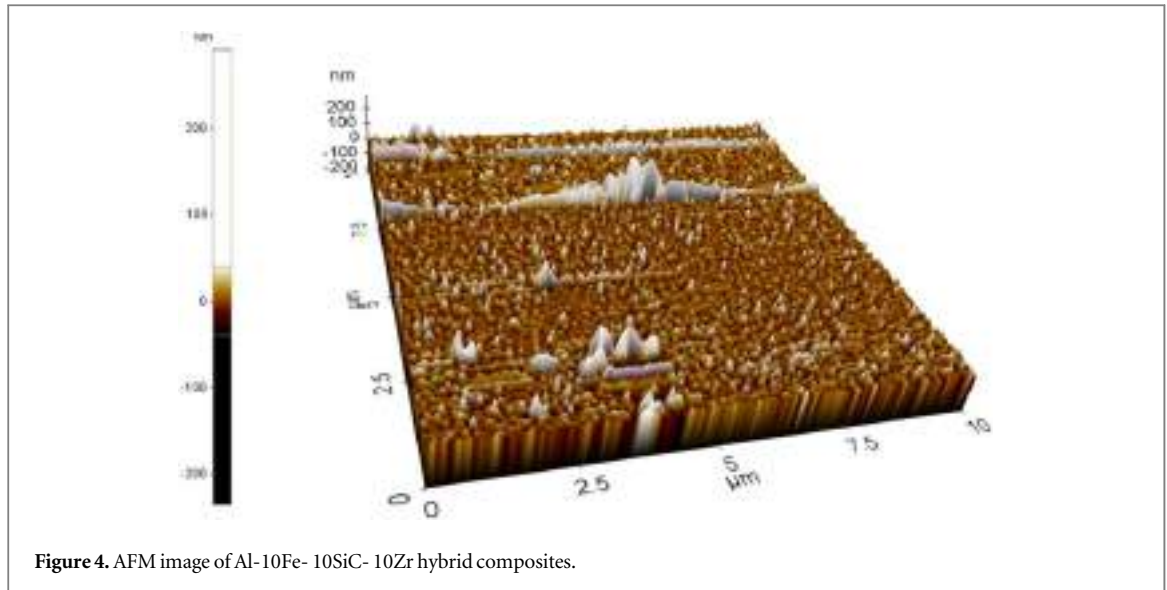


Figure 4. AFM image of Al-10Fe-10SiC-10Zr hybrid composites.

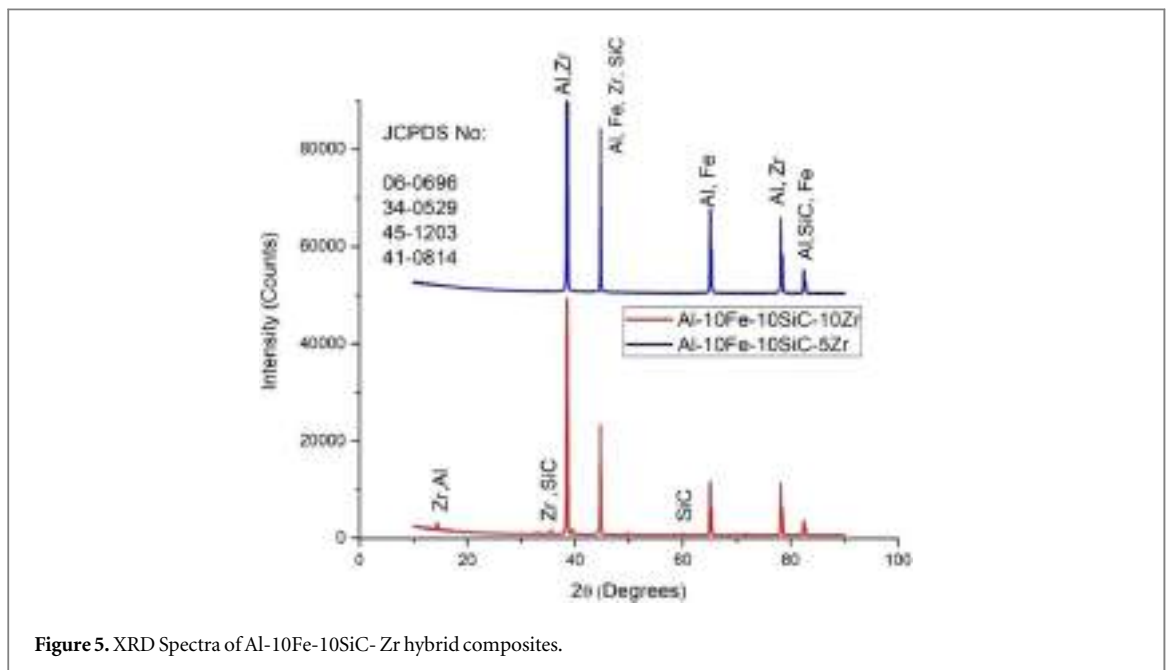


Figure 5. XRD Spectra of Al-10Fe-10SiC-Zr hybrid composites.

Al_3Zr_4 intermetallics and AlFe_3C compound were observed from the XRD analysis (JCPDS No: 45-1203, 41-0814).

3.2. Density and micro hardness

The density, Relative density, porosity and microhardness of the Al-10Fe-10SiC- Zr hybrid composites are represented in table 1. The relationship between density and porosity of Al-10Fe, Al-10Fe-10SiC and Al-10Fe-10SiC- Zr hybrid composites are shown in figure 6. The density of the Al-Fe-SiC ternary composites has improved with the addition of Zr reinforcements. The density of the Al-10Fe composites was found to be 2.98 g cm^{-3} whereas; the density of the Al-10Fe-10SiC-10Zr hybrid composites has increased to 3.44 g cm^{-3} . The porosity of the composite materials decreased with increase in Zr addition. The Al-10Fe-10SiC-10Zr hybrid composites have better density and porosity compared to other combinations. The reason behind this decrease in porosity is due the high density Zr reinforcements and also due the compaction pressure. It was also found that theoretical density of the composites was higher than the actual density of all compositions. The relative density percentage of the Al-10Fe-10SiC-10Zr composites was 94.25% and has increased 1.8% when compared with Al-10Fe composites. The microhardness of the Al-10Fe-10SiC-Zr hybrid composites has increased slightly with the increases in Zr addition. The improvement in microhardness was also due to the reduction in porosity of the composite pellets and also due to the formation of AlFe_3 , Al_3Zr_4 intermetallics.

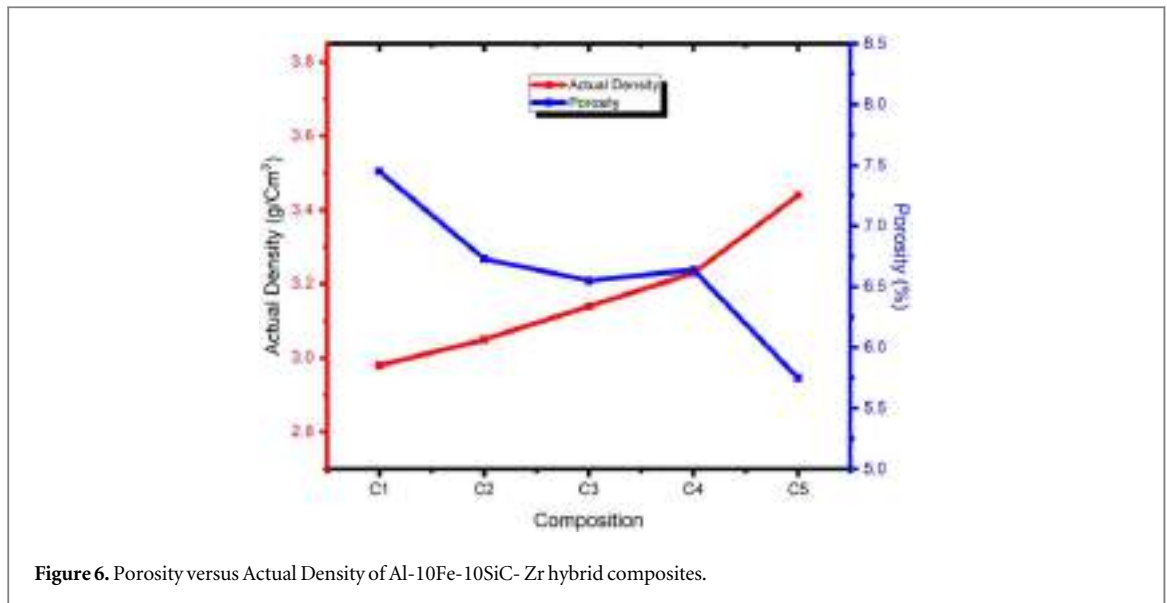


Figure 6. Porosity versus Actual Density of Al-10Fe-10SiC-Zr hybrid composites.

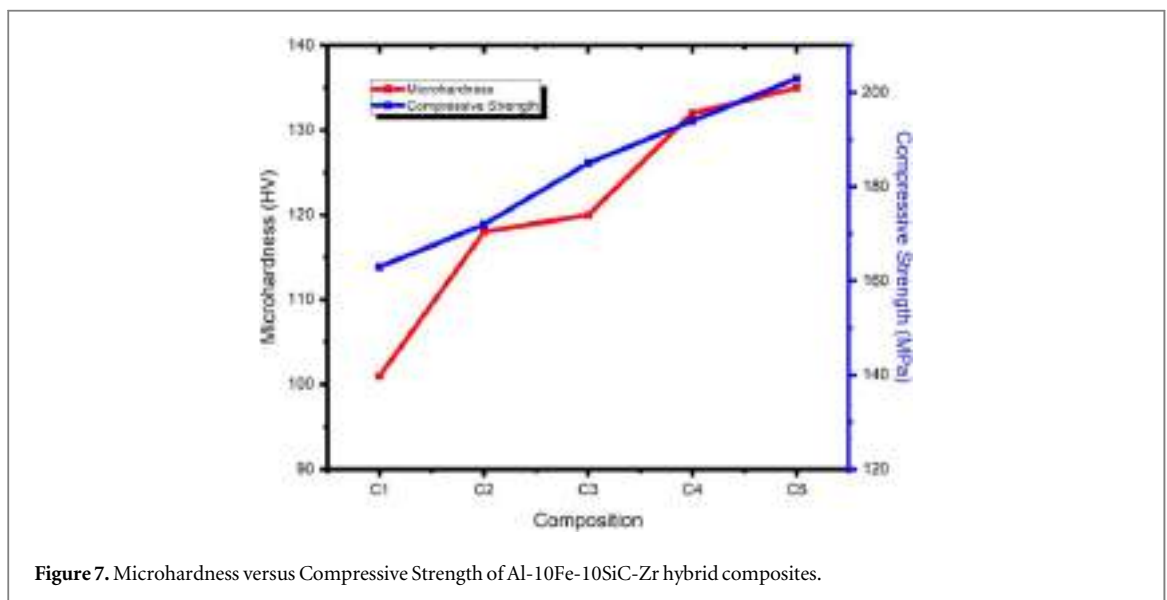


Figure 7. Microhardness versus Compressive Strength of Al-10Fe-10SiC-Zr hybrid composites.

3.3. Compressive strength

The compressive strength of various Al-10Fe-10SiC-Zr hybrid composites is shown in figure 7. The Compressive strength of the Al-10Fe-10SiC-Zr hybrid composites shows betterment with the increase in load bearing Zr reinforcement [30]. The other major reason for the improvement in compressive strength was due to formation of hard $AlFe_3$ intermetallics and oxides as the result of sintering operation. The presence of SiC particles in the composite materials also played a major role in the improvement of compressive strength.

3.4. Wear analysis

The primary concern for any light weight material is long service life and less replacement period, thereby reducing the total expenditure incurred. Hence it is desirable to develop a material which has very less wear loss under sliding wear conditions. The effect of Zr reinforcement on the wear loss of Al-10Fe-10SiC ternary composites is shown in figure 8. From the figure 8(a) it can be understood that the increase in applied load has resulted in increased wear loss irrespective of Zr reinforcement. This phenomenon was due to the increased contact surface between the specimen and rotating disc. Figure 8(b) reveals that the wear loss of the composite materials increases as the distance of sliding increases. The increase in sliding distance increases the contact period of the composite materials with the mating surface thereby increasing the temperature at the interface. The increase in surface temperature further results in softening of materials and the deformation of materials takes place. From the figure 8(c) it is clear that the increase in temperature at the interface due to the increase in Sliding speed has resulted in softening of the composite pellet there by increasing the rate of wear loss. It can be

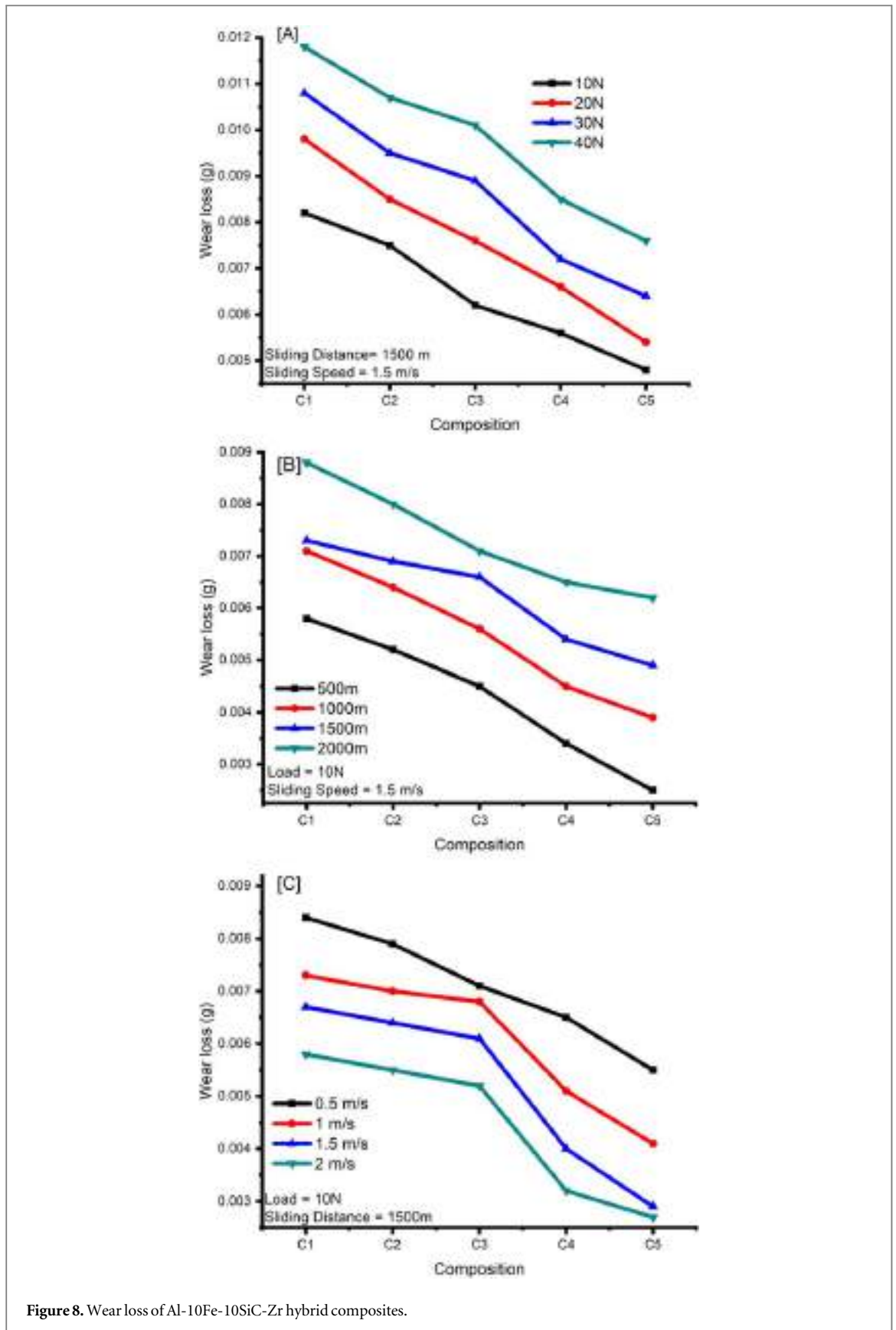


Figure 8. Wear loss of Al-10Fe-10SiC-Zr hybrid composites.

also noted that the wear loss of Al-10Fe and Al-10Fe-10SiC hybrid composites are higher than the Al-10Fe-10SiC-10Zr hybrid composites under all sliding wear conditions. The Coefficient of friction analysis of various Al-10Fe-10SiC-Zr hybrid composites is shown in figure 9. The addition of Zr reinforcements has resulted in reducing the COF values Al-10Fe-10SiC ternary composites. The coefficient of friction of Al-10Fe-10SiC-10Zr composites has improved compared to that of Al-10Fe-10SiC-5Zr, Al-10Fe-10SiC-2.5Zr hybrid composites as

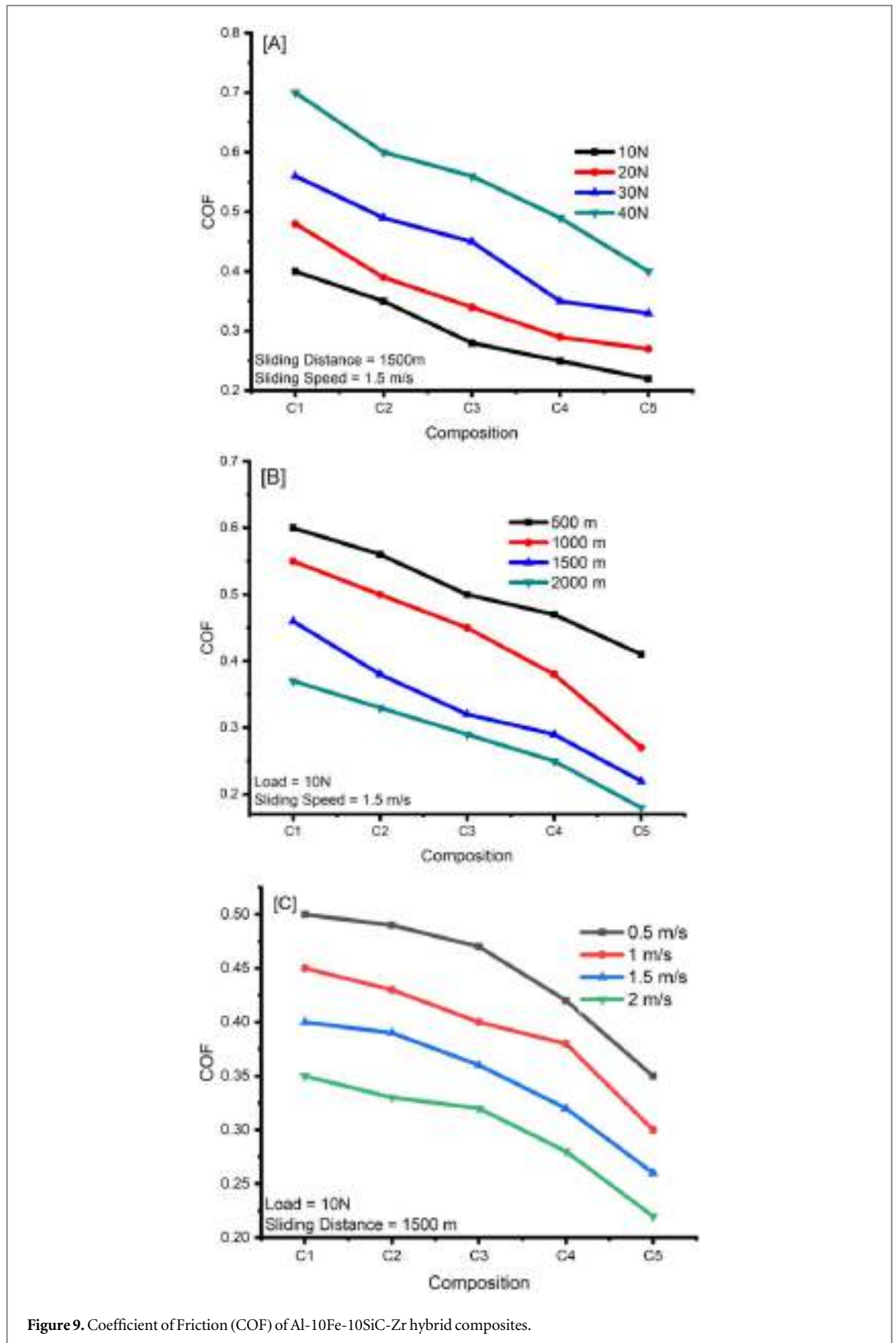


Figure 9. Coefficient of Friction (COF) of Al-10Fe-10SiC-Zr hybrid composites.

well as Al-10Fe and Al-10Fe-10SiC composite materials. This improvement in wear and friction characteristics of the -10Fe-10SiC ternary composites was due to the formation of AlFe_3 and Al_3Zr_4 intermetallics which improved the density and surface hardness of the composite materials. The other major reason was the formation of AlFe_3C compound which increases the hardness and self lubricating property of the composite

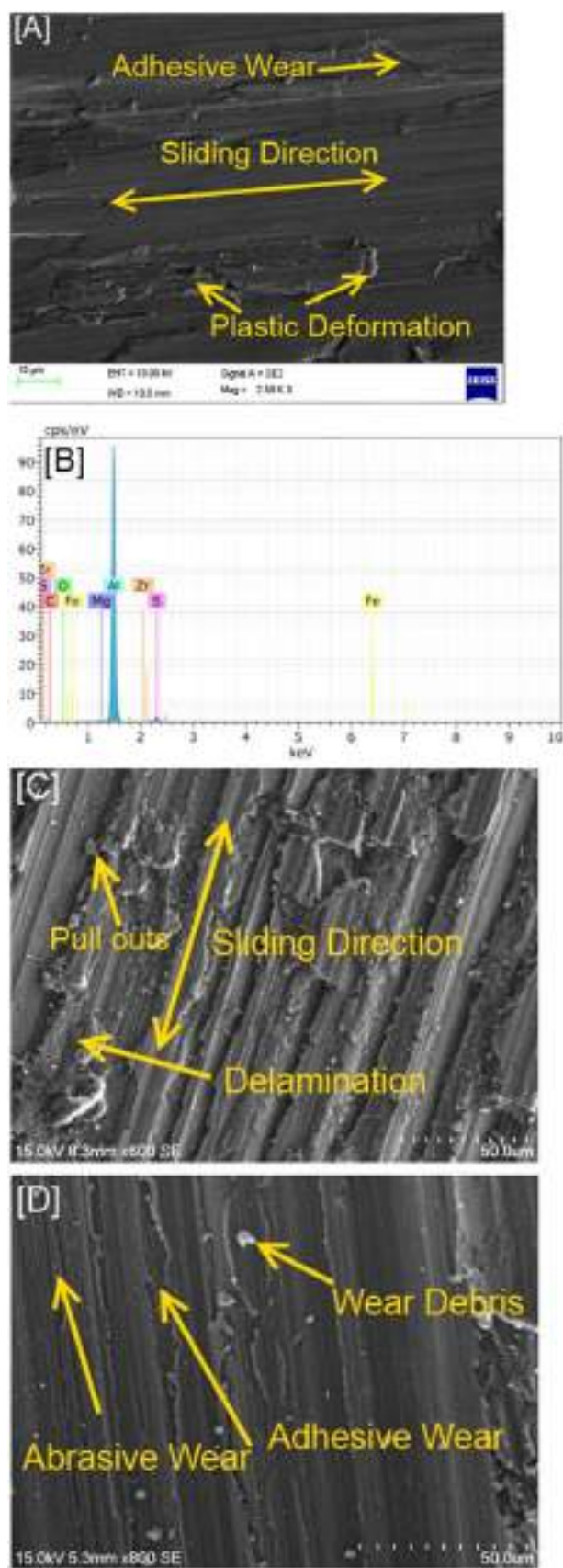


Figure 10. [A] FESEM image of Al-10Fe-10SiC-10Zr worn out Sample, [B] EDAX Spectra of Al-10Fe-10SiC-10Zr worn out Sample [C] SEM image of Al-10Fe worn out sample [D] SEM image of Al-10Fe-10SiC worn out Sample.

materials. Further there is also formation of ZrO_2 , Al_2O_3 and Fe_2O_3 tribo layers which also played a vital role in improving the sliding wear properties of Al-10Fe-10SiC-Zr nanocomposites. The figure 10 exhibits the high resolution FESEM image and EDAX spectra of Al-10Fe-10SiC-10Zr hybrid composites worn out surface after wear analysis. From the FESEM it is evident that the main wear mechanism was adhesive wear with micro cracking which leads to plastic deformation. Figure 10(B) represents the EDAX spectra of Al-10Fe-10SiC-10Zr hybrid composites after wear test, which confirms the presence of Al, Fe, SiC and Zr along with the presence of oxides such as ZrO_2 , Al_2O_3 and Fe_2O_3 at contact surface. Figures 10(C) & (D) shows the SEM images of worn out surfaces of Al-10Fe composites and Al-10Fe-10SiC hybrid composites after wear test. From the spectra's it can be confirmed that the Al-10Fe composites has experienced abrasive wear along with delamination. Whereas the Al-10Fe-10SiC hybrid composites experiences abrasive wear followed by adhesive wear which leads to plastic deformation [31, 32].

4. Conclusions

The Al-10Fe-10SiC-Zr hybrid composites were produced through mechanical alloying process. The mechanical, tribological and corrosion resistance properties of the composites were studied at different conditions.

- The density of the Al-10Fe-10SiC-10Zr hybrid composites has improved to 3.44 g cm^{-3} from 3.14 g cm^{-3} for Al-10Fe-10SiC-2.5Zr hybrid composites.
- The Microhardness of the Al-10Fe-10SiC-10Zr (135 HV) hybrid composites is better to that of Al-10Fe-10SiC-2.5Zr (120 HV) hybrid composites due to the formation of $AlFe_3C$ compound.
- The porosity of the Al-10Fe-10SiC-10Zr (5.75%) hybrid composites has reduced compared to that of Al-10Fe-10SiC-2.5Zr (6.55%) hybrid composites.
- The compressive strength of the the Al-10Fe-10SiC-10Zr hybrid composites has found to be better compared to other combinations.
- The wear resistance and coefficient of friction of the Al-10Fe-10SiC-10Zr hybrid composites has improved significantly compared to other combinations due to the formation of Al_3Zr_4 , $AlFe_3$ intermetallics.
- From the findings of this study, it can be concluded that the Al-10Fe-10SiC-10Zr hybrid composites has better mechanical and tribological properties.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ORCID iDs

G R Raghav  <https://orcid.org/0000-0001-6028-3979>

References

- [1] Mousavi R, Bahrololoom M E and Deflorian F 2016 Preparation, corrosion, and wear resistance of Ni-Mo/Al composite coating reinforced with Al particles *Mater. Des.* **110** 456–65
- [2] Satish J and Satish K G 2018 Preparation of magnesium metal matrix composites by powder metallurgy process *IOP Conf. Ser.: Mater. Sci. Eng.* **310** 012130
- [3] Dyachkova L N, Feldshtein E E, Vityaz P A, Bloch B M and Voronetskaya L Y 2018 Effect of copper content on tribological characteristics of Fe – C – Cu composites *Journal of Friction and Wear* **39** 1–5
- [4] Ajith Kumar K K, Pillai U T S, Pai B C and Chakraborty M 2013 Dry sliding wear behaviour of Mg-Si alloys *Wear* **303** 56–64
- [5] Prasad R V, Jeyasimman D, Parande G, Gupta M and Narayanasamy R 2018 Investigation on dry sliding wear behavior of Mg/BN nanocomposites *Journal of Magnesium and Alloys.* **6** 263–76

- [6] Selvakumar N and Narayanasamy R 2005 Deformation behavior of cold upset forming of sintered Al-Fe composite preforms *J. Eng. Mater. Technol.* **127** 251
- [7] Prakash C, Singh S, Verma K, Sidhu S S and Singh S 2018 Synthesis and characterization of Mg-Zn-Mn-HA composite by spark plasma sintering process for orthopedic applications *Vacuum* **155** 578–84
- [8] Selvakumar N and Vettivel S C 2013 Thermal, electrical and wear behavior of sintered Cu-W nanocomposite *Mater. Des.* **46** 16–25
- [9] Sozhamannan G G, Yusuf M M, Aravind G and Kumaresan G 2018 ScienceDirect effect of applied load on the wear performance of 6061 Al/Nano Ticp/Gr hybrid composites *Materials Today: Proceedings.* **5** 6489–96
- [10] Tong L B, Zhang Q X, Jiang Z H, Zhang J B, Meng J, Cheng L R and Zhang H J 2016 Microstructures, mechanical properties and corrosion resistances of extruded Mg-Zn-Ca-xCe/La alloys *J. Mech. Behav. Biomed. Mater.* **62** 57–70
- [11] Akbarpour M R and Pouresmaeil A 2018 The influence of CNTs on the microstructure and strength of Al-CNT composites produced by flake powder metallurgy and hot pressing method *Diamond & Related Materials.* **88** 6–11
- [12] Zhao X, An Y, Chen J, Zhou H and Yin B 2008 Properties of Al₂O₃-40 wt.% ZrO₂ composite coatings from ultra-fine feedstocks by atmospheric plasma spraying *Wear* **265** 1642–8
- [13] Allaoui A, Bai S, Cheng H M and Bai J B 2002 Mechanical and electrical properties of a MWNT/epoxy composite *Composite Science and Technology* **62** 1993–8
- [14] Sharma P, Khanduja D and Sharma S 2015 Dry sliding wear investigation of Al6082/Gr metal matrix composites by response surface *Integrative Medicine Research.* **5** 29–36
- [15] Narayanasamy P and Selvakumar N 2017 Tensile, compressive and wear behaviour of self-lubricating sintered magnesium based composites *Transactions of Nonferrous Metals Society of China.* **27** 312–23
- [16] Reza S, Golroh S and Mohammadalipour M 2011 Properties of Al₂O₃ nano-particle reinforced copper matrix composite coatings prepared by pulse and direct current electroplating *Mater. Des.* **32** 4478–84
- [17] Goh C S, Wei J, Lee L C and Gupta M 2006 Development of novel carbon nanotube reinforced magnesium nanocomposites using the powder metallurgy technique *Nanotechnology* **17** 7–12
- [18] Raghav G R, Selvakumar N, Jayasubramanian K and Thansekhar M R 2014 Corrosion analysis of copper -TiO₂ nanocomposite coatings on steel using sputtering *International Journal of Innovative Research in Science, Engineering and Technology* **3** 1105–10
- [19] Ashok R, Raghav G R, Nagarajan K J, Rengarajan S, Suganthi P and Vignesh V 2019 Effect of hybrid reinforcement at stirred zone of dissimilar aluminium alloys during friction stir welding *Metall. Res. Technol* **116** 631
- [20] Satish Kumar T, Shalini S and Krishna Kumar K 2020 Effect of friction stir processing and hybrid reinforcement on wear behaviour of AA6082 alloy composite *Mater. Res. Express* **7** 026507
- [21] Kumar T S, Shalini S and Priyadharshini G S 2020 Effect of T6 treatment on wear behavior of Al-7Si/ZrSiO₄ composites *Silicon* **12**
- [22] Xu W, Lu X, Tian J, Huang C, Chen M, Yan Y, Wang L, Qu X and Wen C 2019 Microstructure, wear resistance, and corrosion performance of Ti₃₅Zr₂₈Nb alloy fabricated by powder metallurgy for orthopedic applications *J. Mater. Sci. Technol.* **41** 191–8
- [23] Chand N, Krishna V, Das M and Kumar A 2018 wear and corrosion properties of *in situ* grown zirconium nitride layers for implant applications *Surface & Coatings Technology.* **334** 357–64
- [24] Soorya Prakash K, Balasundar P, Nagaraja S, Gopal P M and Kavimani V 2016 Mechanical and wear behaviour of Mg-SiC-Gr hybrid composites *Journal of Magnesium and Alloys.* **4** 197–206
- [25] Marques F P, Scandian C, Bozzi A C, Fukumasu N K and Tschiptschin A P 2017 Formation of a nanocrystalline recrystallized layer during microabrasive wear of a cobalt-chromium based alloy (Co-30Cr-19Fe) *Tribol. Int.* **116** 105–12
- [26] Osório W R, Peixoto L C, Goulart P R and Garcia A 2010 Electrochemical corrosion parameters of as-cast Al-Fe alloys in a NaCl solution *Corros. Sci.* **52** 2979–93
- [27] Kabir M S, Minhaj T I, Hossain M D and Kurny A 2015 Effect of Mg on the wear behaviour of as-cast Al-4.5Cu-3.4Fe *in situ* composite *American Journal of Materials Engineering and Technology* **3** 7–12 www.sciepub.com
- [28] Raghav G R, Balaji A N, Selvakumar N, Muthukrishnan D and Sajith E 2019 Effect of tungsten reinforcement on mechanical, tribological and corrosion behaviour of mechanically alloyed Co-25C Cermet nanocomposites *Mater. Res. Express* **6** 1165e4
- [29] Muthukrishnan D, Balaji A N and Raghav G R 2018 Effect of Nano-TiO₂ particles on wear and corrosion behaviour of AA6063 surface composite fabricated by friction stir processing *Metallofiz. Noveishie Tekhnol* **409** 397–409
- [30] Satish Kumar T, Shalini S, Kumar K K, Thavamani R and Subramanian R 2018 Bagasse Ash reinforced A356 alloy composite: synthesis and characterization *Materials Today: Proceedings.* **5** 7123–30
- [31] Raghav G R, Balaji A N, Muthukrishnan D and Sruthi V 2018 Preparation of Co-Gr nanocomposites and analysis of their tribological and corrosion characteristics *Metallofiz. Noveishie Tekhnol* **40** 979–92
- [32] Raghav G R, Balaji A N, Muthukrishnan E S D and Sruthi V 2018 An experimental investigation on wear and corrosion characteristics of Mg-Co nanocomposites *Mater. Res. Express* **5** 257–69

Influence of multiwalled carbon nanotubes on the structure and properties of poly(ethylene-co-vinyl acetate-co-carbon monoxide) nanocomposites

Gibin George¹ | Arunjunairaj Mahendran² | Selvakumar Murugesan³ | S. Anandhan³

¹Department of Mechanical Engineering, SCMS School of Engineering and Technology, Ernakulam, Kerala, India

²Kompetenzzentrum Holz GmbH, W3C, Linz, Austria

³Department of Metallurgical and Materials Engineering, National Institute of Technology Karnataka, Mangalore, Karnataka, India

Correspondence

Gibin George, Department of Mechanical Engineering, SCMS School of Engineering and Technology, Karukutty, Ernakulam, Kerala 683576, India.

Email: gg-gibingeorge@scmsgroup.org

Srinivasan Anandhan, Department of Metallurgical and Materials Engineering, National Institute of Technology Karnataka, Srinivas Nagar, Mangalore 575025, Karnataka, India. Email: sa-anandmtg@gmail.com

Abstract

In this work, composites of poly(ethylene-co-vinyl acetate-co-carbon monoxide) (EVACO)/surface-modified multiwalled carbon nanotubes (m-MWCNTs) were prepared using a solution casting technique. Acid treatment was employed for the surface modification of MWCNTs to improve the compatibility between polar EVACO and MWCNTs. The influences of m-MWCNTs on the crystalline, mechanical, thermal, and electrical properties of EVACO at very low filler loading were systematically evaluated. The presence of m-MWCNTs in the EVACO matrix influenced the crystallinity, and the respective changes were determined and quantified using dynamic scanning calorimetry and X-ray diffraction. The mechanical properties of the composites were improved remarkably by the addition of a minute quantity (0.05, 0.1, 0.15, 0.2, and 0.25 wt%) of m-MWCNTs. Additionally, m-MWCNTs in the EVACO matrix improved the thermal stability and electrical properties of EVACO. However, the filler loading is below the threshold loading of the fillers, and there was no drastic improvement in the electrical conductivity of the composite.

KEYWORDS

conductivity, crystallinity, MWCNTs, nanocomposites

1 | INTRODUCTION

Polymer nanocomposites are used in a variety of applications starting from common household to biomedical transplants and space missions. Inorganic nanofillers are the most commonly used fillers in polymer matrices, several unique properties of these fillers can never be reached by organic materials. In many instances, nanofillers exhibit some unique and exceptional properties several orders in magnitude than polymers, and polymers have certain unique properties that cannot be matched by any other materials. In polymer nanocomposites, the properties of

the polymers and nanofillers are compromised and they exhibit superior properties as compared to the virgin polymers in many aspects due to the synergistic action of the nanofiller and the polymer matrix. Due to the high surface area of the nanofillers, a small quantity of the filler is sufficient to make a significant impact on the properties of the polymer matrix alone.

The different nanosized allotropes of carbon as fillers in polymer matrix composites have attracted extensive interest owing to their lightweight, strength, conductivity, and so on. The allotropes of carbon that are commonly used as fillers in polymer composites are carbon

nanotubes (CNTs),^[1–6] graphite,^[7–10] graphene,^[11–13] fullerene,^[14–16] carbon black,^[17–20] and so on, and the resulting nanocomposites can be potentially used in a myriad of applications. In general, the addition of any aforementioned carbon allotropes above the percolation threshold enhances the conductivity of the polymer composites tremendously. The application of carbon nanomaterials as nanofillers in composites is limited to not only polymers but also ceramics^[21–23] and metals.^[24–26] With the introduction of nanofillers in polymer composites, the conventional applications of polymers are widened.

Among the abovementioned carbon-based nanofillers, CNTs, both single and multi-walled, have their own identity starting from their morphology and structure to the properties. The tensile strength of the CNT-filled composites is expected to be higher than the other carbon allotrope-filled composites since CNTs with a high aspect ratio have more entanglements as compared with the latter,^[27] CNTs also exhibit a tensile modulus higher than stainless steel.^[28] The ability of SWCNTs/MWCNTs as nucleating agents to improve the crystallinity in several semi-crystalline polymer matrices has been proven,^[29–33] this increase in crystallinity will also contribute to the enhancement in tensile strength in polymer composites. Similar to carbon black-filled polymers for exterior applications, CNT-filled polymer composites are also resistant to weathering.^[34]

The interfacial bonding of filler and matrix is important in dictating the properties of any polymer composite. Good interfacial interaction is possible by either modifying the filler or matrix of the composite, and the addition of a compatibilizer is an alternative solution. Modifying the polymer is stringent and requires a lot of effort starting from the selection of reagents to the modification of the reaction vessels. Modifying the filler is easier than the modification of polymer and it is a must if there is a large difference in the polarity of the polymer and the filler. The addition of a compatibilizer can have a detrimental effect on the properties of the composite, especially when conducting fillers like CNTs are used, which is capable of improving the electrical properties on the matrix. The filler modification is important irrespective of the composite fabrication routes, such as in situ polymerization, melt blending, solution casting, and so on. Surface modification of CNTs is essential before it is mixed with the organic matrices since pristine CNTs exist as bundles due to their inertness.^[35] These bundles can lead to anomalous properties of the composites, for instance, stress concentration due to these bundles can lead to early failure of the composite under loading.

Poly(ethylene-*co*-vinyl acetate-*co*-carbon monoxide (EVACO) is developed to improve the polarity of poly

(ethylene-*co*-vinyl acetate) (EVA). Polarity in EVA is difficult to enhance just by increasing the vinyl acetate content since excess vinyl acetate can adversely affect the properties of the polymer.^[36] The addition of carbon monoxide to the backbone of EVA increases the polarity of the polymer, thereby improving its adhesion to polar surfaces^[37]; therefore, it is also used as an adhesion booster in coatings. EVACO is semicrystalline and the polyethylene phase imparts crystallization in it.

In this study, EVACO/modified-MWCNTs (m-MWCNTs) composite was prepared through solution casting. Industrial processing of EVACO is mainly in the form of solutions and the method used here is akin to the bulk processing of EVACO. The modification of MWCNTs with polar functional groups by reduction can improve the miscibility of MWCNTs in the polar EVACO matrix. The mechanical properties, electrical conductivity, and crystallizability of EVACO can be improved by the addition of m-MWCNTs in small quantities. The overall improvement in the properties of the composite is attributed to the good interaction of m-MWCNTs with EVACO.

The applications of EVACO are promising as a non-migrating plasticizer in polyvinyl chloride for medical applications and as an adhesion promoter in paints and coating. EVACO can form very thin uniform layers on metallic surfaces due to its high polar nature. Additionally, unlike several other polymers, EVACO exhibits unique properties such as low-temperature impact strength and resistance to environmental degradation. Therefore, by forming EVACO/m-MWCNT nanocomposites, one can achieve better performance in terms of strength and weather resistance in the respective applications.

2 | MATERIALS AND METHODS

EVACO (Elvaloy[®] 4924) provided by Du Pont, USA, MWCNTs (product ID: 677248, purity: >90%,) with 5–15 walls (outer diameter 10–15 nm, inner diameter 2–6 nm, and length 0.1–10 μm and) obtained from Sigma Aldrich Inc., USA, and dichloromethane (DCM) of purity >99% procured from Central Drug House Pvt. Ltd., New Delhi, India, were used for the preparation of composite. Potassium dichromate obtained from Sulab, Baroda, India, and sulfuric acid purchased from Nice chemicals, Cochin, India, was used for the surface modification of MWCNTs.

To modify MWCNTs, 50 mg of MWCNTs was added to 10 N sulfuric acid in which 0.2 g of potassium dichromate was dissolved.^[38] The mixture was ultrasonicated for 1 h and heated for 30 min at 80°C in a constant temperature bath. The mixture was washed repeatedly in distilled water until the pH was neutral. The m-MWCNTs



FIGURE 1 Photographs of unmodified and modified multiwalled carbon nanotubes dispersed in water after 7 days [Color figure can be viewed at wileyonlinelibrary.com]

were centrifuged and dried in a vacuum oven at 100°C. The pristine MWCNTs and m-MWCNTs were dispersed in water by sonication of half an hour and the photographs taken after 7 days are shown in Figure 1. The m-MWCNTs were dispersed well in water even after 1 week, whereas the pristine MWCNTs were settled after 2 h.

The EVACO/m-MWCNT composite was prepared by solution casting a mixture of EVACO and m-MWCNTs in DCM. This mixture was prepared by dissolving 4 g of EVACO in 75 ml of DCM by continuous stirring at room temperature followed by the addition of m-MWCNT dispersion in DCM, in which known quantities of m-MWCNTs were taken. The mixture was then stirred and then ultrasonicated (100 W) for 1 h and poured into a glass Petri dish to cast the sample films, the cast film was dried in a vacuum oven at 50°C for 6 h before further studies. Composite films with 0.05, 0.1, 0.15, 0.2, and 0.25 wt% loading of m-MWCNTs were prepared along with a control EVACO film.

The Raman spectra (inVia, Renishaw, UK) of pristine and modified MWCNTs were collected to understand the effect of acid treatment on CNTs. The scanning electron microscope (SEM) (JSM-6380LA, JEOL, Japan) was used to study the fractured surfaces after the tensile test. The samples were sputtered with gold (JEOL JFC 1600 auto fine coater, JEOL, USA) to make them conductive. A transmission electron microscope (TEM, CM12 PHILIPS, Netherlands) was also used to image the MWCNTs before and after modification, the TEM images of the representative composite sample were also taken. The MWCNTs for TEM imaging are prepared by sonicating the MWCNTs for 30 min in ethanol and then depositing them on 200 mesh carbon-coated Cu TEM grids. X-ray diffraction (XRD) patterns (DX-GE-2P, JEOL, Japan) of the EVACO and EVACO/m-MWCNT composites were recorded under CuK α radiation in a 2θ range of 5–50°.

The degree of crystallinity (X_c) for the samples was calculated by deconvoluting the XRD pattern to separate the amorphous and crystalline contributions to the pattern and the degree of crystallinity was calculated from the ratio of the integrated area of all crystalline peaks to the total integrated area under the X-ray diffractogram.^[39] The degree of crystallinity (X_c) was measured by the following equation:

$$X_c = \frac{I_c}{I_a + I_c}, \quad (1)$$

where I_a and I_c are the integrated intensities corresponding to the amorphous and crystalline phases, respectively. Interplanar distances (d) of the crystallites in the composites are obtained by the following equations:

$$d = \frac{\lambda}{2\sin\theta}, \quad (2)$$

where λ is the wavelength of the X-rays (CuK α = 1.5418 Å) and θ is the Bragg angle.

Fourier transform infrared (FTIR) spectra (Jasco FTIR 4200, Japan) of the pristine MWCNTs, m-MWCNTs, pristine EVACO, and representative nanocomposites were recorded in attenuated total reflection mode in a wavenumber range of 650–4000 cm⁻¹ at an average of 32 scans with a resolution of 0.5 cm⁻¹. In the case of MWCNTs and m-MWCNTs, 128 scans are averaged and the resulting spectra were smoothed using a Savitzky–Golay smoothing algorithm. Thermogravimetric measurements were performed for EVACO and EVACO/m-MWCNT composites under a nitrogen atmosphere flowing at a rate of 100 ml min⁻¹ (Q600 V8.3, TA Instruments). A constant heating rate of 10°C min⁻¹ was maintained and the weight losses versus temperature curves were recorded over a temperature range of 25–700°C.

Differential scanning calorimetric (DSC) measurements were carried out by using about 5 mg of the samples in air-tight aluminum pans in a DSC analyzer (Q1000 V9.9, TA Instruments), under a nitrogen atmosphere with a flow rate of 50 ml min⁻¹ from –50 to 150°C at a heating rate of 10°C min⁻¹.

The % crystallinity of EVACO was determined from the area under the endothermic peak by using the following equation^[40],

$$X_c = \frac{\Delta H_f}{W_i \times \Delta H_{f100\%}} \times 100, \quad (3)$$

where X_c is the crystallinity (%); ΔH_f is the apparent melting enthalpy of crystallinity of EVACO (J/g); $\Delta H_{f100\%}$

is the extrapolated value of the enthalpy of crystallization of a 100% crystalline sample of EVA having a value of 68 J g^{-1} ^[41]; W_i is the weight fraction of EVACO in the composite.

The tensile testing was performed in a universal testing machine (H25KS, Hounsfield, UK), at room temperature as per American Society for Testing Materials (ASTM) standard D 638-10 at a strain rate of 50 mm min^{-1} . The tensile test specimens were punched out by using an ASTM D 412-06a die. The reported values of mechanical parameters are the averages of three values. Maximum deviations in the results of tensile strength, yield strength, M100, and elongation at break were $\pm 5\%$. The electrical direct current (DC) conductivity measurements were carried out on films of $2 \text{ cm} \times 2 \text{ cm}$ samples by using a two-probe method with a digital multimeter (MECO, 81K) under ambient conditions following ASTM D257.

3 | RESULTS AND DISCUSSION

3.1 | Raman analysis

The Raman spectra (Figure 2) of the MWCNTs clearly show the intense D-band at 1354 cm^{-1} (transverse or out-of-plane vibration of graphene walls) and G-band at 1591 cm^{-1} (longitudinal vibration of graphene or disorder of carbon) of typical MWCNTs.^[42] The G-band was split into two modes, G1 at $\sim 1595 \text{ cm}^{-1}$ and G2 at 1619 cm^{-1} , in the deconvoluted image as in the inset of Figure 2. The G2-band of the modified MWCNTs is evident in Figure 2, and it refers to the number of walls in MWCNTs (reduction in ordered arrangement),^[43,44] as the number of walls decreases its intensity will increase. In this case, an increase in the intensity of the peak in m-MWCNTs is may be due to the exfoliation of outer layers of the MWCNTs or due to the removal of amorphous carbon from the nanotubes during the acid treatment. A slight reduction in the intensity of the peaks after modification is attributed to the direct electron charge transfer from the functional groups attached to the surface of the MWCNTs through oxygen.^[45,46] The ratio of the intensity of D-band and G-band (I_D/I_G) gives the number of defects present in the nanotubes, and as the number of defects increases, the D/G ratio also increases. In this study, the ratio is increased from 1.26 to 1.32 after the functionalization of MWCNTs.

3.2 | TEM analysis

The TEM micrographs of the MWCNTs, m-MWCNTs, and EVACO/m-MWCNT composites are shown in

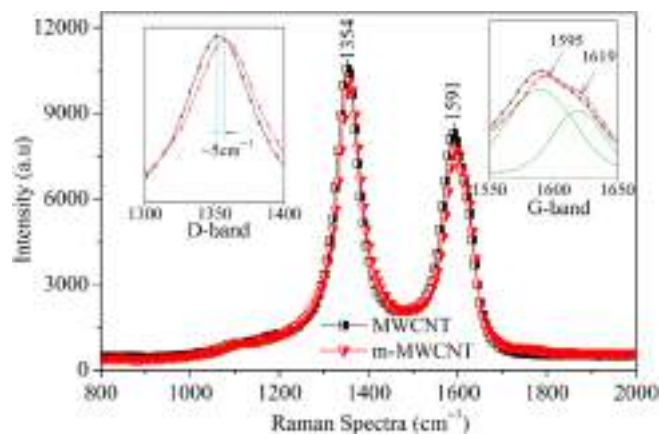


FIGURE 2 Raman spectra of pristine multiwalled carbon nanotube (MWCNT) and modified MWCNT [Color figure can be viewed at wileyonlinelibrary.com]

Figure 3. The average aspect ratio of MWCNTs is ~ 60 , which is calculated from the respective TEM images. In the case of unmodified MWCNTs (Figure 3A), the MWCNTs were with well-defined walls and circular ends and the diameters were less as compared with acid-treated m-MWCNTs. During the acid treatment, the surface and the ends of the MWCNTs were damaged, as clear in Figure 3B. The increase in the diameter of m-MWCNTs can be attributed to the increase in the wall thickness of the nanotubes since the acid treatment can intercalate the functional groups between the layers of the walls. The outer layers of the MWCNTs were severely damaged during the treatment, as in the inset in Figure 3B, which can improve the interfacial adhesion between EVACO and m-MWCNTs. This damage in the outer walls of MWCNTs after modification resulted in the appearance of G2' peak in the Raman spectra, which correspond to the number of graphene layers constituting the wall. In the TEM micrograph of EVACO/m-MWCNT composite, the walls of the MWCNTs are indistinguishable from the matrix, especially at several damaged regions of the m-MWCNTs, which reveals a good interfacial adhesion between MWCNTs and EVACO.

3.3 | FTIR analysis

The FTIR spectra of pristine MWCNT, modified MWCNT, EVACO, and representative EVACO/m-MWCNT composites are shown in Figure 4. In the spectra of pristine MWCNTs, the several peaks at the fingerprint region are attributed to the hexagonal carbon. As the MWCNTs are modified, several strong peaks have appeared in the spectra and the intensity of the peaks corresponding to the hexagonal carbon has reduced. The broad peak at 3245 cm^{-1} is

FIGURE 3 Transmission electron microscopy micrographs of (A) unmodified MWCNTs, (B) m-MWCNTs, and (C) EVACO/m-MWCNT composite with 0.1% m-MWCNT loading. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube

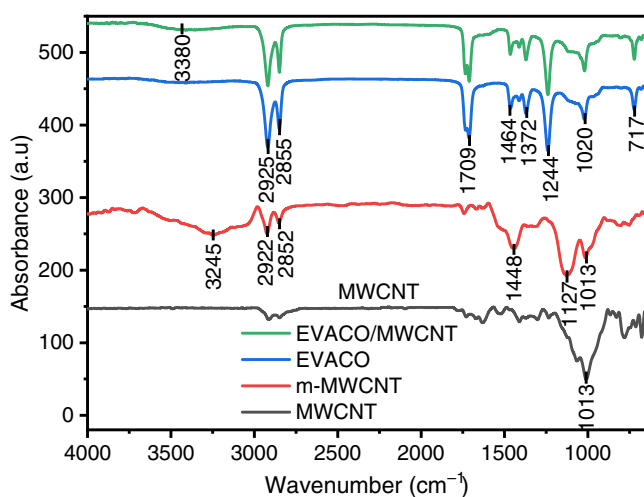
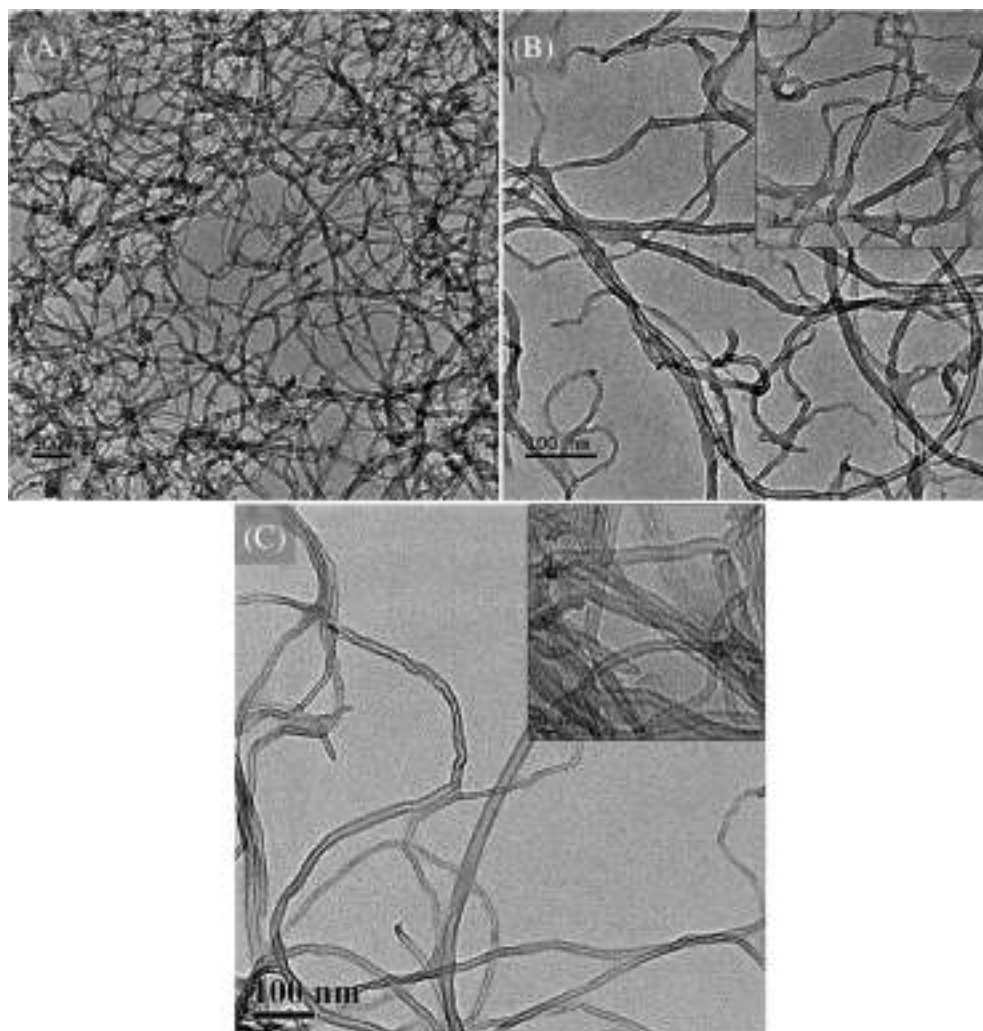


FIGURE 4 Fourier transform infrared spectra of pristine MWCNT, modified MWCNT, neat EVACO, and 0.25% m-MWCNT-loaded EVACO. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

attributed to the overtone of —OH and C=O stretching. The peaks 2925 and 2855 cm^{-1} are assigned to the symmetric and asymmetric stretching of —CH groups formed on the MWCNT surface after modification. The peaks at 1448 cm^{-1} and 1127 cm^{-1} are assigned to —OH deformation and C—O stretching of the carboxylic group, respectively. Thus, it is concluded that the MWCNT surface is attached with —OH and —COOH groups after modification.^[47]

In the spectra of EVACO and EVACO/m-MWCNT composite, the characteristic peak at 3380 cm^{-1} is assigned to OH stretching. The peaks at 2925 and 2855 cm^{-1} are due to symmetric and asymmetric stretching of —CH , respectively. The peak at 1709 cm^{-1} is due to the C=O stretching and 1464 and 1372 cm^{-1} are due to —CH scissoring and —CH deformation respectively. The peaks at 1242 cm^{-1} correspond to C—O stretching and 1019 cm^{-1} is due to C—OH stretching. The peak at 721 cm^{-1} is assigned to the rocking vibration of —CH .^[48]

In comparison with the FTIR spectrum of pristine EVACO, EVACO/m-MWCNT composite spectrum has

several peaks, which are originated from the pristine EVACO, but some peaks are modified in connection with the interaction of EVACO with modified MWCNTs. The intensification of the peak at 3380 cm^{-1} , which corresponds to $-\text{OH}$ stretching, is due to the formation of a hydrogen bond between $-\text{OH}$ and $-\text{COOH}$ groups on the m-MWCNTs and $\text{C}=\text{O}$ groups of EVACO or vice versa. During the formation of the hydrogen bond, the donor hydrogen atom from $-\text{OH}$ forms a bond with a lone pair of electrons in $\text{C}=\text{O}$, which has two lone pairs of electrons.^[49] Therefore, as the OH vibrates, this lone pair also vibrates, which will contribute to more change in the dipole moment and thus an increased intensity of $-\text{OH}$ stretching.^[50] Similarly, this induced dipole moment changed the intensity of the peak at 1709 cm^{-1} . The numerous peaks corresponding to the hexagonal structure of MWCNTs also appear in the composite, which made a downward shift in the spectra of EVACO/m-MWCNT in the fingerprint region. This reveals that a good interaction exists between the EVACO and m-MWCNTs.

3.4 | Thermogravimetric analysis

Thermogravimetric analysis results of EVACO and EVACO/m-MWCNT composites are shown in Figure 5. All the samples exhibit two steps in their degradation process. The first step between 300 and 400°C in the degradation process is the elimination of acetic acid by ester pyrolysis (deacetylation) during which the free acetate radical combines with the β -hydrogen to form acetic acid and this mechanism is akin to the degradation process of the ethylene-vinyl-acetate copolymer, which has an immediate analogy to EVACO. This process is followed by the degradation of the backbone of the polymer chain between 400 and 500°C , which has a polyene structure since the side group is eliminated during ester pyrolysis.^[51]

The thermal degradation temperature is slightly improved in EVACO/m-MWCNT composite as compared with that of pristine EVACO. A filler loading as low as $0.05\text{ wt}\%$ also made a remarkable improvement in the degradation temperature, and it indicates the strong interfacial interaction of the m-MWCNTs with EVACO. As the m-MWCNT loading is increased, the thermal stability of the composites also increases. This increase in the thermal properties may be attributed to the four major reasons, one is the physical adsorption of the polymer chains around the surface-modified nanotube restricting their mobility, thus preventing the sudden degradation of these polymer chains.^[50] The second is the enhanced adsorption of reactive products by the m-MWCNTs. The heterogeneous

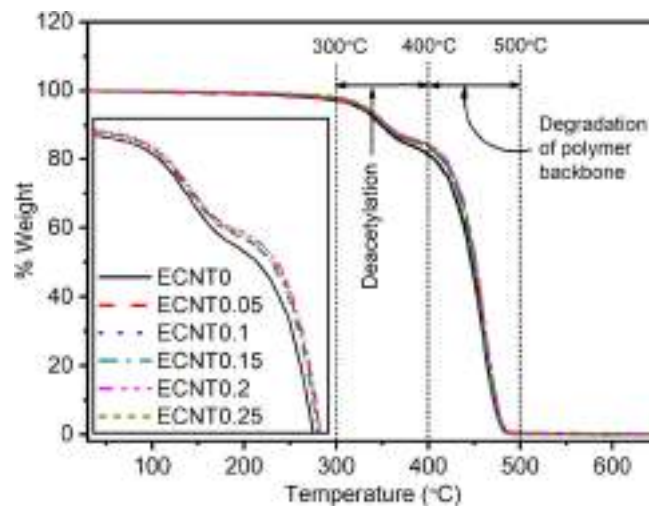


FIGURE 5 Thermogravimetric analysis results of pristine EVACO and EVACO/m-MWCNT (ECNT) composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

adsorption of these organic molecules on the m-MWCNTs is attributed to the high-energy adsorption sites, such as defects, functional groups, and interstitial space between the walls of m-MWCNTs.^[52] This adsorption process can be accelerated at high temperatures. Moreover, the functional groups such as $-\text{OH}$ and $-\text{COOH}$ are capable of trapping the reactive free radicals to form stable molecules.^[53]

The third is the high-temperature stability and good thermal conductivity of the MWCNTs. The high thermal stability of MWCNTs increases the integrity of char residue on the surface, which is formed at the initial stages of degradation, thus preventing the penetration of reactive molecules to the bulk of the composite. The high thermal conductivity helps to distribute the heat uniformly all over the composite.^[53] The fourth is due to the reactive scavenging by capillary condensation,^[54] in which active molecules can be adsorbed to the lumen of the MWCNTs, thus neutralizing the overall degradation process in the presence of MWCNTs.

3.5 | XRD analysis

The structural changes in the composite especially crystallinity are characterized by comparing X-ray diffractograms of EVACO and EVACO/m-MWCNTs with different filler loading as shown in Figure 6. The intense peak at $2\theta = 20.83^\circ$ is due to the (110) plane of polyethylene crystallites, since the polyethylene segments impart crystallinity

in EVACO.^[55] The crystallinity of EVACO/m-MWCNT composite increases as the filler content increases, as presented in Table 1. This increase in crystallinity is observed up to a filler loading of 0.1%, thereafter it decreases. Improvement in the crystallinity at these filler loading is due to the ability of MWCNTs to attract the polymer chains close to each other to form crystallites. But above certain loading, that is, critical loading, these nanotubes may no longer be able to bring the polymer chains together to form crystallites, because, at high filler loading, the high-aspect-ratio nanotube network can hinder the polymeric chain movements. This will be severe as the filler loading is higher since the nanotubes can destroy the coalition of

polymer chains at the spherulite front. Also, at these filler loadings, the bundling of MWCNTs can reduce the possibility of a polymer chain wrapping around the MWCNTs to form the crystalline regions.

3.6 | DSC analysis

The DSC results of EVACO and EVACO/m-MWCNT composites at different MWCNT loadings are shown in Figure 7. The melting of pristine EVACO, as well as composites, occurred in between 30 and 90°C. The melting temperature of EVACO is determined by the segmental mobility of the polyethylene phase, which constitutes the crystalline phase in the terpolymer. As the filler loading has increased, an increase in the crystallinity of the composite over the pristine EVACO is observed. It is due to the ability of modified CNTs to act as a nucleating agent as reported earlier.^[56,57] Nevertheless, the percentage crystallinity is reduced remarkably and comparable to that of pristine EVACO at a filler loading of 0.25 wt%. The dilution of the crystallite growth front by the high-aspect-ratio nanotubes and the arresting of free movement of polymer chains by the networked MWCNTs are expected at this filler loading, which may hinder crystallization of polymer chains that can otherwise undergo crystallization if MWCNTs are absent. Therefore, the modified MWCNTs favor the crystallization for a certain critical filler loading and it decreases after that.

The first heating curves (Figure 7A) of EVACO and EVACO/m-MWCNT composites have two major melting peaks, which correspond to α and β crystallites, whereas during cooling only one melting peak was observed. In

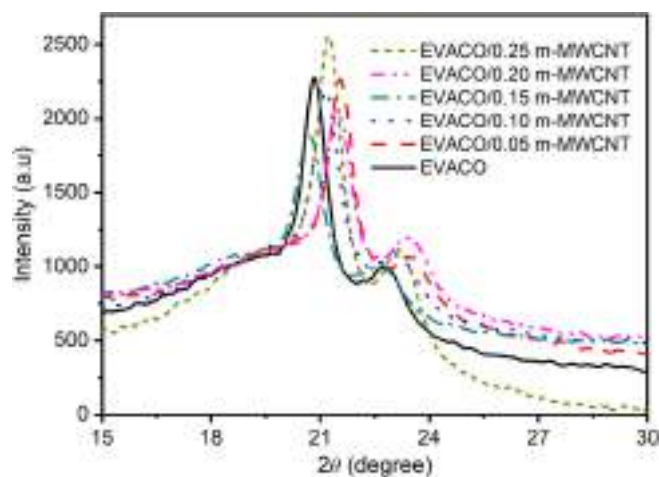


FIGURE 6 X-ray diffractograms of neat EVACO and EVACO/m-MWCNT composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Crystalline properties of EVACO at different m-MWCNT loadings

Filler loading (%)	Peak position (2θ)	d spacing (\AA)	The area under the peaks	Total area	% crystallinity
0.0	20.8	4.34	1499	4715	41.5
	22.8	3.98	461		
0.05	21.6	4.18	1662	4802	43.2
	23.4	3.88	410		
0.1	21.2	4.26	1620	4190	46.3
	23.1	3.92	321		
0.15	21.5	4.20	1486	4378	43.8
	23.5	3.86	433		
0.2	21.2	4.26	2218	7258	42.6
	23.2	3.92	877		
0.25	20.7	4.36	1309	3961	39.9
	22.7	3.98	272		

Abbreviations: EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube.

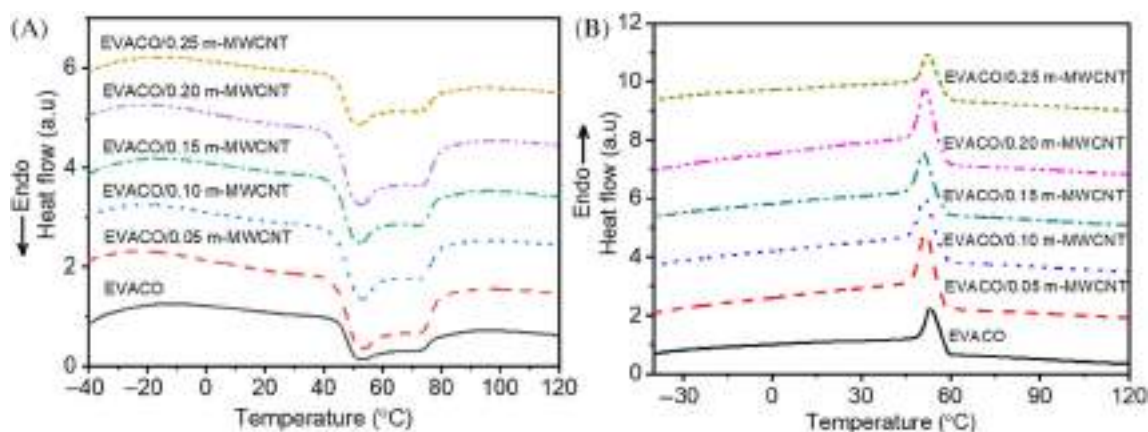


FIGURE 7 DSC curves of neat EVACO and EVACO/m-MWCNT composite (A) heating and (B) cooling. DSC, Differential scanning calorimetry; EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

solution casting, the polymer chains in the solvent are free to move and capable of aligning themselves to a most thermodynamically favorable position before the solvent is completely evaporated, apart from self-crystallization, the presence of m-MWCNTs in the solution also drives the polymer chains to arrange in a preferred order. Ultimately this results in an increased crystallinity in the composite. In melts, the restricted chain movements allow the formation of one type of crystallite (β -crystallite), and the intensity of this melting peak (Figure 7B) is increased in the composite as compared with the pristine EVACO. It is worth noting that there is a shift in melting temperature of the composites to lower values. Therefore, one can scrutinize that the presence of m-MWCNTs in EVACO can impart additional crystallinity in EVACO. Table 2 shows the percentage crystallinity from first heating and second heating DSC curves, glass transition temperature (T_g), and the melting temperature.

The glass transition temperatures (T_g) of the composites are high as compared with the pristine EVACO. The presence of nanotubes in the composite lessens the suppleness of the polymer chain movements and the wrapping of

polymer chains to the nanotubes increases the crystallinity adjacent to the tube surfaces. Besides crystallinity, the interference of m-MWCNTs decreases the polymer chain movements, this interference will be high for a composite with good filler dispersion.^[37] Therefore, a maximum T_g represents a composite with good filler dispersion.

3.7 | Tensile properties

The stress versus strain curves of neat EVACO and EVACO/m-MWCNT composites are shown in Figure 8. The addition of a small quantity of m-MWCNTs, which is as low as 0.05% shows a large enhancement in the tensile strength of the composite. There is a significant improvement in the tensile strength of the other composites also. The elongation at break is the highest for the composite with a good tensile strength, which in turn has good crystallinity as compared with neat EVACO, as observed in DSC and XRD analysis. The number of crystalline block segments in the composite is more than that is in neat EVACO, since the crystallinity is increased in composite

TABLE 2 The crystallinity of EVACO/m-MWCNT with different m-MWCNT loading

Filler loading (wt%)	% crystallinity from the first heating		% crystallinity from cooling		T_g
	Melting temperature (°C)	% crystallinity	Melting temperature (°C)	% crystallinity	
0	51.88	28.2	53.13	18.1	-41.4
0.05	52.05	54.4	51.36	30.2	-40.5
0.1	51.81	51.5	51.31	29.7	-40.07
0.15	52.57	47.9	51.08	26.3	-40.1
0.2	51.32	43.9	50.65	25.1	-40.4
0.25	52.24	33.6	51.19	18.2	-41.3

Abbreviations: EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube.

with filler loading. As the crystallinity increases, the orientation of these segments and the respective tie chains toward the applied force need more effort than the polymer with less crystallinity,^[58] which will increase the tensile strength and more elongation at break in the composites. The simultaneous reduction in the tensile strength with filler loading may be attributed to the less crystallinity in them and also the nodal points in the MWCNT network, which can act as the stress concentrators.

The mechanical properties of the nanocomposite depend on several parameters, they are filler dispersion, crystallinity, filler–matrix interaction, processing methods, and so on.^[59] If the dispersion is poor, even though nanotubes are flexible, the bundled MWCNTs can act as rigid stress concentrators due to their difference in elastic properties compared to the EVACO matrix. The stress concentration leads to the building up of stress around the particles and ultimately results in the debonding of nanotubes at the

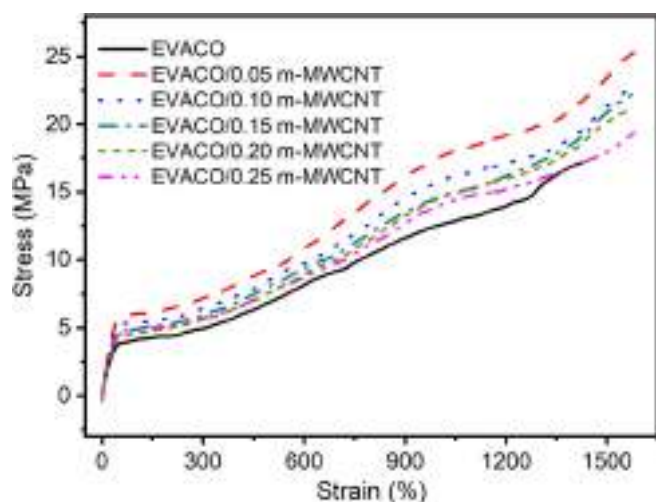


FIGURE 8 Stress versus strain curves of neat EVACO and EVACO/m-MWCNT composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Mechanical properties of virgin EVACO and the composites

wt% of MWCNT	Ultimate tensile strength (MPa)	Yield strength (MPa)	Stress at 100% elongation (MPa)	% elongation at break	Toughness (kJ/m ³)
0	17.2	3.8	4.14	1420	13.4
0.05	24.9	5.8	6.04	1607	22.8
0.1	22.5	5.2	5.45	1573	19.6
0.15	22.1	4.6	4.91	1568	18.6
0.2	21.8	4.8	4.82	1580	17.4
0.25	21.5	4.5	4.59	1573	18.0

Abbreviations: EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); MWCNT, multiwalled carbon nanotube.

MWCNT–EVACO interface. It is true in the case of nanocomposite with 0.25% filler loading, which has the least tensile strength and elongation at break among the composites. The aspect ratio of MWCNTs affects the tensile properties of the EVACO/m-MWCNT composites. On comparing the ultimate tensile strength and toughness of EVACO/nano-alumina trihydrate^[36] and EVACO/halloysite nanotube^[60] composites with EVACO/m-MWCNT composite, one can observe that EVACO/m-MWCNT composites exhibit superior mechanical properties at very low MWCNT loading (Table 3).

3.8 | SEM fractography

The SEM micrographs of tensile fracture surfaces of neat EVACO and EVACO/m-MWCNT composites are shown in Figure 9. All the samples exhibit a typical ductile failure, which is revealed by the continuous crack propagation trajectories. The gradual transformation from ductile to brittle nature is observed on the fracture surfaces. In the composite samples, the stress whitened regions are less intense because of the increase in the crystalline regions in the composite. The tensile and yield strength of the composites are enhanced through filler loading and the elastic recovery zone of the composites is greater than the neat polymer (Figure 8), therefore the composites are resistant to stress whitening.^[61] There are no traces of crazing at the edges of the crack propagation trajectories, but fibrils are present on the fractured surface since the stress in the polymer matrix in the premises of MWCNTs is different from that is away from MWCNTs, which will reduce the sensitivity toward crazing and promote shear yielding, leading to the formation of fibrils.

3.9 | Electrical conductivity

DC volume resistivities of the EVACO/MWCNT composites are shown in Figure 10. The resistivity of the

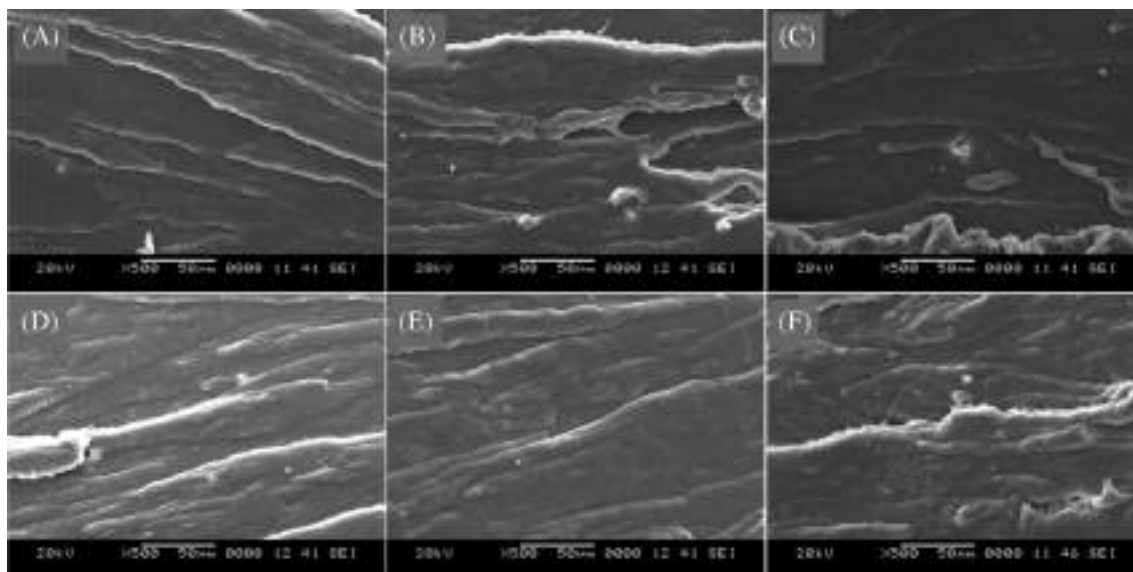


FIGURE 9 The scanning electron microscopy micrographs of EVACO with different m-MWCNT loadings: (A) 0%, (B) 0.05%, (C) 0.1%, (D) 0.15%, (E) 0.2%, and (F) 0.25%. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube

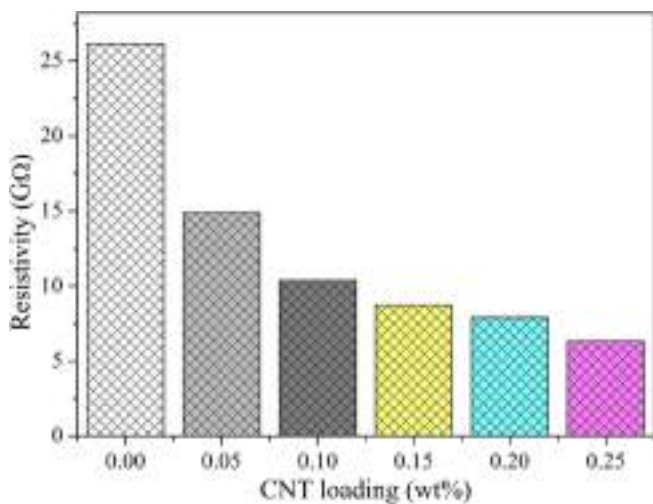


FIGURE 10 Electrical volume resistivity of the composites [Color figure can be viewed at wileyonlinelibrary.com]

composite was reduced as the filler content was increased, but the reduction in the resistivity is not appreciable as the conductivity of the MWCNTs is concerned. The reduction in the resistivity is due to the presence of conductive MWCNTs and its insignificance is due to the fact the MWCNT loadings are far less than that of the percolation threshold. Below the percolation threshold, the MWCNTs are isolated from each other, and no continuous network of MWCNTs is intact for electron transport. The improvement in electron transport properties in the presence of MWCNTs is due to the shortening of

the resistive electron path, which is otherwise completely resistive if EVACO alone is considered.

4 | CONCLUSION

In summary, the addition of modified MWCNTs was an efficient way to improve the strength and crystalline properties of EVACO. A minute quantity of MWCNTs was sufficient enough to make a significant change in the properties of EVACO. Good interaction between m-MWCNTs and EVACO was observed in the FTIR analysis. The thermal stability of the composites was improved with the filler loading. The increase in crystallinity by the addition of m-MWCNTs is observed in the case of all the composites irrespective of the filler loading. Composite with 0.05% loading of MWCNTs exhibited the best crystallinity (30.5%), which in turn resulted in the maximum tensile strength in these composites. The pristine EVACO and its composites exhibited ductile fracture and as the filler loading increased the fracture was approaching brittle nature. The percentage elongation at break for pristine EVACO is 1420%, which is increased to 1607% by the addition of 0.05 wt% of m-MWCNTs, with a subsequent increase in the ultimate strength in 17.2–24.9 MPa. The m-MWCNT loading in the composite was below the percolation threshold; therefore, only a small reduction (26–7 GΩ) in the resistivity was observed among the composite, which may improve the antistatic properties of the composite.

ACKNOWLEDGMENTS

The authors are very grateful to Mr Hariharan, DuPont Dow elastomers, India, for the free supply of EVACO. Ms Reshmi U, Department of Metallurgical and Materials Engineering, NITK, is acknowledged for her assistance in SEM.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Gibin George  <https://orcid.org/0000-0002-4236-4860>

S. Anandhan  <https://orcid.org/0000-0002-4429-7550>

REFERENCES

- [1] O. A. Mendoza Reales, P. A. Carisio, T. C. dos Santos, W. C. Pearl, R. D. Toledo Filho, *Constr. Build. Mater.* **2021**, *281*, 122603.
- [2] Z. Spitalsky, D. Tasis, K. Papagelis, C. Galiotis, *Prog. Polym. Sci.* **2010**, *35*, 357.
- [3] D.-R. Yu, G.-H. Kim, *Polym.-Plast. Technol. Eng* **2013**, *52*, 699.
- [4] M. Park, H. Kim, J. P. Youngblood, *Nanotechnology* **2008**, *19*, 055705.
- [5] G. Konstantopoulos, P. Maroulas, D. A. Dragatogiannis, S. Koutsoumpis, A. Kyritsis, C. A. Charitidis, *Mater. Des.* **2021**, *199*, 109420.
- [6] J. Zhang, W. Yu, X. Zhang, X. Gao, H. Liu, X. Zhang, *J. Appl. Polym. Sci.* **2021**, *138*, 50170.
- [7] R. Sengupta, M. Bhattacharya, S. Bandyopadhyay, A. K. Bhowmick, *Prog. Polym. Sci.* **2011**, *36*, 638.
- [8] N. Saadat, H. N. Dhakal, S. Jaffer, J. Tjong, W. Yang, J. Tan, M. Sain, *Compos. Sci. Technol.* **2021**, *207*, 108654.
- [9] A. Roy, T. Mondal, S. Kar, K. Naskar, R. Ghosal, R. Mukhopadhyay, A. K. Bhowmick, *J. Appl. Polym. Sci.* **2021**, *138*, 49093.
- [10] B. Wei, L. Zhang, S. Yang, *Chem. Eng. J.* **2021**, *404*, 126437.
- [11] A. Tarhini, A. Tehrani-Bagha, M. Kazan, B. Grady, *J. Appl. Polym. Sci.* **2021**, *138*, 49821.
- [12] T. Hu, Y. Song, J. Di, D. Xie, C. Teng, *Carbon* **2018**, *140*, 596.
- [13] A. A. Tarhini, A. R. Tehrani-Bagha, *Compos. Sci. Technol.* **2019**, *184*, 107797.
- [14] R. Izadi, E. Ghavanloo, A. Nayebi, *Phys. B* **2019**, *574*, 311636.
- [15] S. Neelakandan, D. Liu, L. Wang, M. Hu, L. Wang, *Int. J. Energy Res.* **2019**, *43*, 3756.
- [16] A. V. Penkova, S. F. A. Acquah, L. B. Piotrovskiy, D. A. Markelov, A. S. Semisalova, H. W. Kroto, *Russ. Chem. Rev.* **2017**, *86*, 530.
- [17] Q. Zhang, J. Wang, B.-Y. Zhang, B.-H. Guo, J. Yu, Z.-X. Guo, *Compos. Sci. Technol.* **2019**, *179*, 106.
- [18] C. Xiao, W. Liang, Q.-M. Hasi, F. Wang, L. Chen, J. He, F. Liu, H. Sun, Z. Zhu, A. Li, *ACS Appl. Energy Mater.* **2020**, *3*, 11350.
- [19] A. Verma, K. Baurai, M. R. Sanjay, S. Siengchin, *Polym. Compos.* **2020**, *41*, 338.
- [20] R. Ram, V. Soni, D. Khastgir, *Compos. Part B Eng.* **2020**, *185*, 107748.
- [21] V.-H. Nguyen, S. A. Delbari, M. Shahedi Asl, Q. V. Le, M. Shokouhimehr, M. Mohammadi, A. Sabahi Namini, *Ceram. Int.* **2021**, *47*, 12941.
- [22] D. Ding, J. Wang, X. Yu, G. Xiao, C. Feng, W. Xu, B. Bai, N. Yang, Y. Gao, X. Hou, G. He, *Ceram. Int.* **2020**, *46*, 5407.
- [23] O. Popov, J. Vleugels, E. Zeynalov, V. Vishnyakov, *J. Eur. Ceram. Soc.* **2020**, *40*, 5012.
- [24] H. S. Manohar, S. Reddy Mungara, S. N. Anand, K. S. T. Reddy, *Mater. Today Proc* **2020**, *20*, 185.
- [25] K. Aristizabal, A. Katzensteiner, A. Bachmaier, F. Mücklich, S. Suárez, *Sci. Rep.* **2020**, *10*, 857.
- [26] V. Khanna, V. Kumar, S. A. Bansal, *Mater. Res. Bull.* **2021**, *138*, 111224.
- [27] R. Manoj Kumar, S. K. Sharma, B. V. Manoj Kumar, D. Lahiri, *Compos. Part Appl. Sci. Manuf.* **2015**, *76*, 62.
- [28] J. N. Coleman, U. Khan, W. J. Blau, Y. K. Gun'ko, *Carbon* **2006**, *44*, 1624.
- [29] J. Banerjee, K. Dutta, *Polym. Compos.* **2019**, *40*, 4473.
- [30] K. Anoop Anand, U. S. Agarwal, R. Joseph, *Polymer* **2006**, *47*, 3976.
- [31] A. R. Bhattacharyya, T. V. Sreekumar, T. Liu, S. Kumar, L. M. Ericson, R. H. Hauge, R. E. Smalley, *Polymer* **2003**, *44*, 2373.
- [32] L. Li, B. Li, M. A. Hood, C. Y. Li, *Polymer* **2009**, *50*, 953.
- [33] K.-W. Park, G.-H. Kim, *J. Appl. Polym. Sci.* **2009**, *112*, 1845.
- [34] S. Morlat-Therias, E. Fanton, J.-L. Gardette, S. Peeterbroeck, M. Alexandre, P. Dubois, *Polym. Degrad. Stab.* **2007**, *92*, 1873.
- [35] V. Mittal, *Surface Modification of Nanotube Fillers*, Wiley-VCH, Weinheim, Germany **2011**.
- [36] G. George, A. Mahendran, S. Anandhan, *Polym. Bull.* **2014**, *71*, 2081.
- [37] S. Anandhan, H. G. Patil, R. R. Babu, *J. Mater. Sci.* **2011**, *46*, 7423.
- [38] A. M. Shanmugaraj, J. H. Bae, K. Y. Lee, W. H. Noh, S. H. Lee, S. H. Ryu, *Compos. Sci. Technol.* **2007**, *67*, 1813.
- [39] S. Park, J. O. Baker, M. E. Himmel, P. A. Parilla, D. K. Johnson, *Biotechnol. Biofuels* **2010**, *3*, 10.
- [40] Y. Kong, J. N. Hay, *Polymer* **2002**, *43*, 3873.
- [41] S. Chattopadhyay, T. K. Chaki, A. K. Bhowmick, *Radiat. Phys. Chem.* **2000**, *59*, 501.
- [42] L. Bokobza, J. Zhang, *Express Polym. Lett.* **2012**, *6*, 601.
- [43] S. L. H. Rebelo, A. Guedes, M. E. Szeftczyk, A. M. Pereira, J. P. Araújo, C. Freire, *Phys. Chem. Chem. Phys.* **2016**, *18*, 12784.
- [44] L. Bokobza, J.-L. Bruneel, M. Couzi, *J. Carbon Res.* **2015**, *1*, 77.
- [45] S. Costa, E. Borowiak-Palen, *Acta Phys. Pol. A* **2009**, *116*, 32.
- [46] A. Felten, I. Suarez-Martinez, X. Ke, G. Van Tendeloo, J. Ghijsen, J.-J. Pireaux, W. Drube, C. Bittencourt, C. P. Ewels, *ChemPhysChem* **2009**, *10*, 1799.
- [47] V. T. Le, C. L. Ngo, Q. T. Le, T. T. Ngo, D. N. Nguyen, M. T. Vu, *Adv. Nat. Sci. Nanosci. Nanotechnol.* **2013**, *4*, 035017.
- [48] B. D. Mistry, *A Handbook of Spectroscopic Data Chemistry: (UV, IR, PMR, 13CNMR and Mass Spectroscopy)*, Oxford Book Company, Jaipur, India **2009**.
- [49] F. L. A. Khan, P. Sivagurunathan, J. Asghar, *Indian J. Pure Appl. Phys.* **2008**, *46*, 12.
- [50] K. Pielichowski, A. Leszczyńska, J. Njuguna, *Optimization of Polymer Nanocomposite Properties*. (Eds: V. Mittal) John Wiley & Sons, Ltd, Weinheim, Germany **2010**, p. 195.
- [51] R. Wilson, T. S. Plivelic, A. S. Aprem, C. Ranganathaiagh, S. A. Kumar, S. Thomas, *J. Appl. Polym. Sci.* **2012**, *123*, 3806.
- [52] B. Pan, B. Xing, *Environ. Sci. Technol.* **2008**, *42*, 9005.
- [53] S. P. Su, Y. H. Xu, P. R. China, C. A. Wilkie, in *Polymer-Carbon Nanotube Composites: Preparation, Properties and Applications*

- (Eds: T. McNally, P. Pötschke), Woodhead Publishing, Cambridge, UK **2011**, p. 482.
- [54] J. T. W. Yeow, J. P. M. She, *Nanotechnology* **2006**, *17*, 5441.
- [55] M. Selvakumar, A. Mahendran, P. Bhagabati, S. Anandhan, *Adv. Polym. Technol.* **2015**, *34*, 21467.
- [56] J. Y. Kim, H. S. Park, S. H. Kim, *Polymer* **2006**, *47*, 1379.
- [57] A. Funck, W. Kaminsky, *Compos. Sci. Technol.* **2007**, *67*, 906.
- [58] P. J. Flory, *Principles of Polymer Chemistry*, Cornell University Press, Ithaca, USA **1953**.
- [59] S. C. Tjong, *Mater. Sci. Eng. R Rep.* **2006**, *53*, 73.
- [60] G. George, M. Selvakumar, A. Mahendran, S. Anandhan, *J. Thermoplast. Compos. Mater.* **2017**, *30*, 121.
- [61] M. Tanniru, R. D. K. Misra, *Mater. Sci. Eng., A* **2006**, *424*, 53.

How to cite this article: G. George, A. Mahendran, S. Murugesan, S. Anandhan, *Polymer Composites* **2021**, *1*. <https://doi.org/10.1002/pc.26158>

PAPER • OPEN ACCESS

Impact of Ground Nut Shell Ash on Cobalt-Chromium metal matrix composites synthesized using Powder metallurgy process.

To cite this article: G R Raghav *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1166** 012006

View the [article online](#) for updates and enhancements.

You may also like

- [High-rate multi-GNSS attitude determination: experiments, comparisons with inertial measurement units and applications of GNSS rotational seismology to the 2011 Tohoku Mw9.0 earthquake](#)
Peiliang Xu, Yuanming Shu, Xiaoji Niu et al.

- [Using Allan variance to evaluate the relative accuracy on different time scales of GNSS/INS systems](#)
Quan Zhang, Xiaoji Niu, Qijin Chen et al.

- [Investigation on mechanical, wear and corrosion properties of Fe-Co-Cr-W-GNSA hybrid composites synthesized using powder metallurgy process](#)
G R Raghav, D MuthuKrishnan, R Sundar et al.

An advertisement for the ECS Meeting. It features a hand pointing at a glowing globe of the Earth, surrounded by a network of blue icons representing people and connections. The ECS logo is in the top right. The text reads: 'Connect with decision-makers at ECS', 'Accelerate sales with ECS exhibits, sponsorships, and advertising!', and 'Learn more and engage at the 244th ECS Meeting!' with a yellow play button icon.

ECS

Connect with decision-makers at ECS

Accelerate sales with ECS exhibits, sponsorships, and advertising!

▶ Learn more and engage at the 244th ECS Meeting!

Impact of Ground Nut Shell Ash on Cobalt-Chromium metal matrix composites synthesized using Powder metallurgy process.

G R Raghav¹, Suraj R^{2*}, Sheeja Janardhanan³, Vidya Chandran⁴,
K J Nagarajan⁵ and Nikhil Asok⁶

^{1,2,3,4,6}Department of Mechanical Engineering, SCMS School of Engineering and
Technology, Ernakulam, Kerala, India.

⁵Department of Mechanical Engineering, KLN College of Engineering,
Pottapalayam, Sivagangai Dt. Tamil Nadu, India 630612

*Corresponding Author email id: surajr@scmsgroup.org

Abstract. Co-based composites are extensively utilized in the field of prosthesis and dental implants. Hybrid composites made using Powder metallurgy process, Co-10Cr-GNSA were studied. The surface morphology of the hybrid composites were studied using Scanning Electron Microscope. The elemental analysis was carried out using X-Ray Diffraction technique. The hybrid composites were analyzed for its various mechanical properties like microhardness, compressive strength, and density. Value of micro hardness of the composite materials showed slight improvement with addition of GNSA reinforcement. The value of density of the hybrid composites was found to be decreasing linearly with the addition of GNSA. Compressive strength of the materials showed a reasonable increment. Wear analysis to study the tribological characterization of the hybrid composites were done with the help of a pin on disc wear testing machine. The wear and COF studies show that with a rise in GNSA content, wear resistance increases because of the presence of oxides of GNSA particles. From the worn out surfaces of the hybrid composite it is concluded that the deformation of the composites takes places initially due to abrasive wear followed by plastic deformation. An electrochemical workstation was used to understand the corrosion characteristics of the hybrid composites in the presence of 3% NaClelectrolyticsolution. Co-5Cr-5GNSA hybrid composites exhibit better electrochemical corrosion resistance compared to other specimens.

Keywords: Powder metallurgy, Wear, Corrosion, GNSA

1. Introduction

Now a days more and more people suffer from osteoarthritis disorder, which makes them experience severe pain and discomfort. Recent survey suggests that there are nearly 50 million cases worldwide who are suffering from osteoarthritis disorder and in need of joint replacement surgery [1]. Co-Cr-Mo alloy is the extensively used artificial prosthetic material considering its higher value of wear, hardness and corrosion resistance. Even though Co-Cr-Mo alloys are excellent prosthetic material, still there are certain disadvantages such as wear of implants in the hip joints and problems related to bio compatibility since Mo is not a bio degradable material[2–5]. Therefore it is the need of the hour to



produce a composite with much better wear and corrosion resistance which is also bio degradable and compatible to human body.

The ground Nut Shell Ash (GNSA) which is primarily a biological waste and is available in abundance all over the world. Moreover the GNSA particles have presence of $MgSiO_3$ and $AlSiO_3$ in high concentration .Hence it can be used to replace the hazardous Mo reinforcements[6].

There are many conventional methods to produce wear resistance artificial prosthetic implants such as plasma spraying, physical vapor deposition, electro deposition and chemical vapour deposition. Since these manufacturing processes includes more complex steps and requires costly equipments, the cost of the implants is high. The powder metallurgy technique has its own advantages which include uniform dispersion, low processing cost and ability to manufacture high melting point materials. Hence the powder metallurgy process possesses great potential for producing Co-Cr based hybrid composite materials with highly desirable mechanical properties along with wear and corrosion resistance[7–13].

This work aims to develop a Co-Cr-GNSA hybrid composite material with better wear, corrosion resistance and mechanical properties. In this study, four different compositions based on weight percent is formulated as follows Co-10Cr, Co-10Cr-2.5GNSA, Co-10Cr- 3.5 GNSA and Co-10Cr-5GNSA. The composite powders are mechanically milled and compacted and sintered in order to develop specimens of 8mm cylindrical pellets. The hybrid composites are then studied in order to explore their morphological properties using SEM. The mechanical behavior along with tribological and corrosion resistance behavior were studied and their mechanisms were reported.

2. Materials and Method

The materials CoCr (99.5% purity) which is used in this study were purchased from Mepco Ltd Tamil Nadu, India. The ground nut shell ash (GNSA) powder used in this work is prepared using heat treatment method which is discussed in our pervious paper [6]. Mechanical ball milling process was used for alloying the Co-Cr- GNSA hybrid composites. The process was carried out for two hours and was then compacted into 8 mm diameter pellet which is cylindrical in shape. The value of compaction pressure was set to 750 MPa consistently. After this, the soft green compacts were hardened by forcing them to sintering process at 1000oC for 2h.The morphology of the hybrid composites were studied using a Field Emission Scanning Electron Microscope (FE-SEM).ASTM: B962-13 standards were used to calculate the density of the Co-Cr- GNSA hybrid composites. The ASTM E384 standards were used to study the micro hardness of the hybrid composite pellets at a uniform load and dwell time of 1 kgf and 10 seconds respectively. Compressive strength of the hybrid composites were studied at a scan rate of 5 mm/min, with the help of a Universal Testing Machine (UTM). ASTM G99-05 standards were used to study the wear and friction behavior of the composites. EN 32 steel of hardness 65 HRC was used for the analysis. The specimens were cleaned using acetone solution before and after the wear test. The wear analysis of the composites was done at various sliding conditions such as the load, sliding distance and sliding speed. The electrochemical corrosion tests were simulated on a three electrode workstation using 3% NaCl solution as electrolyte[14–16].

3. Results and Discussion

3.1 Field Emission Scanning Electron Microscope Analysis

FE-SEM images of Co-10Cr- 3.5 GNSA & Co-10Cr-5GNSA hybrid Composites respectively are shown in Figure 1. There is a homogenous mixture of GNSA Particles with Co and Cr particles. The wettability of the GNSA particles was the major factor in achieving uniform amalgamation. It can be noted that due the milling operation the size of Cr particles have reduced to around 500 nm in size and are bonded strongly with Co matrix.

3.2 Microhardness

The microhardness test was done using a Vickers Micro Hardness Testing Machine with the test being conducted at five different points. Figure 2 shows the variation in the average value of microhardness

of the composites at different configurations based on its composition, i.e. Co-10Cr, Co-10Cr-2.5GNSA, Co-10Cr- 3.5 GNSA and Co-10Cr-5GNSA. The microhardness of the composites varied from 320 HV to 340 HV. The hardness of Co-10Cr was found to be 320 HV and the introduction of GNSA resulted in an increase in the microhardness. The maximum microhardness was found to be in Co-10Cr-5GNSA composite with a value of 340 HV. The uniform amalgamation of GNSA particles was the major reason for this improvement in microhardness.

3.3 Compressive Strength and Density

With the addition of the GNSA reinforcement the density of the Co-10Cr –GNSA hybrid composites were found to be decreasing. The value of density for Co-10Cr composites was recognized as 8.1 g/cm³ whereas the density of the Co-10Cr-5GNSA hybrid composites were around 7.65 g/cm³ as shown in Figure.3. This reduction in density was attributed by the relatively soft nature of the GNSA particles. With the addition of GNSA particles, the compressive strength of the hybrid composite materials showed slight increase in its value. Figure.3 helps us understand the compressive strength of different combinations of Co-10Cr-GNSA hybrid composites. The compressive strength of Co-10Cr composite was established to be in the region of 380 MPa. The compressive strength has slightly increased to 401 MPa for the Co-10Cr- 5 GNSA hybrid composites which is due the presence of AlSiO₃ particles in the GNSA ash content.

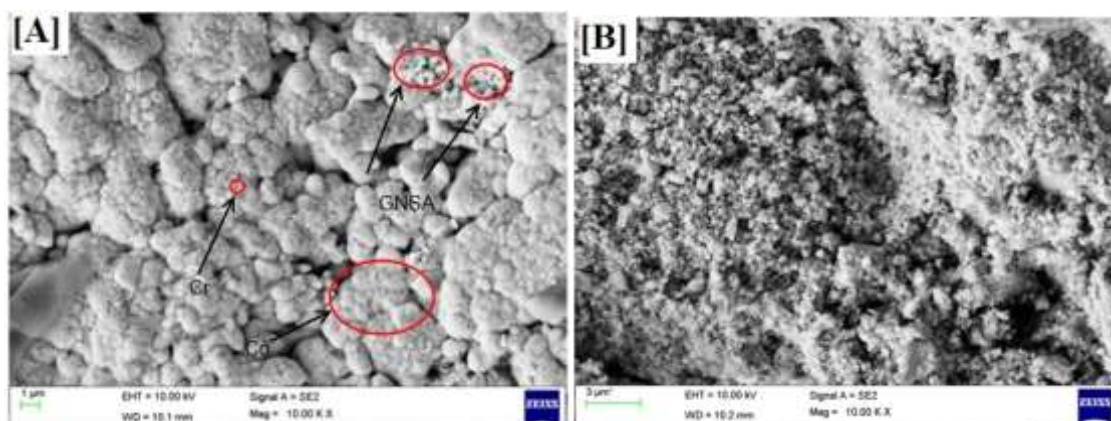


Figure 1.FESEM images of Co-10Cr- 3.5 GNSA & Co-10Cr-5GNSA hybrid Composite.

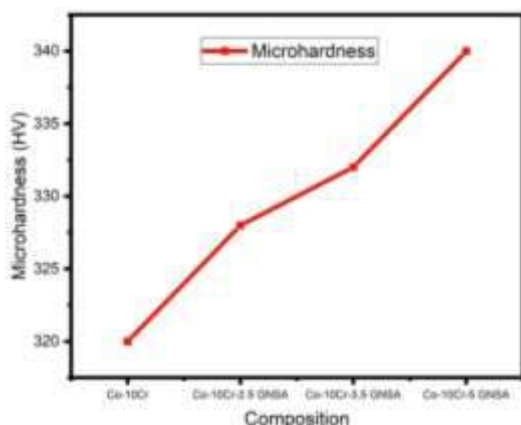


Figure 2.Graphical Representation of Co-10Cr-GNSA hybrid composites.

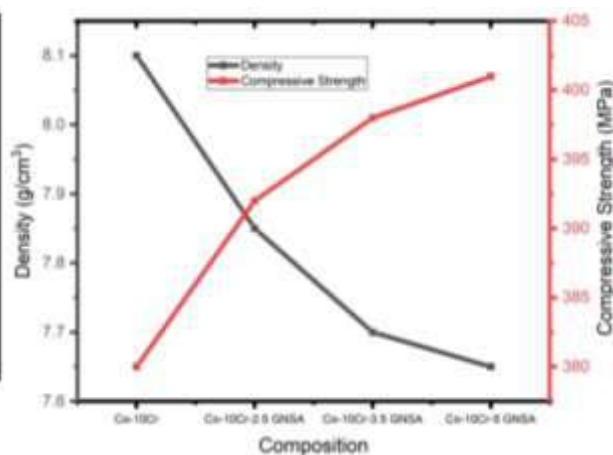


Figure 3.Comparison of Density and Compressive Strength of the Co-10Cr-GNSA hybrid composites.

3.4 Wear and COF Analysis

The loss of material due to wear of the Co-10Cr-GNSA hybrid composites is shown in Figure 4. The variation of wear loss of Co-10Cr-GNSA hybrid composites is depicted as graphical plots. The Figure 4 (A) indicates the wear analysis data of the Co-10Cr-GNSA hybrid composites at different loads (10N, 15N and 20N). The sliding speed (1.5 m/s) and sliding distance (1000 m) were kept constant. The Co-10Cr-5GNSA hybrid composites have witnessed very minimal wear loss at all loading conditions. The wear loss of Co-10Cr-GNSA hybrid composites at various sliding distance and speed is shown in Figure 4 (B&C) respectively. The wear loss has experienced similar trend. With the increase in GNSA concentration in the matrix there is definite resistance to wear and thereby the wear loss is very minimal for the Co-10Cr-5GNSA hybrid composites. The variation in coefficient of friction at different loads, sliding distance and sliding speed for Co-10Cr-GNSA hybrid composites is depicted in Figure 5(A,B&C). It was observed that with an increase in load, the COF of the hybrid composites increased. Whereas, it reduced with an increase in sliding speed. Overall the Co-10Cr-5GNSA hybrid composites displayed better COF value. This improvement in Wear and friction characteristics is may be attributed to the presence of AlSiO₃ compounds in the composite material and also due to the tribo oxide surface layer formation on the surface of the composite specimen. The worn out surface analysis of the Co-10Cr-GNSA hybrid composites after wear analysis is represented in Figure 6. From the worn out surface analysis it can be concluded that there is plastic deformation experienced in hybrid composites which is preceded by abrasive wear.

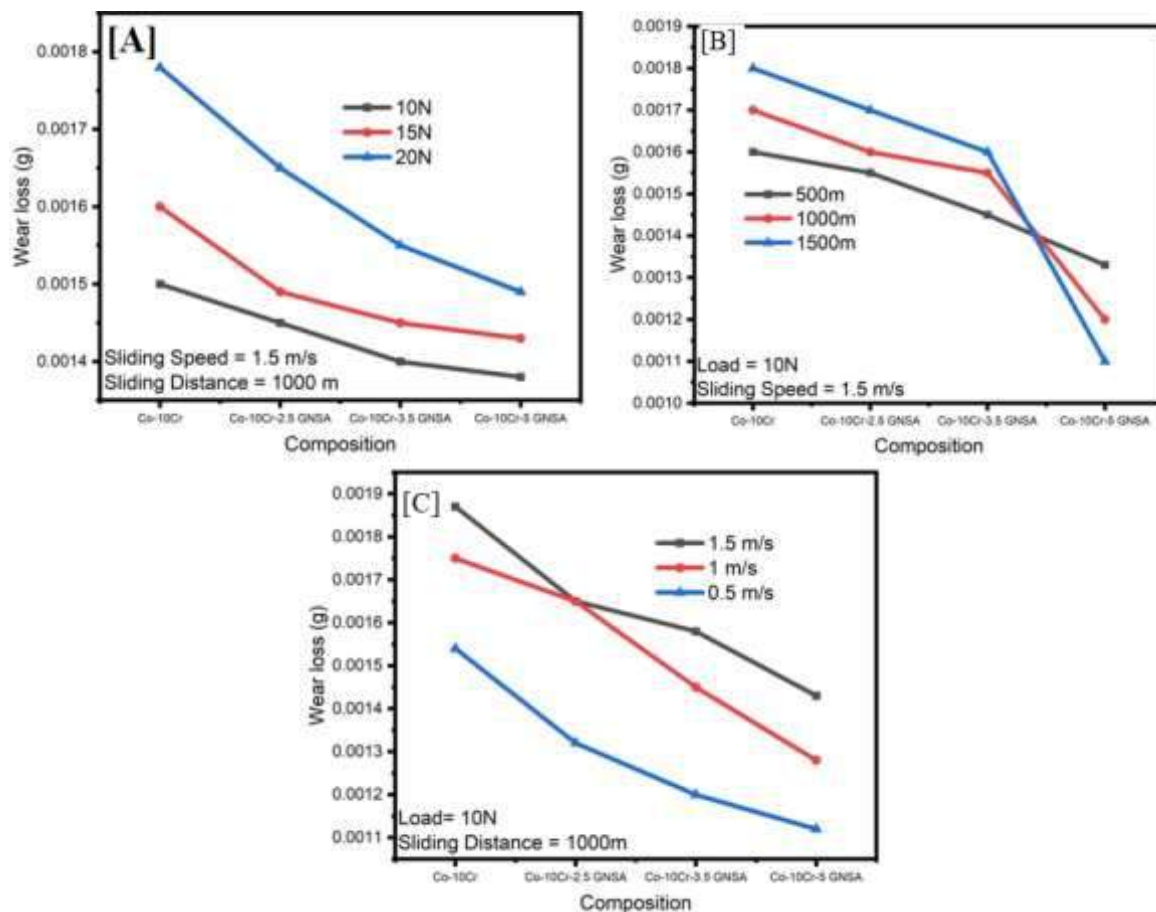


Figure 4. Wear Loss plot of Co-10Cr-GNSA hybrid composites.

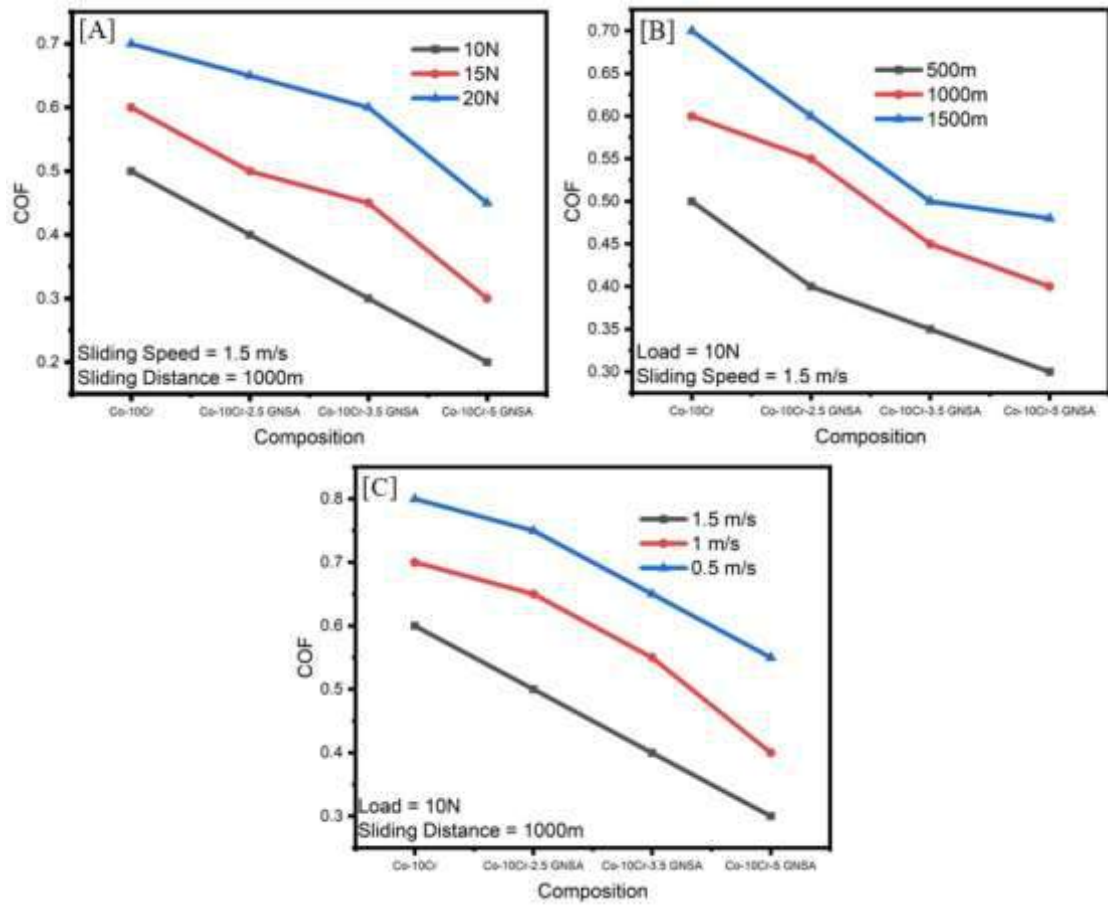
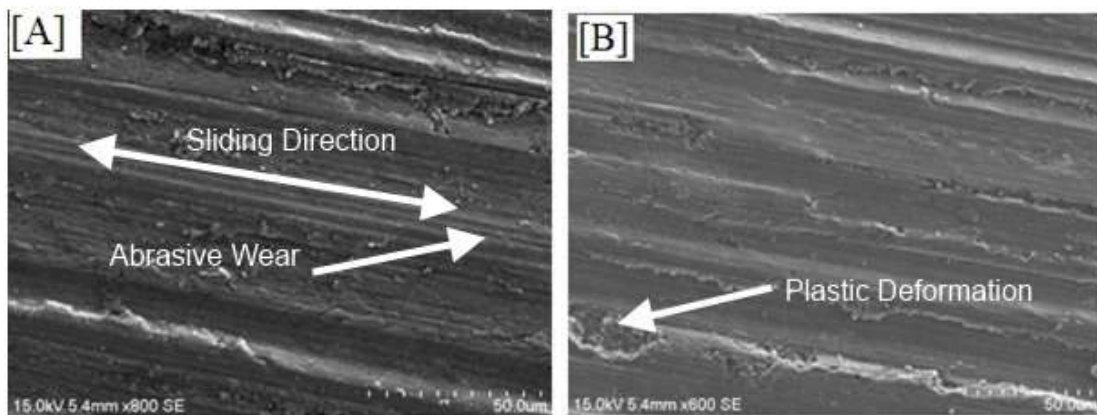


Figure 5. COF plot of Co-10Cr-GNSA hybrid composites.



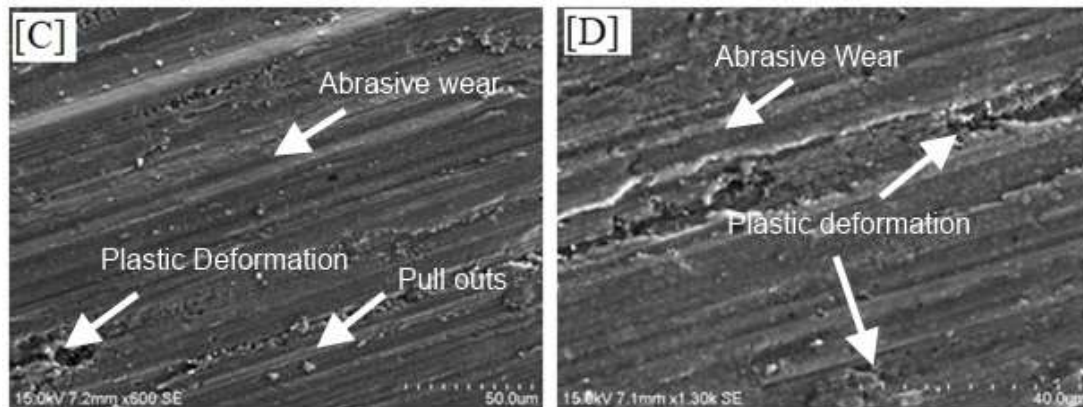


Figure 6. Worn out Surface analysis of Co-10Cr-GNSA hybrid composites.

3.5 Electrochemical Corrosion Analysis.

The corrosion analyses of the Co-10Cr-GNSA hybrid composites were done using an electrochemical work station with three electrodes. The electrolyte which was used in this study is 3% NaCl solution. The polarization curves are obtained by using tafel extrapolation methods as shown in Figure.7. The test results exhibit that the corrosion potential value, E_{corr} and the corrosion current value, I_{corr} of Co-10Cr-5GNSA hybrid composites was found to be better compared to other combinations of hybrid composites. The E_{corr} value of Co-10Cr-5GNSA hybrid composites was found to be -0.419 V and I_{corr} value was around -0.12 mA/cm². The corrosion performance of Co-10Cr-3.5 GNSA was also similar to that of Co-10Cr-5GNSA hybrid composites. The Co-10Cr composite shows lesser corrosion resistance than the hybrid composites as shown in Table.1.

Table 1. Tafel plot fallouts of Co-10Cr-GNSA hybrid composites.

S.No	Specimen	E_{corr} (V)	I_{corr} (mA/cm ²)
1	Co-10Cr	-0.442±0.051	0.5±0.020
2	Co-10Cr-2.5 GNSA	-0.437± 0.044	0.4±0.011
3	Co-10Cr-3.5GNSA	-0.420±0.021	-0.1±0.003
4	Co-10Cr-5GNSA	-0.419±0.0191	-0.1±0.002

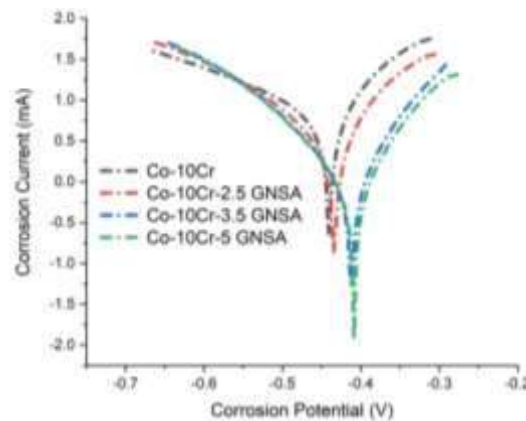


Figure 7. Potentiodynamic polarization plot of Co-10Cr-GNSA hybrid composites

4. Conclusions

The Co-10Cr-GNSA hybrid composites were studied and their mechanical, Wear and corrosion mechanisms were reported.

- The addition of GNSA reinforcement resulted in an increment in the Microhardness of the Co-10Cr-5GNSA hybrid composites (340 HV) compared to Co-10Cr composites.
- The compression strength of the Co-10Cr-5GNSA hybrid composites (401 MPa) has improved considerably than the Co-10Cr composites.
- The value of density of the Co-10Cr-5GNSA hybrid composites showed a considerable decrement due to the addition of less dense GNSA reinforcement.
- The Co-10Cr-5GNSA hybrid composites exhibited a higher resistance to wear.
- Corrosion resistance of Co-10Cr-5GNSA hybrid composites was found to be better than the Co-10Cr composites from the electrochemical corrosion analysis.

References:

- [1] Han Y, Liu F, Zhang K, Huang Q, Guo X, Wang C. *A study on tribological properties of textured Co-Cr-Mo alloy for artificial hip joints*. Int J Refract Met Hard Mater 2021;**95**:105463. <https://doi.org/https://doi.org/10.1016/j.ijrmhm.2020.105463>.
- [2] Marques FP, Scandian C, Bozzi AC, Fukumasu NK, Tschiptschin AP. *Formation of a nanocrystalline recrystallized layer during microabrasive wear of a cobalt-chromium based alloy (Co-30Cr-19Fe)*. Tribol Int 2017;**116**:105–12. <https://doi.org/10.1016/j.triboint.2017.07.006>.
- [3] Yamanaka K, Mori M, Torita Y, Chiba A. *Impact of minor alloying with C and Si on the precipitation behavior and mechanical properties of N-doped Co–Cr alloy dental castings*. Mater Sci Eng C 2018; **92**:112-120.. <https://doi.org/10.1016/j.msec.2018.06.035>.
- [4] Zhou Y, Li N, Yan J, Zeng Q. *Comparative analysis of the microstructures and mechanical properties of Co-Cr dental alloys fabricated by different methods*. J Prosthet Dent 2018:1–7. <https://doi.org/10.1016/j.prosdent.2017.11.015>.
- [5] Rodrigues WC, Broilo LR, Schaeffer L, Knörschild G, Romel F, Espinoza M. *Powder metallurgical processing of Co – 28 % Cr – 6 % Mo for dental implants : Physical , mechanical and electrochemical properties*. Powder Technol 2011;**206**:233–8. <https://doi.org/10.1016/j.powtec.2010.09.024>.
- [6] Raghav GR, Muthu Krishnan D, Sundar R, Ashokkumar R, Nagarajan KJ. *Investigation on mechanical, wear and corrosion properties of Fe-Co-Cr-W-GNSA hybrid composites synthesized using powder metallurgy process*. Eng Res Express 2020;**2**. <https://doi.org/10.1088/2631-8695/ab9517>.
- [7] Gopinath S, Prince M, Raghav GR. *Enhancing the mechanical, wear and corrosion behaviour of stir casted aluminium 6061 hybrid composites through the incorporation of boron nitride and aluminium oxide particles*. Mater Res Express 2020;**7**:016582. <https://doi.org/10.1088/2053-1591/ab6c1d>.
- [8] Prakash C, Singh S, Verma K, Sidhu SS, Singh S. *Synthesis and characterization of Mg-Zn-Mn-HA composite by spark plasma sintering process for orthopedic applications*. Vacuum 2018;**155**:578–84. <https://doi.org/10.1016/j.vacuum.2018.06.063>.
- [9] Elkhoshkhany N, Hafnway A, Khaled A. *Electrodeposition and corrosion behavior of nano-structured Ni-WC and Ni-Co-WC composite coating*. J Alloys Compd 2017;**695**:1505–14. <https://doi.org/10.1016/j.jallcom.2016.10.290>.
- [10] Stewart DA, Shipway PH, McCartney DG. *Abrasive wear behaviour of conventional and nanocomposite HVOF-sprayed WC – Co coatings*, Wear 1999;**225-229**:789–798.
- [11] Liu C, Su F, Liang J. *Nanocrystalline Co-Ni alloy coating produced with supercritical carbon dioxide assisted electrodeposition with excellent wear and corrosion resistance*. Surf Coat Technol 2016;**292**:37–43. <https://doi.org/10.1016/j.surfcoat.2016.03.027>.
- [12] Bajat JB, Vasilic R. *Corrosion Stability of Oxide Coatings Formed by Plasma Electrolytic*

- Oxidation of Aluminum : Optimization of Process Time* 2013;**69**:693–702.
- [13] Gadow R, Killinger a., Stiegler N. *Hydroxyapatite coatings for biomedical applications deposited by different thermal spray techniques*. Surf Coatings Technol 2010;**205**:1157–64. <https://doi.org/10.1016/j.surfcoat.2010.03.059>.
- [14] Raghav GR, Balaji AN, Selvakumar N, Muthukrishnan D, Sajith E. *Effect of tungsten reinforcement on mechanical, tribological and corrosion behaviour of mechanically alloyed Co-25C Cermets nanocomposites*. Mater Res Express 2019;**6**. <https://doi.org/10.1088/2053-1591/ab4f0a>.
- [15] Raghav GR, Balaji AN, Muthukrishnan D, Sruthi V, Sajith E. *An experimental investigation on wear and corrosion characteristics of Mg-Co nanocomposites*. Mater Res Express 2018;**5**:066523. <https://doi.org/10.1088/2053-1591/aac862>.
- [16] Toptan F, Alves AC, Kerti I, Ariza E, Rocha LA. *Corrosion and tribocorrosion behaviour of Al-Si-Cu-Mg alloy and its composites reinforced with B4C particles in 0.05M NaCl solution*. Wear 2013;**306**:27–35. <https://doi.org/10.1016/j.wear.2013.06.026>.



Hardfacing and its effect on wear and corrosion performance of various ferrous welded mild steels

R. Suraj

Department of Mechanical Engineering – SCMS School of Engineering and Technology, Karukutty, Ernakulam, India

ARTICLE INFO

Article history:

Received 5 October 2020

Received in revised form 13 November 2020

Accepted 18 November 2020

Available online 10 January 2021

Keywords:

Life-limiting factor

Hardfacing

Wear

Corrosion

ABSTRACT

Wear and corrosion exist as one of the main important factor of energy and material losses in mechanical and chemical process. This work is about the methods to evaluate the wear and corrosion resistant properties of the mild steel like EN-8, EN-9 and EN-24 by calculating its corrosion rate. All materials have to be analyzed for its wear properties since higher wear can lead to a machine failure. The Pin on Disc apparatus is used for the analysis. Every oil-washed system- engines, hydraulics, transmissions, and final drives- produces wear metals in everyday operation. If wear accelerates, the concentration of wear metal particles increases, signaling a problem. Wear Analysis allows us to find problems before they result in major repairs or machine failure. Prediction of the material behaviour at the increasing load is necessary for a safe working of the machines. The ferrous materials are hardfaced using Tungsten Inert Gas welding process. The wear analysis of ferrous welded materials is carried out. The various forms of mild steel selected are selected are EN 8, EN 9, EN 24. The materials are hardfaced using TIG (Tungsten Inert Gas) welding process and filler material used is same for all the materials. The materials are cut into specific dimensions using Wire cut EDM process. These specimens are tested for its wear properties, microhardness etc. Pin on Disc apparatus is used for wear analysis and Vicker's microhardness tester is used for microhardness. Similarly a corroded component results in reduced life. Corrosion results in unexpected failures of critical components. Corrosion testing is a very time-consuming process; especially in the case of outdoor atmospheric tests. Such long timescales involved in such tests prevent the opportunity for proper materials selection. The very commonly used corrosion tests are measurements of the weight loss or thickness loss. This test can be simply done in laboratory in limited period of time and thereby it's possible to predict the corrosion rate of the materials. By comparing wear and corrosion rates of hardfaced and non hardfaced surface its possible to conclude that the hardfacing improves both the wear and corrosion resistant property of these materials.

© 2020 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the scientific committee of the Second International Conference on Recent Advances in Materials and Manufacturing 2020.

1. Introduction

The word steel is used for many different alloys of iron. These alloys differ both in the way they are made and in the extent of the materials added to the iron. All steels, though, contain minute amounts of carbon and manganese. In other words, it can be said that steel is a crystalline alloy of iron, carbon and several other elements, which hardens above its critical temperature. A study on the wear properties of different grades of steel is done. The selected grades of steels are EN 8, EN 9, and MS. These grades of steels are widely used in manufacturing of different components like struc-

tural beams, car bodies, kitchen appliances, and cans. The materials are welded using hardfacing technique to learn the properties of the material. The welding method used is Tungsten Inert Gas welding technique or the Gas Tungsten Arc Welding. Corrosion is one amongst the life-limiting factor of a component. Unexpected corrosion failure can happen any time to any critical component at the worst possible instant. Corrosion testing is a very time-consuming process; particularly in the case of outside atmospheric tests. Unfortunately, the higher timescales involved in such tests prevent the chance for proper materials selection. In real life situations, the component might already be half way of their lifecycle

when identified with corrosion. A proper accelerated testing for corrosion has to be done before choosing the material for any component. Accelerated testing instead of limiting to the design stage of a system's lifecycle, can also be used to provide support at the time of identification of corrosion. At certain times, the emergence of sudden corrosion problems requires quick answers. Preventing corrosion in critical components, to extend its service life and ensure reliability, is of paramount importance. A clear-cut test data within a short span of time is required to prevent the corrosion and predict the characteristics of the material.

Hardfacing is a type of metal working process where a harder or tougher material is welded over to a base metal. The welding of the tougher material to the base material, usually is in the form of specialized electrodes for arc welding or filler rod for oxy-acetylene and TIG welding. Hardfacing with the help of arc welding is a kind of surfacing operation to extend the operation time of critical industrial components, especially on new components, or during maintenance program.

Hardfacing is a low cost method of depositing wear and corrosion resistant surfaces usually by welding on metal components to extend service life. It is primarily used to restore worn parts to usable condition, but hardfacing is also applied to new components before being placed into service to get a long service life thereby reducing the cost of maintenance.

Welding material selection depends upon three major factors:

1. Base Metal – Primarily affects the choice of build-up materials.
 - a. Manganese steel is used for components subject to high impact loading. Rebuild to size using manganese steel weld deposits.
 - b. Carbon and alloy steel components are rebuilt to size using low alloy steel weld deposits.
2. Type of Wear – The primary consideration in selecting the final hardfacing layers is the type of wear to be encountered in service.
3. Corrosion – Chemical attack.

TIG welding is the currently used method for hardfacing the metal surface. The majority of the of researchers concentrates their work on reducing the wear rate of material by improving the wear resistance by addition of alloy elements in the base material [1]. As we know that the wear is surface phenomena it only occurs on a surface of the material, surface modification is the most common and economical way to improve the wear resistance of a material [1]. Hardfacing is a metalworking process where harder material applied to the base material with the help of different welding processes like Arc welding, TIG welding and plasma Arc welding processes. This process is called Hardfacing because the deposited surfaces are harder than the base metal usually [2]. Hardfacing is generally used to improve the surface property of the material. An alloy is homogeneously deposited onto the surface of a soft material by welding, to increase hardness and wear resistance without significant loss in ductility and toughness of the substrate [3]. The hard-facing alloy is applied to the material to achieve high wear resistance and better properties [4]. Mild steel is the most commonly used steel. It is the combination of carbon, manganese, phosphorus, Sulphur, and silicon. It is low carbon steel; Mild steel is very much suitable as structural steel Mild steel is widely used in bolted, riveted or welded construction of bridges, building it is also used in forming tanks, bearing plate, fixture, sprockets, cams, gears, base plates, forging, brackets, automotive and agricultural equipment, machinery parts. Augustin Gualco, [4] perform hardfacing with help of FCAW on iron-based alloy and conclude that 2-layer welding gives higher wear resistance. G.R.C Pradeep [5] perform hardfacing with 3 different welding processes Tig, Arc and Gas welding and conclude that the Arc and Gas welding sam-

ples yielded better welding property. Harvinder Singh [1] perform hardfacing process with 3 different electrode Hardalloy400, Hard alloy-III and Hard alloy-V and conclude that Hard-alloy V gives better hardness compare to another electrode. Z. Horvat, [6] used SMAW and Induction welding for hardfacing and conclude that the Weight loss due to erosion was lower on both the ploughshares as compared to standard shares. John J. Coronado, [7] used SMAW and FCAW and found that FCAW gives higher Abrasion wear resistance rate than the SMAW. Patrick W. Leech, [8] used the high alloy (SHS9290) & tungsten carbide-Ni-based matrix composite with SMAW welding and found that the SHS9290 alloy has lower wear rate than the WC- Ni-based MMC in the dry sand rubber wheel tests and pin-on- flat tests using garnet abrasive. Amardeep Singh Kang, [9] performed MMAW process on spring steel (EN-45A) with 3HCr, 8HCr 10HCr, 18HCr electrode and found the wear rate of hard-faced material was lower 18HCr hardfacing electrode gives higher hardness and maximum wear resistance. Hülya Durmus., [10] used arc welding on St37 for Hardfacing processes with Fe-Cr-C-B, Fe- Cr-C contains electrode and found out that the wear resistance is not only correlated with hardness but also affected by the morphology of microstructural constituents. S. Sittthipong, [11] used MAG, FCAW SMAW, welding on Propeller Shaft AISI with X111T5- K4, ER110S-G and E11018-GH4R, electrode and conclude that the grain structure of weld metal by FCAW was finer and harder than the other welding also at the weld zone structure are fine then the

HAZ Vickers test FCAW Produced higher hardness value than other welding Resistance of abrasive wear is higher in FCAW Welding. M. Kirchgäßner, [12] used the GMAW process with the help of Fe-Cr- C-Nb hardfacing alloy and found out Fe-Cr-C-Nb alloys provide good wear behaviour under all test. G.P. Rajeev, [13] used AISI H13 die steel with CMT welding and Stellite 21 alloy for hardfacing and found that Stellite coated H13 Plate could be subjected to quenching and tempering heat treatment to restore the properties of the welding layer without defects. In present work, the arc welding process is used to perform hardfacing processes on ASTM A-36 Mild Steel. In this study, two different hardfacing electrode Zed alloy 550, and Nikko steel Hv-600 are used to prepare different samples. With the help of Pin on Disk wear testing machine wear rate of different samples was investigated, also microhardness and microstructure of the samples are investigated simultaneously.

2. Experimental details

2.1. Material selection

The materials selected are EN 8, EN 9, and EN24, whose chemical composition is already discussed. These materials are chosen because of the immense applications of these materials in various engineering application and less amount of study which is done in this field. The properties of these materials also make it special and durable. These materials fall under the category of mild steel. All of the above materials have carbon in the range of 0.4%. The amount of carbon in these materials makes it corrosive and thus the materials are prone to corrosion to a large extent. A study is also made to analyse the properties regarding the hardness of the material. The property of the material is found unaltered after hardfacing the material with the TIG welding filler wire. The filler wire used here is ER70S2. This is usually used for the welding of mild steel.

2.1.1. En 8

EN8 is an unalloyed medium carbon steel with good tensile strength. It is normally supplied in cold drawn or as rolled. Tensile properties can vary but are usually between 500 and 800 N/mm².

EN8 is available from stock in bar and can be cut to our requirements. Table 1 shows the various Composition of EN 8.

2.1.2. En9

EN9 is an unalloyed medium carbon steel. It is supplied at the hardness obtained after hot rolling or cold drawing, with hardness normally within the range of 180 to 230HB. EN9 is available from stock in bar and can be cut to your requirements. Also EN9 can be found in plate form and its flame cut to required sizes and normalised. Table 2 shows the main Composition of EN 9

2.1.3. En 24

EN24 is a high quality, high tensile, alloy steel. Usually supplied readily machine able in 'T' condition, it combines high tensile strength, shock resistance, good ductility and resistance to wear. MS is available from stock in round bar, flat bar and plate Table 3. shows the various Composition of EN 24

2.2. Selection of electrode

ER 70S-2.

ER 70S-2 is a copper coated GTAW rod containing Al, Ti and Zr as strong deoxidants in addition to Mn and Si and is often referred to as triple deoxidised. This has advantages when rimming or semi-killed mild steels are welded or where joint preparations are rusty or contaminated. Fig. 1 show the ER 70S-2 filler rod. ER 70S-2 is primarily used for single pass welding. Table 4 shows the Composition of ER 70S-2.

2.3. Welding details and parameters

The welding process used here is Tungsten inert Gas welding process also known as Gas Tungsten Inert Gas welding process.

The same filler material ER 70S-2 is taken for the entire specimen. The specimen size is 12 X 50 X 50 mm. The hardfacing is to be done on the 12 X 50 mm face. The welding speed doesn't affect the material much as the required are of welding is small. So the speed is not taken into consideration. The material after hardfacing is grinded to get a smooth surface using manual grinding operation. Each specimen has 3 samples making it a total of 9 samples. 3 for EN 8, 3 for EN9 and 3 for EN 24 specimen.

Table 1
Composition of EN 8.

COMPOSITION				
C.	Si.	Mn.	S.	P.
0.40%	0.25%	0.80%	0.015%	0.015%

Table 2
Composition of EN 9.

COMPOSITION				
C.	Si.	Mn.	S.	P.
0.50%	0.25%	0.70%	0.05%	0.05%

Table 3
Composition of EN 24.

COMPOSITION			
C.	Si.	Mn.	Cu
0.19%	0.6%	1.65%	0.6%



Fig. 1. ER 70S-2 filler rod.

Table 4
Composition of ER 70S-2.

Material	Composition
C	0.05
Si	0.5
Mn	1.2

For Corrosion testing, the hardfaced surface was cut for 25 mm X 10 mm X 10 mm using wire cut EDM and it was employed for the test. A hole was drilled near the upper edge of the specimen in order to hook it on to the glass rod for immersion. The Specimen are polished with emery sheet, degreased and washed with distilled water. The specimens are stored in desiccators in the absence of moisture before their use for the investigation.

The welding parameter variation with respect to the specimen are shown in Table 5, 6 and 7.

2.4. Sample preparation and testing

In this Experiment Mainly Three Test was conducted after applying the hard-facing layer,

A. Wear Analysis:

1. Micro hardness test.
2. Wear test Pin on Disk (POD).
3. Microstructure Examination.

B. Corrosion analysis:

1. Weight loss method.

3. Result and discussion

3.1. Wear analysis

The first testing method deals with wear analysis. Many types of wear analysis testing are done.

3.1.1. Microhardness testing

The welded specimen was tested for its microhardness prior to wear analysis. Due to the carbon content in the specimen the specimen is expected to exhibit a high hardness value. The test for microhardness is carried out in a Vicker's Microhardness testing apparatus. The arithmetic mean value of the diameter is automatically measured and displayed. The test is carried out for a load of 1000gms. Tables 8-10 shows Microhardness test results for EN 8, EN9 and EN 24 respectively.

3.1.2. Wear test Pin on Disk (POD)

The wear analysis was done for a time period of 10 min for each specimen and the respective wear of the material was noted with respect to the time. The test was carried out at 200 rpm with a load of 74 N. The mass loss of the material if noted as test proceeds. The track diameter of the disc is 100 mm.

Table 5
Welding Parameters for EN 8 specimen.

PARAMETER	TRIAL 1	TRIAL 2	TRIAL 3	TRIAL 4	TRIAL 5
VOLTAGE (volts)	15	15	15	15	15
CURRENT (amperes)	100	105	110	115	120
GAS FLOW RATE(lps)	6	6.5	8	7.5	6.3
GAS	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium
FILLER MATERIAL	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2
ELECTRODE	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)

Table 6
Welding parameters for EN 9.

PARAMETER	TRIAL 1	TRIAL 2	TRIAL 3	TRIAL 4	TRIAL 5
VOLTAGE (volts)	15	15	15	15	15
CURRENT (amperes)	125	115	110	100	120
GAS FLOW RATE(lps)	6.2	8.5	8	9	7.2
GAS	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium
FILLER MATERIAL	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2
ELECTRODE	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)

Table 7
Welding parameters for EN 24.

PARAMETER	TRIAL 1	TRIAL 2	TRIAL 3	TRIAL 4	TRIAL 5
VOLTAGE (volts)	15	15	15	15	15
CURRENT (amperes)	103	110	115	120	125
GAS FLOW RATE(lps)	7.3	8.1	8	6	9
GAS	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium	Argon + Helium
FILLER MATERIAL	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2	ER 70S-2
ELECTRODE	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)	Thoriated Tungsten (2.4 mm dia)

Table 8
Microhardness test results for EN 8.

SI NO	LOAD (F) kgf	MEAN DIAMETER (d) mm	VICKERS HARDNESSHV = $\frac{2F\sin136/2}{d^2}$
1	1	0.0799	290

Table 9
Microhardness test results for EN 9.

SI NO	LOAD (F) kgf	MEAN DIAMETER (d) mm	VICKERS HARDNESS HV = $\frac{2F\sin136/2}{d^2}$
1	1	0.08123	281

Table 10
Microhardness test results for EN 24.

SI NO	LOAD (F) kgf	MEAN DIAMETER (d)	VICKERS HARDNESS HV = $\frac{2F\sin136/2}{d^2}$
1	1	0.080	285

The observation for the respective specimen is as shown below: Tables 11-13 shows Wear analysis on EN9, EN8 and EN24 respectively. Fig. 2, Fig. 3, Fig. 4 shows the Variation of wear with rest to time for EN8, EN9 and EN 24 respectively.

Wear analysis shows that EN 9 is having maximum wear. The wear is least for EN24.

Table 11
Wear analysis on EN9.

SI NO	TIME (sec)	WEAR (μm)	FRICTION FORCE (N)	COEFFICIENT OF FRICTION
1	0	0.38	3.45	0.53
2	60	102.34	41.98	0.53
3	120	142.34	34.89	0.54
4	180	245.34	34.76	0.53
5	240	250.56	35.09	0.54
6	300	293.55	36.13	0.54
7	360	323.36	37.03	0.56
8	420	405.11	37.86	0.55
9	480	540.22	40.87	0.55
10	540	670.21	38.97	0.56
11	600	700.74	45.67	0.56

Table 12
Wear analysis on EN8.

SI NO	TIME (sec)	WEAR (μm)	FRICTION FORCE (N)	COEFFICIENT OF FRICTION
1	0	-14.67	2.08	0.63
2	60	10.78	36.57	0.62
3	120	41.39	40.03	0.63
4	180	265.32	38.11	0.63
5	240	440.95	45.98	0.64
6	300	367.9	43.67	0.64
7	360	541.88	40.49	0.66
8	420	513.53	41.5	0.65
9	480	540.22	42.76	0.62
10	540	532.82	42.34	0.63
11	600	645.45	40.08	0.63

Table 13
Wear analysis of EN24.

SI NO	TIME (sec)	WEAR (μm)	FRICTION FORCE (N)	COEFFICIENT OF FRICTION
1	0	0	1.3	0.61
2	60	98.12	42.97	0.64
3	120	100.34	43.52	0.64
4	180	110.71	42.19	0.63
5	240	111.04	43.98	0.66
6	300	133.25	45.1	0.68
7	360	259.45	44.97	0.59
8	420	298.22	44.68	0.61
9	480	244.83	43.44	0.63
10	540	353.83	44.51	0.65
11	600	400.89	43.83	0.76

HARDNESS COMPARISON

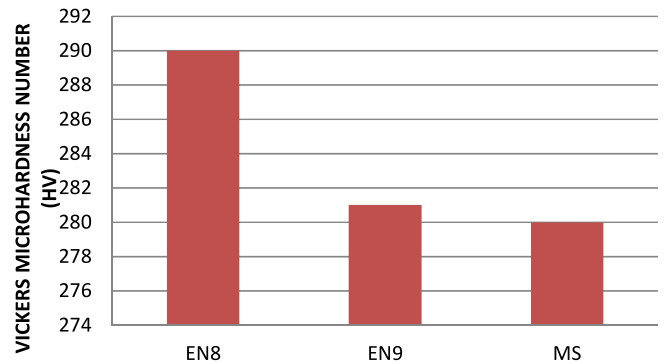


Fig. 5. Variation of hardness with respect to the material.

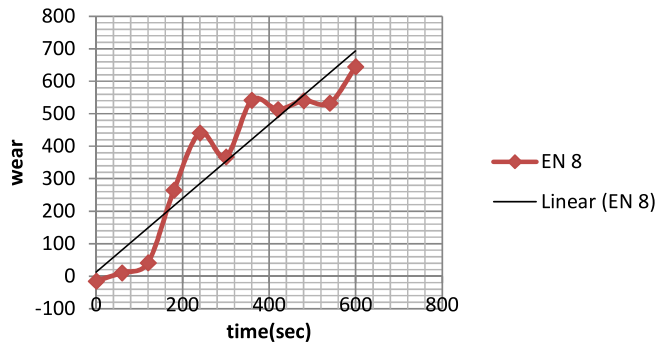


Fig. 2. Variation of wear with rest to time for EN8.

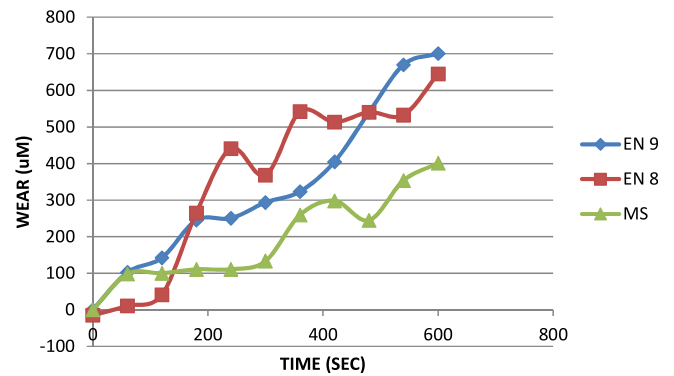


Fig. 6. Comparison of wear for EN 8, EN 9, MS.

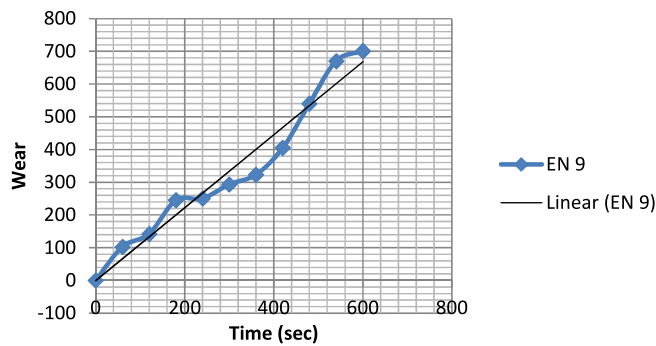


Fig. 3. Variation of wear with respect to time for EN9.

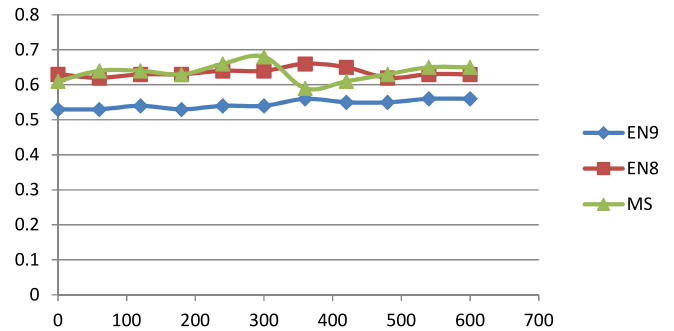


Fig. 7. Comparison of frictional coefficients for the materials..

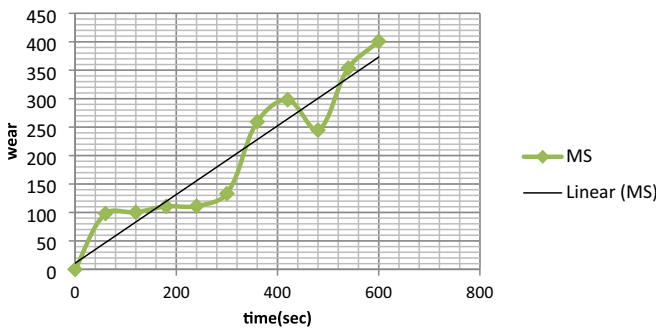


Fig. 4. Variation of wear with respect to time for EN24.

3.1.3. Hardness test

Fig. 5 shows the Variation of hardness with respect to the material. The hardness of EN24 is found to be higher than EN8 and EN9. The microhardness test results prove this. It is expected to have less wear when the hardness is more. The wear analysis suggests

the same. MS has the least wear when compared to the other specimens. The comparison of wear behaviour of the 3 materials is as shown in figure below. The results say that the wear of the three materials selected are increasing with respect to the time. As time increases the wear of the material keep on increasing. The increment is almost linear as shown in the figure. Fig. 6 shows the Comparison of wear for EN 8, EN 9, EN 24.

The comparison of frictional coefficient of the specimens selected is shown below. The variation of frictional coefficient is seen to be almost constant for the materials. MS is found to have the highest frictional coefficient. Fig. 7 shows the Comparison of frictional coefficients for the materials.

3.1.4. Microsturcture analysis

The microstructure of the specimens is studied respectively on the hardfaced area, parent material and the worn area. The image is captured using an image analysis system. The analysis is done using the Quantiment image analysis software.

The images obtained are shown in Fig. 8, Fig. 9, Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, Fig. 16.

The observations from the microstructure analysis are tabulated in Table 14. EN 24 worn surface is having the higher percentage of porosity. The percentage porosity of EN 24 worn surface is 91.11% and the volume fraction is 0.29.



Fig. 8. Microstructure of the base EN8.



Fig. 9. Microstructure of base EN 9.



Fig. 10. Microstructure of base EN 24.



Fig. 11. Microstructure of worn surface of EN 8

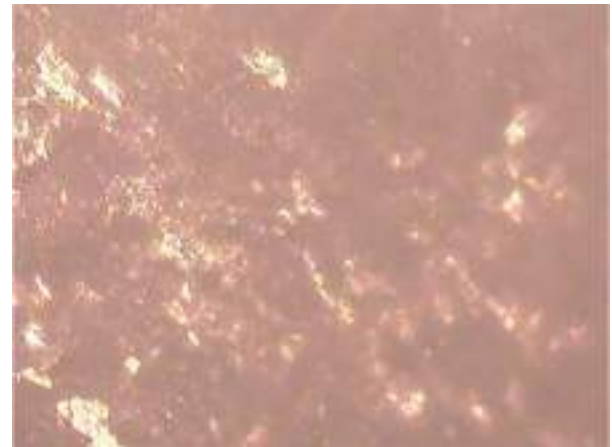


Fig. 12. Microstructure of worn surface of EN 9.



Fig. 13. Microstructure of worn surface of EN 24.

3.2. Corrosion analysis

Sulphuric acid is used directly or indirectly in all industries and is a vital commodity in our national economy. The widespread use of this acid places it in an important position with regard to costs and destruction of corrosion. In some cases, corrosion increases

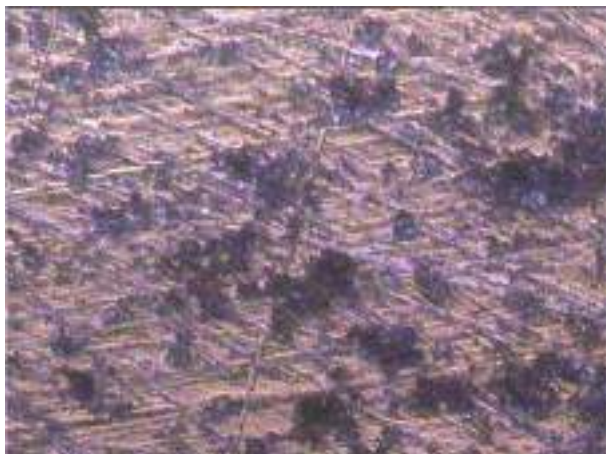


Fig. 14. Microstructure of welded area of EN 8.



Fig. 15. Microstructure of welded area of EN 9.



Fig. 16. Microstructure of welded area of EN 24.

with concentration of the acid and in others it decreases. For these reasons, it is important to have a good picture of corrosion by sulphuric acid. Therefore the present experiment was carried out by using 1 N sulphuric acid of commercial grade. The 1 N sulphuric acid solution was prepared by mixing 28 ml of commercial grade sulphuric acid in 1000 ml of distilled water.

Table 14
Observations from microstructure analysis.

s	MATERIAL	GRAIN SIZE	PERCENTAGE POROSITY	VOLUME FRACTION
1	EN 8 base metal	5.71	78.06	0.22
2	EN 9 base metal	8.23	0.52	0.29
3	EN 24 base metal	9.08	56.82	0.23
4	EN 8 welded area	8.7	55.98	0.25
5	EN 9 welded area	8.52	45.41	0.24
6	EN 24 welded area	9.06	57.16	0.23
7	EN 8 worn surface	6.95	0.37	0.36
8	EN 9 worn surface	9.79	0.38	0.36
9	EN 24 worn surface	4.92	91.11	0.29

3.2.1. Corrosion rate expressions

Metals and nonmetals will be compared on the basis of their corrosion resistance. To make such comparisons meaningful, the rate of attack for each material must be expressed quantitatively. Corrosion rates have been expressed in a variety of ways in the literature; such as percent weight loss, milligrams per square centimeter per day and grams per square cm per hour. These do not express corrosion resistance in terms of penetration. From an engineering viewpoint, the rate of penetration or the thinning of a structural piece can be used to predict the life of a given component.

The expression mil per year (mpy) is the most desirable way of expressing corrosion rates. This expression readily calculated from the weight loss of the metal specimen during the corrosion test by the formula given below:

$$mpy = 534W/DAT \tag{1}$$

Where, W = Weight loss, mgD = Density of specimen, g/cubic cmA = Area of specimen Sq. cmT = Exposure Time, hr

Thus corrosion rate calculation involves whole numbers, which are easily handled.

The corrosion analysis test using weight loss method is carried out according to the ASTM standards. Table 15 shows the Corrosion Rate of different materials.

By comparing the corrosion rates of the three metals before and after the hardfacing using TIG welding. The following results are obtained and it is plotted as graphs below in Fig. 17.

From the above graph it is clear that the two meals (EN-9 and MS) in its original condition has similar corrosion rate and among this EN-8 shows more corrosion rate.

Next part is to compare the corrosion rates of parent and the hardfaced metal. The effect of hardfacing on these three metal shows an improvement in the corrosion resistant properties. The three comparison graphs are shown below in Fig. 18, Fig. 19 and Fig. 20 Fig. 21.

After comparing with its parent material, it need to going to compare the three welded materials and following result are obtained.

4. Conclusion

Steel being an important constituent in industrial applications has to be studied for its durability and worthiness. This study gives

Table 15
Corrosion Rate.

Material	1 h (mpy)	3 h (mpy)
EN-8	108.436	195.787
EN-9	105.945	194.081
MS	192.533	287.695
EN-8 (Weld)	87.484	129.768
EN-9 (Weld)	87.20	141.256
MS (Weld)	169.428	244.708

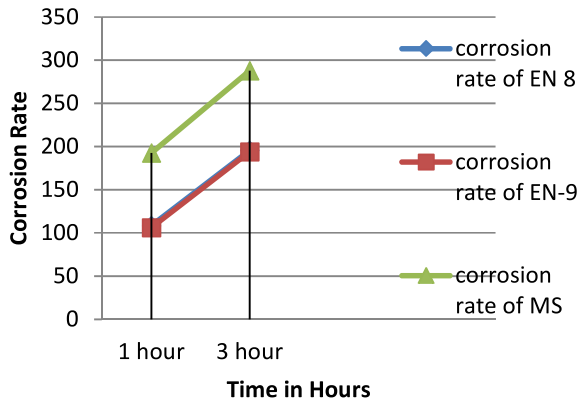


Fig. 17. Corrosion rate.

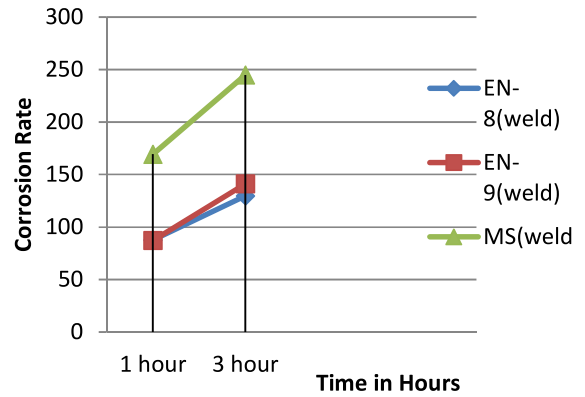


Fig. 21. Comparison of Welded Materials.

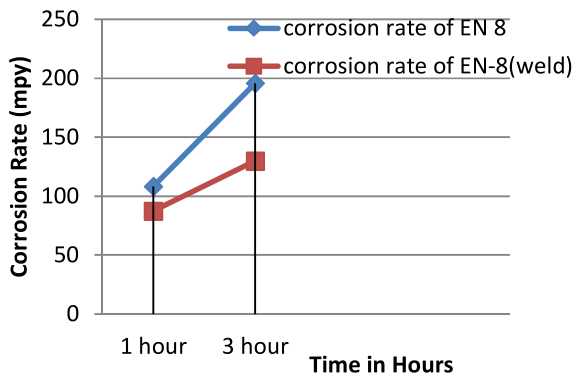


Fig. 18. Comparison of Corrosion Rate of Parent Material and Weld of EN-8.

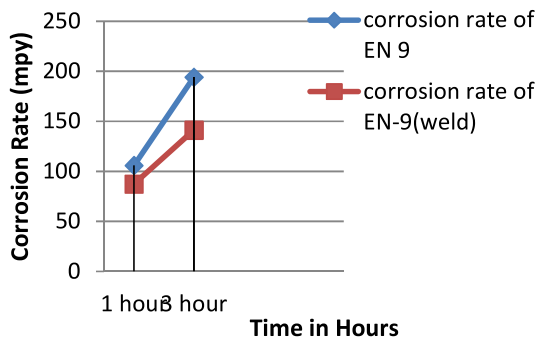


Fig. 19. Comparison of Corrosion Rate of Parent Material and Weld of EN-9.

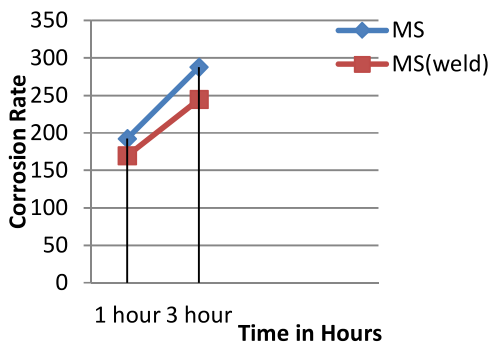


Fig. 20. Comparison of Corrosion Rate of Parent Material and Weld of MS.

an idea regarding the durability of the materials. The properties of the materials are studied. The material due to its wide range of applications plays a vital role in modern day industry so the study suggests a better option among the three selected materials. Further studies on the properties can give a clear idea about the selection.

- The wear properties of ferrous welded materials like EN 8, EN 9, and EN 24 are studied.
- It is found the EN24 has the least wear when compared to EN 8 and EN 9.
- The microhardness of EN24 is higher than EN 8 and EN 9 thus making it wear resistant than EN 8 and EN 9
- The coefficient of friction in dry sliding condition is found to be constant throughout the experiment.
- The hardfaced material has much more corrosion resistant capability than the parent material.
- Hardness of three materials varies accordingly with the chemical composition.

CRedit authorship contribution statement

R. Suraj: Conceptualization, Methodology, Software, Data curation, Writing - original draft, Visualization, Investigation, Supervision, Software, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] Harvinder Singh, Studies the effect of iron based hardfacing electrodes on stainless steel properties using shielded metal arc welding process, Inter. J. Res. Adv. Technol. 2 (4) (April 2014).
- [2] B.V. Cockeram, Some observations of the influence of d -ferrite content on the hardness, galling resistance, and fracture toughness of selected commercially available iron-based hardfacing alloys, Metall. Mater. Trans. 33A (2002) 3403.
- [3] S. Chatterjee, T.K. Pal, Wear behaviour of hardfacing deposits on cast iron, Wear 255 (1-6) (2003) 417–425.
- [4] A. Gualco, C. Marini, H. Svoboda, E. Surian, Wear resistance of Fe-based nanostructured hardfacing, Procedia Mater. Sci. 8 (2015) 934–943.
- [5] G.R.C. Pradeep, A. Ramesh, B. Durga Prasad, Comparative study of hardfacing of aisi 1020 steel by three different welding processes, Global J. Res. Eng. XIII (IV) (2013) 10–16.
- [6] Z. Horvat, D. Filipovic, S. Kosutic, R. Emert, Reduction of mouldboard plough share wear by a combination technique of hardfacing, Tribol. Int. 41 (8) (2008) 778–782.

- [7] J.J. Coronado, H.F. Caicedo, A.L. Gómez, The effects of welding processes on abrasive wear resistance for hardfacing deposits, *Tribol. Int.* 42 (5) (2009) 745–749.
- [8] P.W. Leech, X.S. Li, N. Alam, Comparison of abrasive wear of a complex high alloy hardfacing deposit and WC–Ni based metal matrix composite, *Wear* 294 (2012) 295.
- [9] H. Durmuş, N. Çömez, C. Gül, M. Yurddaşkal, M. Yurddaşkal, Wear performance of Fe–Cr–C–B hardfacing coatings: Dry sand/rubber wheel test and ball-on-disc test, *Int. J. Refract. Metals Hard Mater.* 77 (2018) 37–43.
- [10] S. Sitthipong, P. Towatana, A. Sitticharoenchai, C. Meengam, Abrasive wear behavior of surface hardfacing on propeller Shafts AISI 4140 Alloy steel, *Mater. Today: Proceed.* 4 (2) (2017) 1492–1499.
- [11] M. Kirchgäßner, E. Badisch, F. Franek, Behaviour of iron-based hardfacing alloys under abrasion and impact, *Wear* 265 (2008) 772–779.
- [12] <http://www.matweb.com/search/datasheet.aspx?matguid=d1844977c5c84-40cb9a3a967f8909c3a&ckck=1>.
- [13] <https://www.steeltank.com> (American Welding Society).

Further Reading

- [1] Amardeep Singh Kang, Gurmeet Singh Cheema, Shivali Singla, Wear behaviour of hardfacings on rotary tiller blades, *Procedia Eng.* 97 (2014) 1442–1451.
- [2] G.P. Rajeev, M. Kamaraj, S.R. Bakshi, Comparison of microstructure, dilution and wear behavior of Stellite 21 hardfacing on H13 steel using cold metal transfer and plasma transferred arc welding processes, *Surf. Coat. Technol.* (2019.07.019).
- [3] <http://www.adorwelding.com>.
- [4] <https://www.nikkosteel.com>.
- [5] M.F. Buchely, J.C. Gutierrez, L.M. León, A. Toro, The effect of microstructure on abrasive wear of hardfacing alloys, *Wear* 259 (1–6) (2005) 52–61.
- [6] R. Colaço, R. Vilar, A model for the abrasive wear of metallic matrix particle-reinforced materials, *Wear* 254 (7–8) (2003) 625–634.
- [7] K. Van Acker, D. Vanhoyweghen, R. Persoons, J. Vangrunderbeek, Influence of tungsten carbide particle size and distribution on the wear resistance of laser clad WC/Ni coatings, *Wear* 258 (1–4) (2005) 194–202.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Dispersion analysis of nanofillers and its relationship to the properties of the nanocomposites

Gibin George^{a,*}, Amal P. Dev^b, N. Nikhil Asok^a, M.S. Anoop^b, S. Anandhan^{c,*}

^aDept. of Mechanical Engineering, SCMS School of Engineering and Technology, Pallissery, Ernakulam, Kerala, India

^bDept. of Automobile Engineering, SCMS School of Engineering and Technology, Pallissery, Ernakulam, Kerala, India

^cDept. of Metallurgical and Materials Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India

ARTICLE INFO

Article history:

Received 3 March 2021

Received in revised form 8 May 2021

Accepted 12 May 2021

Available online xxxxx

Keywords:

Nanocomposite

Crystallization

Halloysite nanotube

Image processing

ABSTRACT

The dispersion and distribution characteristics of the reinforcements are the key reasons that influence the mechanical properties of the nanocomposites. In this paper, the dispersion and distribution analysis of nanofillers in a representative polymer is performed and the results are correlated to the crystalline and mechanical properties of the nanocomposite. The nanocomposite used in the present study is Elvaloy[®]4924 (EVACO)/halloysite nanotubes (HNTs) composite. The dispersion of halloysite nanotubes in the EVACO matrix is recorded as aluminum elemental maps obtained from energy dispersive spectroscopy (EDS). The dispersion and distribution of fillers in the composite are quantified using an image processing technique and it is correlated to the crystalline and tensile properties of the composites. The better dispersion and distribution of HNTs at 1wt.% filler loading resulted in a remarkable improvement in the crystallinity of the composite, which is measured by X-ray diffraction (XRD) and differential scanning calorimetry (DSC). The tensile strength was highest for composites loaded with 1 wt.% filler, and the strength decayed as the loading was further increased. Agglomeration of halloysite nanotubes and polymer-filler debonding was the major reason behind the reduction in tensile strength with filler loading, as observed in the scanning electron micrographs of the fractured surfaces.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

1. Introduction

The addition of nanomaterials in polymer matrices is extensively used for refining the thermal, mechanical, flame resistance, and electrical characteristics of polymers. The improvement in these properties depends on the nature of the nanofiller used and their interaction with the respective polymer matrix. The naturally existing nanomaterials which are abundant, low cost, and non-toxic are used in the bulk production of polymer nanocomposites (PNCs) [1–3]. Clay and its minerals are the most commonly used nanofillers in PNCs because of their large availability, good interaction with the matrix, and ability to exfoliate into two-dimensional nanostructured layers [4].

The specific surface area of the filler and the interaction at the interface of the matrix and the filler have a vital role in improving the properties of the composite. In nanocomposites, the interaction between the polymer and the nanofiller is better than their micro-sized counterparts due to the large surface area of the nanoparticles [5]. The high surface area and associated high surface energy of the nanomaterials reduce the number of nanoparticles required to achieve a significant improvement in the properties of the composite [6]. As compared with the micro-sized particles, the quantity of nanoparticles required is only 1/100th to achieve the same properties in the composite [7]. The interaction of the fillers and polymer matrix in a polymer composite is a function of the surface area of the nanoparticles, and the nanoparticles have a large surface area as compared to the micro-sized counterparts of the same weight.

In polymer matrices, the addition of nanofillers exhibits a simultaneous enhancement in several properties of the matrix. Carbon nanotubes, for example, can simultaneously enhance polymer matrices' crystallization [8], tensile properties [9],

* Corresponding authors.

E-mail addresses: gibingeorge@scmsgroup.org (G. George), amaldev@scmsgroup.org (A.P. Dev), nikhil@scmsgroup.org (N.N. Asok), anoopms@scmsgroup.org (M.S. Anoop), anandtmg@gmail.com, anandhan@nitk.edu.in (S. Anandhan).

<https://doi.org/10.1016/j.matpr.2021.05.285>

2214-7853/© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

conductivity [10,11], and UV stability [12]. Similarly, besides mechanical properties, clay and layered hydroxides can improve the flame retardancy [13] and barrier properties [14,15] of the polymers. The properties exerted by a certain nanofiller are unique for a polymer matrix composite that cannot be replicated by another polymer/nanofiller combination.

EVACO is a semi-crystalline polymer and the presence of HNTs can improve the crystallinity of EVACO as the same is observed in many other semi-crystalline polymer matrices. The presence of carbonyl groups in the backbone of EVACO increases the polarity and thereby its affinity for metallic surfaces [16]. Halloysite is a halogen and phosphorous free flame retardant and the water molecules present between the SiO_4 and AlO_6 layers [17] will dilute the free radicals or the reactive species at the flame front to enter the flame as the combustion begins. Halloysite nanotubes have also found applications, in controlled drug release [18] and protective agents [19], fillers [20–23], emulsifiers [24], adsorbents for pollutants [17], etc.

The characteristics of the polymers depend on their chain dynamic [25] and the motions of the chains are also influenced by nanofillers. The nucleating ability of the nanofillers, at low filler loading, improves the crystallinity of polymer matrices [21,26]. The solution cast EVACO/HNT composites were extensively studied using different characterization techniques. In this work, the effect of the position of HNTs on the crystalline and mechanical properties of EVACO matrix is studied by image processing of elemental mapped electron micrographs. The tensile properties and crystallinity of the composites are affected by the dispersion and distribution of HNTs. The presence of HNTs in the EVACO matrix improved the flame retardancy of the composite.

2. Materials and methods

The materials for the present study were Elvaloy[®]4924 (EVACO), HNTs, and dichloromethane (DCM), which are purchased from Du Pont India., Sigma Aldrich, India, (product ID: 685445) and Central Drug House (P) Ltd, India, respectively. To prepare the composites, a known quantity of EVACO was dissolved by constant stirring (at a speed of ~700 rpm) into a fixed quantity of DCM using a magnetic stirrer in a closed beaker. A known quantity of HNTs was well dispersed in a small part of DCM by stirring and subsequent ultrasonication for 30 min. The above solution of HNTs in DCM was combined with the former EVACO solution by stirring followed by ultrasound treatment. The mixture was then poured onto glass petri dishes to create the respective composite films and are allowed to dry at room temperature and then in a vacuum oven at 50 °C for 6 h. Composites films with HNT loadings 1, 3, 5, 7, and 10 wt.%, respectively were prepared.

Energy-dispersive X-ray spectroscopy (EDS) (Link ISIS-300, Oxford Instruments, UK) was used to map the aluminum in the composite, each aluminum dot corresponds to HNTs, and Image-J software [27] to analyse the maps. X-ray diffraction patterns (JEOL, DX-GE-2P, Japan) of the composite sheets were analyzed using CuK_α radiation to determine the crystallinity of nanocomposites. The percentage of crystallinity (X_c) was estimated by deconvoluting the XRD patterns to amorphous and crystalline contributions, and the extend of crystallinity was estimated by the ratio [28].

$$X_c = \frac{I_c}{I_a + I_c} \quad (1)$$

where I_a and I_c represent the integrated intensities of the amorphous and crystalline regions in EVACO, respectively.

Fourier transforms infrared (FTIR) spectra (Jasco FTIR 4200, Japan) of the EVACO and the representative composite were

recorded in ATR mode in a wavenumber range of 650–4000 cm^{-1} . Thermogravimetric measurements (TGA Q5000, TA instruments, USA) were performed under nitrogen atmosphere for the samples under a nitrogen flow of 25 mL min^{-1} and at a constant heating rate of 10 $^\circ\text{C min}^{-1}$. Differential scanning calorimetric measurements (DSC) (Mettler Toledo DSC, USA) were carried out in a nitrogen atmosphere between 0 and 150 $^\circ\text{C}$ at a heating rate of 10 $^\circ\text{C min}^{-1}$. The extend of crystallinity of EVACO and the composites were determined from the area under the endothermic curve, using the equation [29]:

$$X_c = \frac{\Delta H_f}{W_i \times \Delta H_{f100\%}} \times 100 \quad (2)$$

where W_i = weight fraction of the polymer, X_c = crystallinity (%), ΔH_f = enthalpy of melting of completely crystalline EVACO (J/g), $\Delta H_{f100\%}$ = enthalpy of crystallization of a 100% crystalline sample of EVA = 68 Jg^{-1} [30].

The tensile measurements (Hounsfield Universal Testing Machine, H25KS, Hounsfield, UK) at ambient conditions were made for three dumb-bell samples, prepared according to ASTM D 412-B. The fractured surfaces were analyzed using a scanning electron microscope (SEM) (JSM-6380LA, JEOL, Japan). The specimens were sputtered with gold (JEOL JFC 1600) in an auto fine coater before imaging.

3. Results and discussion

3.1. Dispersion and distribution of HNTs in EVACO

To find the dispersion of HNTs, each dot in the aluminum elemental map was considered as an HNT and a sparse sampling technique was employed. The aluminum elemental map of a representative composite is shown in Fig. 1. In sparse sampling, the elemental maps were divided into 20 equal sections and the numbers of particles in each section were counted. The average number of particles per unit area was calculated and the respective standard deviation was estimated for each composite. A large standard deviation shows a poor dispersion and the standard deviation was calculated using the following equation [31]:

$$\sigma = \sqrt{\frac{1}{n} \sum_i (N_{Ai} - \bar{N}_A)^2} \quad (4)$$

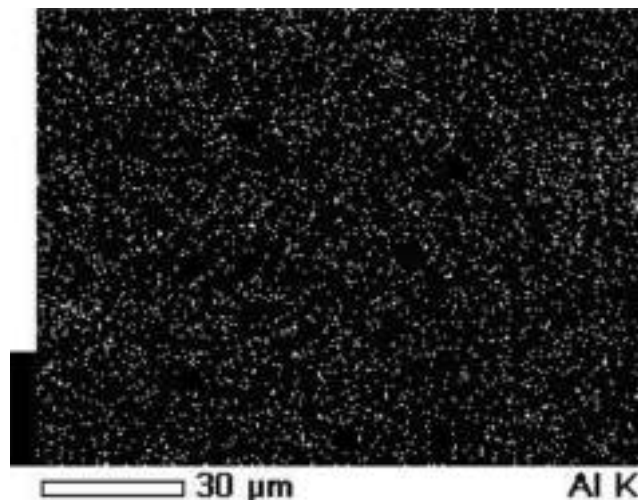


Fig. 1. Representative Aluminium elemental map of the composite with 1 wt.% HNT loading.

where N_{Ai} represents the counts of inclusions per unit area in the i^{th} location, \bar{N}_A is the number of particles in the unit area and σ is the standard deviation.

The sparsely sampled elemental maps of the synthesized composites are shown in Fig. 2. The average number of particles per unit area, standard deviation, and the expected number of particles are presented in Table 1. From Table 1, it is clear that the standard deviation in the number of particles in each section of the composite is increased with filler loading. The increase in standard deviation is due to the agglomeration of the particle that in turn made a significant difference in the average number of particles which was supposed to be increasing akin to the expected number of particles. The expected number of particles was calculated by multiplying the average number of particles in a composite with 1 wt.% filler loading with the higher filler loadings. At 1 wt.% filler loading, HNTs were dispersed uniformly.

To understand the distribution of the particle, the distance between each particle and its nearest neighbor (NND) was calculated using an ImageJ plug-in [32] from the aluminum elemental maps. Fig. 3 shows the distribution of nearest neighbor distances for a representative composite. The nearest neighbor distance compares the position of a particular nanotube with respect to the other nanotubes in the composite. For uniform distribution, the nearest neighbor distribution should be narrow and it was estimated by calculating the FWHM of the Gaussian fit of the distribution and it is presented in Table 2.

The ratio of the average actual neighbor distance (R_k) to the average expected nearest neighbor distance (E_k) [33] of the particles in the composites is another means of optimizing the NND. Higher the R_k/E_k ratio, the better the distribution. R_k and E_k are estimated using the following equation.

$$R_k = \frac{\sum_i^n d_i}{n} \text{ and } E_k = \frac{0.5}{\sqrt{\frac{n}{A}}} \quad (5)$$

where A is the area under study, d_i is the distance between the i^{th} particle and its imminent particle, and n is the number of particles.

In EVACO/HNT composites, the NND of 1 wt.% HNT loading has a wider distribution than that of a 3 wt.% HNT loaded composite. Since 1 wt.% HNT loaded composite has less number of HNTs in it. For filler loadings above 3 wt.%, NND has a broad distribution, which is due to the presence of agglomerates. R_k/E_k ratio is also

Table 1
Sparse sampling.

Filler loading (wt.%)	The average number of particles unit area	Standard deviation in the number of particles	Expected number of particles.
1	85	8	85
3	98	9	255
5	99	14	425
7	89	18	595
10	80	24	850

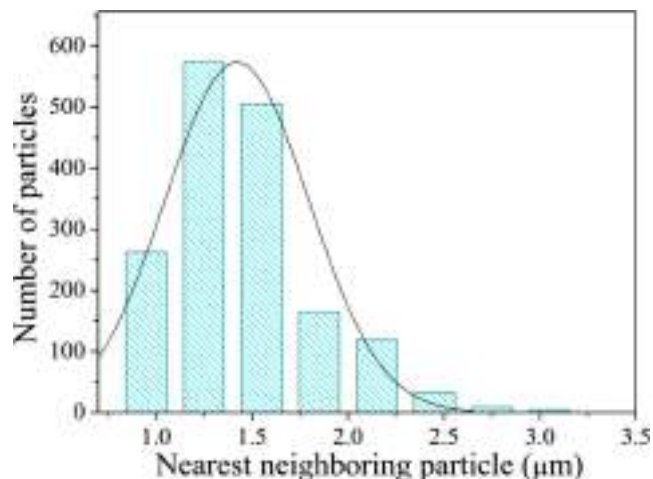


Fig. 3. Nearest neighbor distribution of 1 wt.% HNT filled nanocomposite.

Table 2
Nearest neighbor distance.

HNT loading (wt.%)	FWHM of NND	R_k/E_k
1	0.88	1.365
3	0.74	1.344
5	0.79	1.331
7	0.88	1.292
10	0.94	1.237

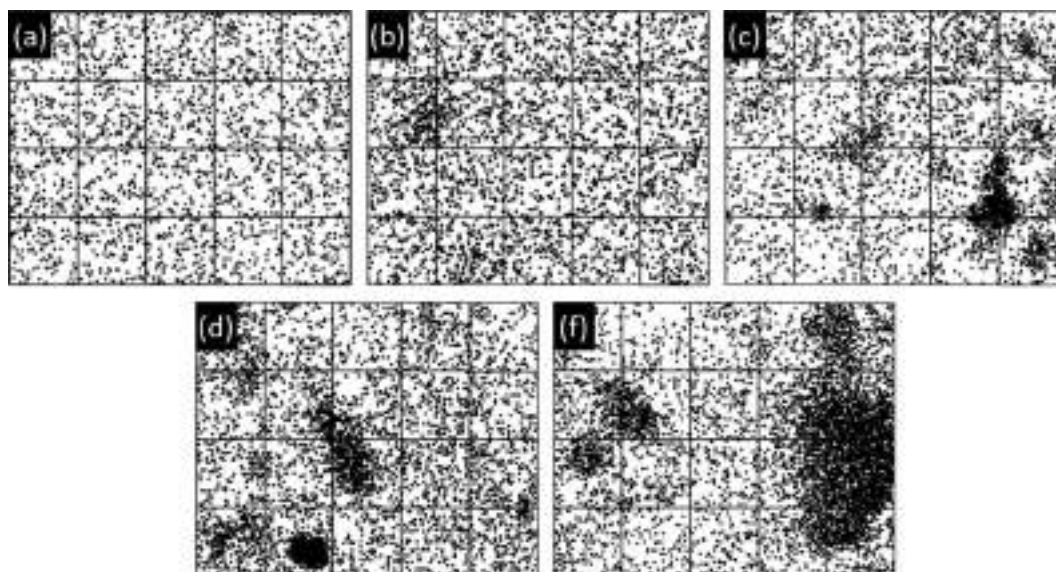


Fig. 2. Sparse sampled aluminum maps of HNT loaded composites, (a) 1 wt.%, (b), 3 wt.% (C), 5 wt.% (d) 7 wt.% and (e) 10 wt.%.

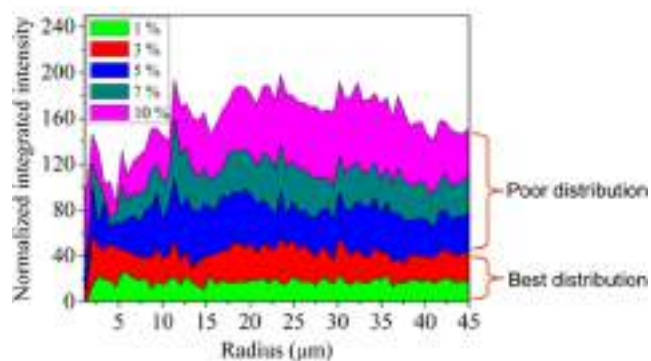


Fig. 4. Radial distribution of HNTs in EVACO/HNT composites.

decreasing with the filler loading since the agglomerates can impact the uniform dispersion and distribution of the fillers in the polymer matrix.

The radial distribution function also specifies the distribution of the nanotubes in the composites. It is a measure of the number density of the particles along the radial direction with respect to a reference particle [34].

The uniformity in the radial distribution of HNTs is reduced as the weight percent of the nanofiller is increased (Fig. 4). The straight radial map shows a uniform distribution, whereas the rough map stands for the non-uniform distribution of HNTs. The indistinguishable boundary of the HNTs in the TEM image (Fig. 5) of the selected composite reveals the interaction between the filler and the matrix. The crystallization of the polymer around HNTs tells the nucleating ability of the HNTs. The major vibrational peaks corresponding to the functional groups of pristine EVACO is obtained through FTIR analysis and it is presented in Table 3.

3.2. The crystallinity of EVACO/HNT composite

The influence of HNTs on the crystallinity of EVACO was evaluated by DSC and XRD analysis. The percentage crystallinity obtained from XRD and DSC analysis is presented in Table 4. The broad melting peak between 20 and 100 °C in Fig. 6 is due to the uneven random crystals in the semi-crystalline EVACO. The addition of small quantities of HNTs i.e., 1 wt.% and 3 wt.% improves

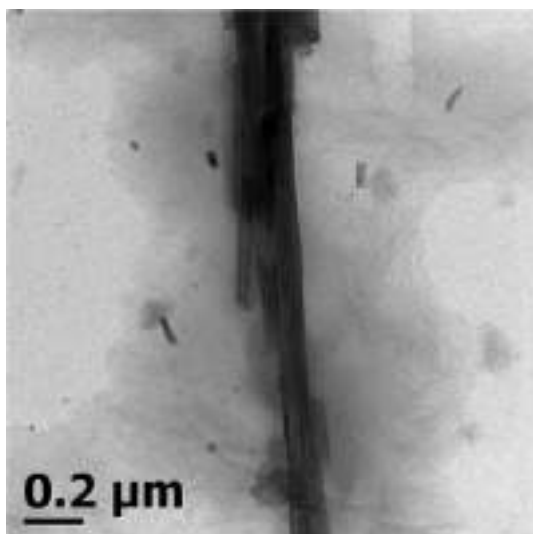


Fig. 5. TEM micrograph of EVACO/CNT composite with 1 wt.% HNT loading.

Table 3
FTIR spectra peak positions.

Peak position (cm ⁻¹)	Assignment
3418	—OH stretching
2919 and 2850	Symmetric and asymmetric CH ₂ stretching
1736	C=O stretching
1467 and 1375	CH scissoring and symmetric deformation
1231	Twisting and wagging of CH
1019	C—OH stretching
721	Rocking vibration of CH

Table 4
Percentage crystallinity of EVACO and EVACO/HNT composites.

Filler loading (wt.%)	% crystallinity from DSC	% crystallinity from XRD
0	46.84	26.59
1	50.33	28.96
3	51.77	30.21
5	46.44	25.75
7	42.13	24.71
10	41.88	18.71

the crystallinity in the composite remarkably. The uniform dispersion and distribution of HNTs can be attributed to this increment. If the filling of the filler is increased above 3 wt.%, the nanotubes may arrest the spherulitic growth front, which originates from the nucleation source, thereby reducing the growth of crystalline regions and ultimately a reduction in the crystallinity. Additionally, a large number of tubes in the matrix can decrease the movement of polymer chains, which can otherwise undergo crystallization in the absence of halloysite nanotubes. The large agglomerates can also adversely affect the crystallinity of the composite with high halloysite nanotube loading.

The XRD results also show the increase in crystallinity for filler loadings of 1 wt.% and 3 wt.% and decrease thereafter. The discrepancy in percentage crystallinity from the study of DSC and XRD is due to the eventual errors that can arise during XRD pattern deconvolution and baseline line correction in DSC curves. [35]. It can be concluded that the dipole–dipole attraction between the nanofiller and the matrix at low nanofiller loading, especially when they are in the solution, can bring the polymer chains close together and align them in an order to favor crystallization.

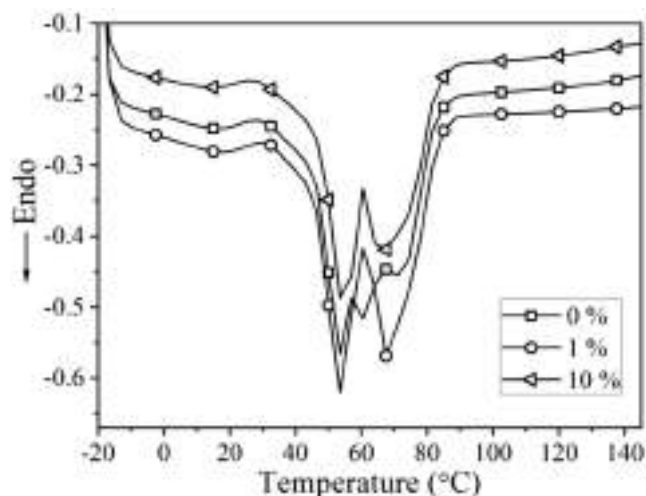


Fig. 6. DSC first heating curves of pristine EVACO and representative composites.

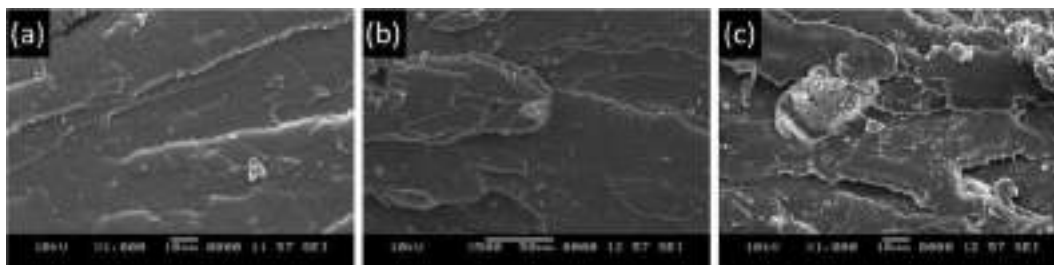


Fig. 7. The fracture surface of (a) EVACO, (b) EVACO/HNT composite with 1 wt.% HNT loading, and (c) 7 wt.% HNT loading.

Table 5

Tensile properties of virgin EVACO and its nanocomposites.

Filler loading (wt.%)	Tensile strength (MPa)	Toughness (kJ/m)
0	31.5	14.3
1	33.2	16.2
3	31.2	14.1
5	29.2	13.5
7	27.2	12.2
10	25.9	8.3

3.3. Mechanical properties

The pristine and composite samples of EVACO exhibited ductile fracture, as it can be identified by the continuous crack propagation trajectories on the fractured surfaces (Fig. 7). Due to the good dispersion and distribution of halloysite nanotubes, the highest tensile strength is observed for the nanocomposite with 1 wt.% percent filler loading (Table 5). In the composite with 3 wt.% halloysite nanotube loading, an overall increase in crystallinity is found, but the dispersion as well as the distribution of the halloysite nanotubes is inferior to 1 wt.% HNT loading. The resulting non-uniform distribution of stress ultimately leads to a small decrease in the ultimate tensile strength. At high filler loading, >3 wt.%, the cluster of halloysite nanotubes and the debonding of these agglomerations from the polymer leads to premature failure, as observed in SEM micrographs of the fracture surface in Fig. 7, and the lessening in the ultimate tensile strength.

4. Conclusions

It is summarised that the dispersion and distribution of the filler play a key role in controlling the crystallizability and mechanical characteristics of the Elvaloy[®]4924 (EVACO)/halloysite nanotube nanocomposites. The image processing of SEM-elemental maps revealed that 1 wt.% HNT loading shows a good dispersion and distribution of the fillers in the matrix. The reduction in mechanical and crystalline characteristics of the composites are in good agreement with dispersion and the uniform spreading of halloysite nanotubes in an array. For 1 wt.% HNT loading, the composite exhibits the best mechanical characteristics and crystallinity. The halloysite nanotubes influence the crystallinity of EVACO at low filler weight fractions, thus discloses the halloysite nanotube's potential as a nucleating agent.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors greatly appreciate the continuous support from the Departments of Mechanical Engineering and Automobile Engineering, SCMS School of Engineering and Technology, Ernakulam, India, and the management of SCMS Group of Educational Institutions, Ernakulam India.

References

- [1] P. Chaudhary, F. Fatima, A. Kumar, Relevance of nanomaterials in food packaging and its advanced future prospects, *J. Inorg. Organomet. Polym. Mater.* 30 (12) (2020) 5180–5192.
- [2] S. Fu, Z. Sun, P. Huang, Y. Li, N. Hu, Some basic aspects of polymer nanocomposites: a critical review, *Nano Mater. Sci.* 1 (1) (2019) 2–30.
- [3] V.K. Sharma, J. Filip, R. Zboril, R.S. Varma, Natural inorganic nanoparticles – formation, fate, and toxicity in the environment, *Chem. Soc. Rev.* 44 (23) (2015) 8410–8423.
- [4] J.L. Suter, D. Groen, P.V. Coveney, Mechanism of exfoliation and prediction of materials properties of clay-polymer nanocomposites from multiscale modeling, *Nano Lett.* 15 (12) (2015) 8108–8113.
- [5] D.R. Paul, L.M. Robeson, Polymer nanotechnology: nanocomposites, *Polymer* 49 (2008) 3187–3204.
- [6] D.R. Baer, M.H. Engelhard, G.E. Johnson, J. Laskin, J. Lai, K. Mueller, P. Munusamy, S. Thevuthasan, H. Wang, N. Washton, A. Elder, B.L. Baisch, A. Karakoti, S.V.N.T. Kuchibhatla, D. Moon, Surface characterization of nanomaterials and nanoparticles: Important needs and challenging opportunities, *J. Vac. Sci. Technol. Vac. Surf. Films Off. J. Am. Vac. Soc.* 31 (5) (2013) 050820, <https://doi.org/10.1116/1.4818423>.
- [7] I.M. Hamouda, Current perspectives of nanoparticles in medical and dental biomaterials, *J. Biomed. Res.* 26 (2012) 143–151.
- [8] R. Andrews, M.C. Weisenberger, Carbon nanotube polymer composites, *Curr. Opin. Solid State Mater. Sci.* 8 (1) (2004) 31–37.
- [9] J.N. Coleman, U. Khan, W.J. Blau, and Y.K. Gun'ko: Small but strong: A review of the mechanical properties of carbon nanotube-polymer composites. *Carbon* 44(9), 1624–1652 (2006).
- [10] C. Min, X. Shen, Z. Shi, L. Chen, Z. Xu, The electrical properties and conducting mechanisms of carbon nanotube/polymer nanocomposites: a review, *Polym.-Plast. Technol. Eng.* 49 (12) (2010) 1172–1181.
- [11] N. Hu, Z. Masuda, C. Yan, G. Yamamoto, H. Fukunaga, T. Hashida, The electrical properties of polymer nanocomposites with carbon nanotube fillers, *Nanotechnology.* 19 (2008) 215701.
- [12] S. Morlat-Therias, E. Fanton, J.-L. Gardette, S. Peeterbroeck, M. Alexandre, P. Dubois, Polymer/carbon nanotube nanocomposites: Influence of carbon nanotubes on EVA photodegradation, *Polym. Degrad. Stab.* 92 (10) (2007) 1873–1882.
- [13] A.B. Morgan, C.A. Wilkie, *Flame Retardant Polymer Nanocomposites*, Wiley-Blackwell, New Jersey, 2007.
- [14] D. Feldman, Polymer nanocomposite barriers, *J. Macromol. Sci. Part A.* 50 (4) (2013) 441–448.
- [15] V. Mittal, *Barrier properties of Polymer Clay Nanocomposites*, Nova Science Pub. Inc., New York, 2010.
- [16] J.O. Emslander: Image receptor medium containing ethylene vinyl acetate carbon monoxide terpolymer. U.S. Patent, WO2000052532 A1 (2000).
- [17] M. Zhao, P. Liu, Adsorption behavior of methylene blue on halloysite nanotubes, *Microporous Mesoporous Mater.* 112 (1-3) (2008) 419–424.
- [18] N.G. Veerabadran, R.R. Price, Y.M. Lvov, Clay nanotubes for encapsulation and sustained release of drugs, *Nano* 02 (02) (2007) 115–120.
- [19] Y.M. Lvov, D.G. Shchukin, H. Möhwald, R.R. Price, Halloysite clay nanotubes for controlled release of protective agents, *ACS Nano* 2 (5) (2008) 814–820.
- [20] H. Ismail, S.Z. Salleh, Z. Ahmad, Properties of halloysite nanotubes-filled natural rubber prepared using different mixing methods, *Mater. Des.* 50 (2013) 790–797.
- [21] M. Liu, B. Guo, M. Du, F. Chen, D. Jia, Halloysite nanotubes as a novel β -nucleating agent for isotactic polypropylene, *Polymer* 50 (13) (2009) 3022–3030.

- [22] E. Abdullayev, V. Abbasov, A. Tursunbayeva, V. Portnov, H. Ibrahimov, G. Mukhtarova, Y. Lvov, Self-healing coatings based on halloysite clay polymer composites for protection of copper alloys, *ACS Appl. Mater. Interfaces*. 5 (10) (2013) 4464–4471.
- [23] L.A. Dobrzański, B. Tomiczek, M. Adamiak, K. Golombek, Mechanically milled aluminium matrix composites reinforced with halloysite nanotubes, *J. Achiev. Mater. Manuf. Eng.* 55 (2012) 7.
- [24] Z. Wei, C. Wang, H. Liu, S. Zou, Z. Tong, Halloysite nanotubes as particulate emulsifier: preparation of biocompatible drug-carrying PLGA microspheres based on pickering emulsion, *J. Appl. Polym. Sci.* 125 (2012) E358–E368.
- [25] M.F. Talbott, G.S. Springer, L.A. Berglund, The effects of crystallinity on the mechanical properties of PEEK polymer and graphite fiber reinforced PEEK, *J. Compos. Mater.* 21 (11) (1987) 1056–1081.
- [26] E. Assouline, A. Lustiger, A.H. Barber, C.A. Cooper, E. Klein, E. Wachtel, H.D. Wagner, Nucleation ability of multiwall carbon nanotubes in polypropylene composites, *J. Polym. Sci. Part B Polym. Phys.* 41 (2003) 520–527.
- [27] J.M.M. Perez, J. Pascau, *Image Processing with ImageJ*, Packt Publishing, 2013.
- [28] S. Park, J.O. Baker, M.E. Himmel, P.A. Parilla, D.K. Johnson, Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance, *Biotechnol. Biofuels* 3 (2010) 10.
- [29] Y. Kong, J.N. Hay, The measurement of the crystallinity of polymers by DSC, *Polymer* 43 (14) (2002) 3873–3878.
- [30] S. Chattopadhyay, T.K. Chaki, A.K. Bhowmick, Heat shrinkability of electron-beam-modified thermoplastic elastomeric films from blends of ethylene-vinyl acetate copolymer and polyethylene, *Radiat. Phys. Chem.* 59 (5-6) (2000) 501–510.
- [31] J.J. Friel, A.S.M. International, *Practical Guide to Image Analysis*, ASM International, 2000.
- [32] Y. Mao, Nearest Neighbor Distances Calculation with ImageJ - EVOCD, (2016). https://icme.hpc.msstate.edu/mediawiki/index.php/Nearest_Neighbor_Distances_Calculation_with_ImageJ.html (accessed February 26, 2021).
- [33] P.J. Clark, F.C. Evans, Distance to nearest neighbor as a measure of spatial relationships in populations, *Ecology* 35 (1954) 445–453.
- [34] P. Baggethun, Radial Profile Plot, (2002). https://imagejdocu.tudor.lu/macro/radial_distribution_function (accessed February 26, 2021).
- [35] M.F.S. Lima, M.A.Z. Vasconcellos, D. Samios, Crystallinity changes in plastically deformed isotactic polypropylene evaluated by x-ray diffraction and differential scanning calorimetry methods, *J. Polym. Sci. Part B Polym. Phys.* 40 (9) (2002) 896–903.



Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Effect of nanofillers on the crystalline and mechanical properties of EVACO polymer nanocomposites

Gibin George^{a,*}, H. Manikandan^a, T.M. Anup Kumar^a, Sam Joshy^a, A.C. Sanju^a, S. Anandhan^{b,*}

^aDept. of Mechanical Engineering, SCMS School of Engineering and Technology, Pallisseri, Ernakulam, Kerala, India

^bDept. of Metallurgical and Materials Engg, National Institute of Technology Karnataka, Surathkal, Karnataka, India

ARTICLE INFO

Article history:

Received 3 March 2021

Received in revised form 12 April 2021

Accepted 15 April 2021

Available online xxxx

Keywords:

Nanocomposite

Crystallization

Nanofillers

Polymer composite

ABSTRACT

In this work, the effect of different fillers on the crystalline and mechanical properties of the poly (ethylene-co-vinyl acetate-co-carbon monoxide) (EVACO) terpolymer composite is studied systematically. Alumina trihydrate nanoparticles (nano-ATH), halloysite nanotubes (HNTs), and the multiwalled carbon nanotubes (MWCNTs) are the representative fillers used in the present study. The surface of MWCNTs are decorated using carbonyl, however, nano-ATH and HNTs are used without any surface treatment. The mechanical properties of the composites are evaluated using a tensile test and the improvement in the mechanical properties can be correlated to the improvement in the crystallinity in the composite. The presence of nanofillers in the EVACO matrix significantly influenced the crystallinity, which was determined by X-ray diffraction. The fractography studies reveal the presence of agglomerates at high filler loading results in the subsequent reduction in the tensile properties. Interestingly, the MWCNTs at very low filler loading significantly enhances the tensile properties of EVACO.

© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

1. Introduction

Polymer nanocomposites are finding new applications every day and replacing the conventional polymers from household to advanced engineering applications [1]. Nanofillers from various sources/origin are commonly used as fillers in polymer nanocomposites. The nanofillers with superior mechanical properties are often used in polymers with poor mechanical characteristics. The addition of nanomaterials in the polymer matrix can impart certain unique properties that cannot be matched by any other material. In the case of nanocomposites, the properties of polymers and nanofillers are often compromised and they exhibit superior properties as a combined material [2]. Additionally, a small quantity of the nanofiller is sufficient to make a significant impact on the properties of the polymer matrix.

Nanomaterials, such as carbon nanotubes [3], clay [4], alumina trihydrate [5], layered double hydroxides [6], halloysite nanotubes

[7], nanocellulose [8], graphene [9], etc. are the common multifunctional nanofillers used in the polymer nanocomposites. The properties such as crystallinity, thermal degradation, tensile strength, permeation resistance, electrical conductivity, flame retardance, etc. are affected by the addition of nanofillers. The unique characteristics of the fillers impart significant property enhancement in the polymer nanocomposite without affecting the processability of the polymer. For instance, the two-dimensional layered nanostructures can influence the permeation characteristics of the polymer [10]. Similarly, the carbon nanotubes increase the electrical conductivity [11], and alumina trihydrate imparts flame retardancy [12]. The aspect ratio (AR) of the nanofillers also impacts the mechanical properties of the polymer nanocomposites [13]. The proven ability of nanofillers as nucleating agents to improve the crystallinity of several semi-crystalline polymer matrices [14–16] that in turn contribute to the enhancement in the tensile strength of polymer composites.

Interfacial bonding between the matrix and the nanofiller plays an important role in determining the properties of the nanocomposites. The interfacial bonding of the filler and the matrix can be improved by modifying the polymer or the filler with suitable functional groups. However, modifying the filler is easier than

* Corresponding authors.

E-mail addresses: gibingeorge@scmsgroup.org (G. George), manikandan@scmsgroup.org (H. Manikandan), anupkumartm@scmsgroup.org (T.M.A. Kumar), samjoshy@scmsgroup.org (S. Joshy), sanju@scmsgroup.org (A.C. Sanju), anandtmg@gmail.com, anandhan@nitk.edu.in (S. Anandhan).

<https://doi.org/10.1016/j.matpr.2021.04.613>

2214-7853/© 2021 Elsevier Ltd. All rights reserved.

Selection and peer-review under responsibility of the International Conference on Sustainable materials, Manufacturing and Renewable Technologies 2021.

the modification of polymer, and the fillers are often modified to match the polarity of the polymer matrix.

A polar filler is modified with a non-polar agent to be used in a non-polar polymer matrix, but it can be directly used in a polar polymer. Additionally, the dispersion of the nanofillers also impacts the mechanical properties of the nanocomposites [17]. Surface modification of carbon nanotubes (MWCNTs) is essential before it is mixed with the organic matrices, since pristine MWCNTs exist as bundles due to their inertness [18]. The polar nanofillers such as ATH and HNTs can be directly used as nanofillers into a polar polymer matrix.

A carbonyl group is introduced to the copolymer poly(ethylene-co-vinyl acetate) (EVA) to form the terpolymer poly(ethylene-co-vinyl acetate-co-carbon monoxide) (EVACO). Such an addition increases the polarity of the new polymer. The polarity of EVA is difficult to improve by increasing the vinyl acetate content alone, as the increase in vinyl acetate content can adversely affect the properties of EVA [19]. The addition of carbon monoxide to the backbone of EVA increases the polarity of the polymer, thereby improves its adhesion to polar surfaces, therefore EVACO is used as an adhesion booster in coatings. EVACO is semi-crystalline in nature, and the polyethylene phase imparts crystallization in it.

In this study, EVACO/nanofiller composites are prepared through solution casting. Industrial processing of EVACO is mainly in the form of solutions, and the method used here is akin to the bulk processing of EVACO. The polar fillers such as ATH and HNTs are directly reinforced to the EVACO matrix, but, MWCNTs are surface modified with polar functional groups before reinforcing them into the EVACO matrix. The mechanical properties and crystallizability of EVACO can be improved by the addition of nanofillers in small quantities. The effect of different nanofillers on the mechanical properties and crystallinity of EVACO is studied here.

2. Materials and methods

Poly(ethylene-co-vinyl acetate-co-carbon monoxide) (Elvaloy® 4924) provided by Du Pont, India, halloysite nanotubes (HNTs) and carbon nanotubes (MWCNTs) procured from Sigma Aldrich, India, nano-ATH obtained from US Research Nanomaterials Inc., USA, and dichloromethane (DCM) procured from Central Drug House (P) Ltd, New Delhi, India were used in the present study.

To fabricate the composites, a predetermined quantity of EVACO is dissolved in DCM by continuous mixing using a magnetic stirrer. To the above solution, the appropriate quantities of nanofillers are slowly added and mixed thoroughly by vigorous stirring and subsequent ultrasonication. The mixture is then poured into Petri-dishes and allowed to dry to get the respective composite films. The composite films are then dried at room temperature and then in a vacuum oven at 50 °C for 6 h.

The tensile measurements (Hounsfield Universal Testing Machine, H25KS, Hounsfield, UK) at ambient conditions were made for three dumb-bell samples, at a crosshead speed of 50 mm/minute, prepared according to ASTM D 412-B. The fractured surfaces were analyzed using a scanning electron microscope (SEM) (JSM-6380LA, JEOL, Japan) and the samples were sputtered with gold (JEOL JFC 1600), in an auto fine coater, prior to imaging. Transmission electron microscope, CM12 PHILIPS, Netherlands, was used to image the morphology of the nanofillers. X-ray diffraction patterns (JEOL, DX-GE-2P, Japan) of the composite sheets were analyzed using CuK_α radiation to determine the crystallinity of nanocomposites. The degree of crystallinity (X_c) was calculated by deconvoluting the XRD patterns to amorphous and crystalline contributions and the degree of crystallinity was calculated by the ratio [20].

$$X_c = \frac{I_c}{I_a + I_c} \quad (1)$$

where I_a and I_c are the integrated intensities corresponding to the amorphous and crystalline phases, respectively.

3. Results and discussion

3.1. Morphology of fillers

The morphology of the nanofillers is compared in Fig. 1 a–c, and one can observe a significant difference in the aspect ratio (AR) of the nanofillers used in the present study. On comparing the TEM images of the nanofillers, the highest AR is observed in the case of MWCNTs (AR≈100) (Fig. 1a), followed by HNTs (AR≈20) (Fig. 1b), and the lowest in the case of nano-ATH (AR≈1.0) (Fig. 1c). Additionally, the surface texture of the MWCNTs is smoother than HNTs and ATH besides the smaller diameter.

3.2. Crystallinity of the nanocomposites

The crystallinity of the semicrystalline polymers can be influenced by the nanofillers. Many nanofillers act as nucleating agents when they are incorporated into different semicrystalline polymer matrices [21,22]. Due to the high polar nature of the nanofillers used in the present study, the crystallizable polymer chains are pulled together to form for more crystallites, that are not formed in the absence of nanofillers [23,24]. The percentage crystallinity is increasing with the filler loading initially but decreases thereafter.

From Table 1, in the case of each nanofiller type, as the filler loading is increased, the corresponding crystallinity (X_c) was initially increased (at 1% loading of ATH and HNTs, and 0.1 wt% loadings in the case of MWCNTs). The crystallinity is decreasing slowly after the above threshold of filler loading in the case of all the fillers. It is likely that, at high filler loading, due to agglomerations, the nanoparticles hinder the polymer chain movements and reduces the crystallinity. One can note that, in the case of MWCNT loading, even though the wt% of MWCNT loading is far less than HNTs or nano-ATH loading, the change in crystallinity is analogous to the other experiments. Apparently, the crystallinity of the present nanocomposites is a function of filler loading besides the nature of the fillers.

3.3. Mechanical properties

In general, all the semicrystalline polymers exhibit a ductile fracture. EVACO is a semicrystalline polymer, therefore, large plastic deformation is expected for pristine EVACO and its nanocomposites. The ductility of the composite decreases as the filler loading exceeds a certain threshold. In general, irrespective of the wt% of the nanofiller, the mechanical properties decrease after an initial surge for all the nanocomposite variants. The maximum value of ultimate tensile strength and toughness is observed for 1% filler loading in both HNT and ATH nanocomposites, whereas for MWCNT reinforced nanocomposites the the highest tensile properties are observed for 0.05% filler loading, as shown in Table 2. The reduction in the mechanical properties with an increase in filler loadings accounts for the agglomeration of nanoparticles leading to non-uniform stress distribution in the composites. The low filler loading of MWCNTs makes a significant impact on the tensile properties of EVACO than HNTs or nano-ATH, since MWCNTs have a significantly higher aspect ratio, as shown in Fig. 1a, and lower density ($\rho = 2.1$ g/cc) as compared to HNTs ($\rho = 2.53$ g/cc) and ATH ($\rho = 2.42$ g/cc). Moreover, the tensile strength of individual strands of MWCNTs is equivalent to that of a steel wire with the same dimensions. Additionally, the dispersion of the nanofillers is uniform at low filler loading, consequently, the effect of stress concentration by agglomeration and the associated premature

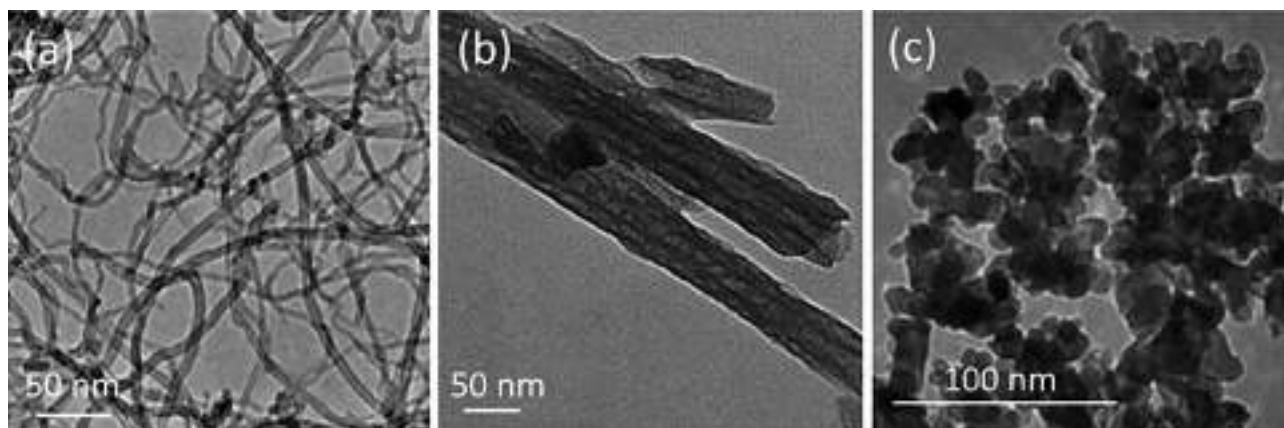


Fig. 1. TEM images of (a) MWCNTs, (b) HNTs and (c) nano-ATH.

Table 1

Normalized percentage of crystallinity for each filler loading.

Filler	Sl. No	Filler Loading (wt%)	X _c
ATH	1	0	46.84
	2	1	50.33
	3	3	51.77
	4	5	46.44
	5	7	42.13
HNTs	1	0	46.84
	2	1	56.98
	3	3	53.69
	4	5	42.73
	5	7	37.69
MWCNTs	1	0	46.84
	2	0.05	48.76
	3	0.1	52.51
	4	0.15	49.67
	5	0.2	48.31
	6	0.25	45.25

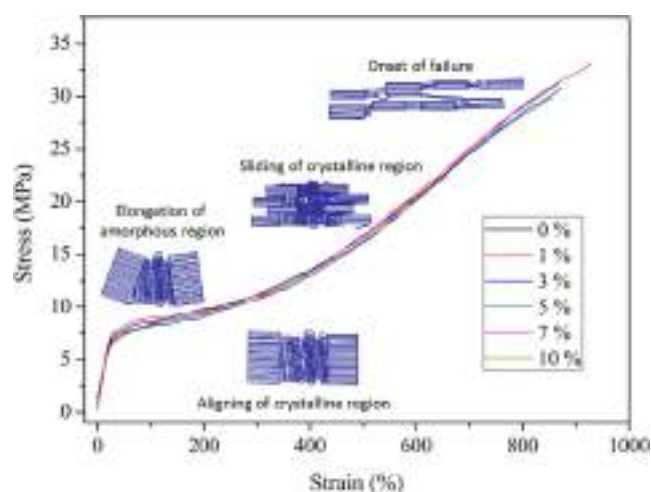


Fig. 2. Representative stress-strain curve of EVACO/nanofiller composite.

failure can be mitigated in those composites. In all the cases, the tensile strength increases initially and decreases thereafter as the filler loading is increased. The representative stress-strain curve of EVACO/HNTs nanocomposites is shown in Fig. 2. There is no apparent change in the profile with filler loading, however, the strain is improved after the filler addition, as compared with the pristine polymer. The schematic diagram in the inset shows the behavior of crystallites in a semi-crystalline polymer as the applied

stress increases. Which in turn indicate the role of crystallinity on the tensile properties of the polymer nanocomposites.

3.4. Fractography analysis

The fracture surfaces of the representative nanocomposites after the tensile test is observed using SEM, as shown in Fig. 3a-c. From

Table 2

Comparison of mechanical properties of EVACO/nanofiller composites.

Type of filler	Filler loading (wt%)	Ultimate tensile strength (MPa)	Toughness (kJ m ⁻¹)	Percentage elongation at break (%)
Nano-ATH	0	31.5	14.3	834
	1	33.2	16.2	893
	3	31.2	14.1	821
	5	29.2	13.5	748
	7	27.2	12.2	711
	10	25.9	8.30	834
HNTs	0	31.5	14.3	930
	1	34.2	16.8	882
	3	32.0	14.5	845
	5	28.5	12.4	802
	7	26.5	9.9	834
	0	31.5	14.3	944
MWCNTs	0.05	45.6	24.3	924
	0.1	41.2	20.9	921
	0.15	40.5	19.8	913
	0.2	39.9	18.6	902
	0.25	39.4	18.0	910

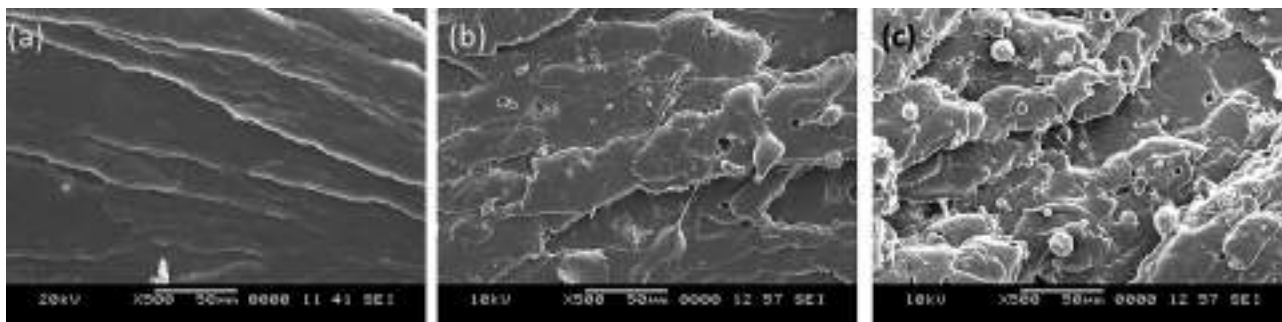


Fig. 3. The fracture surface of (a) EVACO, (b) EVACO/1 wt% ATH composite, and (c) EVACO/10 wt% ATH composite.

the above figures, one can conclude that the fillers significantly affect the mechanism of fracture. In the case of pristine EVACO, the crack propagation trajectories are very clearly indicating a pure ductile failure. As the filler loading is increased to 1 wt% stress whitened regions are observed, which is an indication of more crystalline deformation due to high crystallinity of the composite at 1 wt% loading. In general, the stress whitened regions appears as white regions on the fractured surface of a semicrystalline polymer, which are formed due to the elongation of polymeric chains forming the crystalline part of the polymer. At 10 wt% loading filler loadings, the agglomerated particles and the debonding of these agglomerates from the matrix is clearly visible. The formation of these agglomerates leads to premature failure, resulting in a lower tensile strength than the pristine polymer as observed previously. It is important to note that the roughness of the fractured surfaces are increasing with the filler loading, which conveys the brittle nature of the composites at high filler loading.

4. Conclusions

In summary, the mechanical properties of the EVACO nanocomposites are dependent on the nature of fillers and the wt.% of filler loading. In the case of a semicrystalline polymer, the ability of the nanofillers to form crystallites determines the overall mechanical properties. Thus the filler aspect ratio and the properties of the fillers, in turn, affect the optimal filler loading and the tensile properties. The good dispersion and distribution of fillers also play a major role in controlling the crystallizability and ultimately the mechanical properties. In this study, 1 wt% loading of both HNTs and nano-ATH results in the highest tensile properties in the case of EVACO/HNTs and EVACO/nano-ATH nanocomposites and 0.05 wt% MWCNTs in the case of EVACO/MWCNTs nanocomposites. The improvement in the mechanical properties of EVACO/MWCNTs is way higher at a very low MWCNT loading as compared to HNT and nano-ATH loaded composites. It is assumed that the high aspect ratio and the mechanical properties of MWCNT result in a significant improvement in the tensile properties.

CRediT Author statement

Gibin George: Conceptualization, Data collection, Formal analysis and manuscript drafting. **H. Manikandan:** Manuscript correction, structuring and reviewing. **T.M. Anup Kumar:** Article preparation and data analysis. **Sam Joshy:** Reviewing and editing of manuscript. **A.C.Sanju:** Date interpretation and revision of the manuscript. **S. Anandhan** Designed the experiment, data analysis and article revision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors greatly appreciate the continuous support from the Department of Mechanical Engineering, SCMS School of Engineering and Technology, Ernakulam, India and the management of SCMS Group of Educational Institutions, Ernakulam India.

References

- [1] B. Ates, S. Koytepe, A. Ulu, C. Gurses, V.K. Thakur, Chemistry, structures, and advanced applications of nanocomposites from biorenewable resources, *Chem. Rev.* 120 (17) (2020) 9304–9362.
- [2] F. Hussain, M. Hojjati, M. Okamoto, R.E. Gorga, Review article: polymer–matrix nanocomposites, processing, manufacturing, and application: an overview, *J. Compos. Mater.* 40 (2006) 1511–1575.
- [3] C.A. Hewitt, A.B. Kaiser, S. Roth, M. Craps, R. Czerw, D.L. Carroll, Multilayered carbon nanotube/polymer composite based thermoelectric fabrics, *Nano Lett.* 12 (3) (2012) 1307–1310.
- [4] F. Gao, Clay/polymer composites: the story, *Mater. Today.* 7 (11) (2004) 50–55.
- [5] X. Zhang, F. Guo, J. Chen, G. Wang, H. Liu, Investigation of interfacial modification for flame retardant ethylene vinyl acetate copolymer/alumina trihydrate nanocomposites, *Polym. Degrad. Stab.* 87 (3) (2005) 411–418.
- [6] H. Pang, Y. Wu, X. Wang, B. Hu, X. Wang, Recent advances in composites of graphene and layered double hydroxides for water remediation: a review, *Chem. – Asian J.* 14 (15) (2019) 2542–2552.
- [7] M. Liu, Z. Jia, D. Jia, C. Zhou, Recent advance in research on halloysite nanotubes–polymer nanocomposite, *Prog. Polym. Sci.* 39 (8) (2014) 1498–1525.
- [8] H.-M. Ng, L.T. Sin, S.-T. Bee, T.-T. Tee, A.R. Rahmat, Review of nanocellulose polymer composite characteristics and challenges, *Polym.-Plast. Technol. Eng.* 56 (7) (2017) 687–731.
- [9] X. Ji, Y. Xu, W. Zhang, L. Cui, J. Liu, Review of functionalization, structure and properties of graphene/polymer composite fibers, *Compos. Part Appl. Sci. Manuf.* 87 (2016) 29–45.
- [10] G. Choudalakis, A.D. Gotsis, Permeability of polymer/clay nanocomposites: a review, *Eur. Polym. J.* 45 (4) (2009) 967–984.
- [11] Y. Zeng, P. Liu, J. Du, L. Zhao, P.M. Ajayan, H.-M. Cheng, Increasing the electrical conductivity of carbon nanotube/polymer composites by using weak nanotube–polymer interactions, *Carbon* 48 (12) (2010) 3551–3558.
- [12] P.V. Bonsignore, Flame retardant flexible polyurethane foam by post-treatment with alumina trihydrate/ latex binder dispersion systems, *J. Cell. Plast.* 15 (3) (1979) 163–179.
- [13] G.S. Ananthpadmanabha, V. Deshpande, Influence of aspect ratio of fillers on the properties of acrylonitrile butadiene styrene composites, *J. Appl. Polym. Sci.* 135 (2018) 46023.
- [14] J.E.K. Schawe, P. Pötschke, I. Alig, Nucleation efficiency of fillers in polymer crystallization studied by fast scanning calorimetry: carbon nanotubes in polypropylene, *Polymer* 116 (2017) 160–172.
- [15] X. Shi, G. Zhang, T. Phuong, A. Lazzeri, Synergistic effects of nucleating agents and plasticizers on the crystallization behavior of poly(lactic acid), *Molecules* 20 (1) (2015) 1579–1593.
- [16] J. Njuguna, K. Pielichowski, S. Desai, Nanofiller-reinforced polymer nanocomposites, *Polym. Adv. Technol.* 19 (8) (2008) 947–959.
- [17] M. Šupová, G.S. Martynková, K. Barabaszová, Effect of nanofillers dispersion in polymer matrices: a review, *Sci. Adv. Mater.* 3 (2011) 1–25.

- [18] M. Nurazzi Norizan, M. Harussani Moklis, S.Z.N. Demon, N. Abdul Halim, A. Samsuri, I. Syakir Mohamad, V. Feizal Knight, N. Abdullah, Carbon nanotubes: functionalisation and their application in chemical sensors, *RSC Adv.* 10 (2020) 43704–43732.
- [19] R.J. Crawford, J.L. Throne, Rotational molding polymers, in: R.J. Crawford, J.L. Throne (Eds.), *Rotational Molding Technol.*, William Andrew Publishing, Norwich, NY, 2002, pp. 19–68.
- [20] S. Park, J.O. Baker, M.E. Himmel, P.A. Parilla, D.K. Johnson, Cellulose crystallinity index: measurement techniques and their impact on interpreting cellulase performance, *Biotechnol. Biofuels* 3 (2010) 10.
- [21] D.S. Chaudhary, R. Prasad, R.K. Gupta, S.N. Bhattacharya, Clay intercalation and influence on crystallinity of EVA-based clay nanocomposites, *Thermochim. Acta.* 433 (1-2) (2005) 187–195, <https://doi.org/10.1016/j.tca.2005.02.031>.
- [22] G. George, M. Selvakumar, A. Mahendran, S. Anandhan, Structure–property relationship of halloysite nanotubes/ethylene–vinyl acetate–carbon monoxide terpolymer nanocomposites, *J. Thermoplast. Compos. Mater.* 30 (1) (2017) 121–140.
- [23] A. Zubkiewicz, A. Szymczyk, S. Paszkiewicz, R. Jędrzejewski, E. Piesowicz, J. Siemiński, Ethylene vinyl acetate copolymer/halloysite nanotubes nanocomposites with enhanced mechanical and thermal properties, *J. Appl. Polym. Sci.* 137 (38) (2020) 49135, <https://doi.org/10.1002/app.v137.3810.1002/app.49135>.
- [24] G. George, A. Mahendran, S. Anandhan, Use of nano-ATH as a multi-functional additive for poly(ethylene-co-vinyl acetate-co-carbon monoxide), *Polym. Bull.* 71 (8) (2014) 2081–2102.



IoT-powered deep learning brain network for assisting quadriplegic people

Vinoj P.G.^{a,b}, Sunil Jacob^c, Varun G. Menon^{d,*}, Venki Balasubramanian^c,
Md. Jalil Piran^{f,*}

^a Department of Electronics and Communication Engineering, APJ Abdul Kalam Technological University, Kerala 695016, India

^b Department of Electronics and Communication Engineering, SCMS School of Engineering and Technology, Kerala 683576, India

^c SCMS Centre for Robotics, SCMS School of Engineering and Technology, Kerala 683576, India

^d Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Kerala 683576, India

^e School of Engineering and Information Technology, Federation University, Mount Helen Campus Ballarat, VIC 3350, Australia

^f Department of Computer Science and Engineering, Sejong University, Republic of Korea

ARTICLE INFO

Keywords:

BCI
DBN, Deep learning
EEG
Intelligent system
Rehabilitation

ABSTRACT

Brain-Computer Interface (BCI) systems have recently emerged as a prominent technology for assisting paralyzed people. Recovery from paralysis in most patients using the existing BCI-based assistive devices is hindered due to the lack of training and proper supervision. The system's continuous usage results in mental fatigue, owing to a higher user concentration required to execute the mental commands. Moreover, the false-positive rate and lack of constant control of the BCI systems result in user frustration. The proposed framework integrates BCI with a deep learning network in an efficient manner to reduce mental fatigue and frustration. The Deep learning Brain Network (DBN) recognizes the patient's intention for upper limb movement by a deep learning model based on the features extracted during training. DBN correlates and maps the different Electroencephalogram (EEG) patterns of healthy subjects with the identified pattern's upper limb movement. The stroke-affected muscles of the paralyzed are then activated using the obtained superior pattern. The implemented DBN consisting of four healthy subjects and a quadriplegic patient achieved 94% accuracy for various patient movement intentions. The results show that DBN is an excellent tool for providing rehabilitation, and it delivers sustained assistance, even in the absence of caregivers.

1. Introduction

Quadriplegia results in partial or full mobility impairment and affects nearly 2% of the world population. Primary reasons identified for paralysis are Stroke (33%) and Spinal Cord Injury (SCI) (27.3%) [1]. Rehabilitation is the popular therapy prescribed to fasten the post-paralysis recovery process. In recent years, brain-controlled assistive technologies are employed to provide rehabilitation for quadriplegic patients. Milan et al. [2] demonstrated one of the preliminary works towards non-invasive BCI by controlling a mobile

This paper is for special section VSI-dlls. Reviews processed and recommended for publication by Guest Editor Feiran Huang.

* Corresponding authors.

E-mail addresses: vinojpg@scmsgroup.org (V. P.G.), suniljacob@scmsgroup.org (S. Jacob), varunmenon@ieee.org (V.G. Menon), venki@scms.edu.in (V. Balasubramanian), piran@sejong.ac.kr (Md.J. Piran).

<https://doi.org/10.1016/j.compeleceng.2021.107113>

Received 1 April 2020; Received in revised form 7 December 2020; Accepted 11 March 2021

0045-7906/© 2021 Elsevier Ltd. All rights reserved.



Malware visualization and detection using DenseNets

V. Anandhi¹ · P. Vinod² · Varun G. Menon³Received: 13 March 2021 / Accepted: 28 May 2021
© Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract

Rapid advancement in the sophistication of malware has posed a serious impact on the device connected over the Internet. Malware writing is driven by economic benefits; thus, an alarming increase in malware variants is witnessed. Recently, a large volume of malware attacks are reported on Internet of Things (IoT) networks; as these devices are exposed to insecure segments, further IoT devices reported have hardcoded credentials. To combat malware attacks on mobile devices and desktops, deep learning-based detection approaches have been attempted to detect malware variants. The existing solutions require large computational overhead and also have limited accuracy. In this paper, we visualize malware as Markov images to preserve semantic information of consecutive pixels. We further extract textures from Markov images using Gabor filter (named as Gabor images), and subsequently develop models using VGG-3 and Densely Connected Network (DenseNet). To encourage real-time malware detection and classification, we fine-tune Densely Connected Network. These models are trained and evaluated on two datasets namely Malimg and BIG2015. In our experimental evaluations, we found that DenseNet identifies Malimg and BIG2015 samples with accuracies of 99.94% and 98.98%, respectively. Additionally, the performance of our proposed method in classifying malware files to their respective families is superior compared to the state-of-the approach calibrated using prediction time, F1-score, and accuracy.

Keywords Convolutional neural networks · DenseNet · Feature maps · Malware visualization · Texture

1 Introduction

In recent years, desktops and smart devices are exposed to serious threat due to the presence of malware attacks [1]. Malware or malicious software are evolving at a faster rate [2, 3]; they are designed to disrupt, gain unauthorized

access, and exfiltrate sensitive information from computer systems. A primary motivation for developing new malware is the financial gain associated with it. Hence, it is an industry worth millions of dollars which is increasing every year. According to statistics, data breaches have increased substantially by 40% [4]. Additionally, AV-Test threat report registers [5] more than 350,000 new malware every day, and malware circulation has increased to 114,530 million in 2021; surprisingly, January 2021 alone reported the presence of 607 million malware.

Recently, malware attacks on IoT devices are increasing at an alarming rate. IoT devices have very specific functionalities such as smart healthcare (for monitoring glucose, smart pacemakers, etc), temperature monitoring particularly used in industrial control systems, smart appliances (e.g., smart refrigerator), baby monitoring systems, surveillance system using security cameras, etc. Vulnerable IoT devices of individuals and organizations are largely attacked by hackers primarily due to (a) hardcoded credentials, (b) outdated operating systems, device drivers, and (c) connection of IoT devices to an insecure network and poor web services. All these aforesaid issues transform IoT devices as a pivot to the internal network and expose them to adversary controlled servers. A widely used

✉ V. Anandhi
anandhi@scmsgroup.org

P. Vinod
vinod.p@cusat.ac.in

Varun G. Menon
varunmenon@scmsgroup.org

¹ Department of Computer Science and Engineering,
SCMS School of Engineering and Technology,
Affiliated to APJ Abdul Kalam Technological University,
Thiruvanthapuram, Kerala, India

² Department of Computer Applications,
Cochin University of Science and Technology, Kerala, India

³ Department of Computer Science and Engineering,
SCMS School of Engineering and Technology, Kerala, India

Service Deployment Strategy for Predictive Analysis of FinTech IoT Applications in Edge Networks

Ambigavathi Munusamy¹, Member, IEEE, Mainak Adhikari², Member, IEEE,
 Venki Balasubramanian³, Member IEEE, Mohammad Ayoub Khan⁴, Member, IEEE,
 Varun G. Menon⁵, Senior Member, IEEE, Danda Rawat⁶, Senior Member, IEEE,
 and Satish Narayana Srirama⁷, Senior Member, IEEE

Abstract—The seamless integration of sensors and smart communication technologies has led to the development of various supporting systems for financial technology (FinTech). The emergence of the next-generation Internet of Things (Nx-IoT) for FinTech applications enhances the customer satisfaction ratio. The main research challenge for FinTech applications is to analyze the incoming tasks at the edge of the networks with minimum delay and power consumption while increasing the prediction accuracy. Motivated by the above-mentioned challenge, in this article, we develop a ranked-based service deployment strategy and an artificial intelligence technique for financial data analysis at edge networks. Initially, a risk-based task classification strategy has been developed for classifying the incoming financial tasks and providing the importance to the risk-based task for meeting users' satisfaction ratio. Besides that, an efficient service deployment strategy is developed using *Hall's* theorem to assign the ranked-based financial data to the suitable edge or cloud servers with minimum delay and power consumption. Finally, the standard support vector machines (SVMs) algorithm is used at edge networks for analyzing the financial data with higher accuracy. The experimental results demonstrate the effectiveness of the proposed strategy and SVM model at edge networks over the baseline algorithms and classification models, respectively.

Index Terms—Edge networks, financial technology (FinTech) applications, Internet of Things (IoT), service deployment, support vector machines (SVMs), task classification.

I. INTRODUCTION

THE Internet of Things (IoT) is a promising and emerging technology in the Industrial domain that connects an enormous amount of smart devices, including sensors and actuators, to the network [1]. The smart devices and advanced sensors collect the environmental parameters and transfer the data to remote computing devices for analysis, and take appropriate action [2]. In recent times, IoT-enabled technology has been applied in many real-time applications, including smart transportation, smart industry, smart grid, smart city, etc., in which smart financial technology (FinTech) application has received more attention by leveraging the IoT technology [3], [4]. The emerging phenomenon of the next-generation IoT (Nx-IoT) for the FinTech application is going to reveal one of the most significant moves toward smart worldwide economic diaspora. Using a smart FinTech framework, the Banks and financial institutions can provide quality services to the customers using personalized virtual supervision by optimizing the financial services with advanced artificial intelligence (AI) technology [5]. In such a scenario, the computations and communications become more vulnerable for analyzing the large volume of financial data at remote computing devices by meeting various Quality-of-Service (QoS) parameters [6]–[8].

Nowadays, FinTech applications such as various Banking services, i.e., ATMs, Bank APPs, etc., are relying on Nx-IoT to interface with their customers and require reliable remote computing services for analyzing large-scale financial data. In the past decades, centralized cloud servers provided a plethora of resources for analyzing financial data with advanced AI technologies. However, the major bottleneck faced by the cloud infrastructure is their limited scalability and centralized architecture that increases the latency and drops the overall performance of FinTech applications [9]. The advancement of a new paradigm in the industrial environment such as edge computing plays an important role in FinTech applications by bringing the resources closer to the customers and provides

Manuscript received 28 February 2021; revised 12 April 2021 and 20 April 2021; accepted 3 May 2021. Date of publication 7 May 2021; date of current version 24 January 2023. The work of Satish Narayana Srirama was supported by MHRD through Institution of Eminence status for University of Hyderabad, Grant F11/9/2019-U3(A). (Corresponding author: Satish Narayana Srirama.)

Ambigavathi Munusamy is with the Department of Electronics and Communication Engineering, CEG Campus, Anna University Chennai, Chennai 600025, India (e-mail: ambigaindhu8@gmail.com).

Mainak Adhikari is with the Mobile & Cloud Laboratory, Institute of Computer Science, University of Tartu, Tartu 50090, Estonia (e-mail: mainak.ism@gmail.com).

Venki Balasubramanian is with the School of Science, Engineering and Information Technology, Federation University Australia, Ballarat, VIC 3350, Australia (e-mail: v.balasubramanian@federation.edu.au).

Mohammad Ayoub Khan is with the College of Computing and Information Technology, University of Bisha, Bisha 67714, Saudi Arabia (e-mail: ayoub.khan@ieee.org).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683 576, India (e-mail: varunmenon@scmsgroup.org).

Danda Rawat is with the Department of Electrical Engineering and Computer Science, Howard University, Washington, DC 20059 USA (e-mail: db.rawat@ieee.org).

Satish Narayana Srirama is with the School of Computer and Information Sciences, University of Hyderabad, Hyderabad 500046, India (e-mail: satish.srirama@uohyd.ac.in).

Digital Object Identifier 10.1109/JIOT.2021.3078148

low latency and energy usage as compared to the centralized cloud servers [10]. In practice, Banks use the local edge devices for satisfying personalized customer experience by processing the latency-sensitive applications locally with minimum delay [11], [12]. For example, virtual tellers or facial recognition technology was difficult to analyze in the centralized cloud servers due to the high latency and low transmission speed. In recent times, due to the edge-centric framework of FinTech applications, the customers' faces can be recognized instantly, receive relevant loan offer information, delivering information to the Banking staff, etc., with minimum delay.

A. Motivation

The main focus of the Bank and FinTech institutes is to process or analyze the financial data, mainly the latency-sensitive applications, namely, virtual tellers or facial recognition technology at the edge of the network with minimum delay. Besides that, due to the limited resource capacity of the local edge devices, the computation-intensive financial data need to be transferred to the centralized cloud servers for analysis. Thus, two main research questions for developing an efficient edge-centric framework for FinTech applications are: 1) how to classify the mixed financial data as per their importance, so that the latency-sensitive risk-based data are analyzed at local edge devices? 2) how to provide the services for the classified data, so that the risk-based data are analyzed at local edge devices with minimum delay and energy usage? Besides that, 3) finding a suitable classification model to analyze the financial data at the edge of the networks with the minimum set of data with higher accuracy is another important research challenge? Nowadays, FinTech applications generate a huge volume of financial data at an exponential rate from the Nx-IoT devices, customers, Banks, insurance sectors, etc. One of the major critical tasks in financial industries is to predict the credit risks of legal clients and detect and prevent fraudulent activities. The traditional risk assessment techniques used in the financial sectors are costly and time consuming to process labor-intensive tasks and cannot handle the large volume of financial data.

B. Related Work

To tackle the aforementioned issues, several research works have focused on service deployment and resource provisioning in edge networks. To provide a network service across multiple domains, a chain-based network deployment strategy has been introduced in [13]. This strategy aims to reduce the cost and latency using the virtual network function. Similarly, in [14], a collaborative service deployment and assignment scheme has been proposed in edge networks. The integrated resource provisioning model has been designed to seamlessly provide services across the edge servers and cloud server in [15]. This method effectively considered various service demands from the users and dynamically schedules the incoming tasks to achieve efficient service deployment. In [16], an energy-efficient task allocation scheme for a mobile cloud system has been designed to minimize the power consumption of the computing servers while meeting the deadline.

Hazra *et al.* [6] have developed a 6G-aware fog federation model to effectively schedule the resources in fog networks using a noncooperative Stackelberg game theory with minimum service costs while maximizing the users' satisfaction ratio. To balance the power consumption and delay trade-off between the mobile devices and computing servers, three queuing models have been applied in [17] that find the optimal uploading probability and transmit power for each server. The energy-efficient multitasking strategy has been proposed at multiaccess mobile-edge computing networks in [18] that minimized the total power consumption of the computing devices with a suitable scheduling order. Furthermore, a joint optimization problem has been formulated in [19] to minimize the power consumption and delay of the incoming tasks using a weighted function.

Thennakoon *et al.* [20] have evaluated the series of machine learning (ML) models over credit card fraud detection data sets to find the best classification model concerning the type of frauds. The various ML classification models have been investigated over different financial data sets in [21] to resolve the issue of the data imbalance. Mashrur *et al.* [22] have studied the ML classification models in various financial institutions that include credit rating, bankruptcy prediction, and fraud detection. Dhieb *et al.* [23] have developed an automated insurance prediction system to reduce human interaction, secure insurance activities, notify risky customers, and detect fraudulent claims. Makki *et al.* [24] have revealed the classification models ineffectively only when the financial data are highly imbalanced. Ullah *et al.* [25] have considered the random forest (RF) algorithm to classify the churned customers using two data sets with higher prediction accuracy. Therefore, the critical challenge for analyzing the FinTech applications at the edge level is to distribute the incoming tasks on the local edge devices or centralized cloud servers as per their importance through an efficient service deployment and prediction strategy with higher accuracy. Considering these challenges as a motivation, we design an efficient ranked-based service deployment (RBSD) strategy for predictive analysis of FinTech applications with the support vector machine (SVM) algorithm at edge networks for achieving higher prediction accuracy and minimum delay.

C. Contributions

Our main contributions of the RBSD strategy for predictive analysis of the FinTech applications at edge networks are summarized as follows.

- 1) Design a new ranked-based strategy for classifying the incoming financial tasks at the edge of the network, such as risk-based and nonrisk-based tasks as per their priority. Such a classification aids for analyzing the risk-based financial data at the distributed edge devices with minimum delay and higher accuracy.
- 2) Devise a service deployment strategy with a perfect matching theorem in the graph theory, i.e., *Hall's* theorem for distributing the ranked-based tasks to the remote computing devices. *Hall's* theorem is used to find a perfect matching between the ranked-based tasks and the

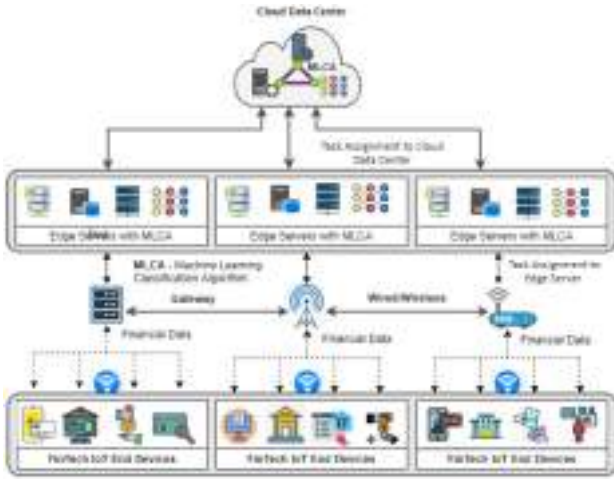


Fig. 1. Edge framework for predictive analysis of FinTech applications.

active set of computing devices for minimizing power consumption at networks.

- 3) Introduce a standard SVM classification model for analyzing the ranked-based tasks at the edge networks using a real data set with higher accuracy and precision. The SVM model uses a small-scale data set for risk prediction at the edge level, whereas a large-scale data set is used for prediction at the cloud level with minimum error.
- 4) Extensive simulation results demonstrate the effectiveness of the proposed RBSD strategy at edge networks for FinTech applications in terms of average delay and power consumption. Besides that, the standard SVM technique demonstrates the effectiveness of analyzing financial tasks with real data sets at edge networks over standard classification models in terms of accuracy and precision.

The remainder of this article is organized as follows. Section II highlights the system model followed by the problem formulation of edge networks for FinTech applications. The proposed service deployment strategy for predictive analysis of FinTech applications is discussed in Section III. The empirical evaluations of the proposed methodology over the existing ones are elaborated in Section IV. Finally, Section V concludes the work and highlights future directions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

This section describes the proposed edge-centric service deployment framework for predictive analysis of FinTech IoT applications followed by the problem formulation.

A. System Model

The proposed edge-centric service deployment framework for FinTech applications is depicted in Fig. 1. This network is constructed with a set of edge servers $\mathcal{S} = \{S_1, S_2, S_3, \dots, S_d\}$ and finite number of remote cloud servers $\mathcal{R} = \{R_1, R_2, R_3, \dots, R_o\}$. The computing servers are highly capable to process the large amount of financial data, collected from the set of FinTech IoT devices $\mathcal{D} =$

$\{D_1, D_2, D_3, \dots, D_f\}$. These devices seamlessly generate the financial tasks $\mathcal{T} = \{T_1, T_2, \dots, T_f\}$ with various degrees of importance, including risk-based (R) and non-risk (NR) financial tasks, i.e., $(\mathcal{T} \in (R \cup NR))$. Furthermore, the financial tasks are processed either locally or transmitted to the remote computing servers for further predictions through a set of gateway devices \mathcal{G} , denoted as $\mathcal{G} = \{G_1, G_2, \dots, G_d\}$. The local gateway devices are responsible for task ranking and service deployment decisions over the received data. Due to inefficient processing capacity ($\tau_{\text{end}}^{\text{CPU}}$) and power consumption ($P_{\text{end}}^{\text{CPU}}$), the efficiency of these two metrics for IoT devices is always less than the edge and cloud servers. Likewise, the CPU capacity and power consumption of an edge device ($\tau_{\text{edge}}^{\text{CPU}}, P_{\text{edge}}^{\text{CPU}}$) should be less than the remote cloud server ($\tau_{\text{cloud}}^{\text{CPU}}, P_{\text{cloud}}^{\text{CPU}}$).

In this network, the set of local edge devices and remote cloud servers is represented as $\mathcal{SR} = (\mathcal{S} \cup \mathcal{R})$. The edge-centric network cogitates that the i th risk-based financial task, referred to as T_i^R , is assigned to the local edge devices. Similarly, the nonrisk-based financial task, referred to as T_i^{NR} , is deployed to the remote cloud servers. The input and output sizes of each task are denoted as T_i^{in} and T_i^{out} , respectively. For instance, the task assignment probability $X(i, j)$ is stated that the assignment of a financial task i to the j th computing device $\forall j \in (\mathcal{D} \cup \mathcal{SR})$. In this scenario, the value of task assignment probability $X(i, j)$ is 1, if the i th task is assigned to the j th computing device, where $\forall j \in (\mathcal{D} \cup \mathcal{SR})$, otherwise, $X(i, j)$ is 0. Therefore, this work mainly focuses to investigate the impact of both power consumption and delay of financial tasks in three different operational modes, including financial task uploading, downloading, and processing.

B. Local Execution Mode

The local FinTech IoT devices have limited power and CPU frequency (τ_i^{CPU}). For instance, the i th task can process locally when the required CPU frequency of the incoming task is less than or equal to the available CPU capacity of the local IoT device. The total time required to execute the i th task in the j th IoT device is expressed as follows:

$$P_{R_{ij}} = X(i, j) \times \frac{T_i^{\text{in}}}{\tau_i^{\text{CPU}}} : \forall i \in \mathcal{T}, j \in \mathcal{D}. \quad (1)$$

Processing the task at local IoT devices depends on CPU frequency instead of the communication delay. Let us consider that the required power to process a 1-bit task at the j th IoT device is defined as P_j^{CPU} . Thus, the overall power consumed by the task i at the j th IoT device is computed as follows:

$$P_{ij}^{\text{proc}} = X(i, j) \times \frac{T_i^{\text{in}}}{\tau_i^{\text{CPU}}} \times P_j^{\text{CPU}} : \forall i \in \mathcal{T}, j \in \mathcal{D}. \quad (2)$$

C. Remote Execution Mode

Due to the limited processing and storage capacity of the FinTech IoT devices, the large volume of financial tasks \mathcal{T} is directly uploaded to the remote edge or cloud servers for further predictions. Therefore, the total time required to process the financial tasks at remote computing devices depends on the uploading, downloading, and processing time. For instance, if

a task i is assigned to the j th computing device, i.e., $\forall i \in \mathcal{T}, j \in (\mathcal{S}, \mathcal{R})$, then the transmission rate of the i th task to j th computing device is defined as $\gamma_{ij}^{\text{up}} = \mathcal{W}_{ij}^{\text{in}} \log(1 + P_j^{\text{up}} \times [\delta_j^{\text{power}} / \alpha_j^2])$. Here, $\mathcal{W}_{ij}^{\text{in}}$ indicates the channel utilization factor between the i th IoT device and the j th computing device. α_j^2 and P_j^{up} represent the additive white Gaussian noise of the local IoT device and the transmission power to offload the task to the j th computing device, respectively. Thus, the total transmission time required to upload the task to the remote computing device can be formulated as follows:

$$T_{ij}^{\text{up}} = X(i, j) \times \frac{T_i^{\text{in}}}{\gamma_{ij}^{\text{up}}} : \forall i \in \mathcal{T}, j \in L(\mathcal{S}, \mathcal{R}). \quad (3)$$

Consequently, the uploading power consumption (P_{ij}^{up}) of the i th financial task to j th remote computing device is expressed as follows:

$$P_{ij}^{\text{up}} = T_{ij}^{\text{up}} \times P_j^{\text{up}} : \forall i \in \mathcal{T}, j \in (\mathcal{S}, \mathcal{R}). \quad (4)$$

The total time required to execute a task $i \forall i \in (\mathcal{T}_i^{\text{R}}, \mathcal{T}_i^{\text{NR}})$ on the j th remote computing device $\forall j \in (\mathcal{S}, \mathcal{R})$ is defined as follows:

$$P_{ij} = \begin{cases} \mu_{kj}^{\text{R}} \times X(i, j) \times \frac{T_i^{\text{in}}}{T_{\text{CPU}}^{\text{in}}} & \text{if, } T_i \in \mathcal{T}_i^{\text{R}}, j \in \mathcal{S} \\ \mu_{kj}^{\text{NR}} \times X(i, j) \times \frac{T_i^{\text{in}}}{T_{\text{CPU}}^{\text{in}}} & \text{if, } T_i \in \mathcal{T}_i^{\text{NR}}, j \in \mathcal{S} \\ (1 - \mu_{kj}^{\text{R}}) \times X(i, j) \times \frac{T_i^{\text{in}}}{T_{\text{CPU}}^{\text{in}}} & \text{if, } T_i \in \mathcal{T}_i^{\text{R}}, j \in \mathcal{R} \\ (1 - \mu_{kj}^{\text{NR}}) \times X(i, j) \times \frac{T_i^{\text{in}}}{T_{\text{CPU}}^{\text{in}}} & \text{if, } T_i \in \mathcal{T}_i^{\text{NR}}, j \in \mathcal{R}. \end{cases} \quad (5)$$

The arrival rate of the financial task on the remote edge and cloud servers is represented as λ_i^{edge} and λ_i^{cloud} , respectively. Furthermore, the waiting time l_{ij} of the i th task before assigning to the j th computing device is defined as follows:

$$l_{ij} = \lambda_i^{\text{edge}} \frac{T_i^{\text{in2}}}{\tau_i^{\text{CPU}}} (\tau_i^{\text{CPU}} - \lambda_i^{\text{edge}} \times T_i^{\text{in}}) : j \in (\mathcal{S}, \mathcal{R}). \quad (6)$$

The total execution delay of the i th task on j th computing device at time t is expressed as $l(t) = \sum_{i=1}^q l_{ij}$. Let P_j^{CPU} represents the processing power to process 1-bit data at remote computing device. Thus, the total consumed power to process the i th task on the j th remote computing device is measured as follows:

$$P_{ij}^{\text{proc}} = P_{ij} \times P_j^{\text{CPU}} : \forall i \in \mathcal{T}, j \in (\mathcal{S}, \mathcal{R}). \quad (7)$$

Let σ_j^{power} represents the channel power gain of the j th computing device. W_{ij}^{out} and δ_j^{power} denote the channel utilization between remote j th computing device to i th IoT device and required transmission power of j th remote computing device. Thus, the power consumption of the i th task during the downloading process ($\gamma_{ji}^{\text{down}}$) is defined as follows:

$$\gamma_{ji}^{\text{down}} = \mathcal{W}_{ij}^{\text{out}} \log \left(1 + P_j^{\text{down}} \times \frac{\delta_j^{\text{power}}}{\alpha_j^2} \right) : \forall i \in \mathcal{T}, j \in (\mathcal{S}, \mathcal{R}) \quad (8)$$

where α_j^2 denotes the Gaussian noise ratio on the j th remote computing device. The downloading time T_{ji}^{down} from the j th

computing device to the i th IoT device is defined as follows:

$$T_{ij}^{\text{down}} = X(j, i) \times \frac{T_i^{\text{out}}}{\gamma_{ji}^{\text{down}}} : \forall i \in \mathcal{T}, j \in (\mathcal{S}, \mathcal{R}). \quad (9)$$

Subsequently, the downloading power consumption of the i th financial task is computed as follows:

$$P_{ij}^{\text{down}} = X(i, j) \times T_{ij}^{\text{out}} \times \frac{P_j^{\text{down}}}{W_{ij}^{\text{out}} \times \log \left(1 + P_j^{\text{down}} \times \frac{\delta_j^{\text{power}}}{\alpha_j^2} \right)}. \quad (10)$$

The total power consumption of a financial task i during computation at j th remote computing device is measured as follows:

$$P_{ij}^{\text{total}} = (P_{ij}^{\text{up}} + P_{ij}^{\text{proc}} + P_{ij}^{\text{down}}). \quad (11)$$

Therefore, the total power consumption ($P_{ij}^{\text{total}}(t)$) of a financial task i during uploading, processing, and downloading to the j th computing device at time t is expressed as follows:

$$P_{ij}^{\text{total}}(t) = (P_{ij}^{\text{up}}(t) + P_{ij}^{\text{proc}}(t) + P_{ij}^{\text{down}}(t)). \quad (12)$$

D. Problem Formulation

The main goal of this work is to minimize the power consumption and delay of the financial tasks in three different modes, such as uploading, processing, and downloading phase. If a financial task is assigned to the local IoT device \mathcal{D} , then the total power consumed (i.e., P_{ij}^{total}) by the i th financial task is equal to the processing power (P_{ij}^{proc}) in the local IoT device. However, if i is assigned to the local edge or remote cloud server j , then the total power consumption (P_{ij}^{total}) by the task i depends on the uploading power P_{ij}^{up} , downloading power P_{ij}^{proc} , and processing power P_{ij}^{proc} , i.e., $P_{ij}^{\text{total}} = (P_{ij}^{\text{up}} + P_{ij}^{\text{proc}} + P_{ij}^{\text{down}})$. The objective function of the work with necessary constraints is formulated as follows:

$$\text{minimize } \sum_{i=1}^n P_{ij}^{\text{total}}(t) \quad (13a)$$

$$\text{subject to } P_{ij}^{\text{total}}(t) \leq \eta_j^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13b)$$

$$l_{ij}(t) \leq l_{ij}^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13c)$$

$$\tau_i^{\text{CPU}}(t) \leq \tau_i^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13d)$$

$$\sum_{i=1}^{(|\mathcal{T}|)} \sum_{j=1}^{(|\mathcal{SR}|)} X(i, j) \leq |\mathcal{S} \cup \mathcal{R}| \quad (13e)$$

$$\sum_{j=1}^{(|\mathcal{T}|)} X(i, j) = 1. \quad (13f)$$

From the above problem formulation, constraints (13b) and (13c) state the total power consumption and delay of a financial task i should be less than or equal to the maximum power consumption η_j^{max} and delay l_{ij}^{max} , respectively. According to constraint (13d), the required CPU frequency of the i th financial task should be less than or equal to the selected computing device j . Equation (13e) represents the active number of remote computing devices in the network. Finally, constraint (13f)

states that each financial task should be assigned at most one computing device at time t .

III. RANKED-BASED SERVICE DEPLOYMENT STRATEGY

This section presents an effective RBSD strategy for FinTech IoT applications at edge networks. Initially, the incoming tasks from various FinTech IoT devices are ranked according to their importance and priorities. Then, the ranked financial tasks are assigned to the suitable computing devices for further analysis.

A. Ranked-Based Task Classification

In the ranked-based classification model, the incoming financial tasks from the IoT devices are classified based on their degrees of importance and service requirements. Subsequently, the ranked tasks are placed into the buffers of a local gateway device for making further decisions. To get instant response from the local edge devices, the rank index (η) factor is introduced to identify the importance of the financial tasks and locate them according to the nondecreasing order. We consider η is a priority threshold value to classify the severity of incoming financial tasks. With the help of (η) value, the financial tasks are effectively categorized into two types: 1) risk-based (R) and 2) nonrisk-based (NR) tasks, represented as T_i^R and T_i^{NR} , respectively. The values 0 and 1 indicates the types of the incoming task, i.e., 0 represents risk-based task T_i^R and 1 represents the nonrisk-based task T_i^{NR} .

In this way, the proposed RBSD strategy satisfies the following two constraints: 1) a task T_i is called a risk-based task if $\eta(T_i) \geq 0.5$ or 2) a nonrisk-based task if $\eta(T_i) < 0.5$. Based on the ranking orders, the risk-based tasks are placed into the risk-based buffer $\omega_i^R(t)$, if $T_i \in T_i^R$ or to the nonrisk-based buffer $\omega_i^{NR}(t)$, if $T_i \in T_i^{NR}$. The systematic workflow of the ranked-based classification model is illustrated in Fig. 2. In this model, the arrival rate of financial tasks is symbolically represented using a Poisson process with the density function $f(t) = \lambda_i^e - \lambda_i^l$. The parameters λ_i and ϕ_{jk} denote the financial task arrival rate and the task uploading probability from the j th IoT device to the k th gateway device, respectively. The offloading decisions at the k th gateway device is defined as $\lambda_{jk}^{rem} = \phi_{jk} \times \lambda_i \forall j \in D$. Thus, the arrival rate of the i th task for processing locally on the j th IoT device is formulated as follows:

$$\lambda_{ij}^{local} = (1 - \phi_{jk}) \times \lambda_i. \quad (14)$$

The arrival rate of the set of financial tasks (σ_{jk}) under a risk-based buffer of the k th local gateway device is defined as follows:

$$\lambda_{jk}^R = \sigma_{jk} \times \lambda_{jk}^{rem}. \quad (15)$$

Similarly, the remaining set of financial tasks that arrive under a nonrisk-based buffer of the k th gateway device is expressed as follows:

$$\lambda_{jk}^{NR} = (1 - \sigma_{jk}) \times \lambda_{jk}^{rem}. \quad (16)$$

The probabilities of assigning risk-based and nonrisk-based financial tasks to the j th computing device are expressed as

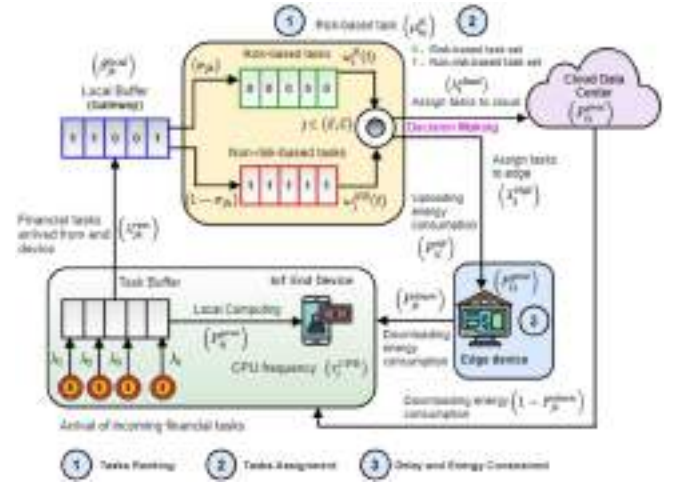


Fig. 2. Workflow of ranked-based task classification.

μ_{kj}^R and μ_{kj}^{NR} , respectively. Thus, the arrival rate of the i th task from the k th gateway device to the j th edge device $\forall j \in S$ is expressed as follows:

$$\lambda_i^{edge} = \mu_{kj}^R \times \lambda_{jk}^R + \mu_{kj}^{NR} \times \lambda_{jk}^{NR} \quad (17)$$

$$= \mu_{kj}^R \times \sigma_{jk} \times \lambda_{jk}^{rem} + \mu_{kj}^{NR} \times (1 - \sigma_{jk}) \times \lambda_{jk}^{rem}. \quad (18)$$

Similarly, the task arrival rate of the i th task to the j th remote cloud server $\forall j \in \mathcal{R}$ from the k th gateway device is represented as follows:

$$\lambda_i^{cloud} = (1 - \mu_{kj}^R) \times \lambda_{jk}^R + (1 - \mu_{kj}^{NR}) \times \lambda_{jk}^{NR} \quad (19)$$

$$= (1 - \mu_{kj}^R) \times \sigma_{jk} \times \lambda_{jk}^{rem} + (1 - \mu_{kj}^{NR}) \times (1 - \sigma_{jk}) \times \lambda_{jk}^{rem}. \quad (20)$$

The total arrival rate of risk-based [i.e., $\sum_{(j \in I)} \omega_j^R(t) \lambda_i^R$] and nonrisk-based financial tasks [i.e., $\sum_{(j \in J)} \omega_j^{NR}(t) \lambda_i^{NR}$], and service rate (μ_{ij}) at the local buffer of the gateway device do not create much impact on financial tasks uploading and downloading decisions at time t . Furthermore, the power-efficient task uploading decisions can be achieved using the following function:

$$\beta_{T_i}^{out}(t) = \text{minimize} \sum_{j \in S, R} \frac{(T_i^{in} \times P_j^{up})}{W_{ij}^{in}} + \frac{P_i^{CPU} \times T_i^{in}}{\tau_i^{CPU}} + \frac{(T_i^{out} \times P_j^{down})}{W_{ji}^{out}} + \sum_i \in I \omega_i^R(t) \times \mu_i(t) - \sum_j \in J \omega_j^{NR}(t) \times \mu_j(t).$$

Based on the above formulation, it is proved that the ranked-based classification model satisfies the power consumption and delay constraints [from (13a)–(13h)] in the edge networks. Next, the classified tasks are assigned to the suitable remote computing devices for further analysis using a perfect matching algorithm.

B. Service Deployment Strategy With Perfect Matching

This section discusses the proposed service deployment strategy with a perfect matching theorem for assigning the ranked-based tasks of the FinTech IoT applications to suitable remote computing devices for further prediction while minimizing the power consumption and delay. To map the ranked financial tasks with the active set of computing servers, a well-known perfect matching theorem in the graph theory, namely, *Hall's* theorem is considered in edge networks. Mathematically, the perfect mapping function is expressed as $P: T_i \rightarrow C$ between the ranked task set T and the computing devices c using a link weight function $F: Q \rightarrow R^+ \cup \infty$. In this model, the weight function F_{ij} between the ranked task T_i and computing server C_j always depends on the total power consumption (P_{ij}^{total}). Constantly, the gateway device produces a new set of ranked financial tasks concerning the availability of the active set of computing devices.

Hall's perfect matching theorem for FinTech IoT applications at the local gateway device is depicted in Fig. 3. The decision making graph is constructed using hall's complete bipartite graph $G(M, N)$, which consists of a set of vertices and dummy edges with a positive link weight ∞ in the form of power consumption. In this graph, the ranked-based task assignment starts with a dual matching solution such that $D_j = 0 \forall j \in C$ and $D_i = \text{MIN}(F_{ij}) : N_{ij} \in K(i) \forall i \in T$. This condition states that the tight edges N' has at least one perfect matching in subgraph G' , defined as $F_{ij} = D_i + D_j$. If there is no matching N' , then the dual value of the corresponding *Hall's* financial tasks set is modified by adding a constant value K to T_i and subtracting the value K from C_j , referred as $D_i = D_i + K$ and $D_j = D_j - K$, respectively.

In a given task assignment graph $G = (M, N)$ with bipartition (T, C) , where $M = (T \cup C)$ and a perfect matching function $P: T \rightarrow C$ such that G assigns set of all ranked-based tasks T in each time frame if and only if $|X| \geq |B(X)|$, where $X \subseteq T$ and $B(X) = \{h \in C | C = (S \cup R), (T, C) \in Q, \text{ and } \forall T \in X\}$. Let us consider that $X = (T_1, T_2, T_3, T_4), X \subseteq T$, then $B(X) = B(T_1) \cup B(T_2) \cup B(T_3) \cup B(T_4) = (C_1, C_2, C_3, C_4)$. Hence, the *Hall's* condition is satisfied with $|X| \leq |B(X)|$, where X is the set of all possible combination of tasks in the financial task set T . The condition $|X| \leq |B(X)|$ denotes that all the subsets of T are mapped when there exists a mapping from financial tasks to the corresponding computing devices. Therefore, *Hall's* condition is satisfied and the graph G has saturated matching from task T to the edge device S .

As shown in Fig. 3, the financial task T_2 is perfectly matched with C_2 , and T_3 is matched with C_3 . However, for task T_3 , there is no tight matching in the set C , which indicates that among the tight edges in N' both the tasks T_2 and T_3 have a perfect matching. Furthermore, for a task T_1 , there is a *Hall's* set, i.e., $T_1 \cup T_3$. Accordingly, the ranked-based task assignment graph needs to be modified using the dual value, so the subgraph G' extends with untight edges until a perfect match is found. For this purpose, the subgraph G' is modified by adding the value of K in the financial task set T and removing K from the set C . Based on the perfect matching theorem, each ranked task T_i is assigned or mapped

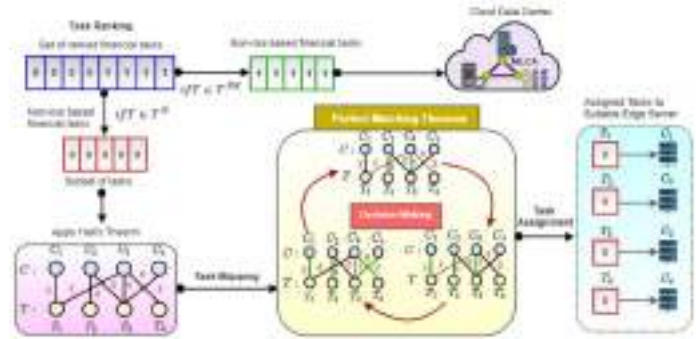


Fig. 3. Service deployment with perfect matching theorem.

to at most one remote computing device C_j , which ensures the financial task assignment constraint (13d). Finally, all the ranked-based financial tasks are assigned to the suitable edge devices based on their perfect matching order. Furthermore, the proposed service deployment strategy decreases the computation and communication overhead of the network by assigning the nonrisk-based tasks to the remote cloud servers while finding a maximum matching between the ranked-based tasks and local edge devices. The systematic procedures of the RBSD strategy are depicted in Algorithm 1.

C. Predictive Analysis at Edge Networks

The huge volume of data, collected from various FinTech applications through Nx-IoT, demands instant decisions and service requirements from the banking or financial sectors. However, most of the financial industries still process customer-related information using traditional or manual screening and analytic tools. Due to the digital transformation of financial data using Nx-IoT, the instant prediction and identification of cybercriminals and frauds are challenging tasks in financial industries. Thus, the financial industries must require an intelligent predictive and analytical model to deal with them. Besides that, the transmission of mixed types of financial data from FinTech IoT devices to the remote cloud server increases the delay and power consumption of the customer service requirements. In such cases, instigating predictive analytic models at the local edge devices helps to analyze, and identify the huge volume of risk-based financial data and provide instant services closer to the customers with minimum delays and errors.

Based on these perceptions, various ML classification models, such as logistic regression (LR), decision trees (DT), SVM, and RF, have been studied and validated using different real-time financial data sets. However, the proposed edge-centric predictive analysis considers the SVM model as the baseline model to effectively analyze and estimate the banking crises with higher accuracy over other classification models. The reason behind selecting the SVM classification model is that the SVM model is capable to handle high-dimensional financial data and improves significant accuracy with less computation power [26], [27]. Furthermore, to estimate the decision function with minimum error, the SVM model uses a linear model with a nonlinear boundaries class based on

Algorithm 1: Ranked-Based Service Deployment

1 **INPUT:** Rank index factor: η , Incoming tasks: \mathcal{T}_i , Set of computing servers: $C \leftarrow (S \cup R)$, Risk based buffer: ω_i^R

2 **OUTPUT:** Classify and assign the incoming tasks to the suitable computing servers using η

1: **for** $i:1$ to n **do**

2: Assign rank index factor η to the incoming tasks

3: **if** A task $\omega_i^{NR} \leftarrow T_i^{NR} \leq \eta$ **then**

4: Assign a $T_i^N R$ to non-risk-based buffer ω_i^{NR}

5: **end if**

6: **if** A task $\omega_i^R \leftarrow T_i^R \geq \eta$ **then**

7: Assign a T_i^R to risk-based buffer ω_i^R

8: **end if**

9: Assign ranked tasks to the suitable C using Perfect matching

10: **if** $|X| \leq |B(X)|$ **then**

11: Graph has a saturated matching of \mathcal{T}_i

12: **end if**

13: **if** $|X| \geq |B(X)|$ **then**

14: Find matching from N' Where $(F_{ij} = D_i - D_j)$;

15: Modify $D_i = D_i + k, \forall i \in \mathcal{T}$

16: Modify $D_j = D_j + k, \forall i \in \mathcal{T}$

17: Update the value of tight edges N' based the matching function F

18: **end if**

19: Assign risk based financial tasks T_i^R to the edge server S_j

20: **end for**

21: **for** All ranked tasks $T_{ij} \in \omega_j^{NR}$ **do**

22: Assign non-risk based financial tasks T_i^{NR} to the remote cloud server R_j

23: **end for**

24: Return a perfect mapping function

support vectors. In the proposed strategy, with the help of the SVM classification model, the ranked-based tasks are analyzed and predicted at the resource-constraints edge devices to get an instant response and enhance the service requirements of the customers. Similarly, the nonrisk-based tasks are analyzed at the remote cloud server for future predictions.

IV. EMPIRICAL EVALUATION

This section briefly discusses the empirical evaluation of the proposed ranked-based classification model and service deployment strategy in edge networks. The proposed edge-centric FinTech framework is quantified and validated concerning average delay and power consumption. To verify the ability of the edge-centric framework, we compare the proposed framework with two baseline schemes, such as CoISDA [14] and OSP [15]. Furthermore, the predictive classification model, i.e., the SVM technique, is applied over the financial tasks at both edge and cloud server to prove the superiority of the proposed framework and the results are compared with the state-of-the-art models, including LR [28], DT [29], and RF, [30]. Furthermore, different validation

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Number of IoT devices (\mathcal{D})	500
Number of Edge devices (S)	20
Number of cloud servers (\mathcal{R})	2
Number of gateway devices (\mathcal{G})	2
Average number of incoming data (λ_i)	500 [tasks/sec]
Maximum channel bandwidth (W)	20 MHz
CPU frequency of IoT devices (τ_i^{CPU})	10×10^5 [cycles/sec]
CPU frequency of edge devices (τ_e^{CPU})	20×110 [cycles/sec]
CPU frequency of cloud servers (τ_c^{CPU})	30×120 [cycles/sec]
CPU processing power usage (P^{CPU})	0.5 Joules
Transmission power of IoT devices (T^I)	1 mW

metrics, including accuracy, precision, recall, and F1 score, are considered to find the effectiveness of the SVM classification models for financial risk predictions.

A. Experimental Setup and Data Set

The proposed strategy has been implemented on Intel Core i7-8550U Quad-Core CPU with 12-GB RAM using the Ubuntu LTS operating system. The simulation test parameters are summarized in Table I. The edge network consists of 500 FinTech IoT devices that generate 500 tasks/s in each timestamp. Here, the maximum data transmission rate is fixed to 2.5 Mb/s, the range of input task size is T_i^{in} is [50 kb–10 Mb], and the financial task arrival rate on the edge devices λ_i^{edge} is 0.125, and the remote cloud server λ_i^{cloud} is 0.25. Here, the ranked-based financial tasks are analyzed using real data sets, such as credit card fraud prediction (D1),¹ credit card risk prediction (D2),² customer churn prediction (D3),³ and insurance claim prediction (D4).⁴ Table II contains the summary of FinTech data sets and their properties for edge-cloud-level analysis.

B. Simulation Results

The simulation results of the proposed service deployment strategy are evaluated in two different phases, such as communication and computation, respectively. In the first phase, the delay and power consumption of the incoming financial tasks have been analyzed in edge networks. Likewise, the prediction accuracy of the classification models has been tested and validated in the computation phase. The quantitative results of the proposed strategy are concisely described in the following subsections.

1) *Analysis of Delay:* Fig. 4 shows the impact of task assignment over the delay in edge networks. The delay of the financial task depends on the processing, uploading, and downloading time while assigning to the remote computing devices. The delay variation of the risk-based tasks is 29.6 ms, which is lower than the nonrisk-based tasks (41.2 ms), as depicted in Fig. 4(a). Moreover, the rank index factor η is introduced to classify the incoming financial tasks based on

¹<https://www.kaggle.com/nandini1999/credit-card-fraud-detection>

²<https://www.kaggle.com/kabure/predicting-credit-risk-model-pipeline>

³<https://www.kaggle.com/kmalit/bank-customer-churn-prediction>

⁴<https://www.kaggle.com/saikrishna223/insuranceclaimprediction>

TABLE II
SUMMARY OF FINTECH DATA SETS AND THEIR PROPERTIES FOR EDGE-CLOUD-LEVEL ANALYSIS

Level of Analysis	Dataset(s)	No of Instances	No of Attributes	Purpose
Edge Server	D1	284808	31	Credit Card Fraud Detection
	D2	1000	20	Credit Card Risk Prediction
Cloud Server	D3	1000	14	Customer Churn Prediction
	D4	1338	8	Insurance Claim Prediction

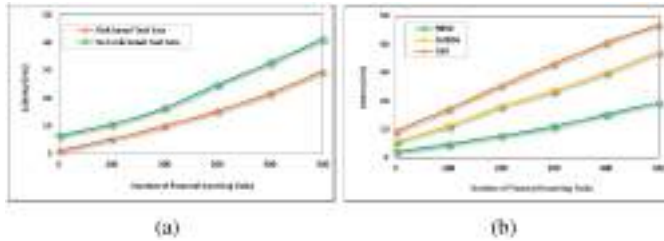


Fig. 4. Impact of task assignment over delay. (a) Various financial tasks. (b) Comparative analysis with baseline schemes.

the different order of severity. Fig. 4(b) presents the comparative analysis of the average delay of the proposed RBSB with the baseline schemes. From the analysis, it is noticed that the average delay of the baseline schemes, i.e., CoISDA (37.2 ms) and OSP (46.9 ms), is increased while varying the task arrival rate, which is higher than the proposed RBSB strategy (19.4 ms). The main reason behind that the existing schemes do not consider any ranking model to classify the incoming financial tasks based on their importance and assign them to suitable computing devices. However, the proposed RBSB method used a ranked-based classification model and an efficient service deployment strategy for analyzing the FinTech tasks at the edge of the networks, which reduces the delay. The proposed RBSB strategy has minimized the delay by 17.8% and 27.5% over CoISDA and OSP, respectively.

2) *Analysis of Power Consumption*: The impact of power consumption during the financial task assignment from the IoT devices to the remote computing devices through a local gateway is shown in Fig. 5. From Fig. 5(a), it is noted that the total required power of the IoT device (24.53 mW) is less than the distributed edge devices (33.67 mW) or remote cloud servers (46.82 mW) while task analysis. However, the total power consumption of the financial tasks depends on the uploading, downloading, and processing power. Besides that, the long communication distance between the IoT devices and remote computing devices can increase the uploading and downloading time of the financial tasks, which further increases the total power consumption. The proposed RBSB strategy distributes the ranked-based tasks on the local edge devices (mainly risk-based tasks), which causes communication distance and required power consumption of the FinTech tasks. Fig. 5(b) presents the comparative analysis of average power consumption of the proposed strategy with baseline schemes. From the analysis, it is observed that the proposed strategy consumes low power (29.93 mW), while the existing schemes CoISDA and OSP consume 37.71 and 43.59 mW, respectively. Moreover, the quantitative analysis results show that RBSB outperforms over CoISDA and OSP schemes, which reduces the power consumption by 7.7% and 13.6%, respectively.

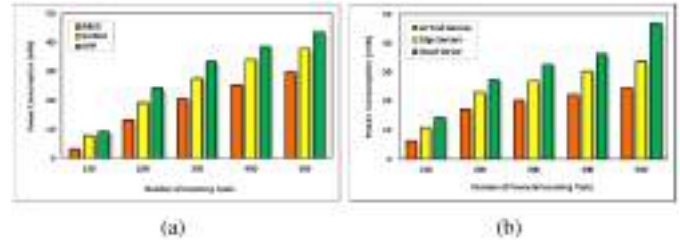


Fig. 5. Impact of task assignment over power consumption. (a) Various financial tasks. (b) Comparative analysis with baseline schemes.

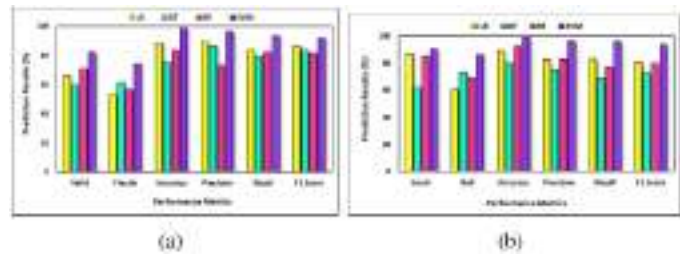


Fig. 6. Edge-level analysis using MLCAs. (a) Prediction results of D1. (b) Prediction results of D2.

3) *Predictive Analysis at Edge Level*: The predictive analysis results of various classification models at the edge devices are listed in Table III. After uploading the risk-based financial tasks to the local edge devices, the standard classification models have been applied over the risk-based data sets. In the edge-based analysis, two different types of risk-based financial data sets (i.e., D1 and D2) are considered to validate and test the classification models. The prediction results of standard classification models with respect to the various performance metrics over D1 and D2 are shown in Fig. 6(a) and (b), respectively. From the analysis, it is evident that the SVM model provides better accuracy over the standard classification models, such as LR, DT, and RF models. The SVM model achieves 98.49% accuracy while predicting the valid and fraud customers using the D1 data set. However, the accuracy result of this model is different when considering the D2 data set to predict the good and bad credit risk assessments. In this case, the accuracy rate of the SVM classifier achieves 99.02%, which is much higher than other standard classification models. Thus, SVM yields a minimum mean absolute error of 0.27 at edge level, which is less than the standard baseline models. This is achieved by ranking and selecting more critical features from the data set before training the models at edge networks.

4) *Predictive Analysis at Cloud Level*: The predictive analysis results of various classification models at the cloud server are summarized in Table IV. The proposed service deployment

TABLE III
PREDICTION ACCURACY OF VARIOUS CLASSIFICATION MODELS IN EDGE SERVER

Edge Level Analysis							
Dataset	MLCA Models	Fraud Detection		Accuracy	Precision	Recall	F1 Score
		Valid	Frauds				
D1	LR	0.6645	0.5331	0.8867	0.8959	0.8362	0.8636
	DT	0.5993	0.6148	0.7532	0.8642	0.7925	0.8386
	RF	0.7076	0.5637	0.8387	0.7306	0.8254	0.8159
	SVM	0.8228	0.7406	0.9849	0.9639	0.9356	0.9211
Dataset	MLCA Models	Risk Prediction		Accuracy	Precision	Recall	F1 Score
		Good	Bad				
D2	LR	0.8711	0.6039	0.8946	0.8273	0.8306	0.8093
	DT	0.6203	0.7321	0.7997	0.7527	0.6914	0.7236
	RF	0.8511	0.6939	0.9246	0.8273	0.7706	0.7993
	SVM	0.9062	0.8657	0.9902	0.9615	0.9558	0.9381

TABLE IV
PREDICTION ACCURACY OF VARIOUS CLASSIFICATION MODELS IN CLOUD SERVER

Cloud Level Analysis							
Dataset	MLCA Models	Churn Prediction		Accuracy	Precision	Recall	F1 Score
		Churned	Retained				
D3	LR	0.7939	0.6133	0.8618	0.7457	0.7822	0.7635
	DT	0.5846	0.6674	0.9465	0.8769	0.9031	0.8328
	RF	0.6382	0.7092	0.9013	0.7643	0.8429	0.7976
	SVM	0.8915	0.8365	0.9964	0.9523	0.9241	0.9354
Dataset	MLCA Models	Insurance Prediction		Accuracy	Precision	Recall	F1 Score
		Claimed	Unclaimed				
D4	LR	0.4835	0.5960	0.7953	0.8067	0.7714	0.7602
	DT	0.6167	0.4928	0.8802	0.7561	0.6992	0.7353
	RF	0.7522	0.6239	0.9350	0.8134	0.8519	0.8225
	SVM	0.8908	0.7014	0.9626	0.9257	0.8911	0.9076

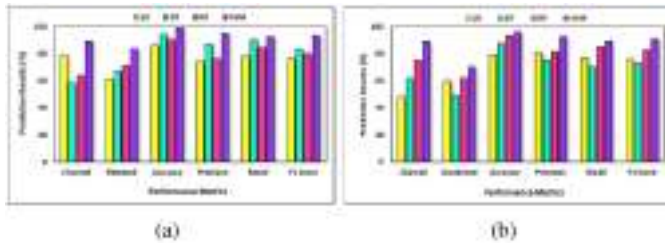


Fig. 7. Cloud-level analysis using MLCAs. (a) Prediction results of D3. (b) Prediction results of D4.

strategy deployed the nonrisk-based financial tasks to the cloud server and the standard classification models have been applied over the nonrisk-based financial data sets for further analysis. In the cloud-based analysis, two different types of nonrisk-based financial data sets (i.e., D3 and D4) are considered to validate and test the classification models. The prediction results of the standards classification models over D3 and D4 are shown in Fig. 7(a) and (b), respectively. From the analysis, it is observed that the accuracy of the SVM classification model is greatly increased than the other standard classification models. The SVM classification model achieves 99.64% accuracy while predicting the churned and retained banking customers using the D3 data set. However, the accuracy of the same model for the D4 data set is improved by 99.26%, which predicts the status of claimed and unclaimed insurance of the customers, which is higher than the standard classification models. Thus, SVM yields a minimum mean absolute error of 0.36 at cloud level, which is less than the standard baseline models.

Also, it is noticed that the values of precision, recall, and F1 score for all the data sets (i.e., D1–D4) show higher variations in the SVM model, whereas other classification models yield fewer variations for the same set of performance metrics. Thus, the proposed RBSD strategy along with the SVM classification model improves the risk prediction accuracy of the financial tasks and power consumption of the edge networks.

V. CONCLUSION

In this article, we have proposed an RBSD strategy for predictive financial data analysis at the edge networks. The main aim of this work is to analyze the risk-based financial task at the local edge devices with a standard SVM algorithm for minimizing the average delay and power consumption while maximizing the prediction accuracy. To achieve this, a ranked-based strategy has been designed for classifying the incoming financial tasks based on their priorities. Furthermore, a service deployment strategy has been developed using a perfect matching theorem, i.e., Hall theorem for assigning the classified task on the suitable remote computing devices as per their importance. Extensive simulation results exhibit the effectiveness of the proposed RBSD strategy and the SVM algorithm at edge networks over baseline algorithms and standard classification models, respectively. The proposed strategy minimizes 17.8%–27.5% average delay and 7.7%–13.6% power consumption over the baseline algorithms. Furthermore, the SVM algorithm achieves 98.49%, and 99.02% accuracy while analyzing the data at the edge level of the network. In the future, we will enhance the proposed strategy for FinTech application by introducing various data aggregation and data

fusion techniques at edge networks for minimizing network overhead and achieving higher prediction accuracy.

REFERENCES

- [1] M. Abbasi, H. Rezaei, V. G. Menon, L. Qi, and M. R. Khosravi, "Enhancing the performance of flow classification in SDN-based intelligent vehicular networks," *IEEE Trans. Intell. Transp. Syst.*, early access, Aug. 13, 2020, doi: [10.1109/TITS.2020.3014044](https://doi.org/10.1109/TITS.2020.3014044).
- [2] S. N. Srirama, F. M. S. Dick, and M. Adhikari, "Akka framework based on the actor model for executing distributed fog computing applications," *Future Gener. Comput. Syst.*, vol. 117, pp. 439–452, Apr. 2021.
- [3] A. Mukherjee, M. Li, P. Goswami, L. Yang, S. Garg, and M. J. Piran, "Hybrid NN-based green cognitive radio sensor networks for next-generation IoT," *Neural Comput. Appl.*, to be published.
- [4] S. Mostafi, F. Khan, A. Chakrabarty, D. Y. Suh, and M. J. Piran, "An algorithm for mapping a traffic domain into a complex network: A social Internet of Things approach," *IEEE Access*, vol. 7, pp. 40925–40940, 2019.
- [5] B. Ji *et al.*, "A survey of computational intelligence for 6G: Key technologies, applications and trends," *IEEE Trans. Ind. Informat.*, early access, Jan. 18, 2021, doi: [10.1109/TII.2021.3052531](https://doi.org/10.1109/TII.2021.3052531).
- [6] A. Hazra, M. Adhikari, T. Amgoth, and S. N. Srirama, "Stackelberg game for service deployment of IoT-enabled applications in 6G-aware fog networks," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5185–5193, Apr. 2021.
- [7] M. J. Piran *et al.*, "Multimedia communication over cognitive radio networks from QoS/QoE perspective: A comprehensive survey," *J. Netw. Comput. Appl.*, vol. 172, Dec. 2020, Art. no. 102759.
- [8] D. Thomas *et al.*, "QoS-aware energy management and node scheduling schemes for sensor network-based surveillance applications," *IEEE Access*, vol. 9, pp. 3065–3096, 2021.
- [9] A. Mukherjee, P. Goswami, M. A. Khan, L. Manman, L. Yang, and P. Pillai, "Energy efficient resource allocation strategy in massive IoT for industrial 6G applications," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5194–5201, Apr. 2021.
- [10] S. R. Pokhrel, S. Verma, S. Garg, A. K. Sharma, and J. Choi, "An efficient clustering framework for massive sensor networking in industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 4917–4924, Jul. 2021.
- [11] C. Gong, F. Lin, X. Gong, and Y. Lu, "Intelligent cooperative edge computing in Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 9372–9382, Oct. 2020.
- [12] S. Verma, S. Kaur, M. A. Khan, and P. S. Sehdev, "Toward green communication in 6G-enabled massive Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5408–5415, Apr. 2021.
- [13] C. Zhang, X. Wang, Y. Zhao, A. Dong, F. Li, and M. Huang, "Cost efficient and low-latency network service chain deployment across multiple domains for SDN," *IEEE Access*, vol. 7, pp. 143454–143470, 2019.
- [14] Y. Chen, Y. Sun, T. Feng, and S. Li, "A collaborative service deployment and application assignment method for regional edge computing enabled IoT," *IEEE Access*, vol. 8, pp. 112659–112673, 2020.
- [15] X. Cao, G. Tang, D. Guo, Y. Li, and W. Zhang, "Edge federation: Towards an integrated service provisioning model," *IEEE/ACM Trans. Netw.*, vol. 28, no. 3, pp. 1116–1129, Jun. 2020.
- [16] S. K. Mishra, D. Puthal, B. Sahoo, S. Sharma, Z. Xue, and A. Y. Zomaya, "Energy-efficient deployment of edge datacenters for mobile clouds in sustainable IoT," *IEEE Access*, vol. 6, pp. 56587–56597, 2018.
- [17] L. Liu, Z. Chang, X. Guo, S. Mao, and T. Ristaniemi, "Multiobjective optimization for computation offloading in fog computing," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 283–294, Feb. 2018.
- [18] Y. Wu, B. Shi, L. P. Qian, F. Hou, J. Cai, and X. S. Shen, "Energy-efficient multi-task multi-access computation offloading via noma transmission for IoTs," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4811–4822, Jul. 2020.
- [19] X. Wei, C. Tang, J. Fan, and S. Subramaniam, "Joint optimization of energy consumption and delay in Cloud-to-Things continuum," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2325–2337, Apr. 2019.
- [20] A. Thennakoon, C. Bhagyan, S. Premadasa, S. Mihiranga, and N. Kuruwitaarachchi, "Real-time credit card fraud detection using machine learning," in *Proc. IEEE 9th Int. Conf. Cloud Comput. Data Sci. Eng. (Confluence)*, 2019, pp. 488–493.
- [21] T. M. Alam *et al.*, "An investigation of credit card default prediction in the imbalanced datasets," *IEEE Access*, vol. 8, pp. 201173–201198, 2020.
- [22] A. Mashrur, W. Luo, N. A. Zaidi, and A. Robles-Kelly, "Machine learning for financial risk management: A survey," *IEEE Access*, vol. 8, pp. 203203–203223, 2020.
- [23] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A secure AI-driven architecture for automated insurance systems: Fraud detection and risk measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020.
- [24] S. Makki, Z. Assaghir, Y. Taher, R. Haque, M.-S. Hacid, and H. Zeineddine, "An experimental study with imbalanced classification approaches for credit card fraud detection," *IEEE Access*, vol. 7, pp. 93010–93022, 2019.
- [25] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: Analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE Access*, vol. 7, pp. 60134–60149, 2019.
- [26] S. García-Méndez, M. Fernández-Gavilanes, J. Juncal-Martínez, F. J. González-Castaño, and Ó. B. Seara, "Identifying banking transaction descriptions via support vector machine short-text classification based on a specialized labelled corpus," *IEEE Access*, vol. 8, pp. 61642–61655, 2020.
- [27] B. N. Pambudi, I. Hidayah, and S. Fauziati, "Improving money laundering detection using optimized support vector machine," in *Proc. IEEE Int. Seminar Res. Inf. Technol. Intell. Syst. (ISRITI)*, 2019, pp. 273–278.
- [28] Y. Li, "Credit risk prediction based on machine learning methods," in *Proc. IEEE 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, 2019, pp. 1011–1013.
- [29] A. A. Khine and H. W. Khin, "Credit card fraud detection using online boosting with extremely fast decision tree," in *Proc. IEEE Conf. Comput. Appl. (ICCA)*, 2020, pp. 1–4.
- [30] S. Xuan, G. Liu, Z. Li, L. Zheng, S. Wang, and C. Jiang, "Random forest for credit card fraud detection," in *Proc. IEEE 15th Int. Conf. Netw. Sens. Control (ICNSC)*, 2018, pp. 1–6.

Optimal Distribution of Workloads in Cloud-Fog Architecture in Intelligent Vehicular Networks

Mahdi Abbasi¹, Mina Yaghoobikia, Milad Rafiee², Mohammad R. Khosravi³,
and Varun G. Menon⁴, *Senior Member, IEEE*

Abstract—With the fast growth in network-connected vehicular devices, the Internet of Vehicles (IoV) has many advances in terms of size and speed for Intelligent Transportation System (ITS) applications. As a result, the amount of produced data and computational loads has increased intensely. A solution to handle the vast volume of workload has been traditionally cloud computing such that a substantial delay is encountered in the processing of workload, and this has made a serious challenge in the ITS management and workload distribution. Processing a part of workloads at the edge-systems of the vehicular network can reduce the processing delay while striking energy restrictions by migrating the mission of handling workloads from powerful servers of the cloud to the edge systems with limited computing resources at the same time. Therefore, a fair distribution method is required that can evenly distribute the workloads between the powerful data centers and the light computing systems at the edge of the vehicular network. In this paper, a kind of Genetic Algorithm (GA) is exploited to optimize the power consumption of edge systems and reduce delays in the processing of workloads simultaneously. By considering the battery depreciation, the supporting power supply, and the delay, the proposed method can distribute the workloads more evenly between cloud and fog servers so that the processing delay decreases significantly. Also, in comparison with the existing methods, the proposed algorithm performs significantly better in both using green energy for recharging the fog server batteries and reducing the delay in processing data.

Index Terms—Cloud, fog, genetic algorithm, Internet of vehicles, workload allocation.

I. INTRODUCTION

AS a result of the tremendous growth in the number of smart vehicular devices, the Internet of Vehicles (IoV) has experienced rapid expansion. The increase in the number of devices has caused a multiplication of data and large-scale computation loads [1], [2]. Cloud computing has been proposed as a solution to manage these loads [3]. However,

Manuscript received May 7, 2020; revised October 13, 2020 and February 13, 2021; accepted April 2, 2021. Date of publication April 19, 2021; date of current version July 12, 2021. The Associate Editor for this article was A. Jolfaei. (*Corresponding author: Mahdi Abbasi.*)

Mahdi Abbasi, Mina Yaghoobikia, and Milad Rafiee are with the Department of Computer Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamedan 65178-38695, Iran (e-mail: abbasi@basu.ac.ir; m.yaghoobikia@eng.basu.ac.ir; m.rafee@alumni.basu.ac.ir).

Mohammad R. Khosravi is with the Department of Computer Engineering, Persian Gulf University, Bushehr 75169-13817, Iran, and also with the Department of Electrical and Electronics Engineering, Shiraz University of Technology, Shiraz 71557-13876, Iran (e-mail: m.khosravi@sutech.ac.ir).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Kochi 683582, India (e-mail: varunmenon@scmsgroup.org).

Digital Object Identifier 10.1109/TITS.2021.3071328

1558-0016 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See <https://www.ieee.org/publications/rights/index.html> for more information.

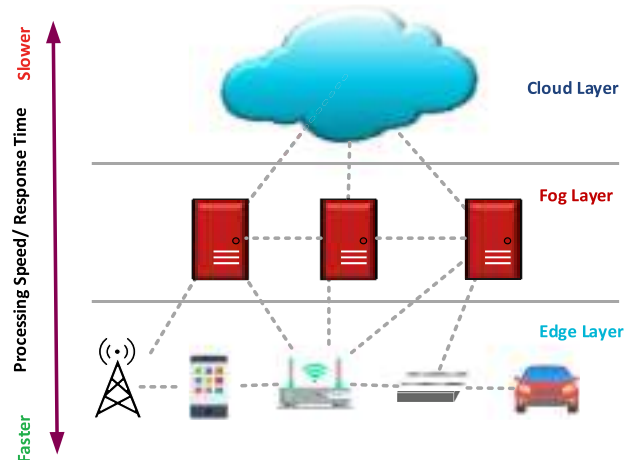


Fig. 1. IoV data processing layer stack.

the time-consuming nature of the processing of workloads in clouds is still a major issue in the field of distributed vehicular networks [4]. Processing the workloads at the edges of the vehicular network can reduce the processing time, but the transmission of workloads from the data centers (which are equipped with sustainable electric power) to the edges leads to serious limitations in terms of supplying the required power for computing [5], [6]. Thus, we need to achieve a balance in distributing the processing requests between the cloud and the edge [7].

Fig. 1 shows the layers of data processing in a cloud-fog architecture for IoV. As can be seen in the figure, the lowest layer contains vehicular devices that produce the data. These devices can use their own processing resources and process the data in positions close to the user. Although the proximity of edge devices to the end-user remarkably reduces the delay in request transmission and increases the response rate, these devices have a lower processing power than the cloud. In the next layer lie powerful routers and servers which are close to the edge and can process the workloads without transferring them to the cloud [8]. By moving away from the edge of the Internet to towards the data centers, the transmission delay will increase. In the highest layer, large data centers that provide the enhanced capability of processing and storage are distributed as clouds around the world. As they are very far from the end-users, these resources usually impose long delays in the processing of requests [9]. Also, they also consume high amounts of electric power, whereas most edge devices can function with small amounts of power or even with batteries.

Fog computing is a kind of distributed computing that can replace cloud computing by using several devices near the edge of the vehicular network (see Fig. 1). Fog computing is more efficient than edge computing in terms of processing, while it is less potent than cloud computing. The chief issue in fog computing is the high costs of the required electric power. Nowadays, a more challenging problem is providing sustainable energy resources that can afford the long-term energy requirements of fog nodes in IoV [10], [11]. The processing nodes chiefly receive their required power from rechargeable batteries [12]. As this type of power source is extremely limited and should be frequently recharged, the use of renewable energies as a secondary or even the only power supply at the network edge is necessary [13]–[15]. Thus, we need to develop a method for striking a balance in distributing the workloads among fog nodes and cloud data centers so that both delay and power consumption could be optimized. As a result, the energy resources of the IoV become more sustainable.

To achieve this goal, the present study makes use of a genetic algorithm in finding the best distribution for the workloads. A review of the literature indicates that few studies have addressed the issue of finding the best cost function and the effect of the coefficients of this function on the algorithm's decision-making. Given this, this study first introduces a cost function of distribution based on two parameters, i.e., power and delay, and then attempts to modify the coefficients corresponding to these parameters according to a genetic algorithm in order to attain the best coefficients of workload distribution in a way that the workloads could be processed with the least delay and the least amount of power consumption.

The genetic algorithm is a method for finding approximate solutions to search and optimization problems. This algorithm is considered as a kind of evolutionary algorithm due to its use of biological concepts such as inheritance and mutation. Genetics addresses inheritance and the transfer of attributes from one generation to the next. In living creatures, chromosomes and genes are responsible for this transfer. This mechanism acts in a way that superior and stronger chromosomes will survive. The final result is that stronger creatures would be able to survive. Genetic programming is a technique of programming that uses genetic evolution as a model for problem-solving. Over time, the genetic algorithm has grown in popularity in a diversity of problems such as optimization, image processing, topology, artificial neural network training, and decision-making systems [16].

A genetic algorithm begins with initializing a random population, which is composed of the possible solutions to the problem. Each solution is a chromosome, and the entire chromosomes form the initial population. In the first step, the value of each chromosome in the population is specified by the fitness function. During the execution of the algorithm, parents with more fitness are selected for reproduction, and the next generation is generated using genetic operators. Crossover, mutation, and selection are the three main operators in genetic algorithms [17].

By using a genetic algorithm, the present paper seeks to obtain the best value function so that we could strike a balance between power consumption at the edge of network and delay

in the transmission of workloads as well as minimize these two parameters. Also, we use renewable energies in the processing and transmission of workloads at the network edge due to the significance of sustainable energy resources in computing tasks. Another innovation of this study is its use of renewable energy as an input parameter in the genetic algorithm. For this purpose, green energy is used in our proposed method to calculate the value function of the algorithm. The main reason for using renewable energies is the limitation of edge devices on power consumption. As a result, these devices need to be regularly connected to a power source for being recharged, which limits their mobility. Also, changing the battery in IoV devices may impose high costs and is sometimes dangerous. For this reason, IoV devices should be able to maintain their independence and sustainability by using green energies and wireless charging ability [18]. In this vein, we aim to utilize renewable energies to minimize the number of batteries at the network edge.

The structure of the paper is as follows. Section 2 is a review of the related literature. In Section 3, the workload allocation model is formulated. Next, after a brief description of the structure of the genetic algorithm, the optimization of delay and power consumption in a cloud-fog environment is discussed. Section 5 describes the implementation of the proposed method and evaluates it in terms of the parameters of the algorithm. The method is also compared with other existing methods. Finally, some concluding remarks are made and ideas for further research are suggested.

II. REVIEW OF LITERATURE

In recent years, many researchers have studied the methodologies to orchestrate the distribution of workloads and reduce the overall processing delay in IoV. We briefly review some of the recent studies that have investigated the optimization of energy usage and delay reduction in fog computing.

Pioneering work has been presented by Xu and Ren [19]. In this work, they inspect the possibility of using renewable energies as backup energy sources in mobile edge networks. Their method uses machine learning algorithms to manage the energy resources and distribute the computation workloads. Their method aims to minimize the prominent costs of the processing requests that include the processing delay and consuming energy. The consequence of using a slow learning mechanism in their method is the weak results in controlling the power consumption of edge nodes.

Unfortunately, in many of the recently proposed methods, only one aspect of the problem is considered. That is, some of them sacrifice the processing time for optimizing the power consumption at edge systems, and vice versa. Hence, in any method to be developed, the processing delay and consuming power should be optimized simultaneously.

Regarding abovementioned challenge, Xu *et al.* presented a reinforcement learning-based method [20]. Their algorithm was able to learn and adapt itself to any system with unknown modeling parameters. Despite the undeniable results of their method in achieving acceptable performance in orchestrating the edge computations and using renewable energy sources for mobile edge nodes, their algorithm failed to fairly distribute the workloads among the computing nodes.

The GLOBE method of Wu *et al.* [12] tries to optimize the performance processing nodes at the network edge by geographically balancing the distribution of loads, and at the same time, controlling the input load of any edge nodes. This method can handle stochastic events concerning the battery status and power limitations. Although the GLOBE is slightly successful in optimizing the battery energy level, it is still so far from perfect.

In 2019, Dalvand and Zamanifar [21] proposed a new model for processing data in the Internet of things (IoT) and developed an IoT-Fog-Cloud in which the fog layer is geographically close to the IoT edge devices. In their system, a multi-purpose dynamic service is created to achieve a balance between delay and resource costs. This service is formulated by MILP and solved through weighted goal programming. The method controls and minimizes only one goal as a compromise between time and power consumption.

None of the above studies have been able to strike a balance between cost and power consumption in distributing workloads among network nodes. Our work is aimed at developing a mechanism for the balanced distribution of workloads between the cloud and edge nodes. This mechanism is supposed to attain a desirable tradeoff between power consumption and workload delay. It will make use of renewable energy sources as the power supply for edge computations. These sources are expected to preserve the battery charge level.

III. THE PROPOSED METHOD

Our point of departure is the fact that none of the previous works have offered an optimum solution to reduce the costs of fog computing. To elaborate on our proposed method and evaluate it, a cloud-fog environment is simulated as in [10]. This environment involves four main models: workload, delay, power consumption, and battery status. In the following, we shall first examine these models as described in the reference and then present our proposed algorithm. Next, we will define a new scenario to study how the algorithm functions. The proposed scenario will be simulated for evaluation.

A. Formulation of the Problem

For the chief scenario, the edge system includes a base station and a set of edge servers that are set geographically close to each other. A battery with a limited capacity is used in each computing resource at the network edge (i.e., fog servers). The shared power supply mechanism used in the network lets the workloads be sent to the cloud especially when the edge servers require battery charge. The workload sent by the users to the edge is first received by the base station. The base station manages and decides on the amount of workload that must be allocated to the edge or transmitted to the cloud. The definition of the formulation parameters is presented in Table I.

The proposed system is modeled here by considering it from four aspects.

1) *Workload*: The equal time intervals $t = 0, 1, 2, \dots$ are used to model the time. The computation capacity of each edge device is specified in each time interval in terms of the number of active servers; $\lambda(t) \in [0, \lambda_{max}]$ is the rate of assigning the workloads to edge nodes and $\mu(t)$

TABLE I
THE MAIN SYMBOLS

Symbol	Meaning	Symbol	Meaning
$\lambda(t)$	The total rate of the input load	$c_{delay}(t)$	The total cost of delays
$\mu(t)$	The amount of workload processed locally	$c_{back}(t)$	The cost of using the supporting power supply
$m(t)$	The number of active servers at the network edge	$d_{op}(t)$	Power consumption for operational tasks at the network edge
$c_{lo}(t)$	The cost of delay in processing workloads at the network edge	$c_{comp}(t)$	Power consumption for processing loads at the network edge
$c_{off}(t)$	The cost of delay in sending workloads to the cloud	$d(t)$	Total power consumption
$g(t)$	The total renewable power received	$b(t)$	Battery level at the network edge
$h(t)$	The congestion status of the network	$s(t)$	System status

is the rate by which the edge nodes process the assigned workload. Finally, $\lambda(t) - \mu(t)$ denotes the remaining part of the workload transmitted to the cloud. The number of dynamic servers in any time interval is $m(t) \in [0, M]$. This number may change in different time intervals.

2) *Delay*: We consider three different delays in the system model:

2-1. The delay in communicating workloads on the wireless network, which is shown by $c_{wi}(t)$. This delay depends on the input load of the network (i.e., $\lambda(t)$). In our model, it is assumed as 0 due to the physical closeness of the active nodes.

2-2. The delay of processing workloads at local subregions of the network edge, which is shown by $c_{lo}(t)$. The amount of this delay directly depends on the number of active servers, the processing rate of them, and the model of managing queues in each of them. In our experiments, the M/G/1 mechanism models the queue management in any active server running on edge nodes. As a result, the delay in processing at the network edge is estimated using the following equation [22]:

$$c_{lo}(\mu(t), m(t)) = \frac{\mu(t)}{m(t) \cdot k - \mu(t)} \quad (1)$$

In this equation, kk represents the processing capacity of each active server.

2-3. The delay in communicating the residual workload to the cloud is shown by $c_{off}(t)$ and is estimated based on the congestion status of the network. This status is represented by $h(t)$, and is computed by adding the round-trip time (RTT) delay and the processing delay of the cloud. As a result, this delay is calculated based on $h(t)$ according to the following equation [22]:

$$c_{off}(h(t), \mu(t), \lambda(t)) = (\lambda(t) - \mu(t)) h(t) \quad (2)$$

Finally, the cost of overall delay of the aggregate input workload is estimated by adding the three above delays [22]:

$$\begin{aligned} c_{delay}(h(t), \lambda(t), \mu(t), m(t)) \\ = c_{wi}(\lambda(t)) \\ + c_{lo}(\mu(t), m(t)) + c_{off}(h(t), \mu(t), \lambda(t)) \end{aligned} \quad (3)$$

Note that, $c_{wi}(\lambda(t))$ is negligible.

3) *Power Consumption*: The total consumed power is composed of two parts:

3-1. A part of the power is used for basic operations and communicating the loads. This part is represented by $d_{op}(t)$ and is independent of any operations regarding

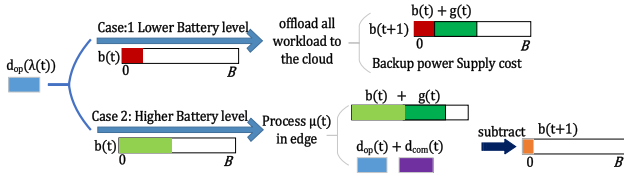


Fig. 2. Two modes of battery status.

processing the loads but merely depends on the input load of the network ($\lambda(t)$). In our model, $d_{op}(t)$ is composed of two power types [22]:

$$d_{op}(\lambda(t)) = d_{sta} + d_{dyn}(\lambda(t)) \quad (4)$$

The d_{sta} and $d_{dyn}(\lambda(t))$ represent the static power consumption of the network edge and the dynamic power consumption, respectively. The latter differs from the input load of the network and is set to 0 in our model due to the physical closeness of the computing nodes.

3-2. The power required for the processing of workloads at the edge is shown by $d_{com}(t)$. To estimate this parameter, the amount of the workload allocated to the edge ($\mu(t)$) and the number of active edge servers ($m(t)$) are required. Finally, the total required power is obtained by the following equation [22]:

$$d(\lambda(t), \mu(t), m(t)) = d_{op}(\lambda(t)) + d_{com}(\mu(t), m(t)) \quad (5)$$

In this model, $g(t)$ denotes any renewable energy source that can be used as the power supply $g(t)$.

4) Battery Status: As formerly explained, one battery with limited charge is used to supply each active edge server. Overall, the total battery charge at the network edge is $b(t) \in [0, B]$, where B denotes a predefined maximum capacity. The renewable energy sources can recharge these batteries. The initial battery level is set to 0. To control the battery level at the network edge, we should control the rate of processing of workloads at the edge servers. Hence, the state of the battery is determined by the following conditions:

D-1. When $b(t) \leq d_{op}(\lambda(t))$, no processing is allowed at the network edge. In this state, since the battery charge is not sufficient, the whole workload $\lambda(t)$ is transmitted to the cloud. In this state, the renewable energy sources recharge the battery. The overall cost of communicating the workload to the cloud is calculated by the following equation [22]:

$$c_{bak}(\lambda(t)) = \varphi \cdot d_{op}(\lambda(t)) \quad (6)$$

where φ is the coefficient reflecting the cost of consuming the supporting power supply. In the next interval, the renewable power source will charge the battery according to equation (7) [22].

$$b(t+1) = b(t) + g(t) \quad (7)$$

The first state in Fig. 2 shows this state.

D-2. If the battery level is sufficiently more than the required power for processing a part of workload, that part of the workload ($\mu(t)$) is processed at the edge, and the remaining part ($\lambda(t) - \mu(t)$) is transmitted to the cloud. Thus,

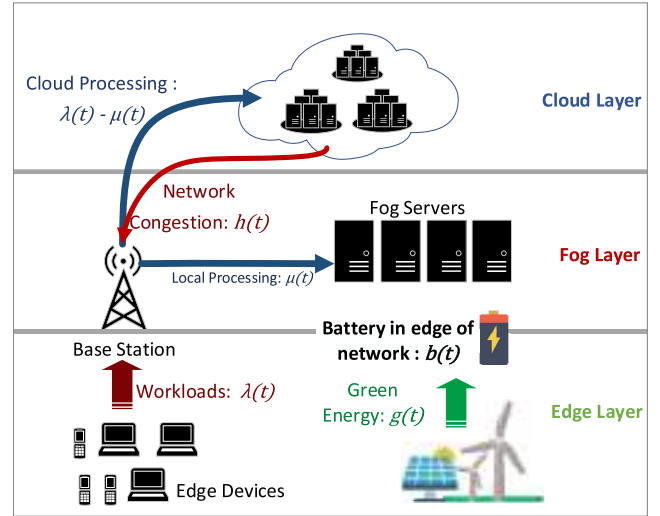


Fig. 3. The cloud-fog architecture.

the following equation calculates the battery level in the next interval as:

$$b(t+1) = b(t) + g(t) - d(\lambda(t), \mu(t), m(t)) \quad (8)$$

The operational cost of the battery in this state is:

$$c_{battery}(t) = \omega \cdot \max\{d(\lambda(t), \mu(t), m(t)) - g(t), 0\} \quad (9)$$

where $\omega > 0$ is the operating cost of one battery unit.

This state is shown by the second state in Fig. 2. Based on the above four models, the architecture of the proposed system can be illustrated as in Fig. 3.

In this figure, the set of requests $\lambda(t)$ which have been sent by the users enter the base station. The base station is responsible for distributing the loads between edge servers (i.e., fog servers) and the cloud. The base station uses the evolutionary algorithm to calculate the amount of workload that can be processed by edge servers ($\mu(t)$). Then, the excessive requests are transmitted to the cloud. The transmission of workloads to the cloud creates congestion in the network and imposes longer delays on the loads. Therefore, the congestion is measured in every interval ($h(t)$) to be taken into account in subsequent decisions. In the meantime, renewable energy sources ($g(t)$) provide the power required for edge computations in each interval. If renewable sources produce more energy than is needed by the servers, the surplus is stored in network batteries $b(t)$. Conversely, if the produced renewable energy is inadequate, the batteries will be used.

IV. USING GENETIC ALGORITHM IN THE OPTIMIZATION OF WORKLOAD DISTRIBUTION

This section describes how a genetic algorithm can be used to distribute the workloads more efficiently. The aim of using a genetic algorithm is to minimize system costs. The solution to this problem using a genetic algorithm is presented in Algorithm 1. Below is the description of the algorithm.

In the beginning, the battery level is checked. If the battery level is not sufficient for the basic operation, the supporting power supply is used, and the entire input workload is transmitted to the cloud. In this case, $\mu(t) = 0$, and the genetic

algorithm is not executed (lines 1 and 2). However, if the battery level is high enough for the basic operation to run, all or part of the input load can be processed at the network edge. In this case, the genetic algorithm is used to calculate $\mu(t)$ (lines 3 to 29). In the first step of the genetic algorithm, the initial population is generated (line 4). This population consists of a set of chromosomes. Each chromosome indicates the amount of workload that can be computed at the edge. Next, the fitness of the initial population is calculated (line 5).

The fitness function returns a non-negative value for each chromosome which is indicative of the individual capacity of that chromosome to reduce the costs. The cost function [20] can be used to calculate the fitness of a chromosome. The proposed algorithm attempts to reduce this amount in order to minimize system costs. Given the battery status of the system, the cost function can be calculated in two ways:

$$c(t) = c_{delay}(h(t), \lambda(t), 0, 0) + c_{bak}(\lambda(t)),$$

$$if(b(t) \leq d_p(\lambda(t)))$$

$$c(t) = c_{delay}(h(t), \lambda(t), \mu(t), m(t)) + c_{battery}(t),$$

$$else \quad (10)$$

This equation is composed of two parts: delay cost and power cost. The following two coefficients are used for the power cost part:

- 1) Battery depreciation coefficient (ω)
- 2) Cost coefficient of the supporting power supply (φ)

As the effect of delay is directly involved in the cost function, a new coefficient called delay cost coefficient (θ) is introduced. The proposed algorithm modifies these coefficients to examine their effect on power consumption and workload delay and to find the optimum state on the network.

Another important genetic operator is crossover. Crossover is used to exchange information between two chromosomes, which accelerates convergence in the genetic algorithm. The probability of the effectiveness of this operator lies in the range of 0.6-0.9. This value is called a crossover rate and denoted by $P_{crossover}$. In this problem, two parents and a random position in the parents' genes are selected. Next, the genes on the right side of the random position of the first parent and those on the left side of the random position of the second parent are selected to produce a new chromosome (lines 9 to 15). Another operator is the mutation, which is responsible for producing new information. This operator randomly changes one of the genes of the child with a low probability, such as 0.01. The probability of mutating any chromosome is called the mutation rate and is denoted by $P_{mutation}$. In the proposed algorithm, one gene from the chromosome is randomly selected and changed (lines 16-19). In this algorithm, the number of children produced by crossover and mutation is set by the variable N_c . In each step of this operation, a new child is added to the set P (line 20). Then the fitness function of the generated population is obtained by crossover and mutation operators as was done for the initial population (line 22).

There are different methods in genetic algorithms to select the superior chromosome and transfer it to the next generation. One of the common methods is tournament selection [23]. In this method, two chromosomes are randomly selected from the population. Next, a random number r between 0 and 1

Algorithm 1 Using a Genetic Algorithm in the Optimization of the Workload Distribution

Input : $\lambda, g, h, b, N_c, N_g, P_{crossover}, P_{mutation}, P_{selection}$
Output : μ

```

1: if  $b(t) \leq d_p(\lambda(t))$ 
2:    $\mu(t) \leftarrow 0$ 
3: else
4:    $P \leftarrow Create\ Population()$ 
5:    $fitness(P)$ 
6:   do
7:     for  $i \leftarrow 0, 1, \dots, N_c // crossover\ and\ mutation$ 
8:        $parent1 \leftarrow random(P)$ 
9:        $parent2 \leftarrow random(P)$ 
10:       $child \leftarrow parent1$ 
11:      if  $(random() > P_{crossover})$ 
12:         $point \leftarrow$ 
13:           $random(length\ of\ chromosome)$ 
14:           $child \leftarrow$ 
15:             $crossover(parent1, parent2, point)$ 
16:      End if
17:      if  $(random() > P_{mutation})$ 
18:         $gen \leftarrow random(length\ of\ chromosome)$ 
19:         $child(gen) \leftarrow mutation()$ 
20:      End if
21:       $Add\ a\ child\ to\ P$ 
22:    end for
23:     $fitness(children\ created\ by\ crossover\ section)$ 
24:     $P \leftarrow selection(P, P_{selection})$ 
25:  while
26:     $\mu(t) \leftarrow best\_chromosome(P)$ 
27:    while
28:       $(battery + green < PowerConsumption(\mu(t)))$ 
29:       $\mu(t) \leftarrow next\_best\_chromosome(P)$ 
30:    end while
31: End if

```

is generated. If $r < P_{selection}$ ($P_{selection}$ is a parameter, e.g., 0.8), the fitter individual will be selected as the parent; otherwise, the less fit individual will be selected. These two are again returned to the population and involved in the selection process. After the selection process, the selected chromosomes are introduced as the new generation and sent to the next iteration of the algorithm (line 23). In the proposed genetic algorithm, the child generation operators such as crossover and mutation as well as fitness calculation and selection are executed for N_g times, which is indicative of the number of generations (line 24). When all generations have been executed, the first element of the population will be put in $\mu(t)$ as the final result (line 25).

If the selected chromosome (which indicates the distribution of processable load at the network edge $\mu(t)$) faces battery limitations, the next chromosome in the population should be selected. The process will continue until the power consumption for $\mu(t)$ becomes proportional to the edge batteries (lines 26-28). At the end of the algorithm, the best value is selected for $\mu(t)$, which in addition to minimizing the cost of delay and

Algorithm 2 The Effect of ω and θ on the Proposed Method

Input : λ, g, h

Output : average delay, average power consumption

```

1: for  $\theta \leftarrow 0.01$  to 1 step 0.01 do
2:   for  $\omega \leftarrow 0.01$  to 1 step 0.01 do
3:     | GA_Algorithm( $\lambda, g, h, \theta, \omega$ )
4:   End for
5: End for
  
```

power consumption regulates power consumption according to the level of edge batteries.

V. IMPLEMENTATION AND EVALUATION

This section describes the implementation and evaluation of the proposed method for optimum distribution of workloads between the cloud and the fog. For this purpose, the evaluation parameters of the problem, the parameters of the different genetic operators, and the implementation environment are examined. Next, the effect of the variations in ω , and θ on the distribution of workloads is studied and the optimum value of these two parameters is obtained. Finally, the proposed method with the optimum values of ω and θ is compared with other existing methods.

A. Simulation Parameters

This section describes the simulation of a cloud-fog environment in order to evaluate the proposed method. In this environment, the genetic algorithm described above is used in the base station as the distributor of workloads between the cloud and fog servers. The simulation aims to examine the effect of the delay cost coefficient and battery depreciation coefficient on the fitness function as well as on the average delay in workload transmission and the power consumption at the network edge. To narrow down the search space in the genetic algorithm, we assume the cost coefficient of the supporting power supply (0.15) as constant and only study the variations in ω , and θ . The process is shown in Algorithm 2. According to this algorithm, with changing the value of ω and θ , the genetic algorithm runs 10000 times in each experiment and the average energy consumption and the delay are measured. In these experiments, $0.01 \leq \omega \leq 1$ and $0.01 \leq \theta \leq 1$, and their values are changed by 0.01 in each experiment.

The proposed method was examined on a system with an 8-core 1.8 GHz CPU and 12GB RAM. In the following, we first initialize the parameters and then discuss the results. The amount of input workload in each interval is specified by a random number that uniformly varies between 10 and 100 requests per second. The renewable energy fluctuates according to a normal distribution of $N(520W, 150)$ [20]. The maximum capacity of each battery is $B = 2kWh$. Also, we assume that the initial charge of battery $b(0) = 0$. The static power consumption of the base station is $d_{sta} = 300W$. We set the maximum number of edge servers $M = 10$. Also, each active server consumes 150W of electricity. The maximum processing rate of each server is 20 requests

per second. We restrict the maximum number of generations of our evolutionary algorithm to 100.

B. The Effect of ω and θ on Workload Distribution

In this section, the results of the experiments are presented using graphs. Then the graphs are analyzed and, by normalizing the values of delay and power consumption, the best coefficients of the fitness function to minimize the costs are obtained.

Fig. 4 illustrates the average delay cost in different experiments for ω and θ . As can be seen in Fig. 4(a), the increased delay coefficient decreases the average delay cost. The reason behind this decrease is the stronger effect of θ on the cost function, which the genetic algorithm seeks to reduce. In fact, the system attempts to reduce the delay cost so that more workloads could be processed locally. For example, Fig. 4(b) shows the variations in the average delay depending on the varying values of θ . In this figure, assuming a constant coefficient of battery depreciation ($\omega = 1$), an increase in the delay coefficient results in a decrease in the delay cost. The most important reason behind the decrease in the delay is the increased value of this parameter in the fitness function as well as the processing of increased amounts of workloads in the fog servers.

Fig. 4(c) depicts the average delay according to the variations of ω for two constant values of θ . When $\theta = 0.01$ (the minimum value), the majority of processes are conducted in the cloud, and the average delay is maximized due to the minimal effect of this parameter on the fitness function and the decision-making. It can be seen that when the delay coefficient is constant, by increasing the battery depreciation coefficient (ω) from 0.01 to 1, the power consumption part in the cost function becomes more significant. Therefore, the system attempts to send more workloads to the cloud to reduce power consumption. Consequently, with the transmission of the loads to the cloud, the average delay begins to escalate. Also, a comparison of the two lines depicted in the figure would show that the average delay with $\theta = 0.08$ is less than with $\theta = 0.01$, which can be explained by its increased effect on the fitness function. Given the above discussion, the greater the coefficient of battery depreciation (ω) and the greater the delay coefficient (θ), the less the average delay.

Fig. 5(a) shows the average power consumption with ω and θ in each experiment. In general, an increase in the coefficient of battery depreciation will reduce power consumption. The reason for this reduction is the increased effect of battery depreciation on the cost function. In fact, the algorithm tries to allocate most of the processes to the cloud to reduce power consumption in the fog servers. Fig. 5(b) depicts the average power consumption with three constant values of θ according to the variations of ω . It can be observed that, as the coefficient of battery depreciation increases, the average power consumption with $\theta = 0.15$ and $\theta = 0.01$ is reduced from 550 w to 450 w. The reason for this reduction is the system's attempt to send more workloads to the cloud and decrease power consumption in the fog servers. As the figure shows, in points where $\theta = 0.01$ (i.e., the minimum value), the majority of workloads are sent to the cloud, and the average power consumption decreases at a higher rate to

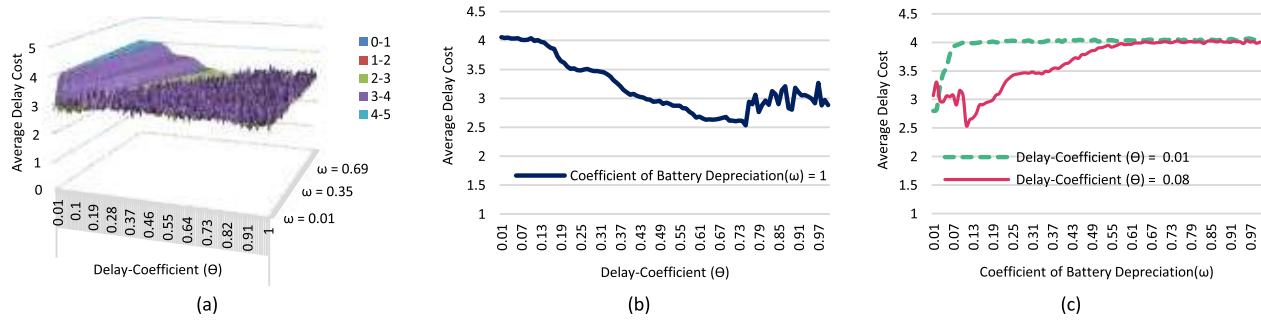


Fig. 4. The average delay cost based on the delay coefficient (θ) and the coefficient of the battery depreciation (ω). (a) The average delay cost by changing the coefficients ω and θ . (b) The effect of delay coefficient (θ) on delay cost. (c) The effect of the coefficient of battery depreciation (ω) on delay cost.

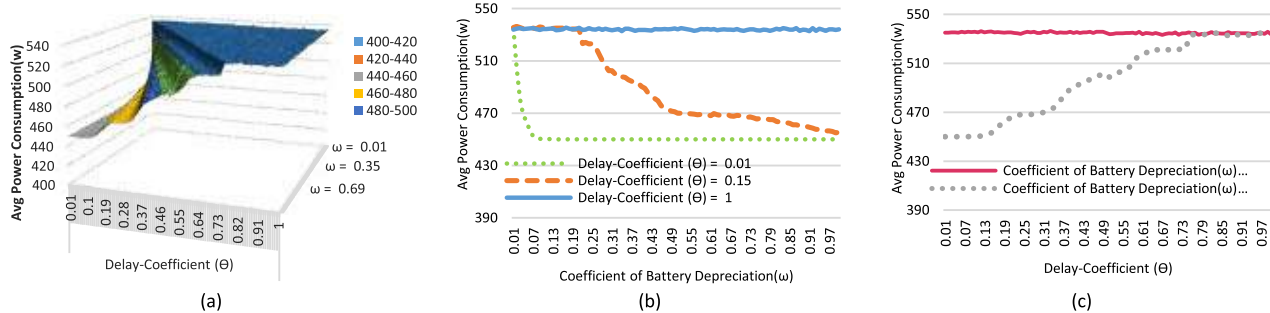


Fig. 5. The average power consumption based on the delay coefficient (θ) and coefficient of battery depreciation (ω). (a) The average power consumption by changing ω and θ . (b) The effect of the coefficient of battery depreciation (ω) on power consumption. (c) The effect of delay coefficient (θ) on power consumption.

achieve its final value (i.e., 450 w). Also, it can be concluded that the rapid decrease in power consumption is due to the minimal effect of delay and the stronger effect of battery power consumption on the fitness function. Another point to mention in this figure is the points on which the delay coefficient $\theta = 1$ is maximum. On these points, due to the strong effect of delay on the cost function, the algorithm sends the majority of processes to the fog server so that they would be conducted locally and the power consumption would not decrease. In this case, the battery level reaches its maximum.

A comparison of the three lines in this graph indicates that the average power consumption of $\theta = 1$ is greater than $\theta = 0.15$ and the average power consumption of $\theta = 0.15$ is greater than $\theta = 0.01$. The high level of power consumption is due to the greater significance of the delay part in the cost function.

Fig. 5(a) shows that, on points with a delay coefficient greater than 0.7, the average power consumption reaches its maximum and remains constant for each state ω . Also, given that $\theta < 0.7$, as the coefficient of battery depreciation increases, attempts are made to send the loads to the cloud and decrease power consumption. As θ decreases, the power consumption part becomes more significant, and the average power consumption is reduced. Fig. 5(c) shows the power consumption graph based on the variations of θ for two values of ω . As can be seen in the figure, when the coefficient of battery depreciation is $\omega = 1$ (maximum), power consumption will increase as the delay coefficient increases and becomes more significant in the cost function. In addition, when the coefficient of battery depreciation is $\omega = 0.01$ (minimum), power consumption will not change with the increase in θ .

This is due to the minimal effect of the coefficient of battery depreciation on the cost function.

Given this, we seek out a state in which the average power consumption is minimized so that the least amount of depreciation could be achieved. As discussed earlier in the formulation of the problem, green energy enters the system as normal distribution according to the equation: $N(520W, 150)$. According to Fig. 5(a), power consumption is almost equal to the average green energy received by the system. It can be thus concluded that this algorithm tries to distribute the workloads in a way that the required power for processing could become almost equal to the green energy and the coefficient of battery depreciation as well as the power consumption at the edge of the network could be minimized. Also, with the decrease in power consumption at the network edge, more green energy could be stored in the batteries.

Fig. 6(a) illustrates the network edge battery levels in different experiments for ω and θ . With the increase in the coefficient of battery depreciation (on points where θ is less than 0.7), more workloads are transmitted to the cloud, which will increase the battery level. Fig. 6 (b) shows the battery level for three constant states of θ . When $\theta = 0.15$ and $\theta = 0.01$, with the increase in the coefficient of battery depreciation (ω), the battery level will increase from 700 w to 2000 w (charging mode). When $\theta = 0.01$ (minimum), due to the transmission of all workloads to the cloud, the green energy not consumed is stored in the batteries and the battery level rises more quickly. However, when $\theta = 1$ (maximum), the loads are maintained at the network edge, thereby leading to the remarkably high power consumption and keeping the battery level at 800 w. It should be mentioned that the proposed algorithm could maintain a full battery in most cases.

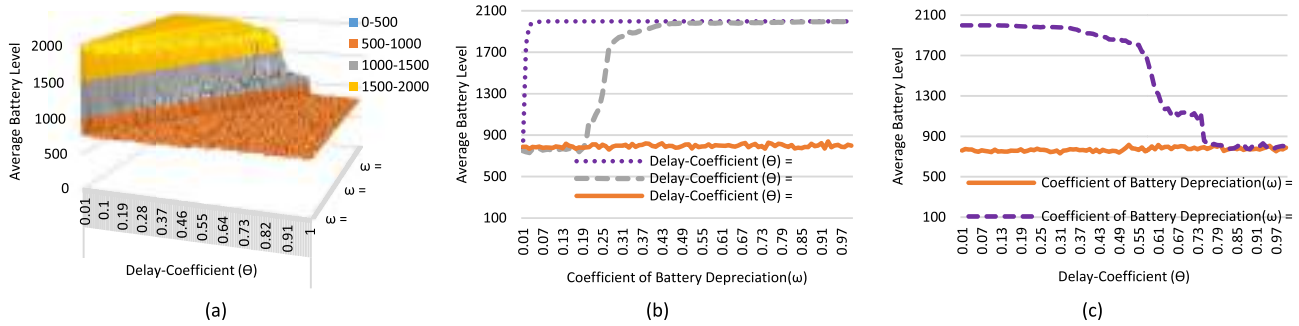


Fig. 6. The average battery level based on the delay coefficient (θ) and the coefficient of battery depreciation (ω). (a) The average battery level by changing ω and θ . (b) The effect of the coefficient of battery depreciation (ω) on battery level. (c) The effect of delay coefficient (θ) on the battery level.

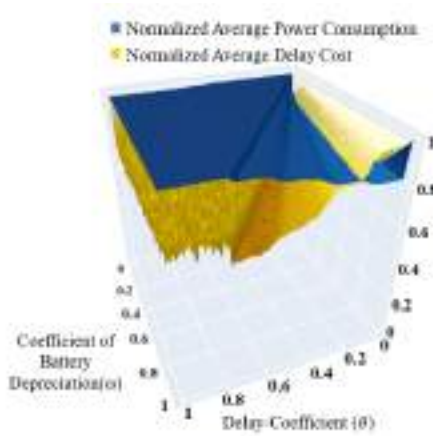


Fig. 7. Normalized values of delay cost and average power consumption.

Comparing the corresponding graphs in Fig. 5 and 6 indicates that, by sending more workloads to the cloud and decreasing power consumption at the network edge, more green energy could be stored in the batteries. This process at the network edge will increase the battery levels. To illustrate this point, let us compare Fig. 5 (c) and 6 (c) in terms of power consumption for the delay coefficient θ and two values of ω . It can be observed that, when the coefficient of battery depreciation is $\omega = 1$ (maximum), power consumption increases with the increase in the delay coefficient as well as its effect on the cost function, thus reducing the average battery level. Also, when the coefficient of battery depreciation is $\omega = 0.01$ (minimum), power consumption will not change with the increase in θ , and the average battery level at the network edge will remain constant. This is not desirable for us because we seek out circumstances in which the average battery level would be maximized. On the other hand, the corresponding graphs in Fig. 4 and 6 indicate that the battery level decreases with the reduction in the delay. The reason is that, in order to reduce the delay costs, the system attempts to process most of the workload in the fog servers, which leads to more battery consumption. Given what was discussed above, we need to reduce the average delay while maintaining the maximum battery level.

C. The Optimum Point of ω and θ

To reach a balance between power consumption and delay in workload distribution, average power consumption and delay cost were normalized to find the optimum state of ω , and θ .

TABLE II
OPTIMUM POINTS BASED ON THE VALUES OF θ AND ω

	1	2	3
Delay Coefficient (θ)	0.05	0.08	0.1
Coefficient of Battery Depreciation (ω)	0.23	0.35	0.47
Average Delay	3.4958	3.49302	3.49707
Average Power	466.59	467.1	466.62

Fig. 7 illustrates the normalized levels of average power consumption and delay cost for every value of ω and θ .

It can be observed that these two parameters have a negative relationship. That is, an increased delay means decreased power consumption and vice versa. As a result, a balance between ω and θ can be attained when the normalized values of delay and power consumption are equal. In other words, the intersection points of these levels in this figure forms a line. Those values of ω and θ that lie on this line are indicative of a balanced state. Of these points, however, only those points provide an optimum state in which the sum of the two normalized parameters is minimal.

On this basis, the three optimum points from Fig. 7 are described in Table II. This table lists the average power consumption and the delay for each of the points in the parametric space (θ, ω). For a better comparison of the three points, six cross-sectional cuts have been made in the graph in Fig. 7 (Fig. 8 (a) to 8 (f)). In Fig. 8(a), the normalized values for $\theta = 0.05$ can be observed. At the intersection point where the sum of the two parameters is minimal, the cost of battery depreciation should be $\omega = 0.23$. For $\omega = 0.23$, Fig. 8(b) shows that the intersection point at which power and delay are minimal is $\theta = 0.05$. Similarly, for the second optimum point (Fig. 8(c) and 8 (d)), $\theta = 0.08$ and $\omega = 0.35$ achieve a balance, and their sum is the minimum amount. Fig. 8 (e) and 8 (f) depict the third optimum point for ω , and θ . At this point, too, the sum of delay cost and power consumption is minimal with $\theta = 0.1$ and $\omega = 0.47$.

The following are the results of the execution of the proposed workload distribution at the optimum points ($\theta = 0.05$ and $\omega = 0.23$). Fig. 9 illustrates the amount of data processed in the fog, the battery consumption of servers, and the amount of workload sent to the cloud.

In this figure, the ratio of offloading in the cloud and the fog to the total input workload is shown in intervals of 1000. On average, in each interval, 64 percent of the total input load has been processed in fog servers, and the rest

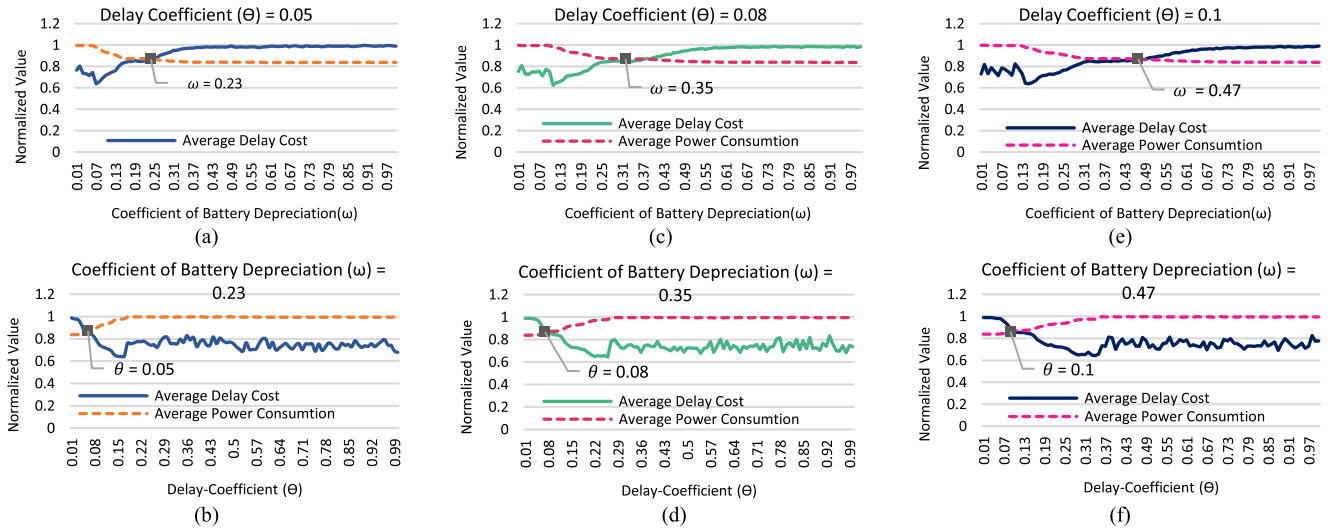


Fig. 8. Normalized values of average delay cost and power consumption for the optimum points. (a) Variations of the coefficient of battery depreciation (ω) for the first point. (b) Variations of delay cost (θ) for the first point. (c) Variations of the coefficient of battery depreciation (ω) for the second point. (d) Variations of delay cost (θ) for the second point. (e) Variations of the coefficient of battery depreciation (ω) for the third point. (f) Variations of delay cost (θ) for the third point.

(around 35 percent) has been sent to the cloud. As most of the loads have been processed locally, it is expected that battery consumption should be high. However, the battery level graph shows that an average of 12 percent of the battery has been consumed in each interval. This can be explained by the optimum use of renewable energy. The system distributes the loads in a way that the power consumed for processing at the network edge be equal to renewable energy. Also, in intervals where more load has been processed locally, there is a rise in battery consumption. For example, battery consumption in the interval 5000-6000 is 3 percent more than in the interval 4000-5000. On the other hand, local processing reduces the delay in the handling workloads.

D. Comparison With Other Methods

In this section, the results of the proposed method at its optimum point are compared with other methods to confirm the decrease achieved in the delay in workload transmission. These methods are briefly described below.

1) *Fixed Power*: In this method, a fixed amount of power is considered for edge computations at each interval of time [24].

2) *Post Decision State (PDS) Algorithm* [20]: The PDS algorithm grabs the state of the system instantly after making the decision at the end of each time interval. The state of the system after making a decision at the end of the interval is an important data that is named the *after-state* variable. The PDS is mainly used as a decision-tree based optimization algorithm. In this algorithm, to find the optimum solution, the problem is broken down into decision nodes and outcome nodes, which correspondingly denote pre-decision and post-decision states. For finding the optimum decision for the vector-valued problem of workload allocation, the PDS tries to find a state that minimizes the long-term costs of the system.

3) *Q-Learning* [25]: Q-learning is considered as a reinforcement learning algorithm that is independent of the type of system model. In this agent-based algorithm, the agent tries to learn a strategy, which results in the best action for each

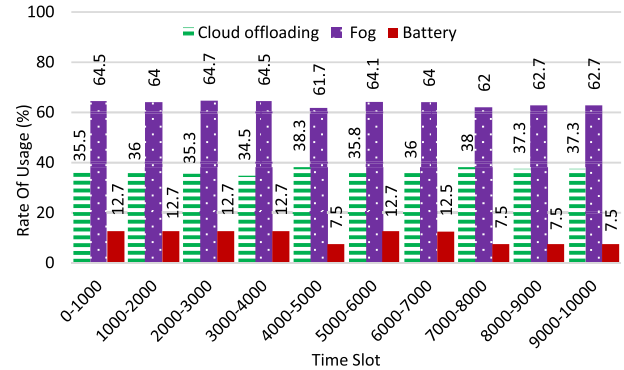


Fig. 9. The rate of usage of cloud and fog resources and power supply over time.

state of the system. Since this algorithm does not need a model of the environment, it can solve the problems with stochastic transitions and payoffs without needing any regulation.

4) *Myopic Optimization* [26]: In this algorithm, regardless of any relationship between the system states and corresponding decisions, the cost function of each state is minimized by only considering the present input information of the system. That is, in the Myopic optimization model, the present knowledge of the workload allocation is densely presented by a Myopic window which represent the knowledge of system in a limited number of time frames. The content of this window may be repeated in different times. As a result, the outcome of the system may be seen repeatedly.

Fig. 10 shows the average delay cost for different methods. As can be observed, learning-based methods perform better and have a lower average delay when run on the battery than when using the electricity network. On the other hand, the proposed algorithm has a lower delay than the other methods. In this figure, the delay cost of all the methods is greater than five, whereas the genetic algorithm used in the proposed method has reduced this cost down to 3.5.

The main point that is clear in both figures is the reduction of the average energy consumption and the reduction of

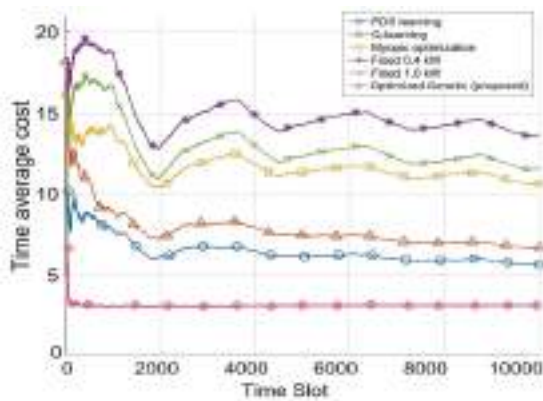


Fig. 10. The average delay cost.

the average delay in successive intervals of time. Reducing the average processing latency for the proposed method in Figure 10 means that workloads are processed more on the fog side, and a smaller percentage of them are sent to the cloud, as evidenced by Figure 9. In Fig.9, for the first 1000 time slots, more percentage of workloads are processed in fog, respectively reducing the average delay. One of the strengths of the proposed method is that it does not have many fluctuations in time slots, especially in the first 2000 time slots.

VI. CONCLUSION

In this paper, we tried to achieve a balance between power consumption at the intelligent vehicular network edge and delay in workload transmission in the clouds by using a genetic algorithm and finding the optimum modes of workload distribution. We also showed that workload distribution at the edge of the vehicular network using renewable energy sources is suitable for vehicular networks in which the processing resources do not have access to the electrical grid and depend on batteries for operation. By utilizing parameters such as the input load and the proportion of green energy as the input parameters of the genetic algorithm, this paper calculated for the first time the optimum number of workloads to be processed locally. Also, by changing the coefficients of the parameters of the cost function of the genetic algorithm, we determined the optimum coefficients for processing the workloads with the least amount of delay and the least power consumption. The simulation results suggest that the proposed method can achieve a better balance in workload distribution than the other existing methods do. While reducing the workload delay by 40 percent and decreasing power consumption at the edge of the vehicular network, this method also seeks to minimize battery consumption by making use of renewable energies.

In future work, other machine learning methods such as neural networks can be used for selecting the optimum parameters.

REFERENCES

- [1] Z. E. Ahmed, R. A. Saeed, and A. Mukherjee, "Challenges and opportunities in vehicular cloud computing," in *Cloud Security: Concepts, Methodologies, Tools, and Applications*. Hershey, PA, USA: IGI Global, 2019, pp. 2168–2185.
- [2] T. Islam and M. M. A. Hashem, "A big data management system for providing real time services using fog infrastructure," in *Proc. IEEE Symp. Comput. Appl. Ind. Electron. (ISCAIE)*, Apr. 2018, pp. 85–89.

- [3] A. Yousefpour *et al.*, "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *J. Syst. Archit.*, vol. 98, pp. 289–330, Sep. 2019.
- [4] M. Shojafar, N. Cordeschi, and E. Baccarelli, "Energy-efficient adaptive resource management for real-time vehicular cloud services," *IEEE Trans. Cloud Comput.*, vol. 7, no. 1, pp. 196–209, Jan. 2019.
- [5] F. S. Abkenar and A. Jamalipour, "EBA: Energy balancing algorithm for fog-IoT networks," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6843–6849, Aug. 2019.
- [6] W. Zhang, Z. Zhang, and H.-C. Chao, "Cooperative fog computing for dealing with big data in the Internet of vehicles: Architecture and hierarchical resource management," *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 60–67, Dec. 2017.
- [7] R. Deng, R. Lu, C. Lai, T. H. Luan, and H. Liang, "Optimal workload allocation in fog-cloud computing toward balanced delay and power consumption," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 1171–1181, Dec. 2016.
- [8] M. Ghobaei-Arani, A. Souri, and A. A. Rahmanian, "Resource management approaches in fog computing: A comprehensive review," *J. Grid Comput.*, vol. 18, no. 1, pp. 1–42, Mar. 2020.
- [9] R. Basir *et al.*, "Fog computing enabling industrial Internet of Things: State-of-the-art and research challenges," *Sensors*, vol. 19, no. 21, p. 4807, Nov. 2019.
- [10] S. Nižetić, N. Djilali, A. Papadopoulos, and J. J. P. C. Rodrigues, "Smart technologies for promotion of energy efficiency, utilization of sustainable resources and waste management," *J. Cleaner Prod.*, vol. 231, pp. 565–591, Sep. 2019.
- [11] M. Aloqaily, A. Boukerche, O. Bouachir, F. Khalid, and S. Jangsher, "An energy trade framework using smart contracts: Overview and challenges," *IEEE Netw.*, vol. 34, no. 4, pp. 119–125, Jul. 2020.
- [12] H. Wu, L. Chen, C. Shen, W. Wen, and J. Xu, "Online geographical load balancing for energy-harvesting mobile edge computing," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.
- [13] Z. Ning, J. Huang, X. Wang, J. J. P. C. Rodrigues, and L. Guo, "Mobile edge computing-enabled Internet of vehicles: Toward energy-efficient scheduling," *IEEE Netw.*, vol. 33, no. 5, pp. 198–205, Sep. 2019.
- [14] X. Wang *et al.*, "Future communications and energy management in the Internet of vehicles: Toward intelligent energy-harvesting," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 87–93, Dec. 2019.
- [15] H. Chen, T. Zhao, C. Li, and Y. Guo, "Green Internet of vehicles: Architecture, enabling technologies, and applications," *IEEE Access*, vol. 7, pp. 179185–179198, 2019.
- [16] F. Ahmadizar, K. Soltanian, F. AkhlaghianTab, and I. Tsoulos, "Artificial neural network development by means of a novel combination of grammatical evolution and genetic algorithm," *Eng. Appl. Artif. Intell.*, vol. 39, pp. 1–13, Mar. 2015.
- [17] S. Verma, N. Sood, and A. K. Sharma, "Genetic algorithm-based optimized cluster head selection for single and multiple data sinks in heterogeneous wireless sensor network," *Appl. Soft Comput.*, vol. 85, Dec. 2019, Art. no. 105788.
- [18] X. Liu and N. Ansari, "Toward green IoT: Energy solutions and key challenges," *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 104–110, Mar. 2019.
- [19] J. Xu and S. Ren, "Online learning for offloading and autoscaling in renewable-powered mobile edge computing," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
- [20] J. Xu, L. Chen, and S. Ren, "Online learning for offloading and autoscaling in energy harvesting mobile edge computing," *IEEE Trans. Cognit. Commun. Netw.*, vol. 3, no. 3, pp. 361–373, Sep. 2017.
- [21] F. M. Dalvand and K. Zamanifar, "Multi-objective service provisioning in fog: A trade-off between delay and cost using goal programming," in *Proc. 27th Iranian Conf. Electr. Eng. (ICEE)*, Apr. 2019, pp. 2050–2056.
- [22] M. Abbasi, M. Yaghoobikia, M. Rafiee, A. Jolfaei, and M. R. Khosravi, "Efficient resource management and workload allocation in fog-cloud computing paradigm in IoT using learning classifier systems," *Comput. Commun.*, vol. 153, pp. 217–228, Mar. 2020.
- [23] C. N. Giap and D. T. Ha, "Parallel genetic algorithm for minimum dominating set problem," in *Proc. Int. Conf. Comput., Manage. Telecommun. (ComManTel)*, Apr. 2014, pp. 165–169.
- [24] K. Kaur, S. Garg, G. S. Aujla, N. Kumar, J. J. P. C. Rodrigues, and M. Guizani, "Edge computing in the industrial Internet of Things environment: Software-defined-networks-based edge-cloud interplay," *IEEE Commun. Mag.*, vol. 56, no. 2, pp. 44–51, Feb. 2018.
- [25] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning* vol. 2. Cambridge, MA, USA: MIT Press, 1998.
- [26] K. Poncelet, E. Delarue, D. Six, and W. D'haeseleer, "Myopic optimization models for simulation of investment decisions in the electric power sector," in *Proc. 13th Int. Conf. Eur. Energy Market (EEM)*, Jun. 2016, pp. 1–9.

Linked Data Processing for Human-in-the-Loop in Cyber–Physical Systems

Zhigao Zheng¹, *Member, IEEE*, Shahid Mumtaz², *Senior Member, IEEE*,
 Mohammad R. Khosravi³, and Varun G. Menon⁴, *Senior Member, IEEE*

Abstract—There are several kinds of smart devices, such as smartphones, sensors, and smart wearable devices, included in the Human-in-the-Loop (HITL) system, but different devices have their own data processing and programming paradigm. Programmers usually need to design the same data processing logic for different devices by using a different programming model. How to mapping the same code to different devices without any change is an emerging topic in the HITL system. Furthermore, the intelligent data processing for the smart CPS sector is experiencing significant growth in data volume, driven by a large number of smart devices that are anticipated in the near future. All these smart devices are expected to improve the overall HITL system performance marvelously. A large number of devices can also outstandingly increase the data volume, which needs to be processed in real time. How to process large-scale data on a smart device in real time is another challenge. Focused on these challenges, this article proposed a computing device-aware HITL CPS data processing framework, named Barge, aiming to map the regular code to the different hardware without any change. In Barge, a semantic model, an architecture-driven programming model, and a graph partition scheme are included. The semantic model is used to express the user-defined graph algorithms by using the domain-specific language. The architecture-driven programming model will execute the graph algorithms on a different device in parallel. Furthermore, the graph partition scheme will partition the large-scale graphs into suitable partitions by aware of the topology to make the partitioned data suitable for kinds of smart devices. We believe that our work would open a wide range of opportunities to improve the performance of large-scale graph processing for HITL systems.

Index Terms—Cyber–physical systems (CPSs), data partition, graph computing, Human-in-the-Loop (HITL), new architecture, programming model, semantic model.

I. INTRODUCTION

THE tremendous amount of smart devices, such as smartphones, radio frequency identification (RFID), sensors,

Manuscript received March 30, 2020; revised June 30, 2020 and August 19, 2020; accepted September 9, 2020. Date of publication April 9, 2021; date of current version September 30, 2021. (*Corresponding author: Zhigao Zheng.*)

Zhigao Zheng is with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: zhengzhigao@hust.edu.cn).

Shahid Mumtaz is with the Instituto de Telecomunicações, Universitário de Santiago, P-3810-193 Aveiro, Portugal (e-mail: smumtaz@av.it.pt).

Mohammad R. Khosravi is with the Department of Computer Engineering, Persian Gulf University, Bushehr 75169-13817, Iran, and also with the Telecommunications Group, Shiraz University of Technology, Shiraz 71557-13876, Iran (e-mail: m.r.khosravi.taut@gmail.com).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India (e-mail: varunmenon@ieee.org).

Digital Object Identifier 10.1109/TCSS.2020.3029569

and embedded devices, have revolutionized both the physical and digital world through the integrated interactions to create the global smart cyber–physical systems (CPSs) [1]–[3]. To process the large scale of the complex linked data between different kinds of devices, both data-intensive and memory-intensive data processing frameworks are included in CPS, i.e., CPS is a software-intensive decentralized system that autonomously perceives its operational context [4], [5]. A CPS system consists of the hardware infrastructure (physical components) and software model [6]–[8]. The most important software is the cyber twin, which is used to simulate the physical things. Internet of Things (IoT), on the other hand, connected kinds of sensors and some other physical things. This means that IoT acts as a connection bridge to network different cyber–physical things. CPS, also known as big data processing technologies, is a hot topic, which leads to a set of new research interests, and it was widely used in many services, such as customer behavior prediction and weather and environment monitoring. However, most of the data processing logic for the same application is the same, but programmers need to develop multiple different copies of code for an application and deploy them on different devices. How to release programmers from the strenuous repetitive work is an emerging topic in the HITL CPS system. Furthermore, all these services will generate an enormous amount of data in real time, which makes it is not easy to store and process such kind of large-scale data. All these difficulties drive the scientist to propose cloud computing technologies along with machine tools, data mining, artificial intelligence, and fog computing technologies to store, process, and analyze large-scale data. By using all these technologies, we can try to uncover the hidden patterns, unknown correlations, and other useful information [9], [10]. The characteristics of big data were well summarized in the Introduction Section of [11]. The relevance of the big data era and CPS are also highly relevant to global sustainable development goals recently discussed in [12].

Graph computing is one of the most famous big data processing technology, which was widely used to process the linked data. In the graph computing paradigm, the graph data model, which is a fundamental mathematical structure used to model pairwise relations between objects, is widely used in machine learning [13], [14] and deep learning [15] technologies to express the connections between different objects. The context in a graph is called vertices (also called nodes), while the links are called edges. Graph theory has

been widely used in Human-in-the-Loop (HITL) data processing [16]–[18], a knowledge graph programming with an HITL discussed in [16]. In Lou’s work [16], the authors examined the advantages of the knowledge graph programming for HITL, such as the flexible programming interface and kinds of “data compiler” method. Then, the authors proposed a knowledge graph programming prototype for HITL. Holzinger *et al.* [17] provided new experimental insights on how to improve computational intelligence by complementing HITL with human intelligence in an interactive machine learning approach. The article [18] described a “big picture” of HITL data analysis, including the user communities’ tools and algorithms, and also the HITL data analysis framework developing technologies and theories. However, how to use graph theory and algorithms to support distinctive characteristics of HITL CPS data analysis and provide high-performance and real-time decision-making policy remains challenging and represents a promising research direction.

With the development of hardware manufacturing, there are several kinds of new computing devices proposed for large-scale data processing, and traditional large-scale graph computing is facing new opportunities and challenges. Graph applications have poor locality and poor cache hit rate and are largely stalled on memory accesses since there are complex connections between graph nodes and the working set of realistic graphs is much larger than the last level cache (LLC) of current machines. The conventional computing architecture is computing-centric, which focuses on memory sharing and message communications; this processing fashion is unable to handle graph applications. In this article, we focus on the key technologies and methods of a new computing architecture for large-scale HITL CPS graph data processing. To solve the poor locality and poor cache hit rate problem, we proposed new computing device-based HITL CPS data processing architecture, named Barge. We made the following contributions to the proposed architecture.

- 1) We conduct an extensive set of comprehensive experiments to explore the parallelism and memory operations of the graph processing systems for HITL CPS data processing.
- 2) We proposed a semantic model for large-scale graph data processing under new computing architecture to mapping the same code to different devices without any change. The semantic model includes semantic rule and graph data interpretation method.
- 3) We proposed an architecture-driven programming model, which is suitable for a different architecture.
- 4) We proposed a data- and topology-aware graph data partition scheme that can partition the large-scale graph data quickly by consider the data structure and also the feature of the computing device.

The rest of this article is organized as follows. We will introduce the characteristics and research challenges of HITL CPS data processing and the motive of using the graph processing method to process the large-scale HITL CPS data in Section II. Section III provides our proposal for applying the proposed Barge model to improve the performance of HITL CPS data processing. Then, we will introduce the design methodologies

and principles of Barge in Section IV. Section V introduces the related work, and we will conclude this article in Section VI.

II. CHALLENGES OF HITL CPS DATA PROCESSING

In this section, we describe our experimental settings, including the graph data sets and algorithms and the graph processing frameworks (GPFs) for our empirical study. We also try to explore the parallelism issues and memory operation characteristics of GPFs for HITL CPS data processing.

A. Experimental Settings

1) *Graph Algorithms*: Most of the graph algorithms can be classified as the iterative algorithm and the traversal algorithm. All vertices will be updated in each iteration of the iterative algorithm, but only active vertices will be updated in each iteration of the traversal algorithm. The most representative algorithms of iterative and traversal algorithms are PageRank (PR) and Breadth-First Search (BFS), respectively. We select these two most representative graph algorithms for the performance evaluation, as well as much prior work [19], [20] do.

- 1) *BFS*: An algorithm that traverses the whole graph in search of one or more vertices, which is a basic component of many other complex graph mining algorithms.
- 2) *PR*: An algorithm that was used to evaluate the influence of vertex within a graph, which was proposed by Google first, and it was initially used to evaluate the importance of a web page.

2) *Graph Processing Frameworks*: Many GPFs have been developed by both academic and industry researchers, such as TOTEM [21], CuSha [22], and some other frameworks. In this section, we select four representative state-of-the-art GPFs to run the graph algorithms and profiling the runtime details.

- 1) *GunRock [19]*: A high-performance graph processing library on GPU with a high-level bulk-synchronous processing scheme. Gunrock provides a data-centric abstraction centered on operations on a vertex or edge frontier. Gunrock achieves a balance between performance and expressiveness by coupling high-performance GPU computing primitives and optimization strategies. Gunrock also proposed a high-level programming model that allows programmers to quickly develop new graph primitives with small code size and minimal GPU programming knowledge.
- 2) *Sep-Graph [20]*: A highly efficient software framework for graph processing on GPU. It provides a hybrid execution mode that can automatically switch among synchronous or asynchronous execution mode, Push or Pull communication mechanism (Push or Pull), and Data-driven or Topology-driven traversing scheme (DD or TD), according to the parameters.
- 3) *Tigr [23]*: A graph transformation framework that can effectively reduce the irregularity of real-world graphs with correctness guarantees for a wide range of graph analytics.

TABLE I
DATA SETS USED IN THE EXPERIMENTS

Datasets	Vertices	Edges	Avg. Degree	Max Degree	Diameter	Exponent(α)	x_{min}	Fitness (p)
web-Stanford	281,904	2,312,497	8.20	38,626	164	2.1310	5	0.894
dblp-2011	986,208	6,707,236	6.80	979	23	3.9736	119	0.4
youtube	1,157,829	2,987,624	5.27	28,754	24	2.1410	8	0.877
RoadNet-CA	1,971,282	5,533,214	2.81	12	8,440	15.5587	4	0
Wiki-Talk	2,394,386	5,021,410	4.19	100,032	11	2.4610	1	0.787
soc-LiveJournal	4,847,571	86,220,856	28.25	22,887	20	2.6510	59	0.930
twitter-2010	41,652,230	1,468,365,182	70.51	3,081,112	23	1.54	12	0.96
webbase-2001	118,142,155	1,019,903,190	8.63	816,127	379	2.2	6	0.538

4) *GSWITCH* [24]: A machine learning model-based graph processing system that dynamically chosen the optimization variants by monitoring the system overhead. In *GSWITCH*, the authors trained 644 real graphs to learn the algorithm pattern, and *GSWITCH* changes the optimization variants by using the pattern information to achieve high parallel system performance.

3) *Graph Data Sets*: All the data sets of the experiments conducts in this article are follow the classic graph formalism [25]. We use V and E to present the vertices and edges of the graph, respectively. $G = (V, E)$ present the graph. The edge presented as e , where $e = (u, v)$ and $e = \langle u, v \rangle$ are the undirected and directed edges. In this article, both directed and undirected graphs are used in our experiments. In order to show the performance of different kinds of graphs, we include both power-law and large diameter graphs in all our experiments. We select eight graphs from different real-world applications, such as e-business, social network, and some other source networks, and all these graphs are with different structures and a varying number of vertices and edges. The graphs are shown in Table I. All these eight graphs can be downloaded from the Stanford Network Analysis Project (SNAP) [26]. As the power-law graph follow a distribution, as shown in formula (1) [27], [28], we list the exponent and the fitness in Table I to compare the power-law attribution. In this article, the graphs are stored in a plain text file with an edge-list format, which is easy for us to locate the edges

$$\mathbb{P}(x) \propto x^{-\alpha}. \quad (1)$$

4) *Hardware Platforms*: All the experiments presented in this section are conducted on NVIDIA Tesla V100 GPU, which is a Volta architecture-based GPU with 5120 CUDA cores and 32-GB onboard memory. The GPU is coupled with a host machine equipped with 28-Ksyun Virtual Senior CPU cores, each at 2.60 GHz, and 12-GB memory. The host machine is running Ubuntu OS version 16.04.10. The algorithm is implemented with C++ and CUDA 10.02 using the “-arch=sm35” compute compatibility flag.

B. Exposing More Parallelism

To improve the parallelism of hardware, the Single Instruction Multiple Data (SIMD) architecture is introduced to the out-of-order (OOO) manner to enable simultaneous execution of multiple independent instructions. In this design philosophy, each core could issue more than one Micro-Operations (μ -op).

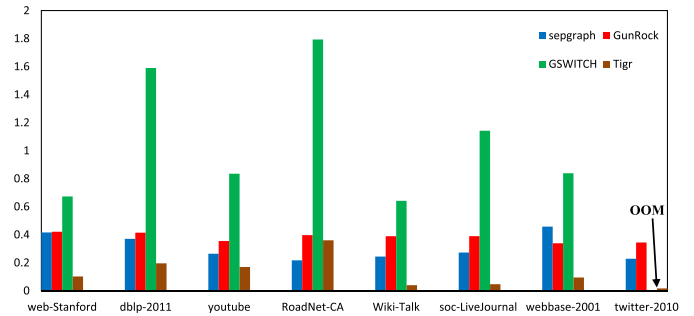


Fig. 1. Average IPC of PR on different data sets with different implementations.

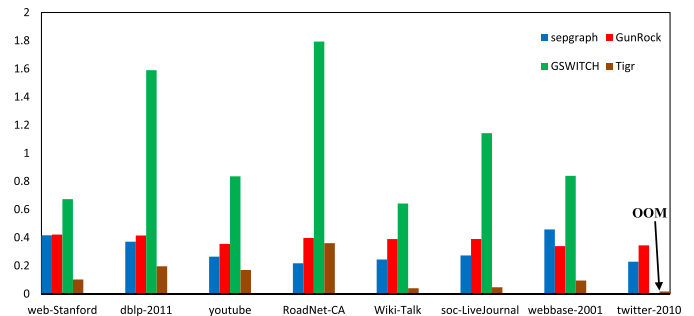


Fig. 2. Average IPC of BFS on different data sets with different implementations.

We illustrate the instruction-level parallelism (ILP) of all the evaluated frameworks. Figs. 1 and 2 show that the average instructions per cycle (IPC) of PR on *GSWITCH* is about 1.793, while the IPC of PR on the other three frameworks is no more than 0.458. This experiment indicates that only about one-eighth of the core’s ability is used for most existed GPFs (except *GSWITCH*). One main reason for low IPC is the long instruction latency [29], and there is a large number of GPU cycles consumed by some instructions.

In order to identify the exact reason for the low ILP phenomena, we also evaluated the Max/Min IPC of different implementations of PR and BFS on different data sets. Figs. 3 and 4 show that the Max IPC of PR and BFS on Sepgraph, Tigr, and GunRock are no more than 0.543, while the MAX IPC of PR on *GSWITCH* is 1.823. This phenomenon indicates the heavy dependence on graph processing algorithms. The execution of one instruction may relate to many other instructions.

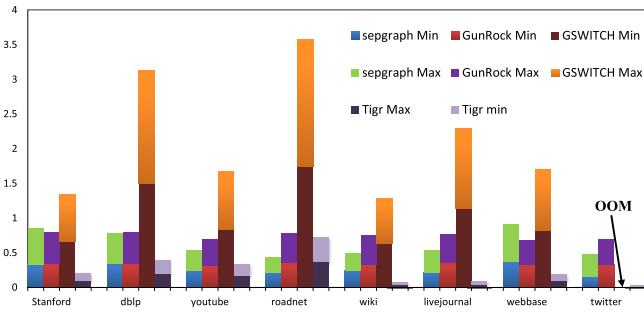


Fig. 3. Max/Min IPC of PR on different data sets with different implementations.

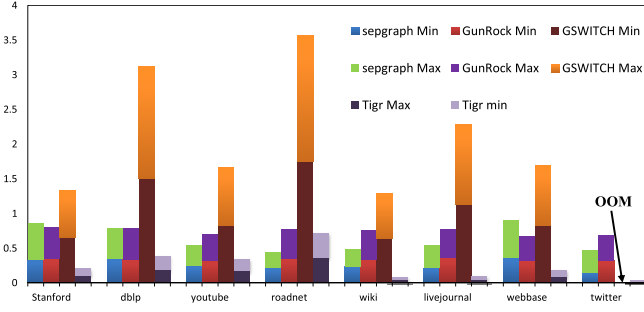


Fig. 4. Max/Min IPC of BFS on different data sets with different implementations.

C. Warp Issue Efficiency

To look deeper into the reason for low ILP, we evaluated the occupancy of GPU warps. In GPU performance profiling, the achieved occupancy is used to measure the warp scheduler’s efficiency by using the ratio of the average active warps of each clock cycle to the maximum warps supported by each multiprocessor (SM), which defined as for the following formula:

$$\text{occupancy} = \frac{\text{active warps}}{\text{maximum warps}}. \quad (2)$$

From formula (2), we can conclude that low occupancy interferes with the ability to hide memory latency, which can degrade the performance. In contrast, higher occupancy does not always indicate higher performance. Many factors can affect occupancy, such as register availability. The register storage differs for different devices, and it enables the threads to keep local variables nearby for low-latency access. However, the threads must share the register file due to a limited commodity. In modern GPU architecture, all the registers are allocated to a block at the beginning of the program. Hence, a supported block of a multiprocessor (SM) will be reduced if each block uses more than one register, and this thread assignment will further reduce the occupancy of the SM.

Figs. 5 and 6 show the occupancy is very low of both PR and BFS implementation of GunRock on all the eight graph data sets, while both the iterative and traversal algorithm can achieve high occupancy of Tigr and GSWITCH implementation.

D. Memory Operations

In this section, we explore the memory operation efficiency by checking the global memory efficiency and throughput.

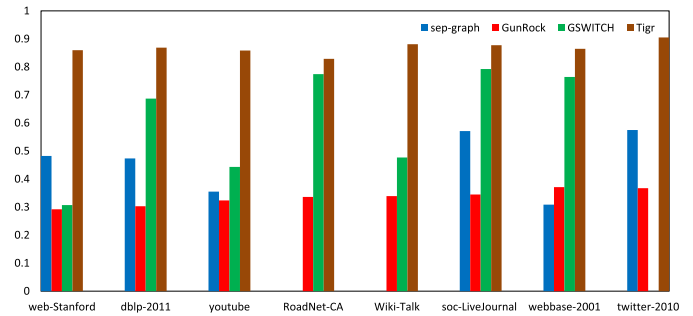


Fig. 5. Achieved occupancy of PR on different data sets with different implementations.

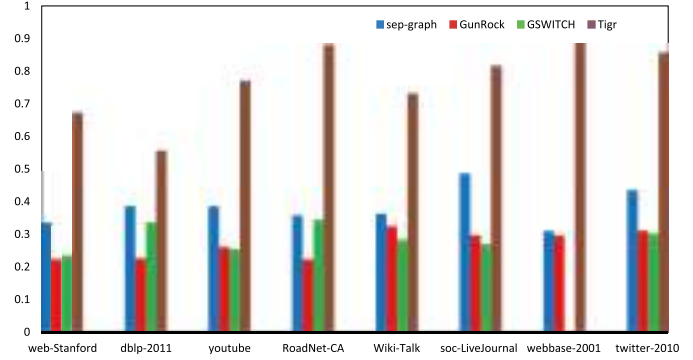


Fig. 6. Achieved occupancy of BFS on different data sets with different implementations.

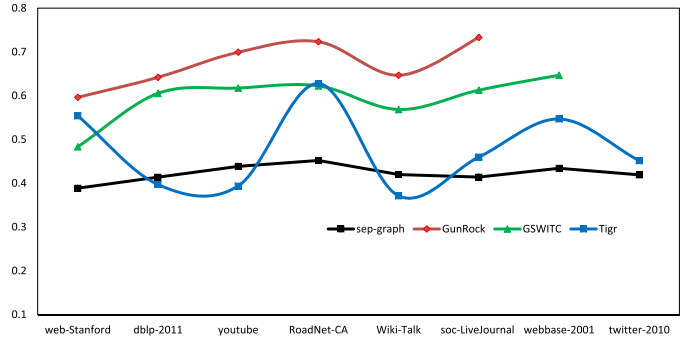


Fig. 7. Global memory efficiency for BFS on different data sets with different implementations.

Figs. 7 and 8 show the global memory efficiency for BFS and PR on different data sets with different implementations. We can conclude from Figs. 7 and 8 that GunRock can achieve the highest memory operation efficiency for both the traversal and iterative algorithms while Tigr and GSWITCH tending to vary widely on different data sets.

Table II shows the global memory throughput for BFS and PR on different data sets, and Tigr can achieve the highest global memory throughput on both BFS and PR. GunRock and GSWITCH achieve the lowest throughput on most power-law graphs, while Sep-Graph can achieve higher throughput on the large-scale graphs (RoadNet-CA) than the power-law graphs.

III. APPLYING BARGE IN GRAPH-BASED HITL CPS DATA PROCESSING

As we discussed in Section II, there are several challenges for graph processing on GPU, such as the control and memory

TABLE II
GLOBAL MEMORY THROUGHPUT FOR BFS AND PR ON DIFFERENT DATA SETS

Dataset	Algorithm	sep-graph	GunRock	GSWITCH	Tigr
web-Stanford	BFS	129.282	20.455	46.950	574.200
	PR	239.054	218.939	141.676	385.350
dblp-2011	BFS	211.695	31.467	57.259	303.110
	PR	317.688	230.4609	180.500	650.760
youtube	BFS	127.786	105.280	67.061	388.960
	PR	86.187	224.371	314.670	519.400
RoadNet-CA	BFS	234.096	11.830	21.154	632.380
	PR	NULL	285.526	446.260	876.900
Wiki-Talk	BFS	128.983	210.178	93.004	387.250
	PR	NULL	255.578	217.316	82.537
soc-LiveJournal	BFS	195.8851	154.3607	107.335	692.410
	PR	189.973	290.041	377.910	143.770
webbase-2001	BFS	276.6052	221.7952	NULL	532.160
	PR	318.656	313.666	210.889	383.780
twitter-2010	BFS	237.222	173.589	197.400	820.510
	PR	181.970	325.354	NULL	48.873

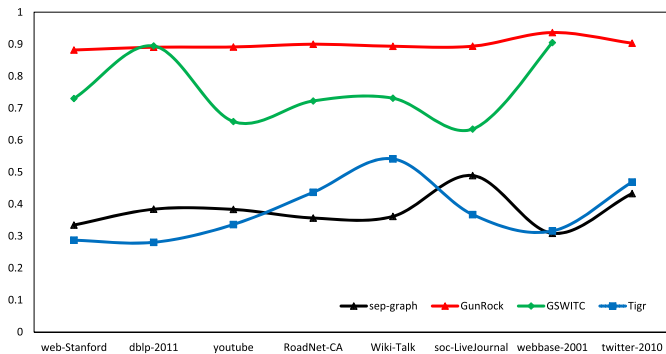


Fig. 8. Global memory efficiency for PR on different data sets with different implementations.

divergence, load imbalance, and global memory access overhead [30]. To achieve an efficient performance of large-scale HITL CPS graph data on new architecture devices, this section introduces the Barge framework for HITL CPS graph data processing.

We design Barge by considering the three primary aspects of graph processing: algorithm semantic expression, programming model, graph partition, and data placement.

- 1) Existing graph processing systems are designed for single hardware by considering the hardware feature to improve the system performance. While this design philosophy cannot achieve expected performance on new hardware architecture, in some cases, the existed frameworks even cannot run on the new device. This design philosophy is easy to limit the scalability of the framework and will lead to strenuous and repetitive work for programmers. In order to solve this problem, this article proposed a semantic model to represent the graph algorithm and make the graph algorithm suitable for the new architecture without reprogram the algorithm. The semantic model provides a set of unified semantics to define the graph structure and a set of unified API for different graph processing systems.
- 2) In order to remove the conflicts between the heterogeneous parallelization of kinds of new hardware devices

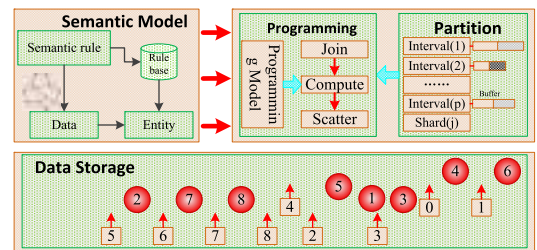


Fig. 9. Illustration of the proposed HITL CPS data processing framework.

and the scalability requirements of graph processing, this article proposes an architecture-aware programming model that can be suitable for multiarchitecture. The proposed model provides a reasonable run-time abstraction of hardware parallel features and the rich programming interfaces for graph computing applications.

- 3) The new architectures devices usually have very high local memory access bandwidth but are just equipped with a small capacity of onboard memory. How to use the limited memory of new architectures devices to process a large-scale graph is a great challenge. In order to solve this problem, this article proposes a new graph partition scheme by considering how to balance the computing workload and communication overhead and the memory access efficiency and also the redundant computation overhead.

Fig. 9 shows an illustration of the proposed HITL CPS data processing framework. In this framework, we propose a semantic model to represent the graph algorithms, a new programming model to fit multiarchitectures, and also a new graph partition scheme.

A. Graph Algorithm Semantic Model for New Architecture

Nowadays, with the proliferation of mobile devices and wearable devices, the HITL CPS graph data are proliferating. The efficient parallelism graph processing algorithm is the most essential aspect to improve the performance of graph processing systems. Combining the features of the

new architecture with the characteristics of graph processing algorithms is the key to improving the performance of graph processing systems. Traditional work usually makes different implementations for each architecture. This method has poor portability and does not meet the needs of multiple graph processing algorithms. It also brings new challenges to implement the algorithms for different kinds of applications and also code management. However, some common operations, similar optimization techniques, and even the same software components are included in various graph algorithms. Considering the abovementioned architecture features and graph processing algorithm characteristics, this article proposes a new graph algorithm description semantic model. This model provides users with productive semantics operations, such as to define graph data structures, describe architecture-aware parallel operations, and fit for different graph algorithms.

The semantic model is an abstraction of the common operations of different algorithms; hence, how to extract the ordinary operations of different algorithms, define the specific operations, and provide a set semantics to describe the parallelizable operations are the base of the semantic model. On the other hand, the semantic model should clearly define parallel operation rules to ensure that user code can be executed correctly and efficiently under the new architecture. At the same time, the semantic model should provide productive parallel operation methods to ensure that users can obtain sufficient expression ability to intuitively, accurately, and completely and describe existing graph algorithms and graph algorithm flows that may appear in the future. How to ensure the efficiency of the graph algorithm by considering the characteristics of the new architecture is the main content of the semantic model. Different architectures, such as GPUs, FPGAs, the SIMD acceleration components, and specialized devices with complicated local storage (such as IBMCELL/Intel SCC), have their parallel execution models. The difference in the execution model of different devices can lead to substantial performance gaps for the same code. In order to achieve excellent overall system performance on different devices, this article first studies how to ensure that the graph algorithm description, which provided by the user, can run on different devices. Furthermore, this article introduced how to determine a parallel operation execution mode to adapt the execution characteristics of the hardware.

B. Architecture-Aware Graph Computing Programming Model

The graph computing programming model is a bridge between the architecture and the graph processing application. On the one hand, it abstracts the details of the hardware characteristics and provides programmers with rich interfaces to implement various graph algorithms. On the other hand, it makes full use of hardware resources efficiently and correctly completes the application requirements, which provides an optimized run-time environment for programs. Graph processing is a strong data dependence application that is hard to parallelism, while most of the new architecture devices are highly parallelism. Hence, how to concisely implement

various graph algorithms while ensuring the efficiency of parallel computing is the first problem to be studied in graph computing programming models.

Although various new-type processors are highly parallelism, their structures are very different. For example, the hardware logic of the computing core of GPU is straightforward, which is suitable for sequence programs without complex logic, while KNL has fewer computing cores than GPU, but the processing logic is relatively complex, which can support more instructions, such as AVX512ER and AVX512CD. Furthermore, the FPGA achieve hardware-level programming by changing the state and combination of gate circuits, which supports some complex logic with high memory bandwidth. These heterogeneous parallelisms make it challenging to achieve excellent system performance in a unified programming manner. Therefore, how to design a run-time optimization mechanism by considering the hardware characteristics, which can be suitable for multiple architectures, is another research challenge for HITL CPS graph processing.

In order to solve the abovementioned two problems, this article proposes an architecture-aware graph computing programming model to match the different architectures. The proposed programming model can effectively and concisely implement various graph algorithms by considering the device characteristics. We can also provide a set of efficient programming methods for developers by using this programming model.

C. Feature-Aware Data Partitioning and Placement Strategy

New accelerator architecture-type processors have large on-chip memory bandwidth, such as GPU and MIC's HBM, which provides very favorable processing conditions for memory-intensive graph computing applications. However, the scale of the HITL CPS graph data shows an explosive growth trend. The growth of the on-chip storage of processors with accelerated structure types is far from meeting the demand for data growth. At the same time, graph processing has some unique features, such as the most graph applications that can be processed by using the iteration processing fashion; on the other hand, the data placement also has a huge impact on the performance of graph algorithms. How to design a suitable data placement strategy and an efficient data partitioning scheme, by considering both the features of the graphs and also the new architecture devices, allocate the partitioned graph data are the crucial points to improve the performance of HITL CPS GPF.

Allocate the partitioned graph data for both single and multimachine, which will cause a large amount of network overhead and bus data transmission overhead on new architecture devices, due to the association of the graph data. It will lead to some other problems, such as load imbalance and low resource utilization, if we only focus on communication and transmission overheads. A high-quality partitioning algorithm itself will cause huge overhead; hence, it is essential for us to study how to reduce the computational complexity of the partitioning algorithm with acceptable partitioning quality.

The graph diameter becomes more extensive due to the rapid growth of the graph size, which leads to lots of redundant

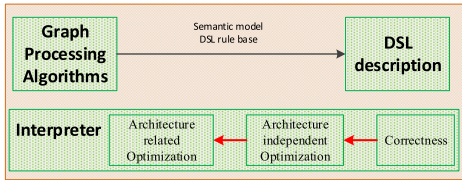


Fig. 10. Graph computing semantic model adapted to new architecture.

computations under the iterative execution fashion on a new architecture device. On the one hand, in the edge centric processing model, the widely used CSR/CSC format easily leads to scattered and irregular memory access, as well as the low memory bandwidth efficiency. On the other hand, the power-law characteristics will lead to a load imbalance problem. In order to solve these problems caused by the data format, this article needs to consider how to improve reading efficiency and how to reduce reading times. In detail, one is how to read the necessary data from the global memory quickly, and the other one is how to implement a heuristic method to read the data, which will be executed in the next iteration, in one memory access to reduce the I/O overhead.

IV. GRAPH-BASED HITL CPS DATA PROCESSING FRAMEWORK

In this section, we first analyze the characteristics of different computing devices and then design a graph algorithm semantic model for these new architectures, and we also abstract the parallel operations for all these architectures in this section. Furthermore, we summarize the data access of large-scale graph processing pattern, and then, we design a multiarchitecture supported graph computing programming model to improve the system performance of large-scale graph processing on new architectures, by simplifying the semantics of the graph processing programming model. Finally, we design a structure-aware data partitioning and placement strategy to make the graph processing system meet the architectures' feature.

A. Graph Algorithm Semantic Model for New Architecture

This article proposes a semantic model of graph algorithms on new architectures by using a domain-specific language (DSL).

The semantic model uses DSL as its entity and uses DSL to express the model elements, such as variable definitions, data definitions, operation definitions, and parallelization flags required in the semantic model. Based on DSL, this article proposes an interpreter for the semantic model. The DSL provides users with a clear and intuitive description of graph algorithms, and the semantic model interpreter explains the description of user-provided graph algorithms. The proposed semantic model is shown in Fig. 10.

This article will analyze the graph processing from a mathematical perspective and then design the semantic model for graph algorithms. A graph $G(V, E)$ is a set of vertices V and an edge set E . The edge and vertex related data can be defined

as a mapping P , P maps the vertex or edge to a particular domain R , and the mapping can be represented as $P : E \rightarrow R$ or $P : V \rightarrow R$. This article uses the mapping to represent the attributes of the edges or vertices. For a given graph $G = (V, E)$ and a series of attributes $\Pi = P_1, P_2, \dots, P_n$, the proposed semantic model should satisfy the following types of graph algorithms: 1) calculate a scalar value from a given graph G and the attribute set, such as the calculating the conductivity of the subgraph; 2) calculate the new attribute from Π , such as calculating the PR value to sort the vertices of the PR algorithm; and 3) select the interested subgraphs, such as the strongly connected components finding algorithm.

Furthermore, the proposed semantic model tries to provide three ways to describe parallel algorithms. The three description ways are as follows: 1) implicit parallelism semantic structure; 2) allow users to distinguish parallel regions accurately; and 3) Well-defined parallel operations. For example, the *for-each* statement is precisely a parallel execution area specified by the user. At the same time, the graph vertex value assignment is an implicit parallel operation, and the widely used reduction operation in graph algorithms is an explicit parallel operation.

This article designs the semantic model interpreter from the following three aspects.

- 1) Check whether the description of the graph algorithm by the user meets the model's requirement. The basic condition is that the user's description should meet the semantic model's grammatical requirements. The interpreter checks the user's input by checking the syntax, data type, and parallel semantics three aspects.
- 2) *Architecture Independent Optimization*: The interpreter converts the code that meets the syntax requirements into a detailed loop or iterative operation and then optimizes the code by cyclic fusion, statement upward and slack protocol boundaries operations.
- 3) *Architecture-Related Optimization*: As a different architecture has its execution fashion and data access method, some architecture-related optimizations should be added through the interpreter by considering the architecture's feature. For example, the GPU uses SIMD fashion to execute the codes in parallel, and the continuous memory access operation can reduce the memory access overhead. Therefore, some data-level parallelism and memory accessing optimization methods should be added through the interpreter. All these optimization methods have a strong correlation with the programming model. Hence, we will introduce them in Section IV-B.

B. JCS Programming Model

Most of the existed GPFs adopt the vertex-centric programming model since this programming model is easy to express most of the graph algorithms, and it can provide high scalability of the graph algorithms by partition the graphs. However, this programming model is also easy to lead random memory access and load imbalance problem due to the skewed degree distribution of real-world graphs. While the edge-centric programming model will introduce

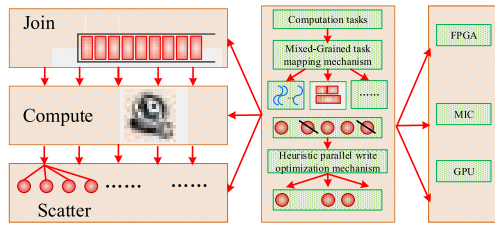


Fig. 11. Graph computing programming model adapted to multiple hardware characteristics.

lots of redundant computation because there are many more edges than the vertices in real-world graphs, it can provide a continuous memory access fashion. The Gather–Apply–Scatter (GAS) programming model proposed in PowerGraph [31] is a fine-grained vertex-centric programming model. In GAS, the computation process is subdivided to increase the degree of parallelism. Previous research shows that there are a number of active vertices in each iteration, which is far less than the total number of vertices in a graph [31]. Hence, this article proposes a queue-based vertex-centric Join–Compute–Scatter (JCS) programming model, which is shown in Fig. 11. In the JCS programming model, the operation on the vertex is divided into the join, compute, and scatter three steps. The join operation adds the active vertices into the worklist, and the compute operation updates the vertex’s value according to the user-defined function, while the scatter operation scatters the vertex’s value to its neighbors, just similar to the scatter operation in the GAS model. In the JCS model, each iteration cares about the vertices that need to be updated. This execution fashion can provide a unified and concise implementation fashion for different algorithms, and it can provide high scalability for the vertex-centric method.

In order to make three phases of the JCS model suitable for different hardware, this article provides a mixed granularity task mapping mechanism and a heuristic parallel optimization mechanism. We introduce the two optimization mechanisms as follows.

- 1) **Mixed-Granularity Task Mapping Mechanism:** Here, we take GPU as an example to introduce the mixed-granularity task mapping mechanism. We assign the vertices to thread, warp, CTA, and kernel according to the number of the neighbor of the active vertex, and the virtual-warp is used to solve the load imbalance problem. While, on KNL, the proposed mechanism not only supports regular round-level parallelism for different task size, including the *for-all* parallelism, reduction parallelism, and scan parallelism, it also supports the irregular parallelism, which can achieve excellent load balancing through reasonable task stealing scheduling.
- 2) **Heuristic Parallel Writes Optimization Mechanism:** Since there are some duplicate vertices in the worklist and some vertices connected with the same vertex, this situation will lead to conflicts in updating operation. Atomic operations and locks are used to ensure data consistency. However, many existed researches show that some updating operations are idempotent (i.e., the updating order does not affect data consistency).

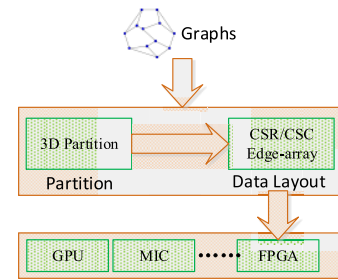


Fig. 12. Data partitioning and placement strategies for data-aware and structure-aware.

Hence, there is no need to design a specific algorithm by using atomic operation or lock to remove the duplicate vertices from the worklist, and just a runtime lightweight heuristic method is enough to remove the redundant computation, which will be more efficient.

C. Feature-Aware Data Partitioning and Placement Strategy

In order to meet the architecture’s feature to unleash the device performance, this article proposes a mixed 3-D graph partition scheme. The proposed graph partition scheme will load the graph blocks into the device on broad memory to make sure the locality of data access and, hence, to reduce the I/O overhead by considering the communication overhead. Fig. 12 shows the essential operation of the proposed graph partition scheme.

The existed graph processing system applied the 1-D or 2-D partition strategy, which is the vertex-centric and edge-centric partition method, respectively. The vertex-centric partition strategy will lead to a load imbalance problem, since the vertex degree distribution of most real world graphs is power-law. While the edge-centric partition strategy will lead to a large amount of communication between the master node and the replicas. Recent research proposed a 3-D partition strategy, which partitions the vertex attribution as the 3-D partition object, and this partition can achieve an excellent system performance in machine learning applications but cannot suitable for full broad applications [32]. While the traversal tree-based partition method can maintain good locality, the partition operation can be executed after traverse the whole graph, and the renumber operation is needed for the subgraphs. The overhead of this partitioning method is huge, and the overhead will increase growth with the graph size. In order to solve the problems of the existed partition methods, this article proposes a hybrid partition method. The hybrid partition strategy will partition the vertices that have similar degrees together by using a 2-D partition method and then partition the subgraphs again by using a 1-D partition method. While, for some specific graph algorithms, the hybrid partition strategy will take the 3-D partition as the first round partition, partition the results again by using a 2-D partition method. This hybrid partition method makes the proposed partition strategy achieve load balance and low communication computation ratio for different algorithms and architectures.

The data placement is closely related to the representation of the graph, and it has a serious effect on system performance.

The upper level framework requires the graph representation method to provide a high memory bandwidth utilization and ensure the locality of memory access as much as possible. While the lower level storage requires high space utilization, avoid the space-wasting for sparsity graph. The storage level also requires the graph that can be loaded into the memory during the I/O operation in graph processing. In order to meet both the upper level and lower level requirements, this article provides a hybrid CSR/CSC graph representation. Then, we further mixed the edge-list representation into the hybrid representation according to the characteristics of the graph. The mixed CSR/CSC graph representation is a benefit for the Scatter/Gather operation. For example, some implementation of the BFS algorithm will change the traversal direction from bottom-up to top-down (top-down to bottom-up), and this hybrid representation will improve the memory access efficiency in this kind of operation. Community is another essential characteristic of the real world graphs, i.e. the vertices in a community are connected but few vertices connect with the vertices outside the community. In this kind of graph, the community can be processed by some SIMD devices, such as GPU, by using the edge-list representation will achieve an excellent performance. Hence, edge-list representation can be an optional method for users.

V. RELATED WORK

We introduce state-of-the-art works for graph computing in this section. Most recent works can be classified into the storage model, the programming model, and the execution model three aspects. We will introduce the related works from these three aspects, receptively.

A. Storage Model

Since most graph applications are memory intensive with a random memory access model, on the other hand, the real-world graphs are with huge size. Both these characteristics will lead to high overhead for both memory and disk access. In order to solve this problem, many researchers proposed a set of state-of-the-art optimization methods. For example, some researchers proposed to use a new storage device, such as Flash-based SSD, to reduce data access overhead. There are also some other optimization methods. In GraphChi [33], the authors proposed a sliding window method to load the graph blocks into memory on demand; by using this method, GraphChi can significantly reduce the disk access overhead. GridGraph [32] proposed a 2-D edge partition method to selectively schedule the graph blocks to reduce the I/O overhead, and the experimental results show the proposed method can achieve a useful data accessing performance. In EC-VHP [34], the authors designed both a simple hash index structure and a multiqueue parallel sequential index structure to improve the processing efficiency of message communication. GraphX [35], which is the core component of Spark [36], providing a stack data solution on top of Spark, can conveniently and efficiently complete a complete set of pipeline operations for graph calculation. Chaos [37] used the secondary memory and graph partitioning scheme to maximize the degree of parallelism and reduce communication overhead.

B. Programming Model

Most of existed research, such Pregel [38], which is the first graph processing system developed by Google, as well as PowerGraph [31], GraphLab [39], and PowerLyra [40], are adopt the vertex-centric. The vertex-centric model-based GPFs are using the Think Like A Vertex (TLAV) paradigm [41]. Compared with the traditional MapReduce model, the TLAV provides a better locality of data access and better scalability, which makes it is more fixable to implement the graph algorithms. However, the overhead of a large amount of random memory access to the TLAV frameworks limits the system performance. In order to solve this problem, X-Stream [42] proposed an edge-centric programming method. In the edge-centric programming model, the edges can be visited in a sequential fashion, which can significantly improve the data access efficiency. In addition, there are some other kinds of the programming model, such as the path-centric programming model proposed in PathGraph [43] and the data-centric programming method [19].

C. Execution Model

Many recent research, such as Gemini [44], Cyclops [45], and Hama [46], are adopt the bulk synchronous parallel (BSP) [38] execution model. In the BSP execution model, a global synchronization is required at the end of each iteration. Due to the uneven degree distribution of the real graph, it is easy to lead to the straggler problem for the vertex-centric programming model because the small degree vertices need to wait for the large degree vertices in the synchronization operation. In order to solve this problem, some other recent state-of-the-art works, such as Chaos [37], PowerGraph [31], PowerLyra [40], and GPS [47], adopt the GAS [31] execution model. In the GAS model, all the operations have been divided into three phases: the information collection phase (Gather), the application phase (Apply), and the distribution phase (Scatter). Since the GAS is an asynchronous execution model, the read-write and write-write conflicts should be considered. There are also some other types of execution models, such as the similar variants of GAS, and GunRock [19] divided the execution operation into Advance, Computer, and Filter phases, while SC-BSP divided the operation as Update, Push, Pull, and Sink (UPPS) [48].

D. Graph Processing on New Architectures

In recent years, there are also some researchers proposed some works, which focused on GPU, FPGA, and some other new architectures. For example, the Medusa [30], GunRock [19], CuSha [22], and Frog [49] are designed on GPU. These works attempt to use the high memory bandwidth to improve the memory access efficiency of GPU, and they also proposed some insights on the programming model, execution model, and data storage model. FPGP [50] and Graphops [51] are designed by using the FPGA. There are also some frameworks designed for the hybrid architectures, such as the CPU and GPU hybrid GPF, i.e., Totem [21]. There are also some attempts to design the hardware-based accelerator for graph processing, such as Graphicionado [52]

that is designed by combating the application execution inefficiencies on general-purpose CPUs, while article [53] proposed a hardware-driven solution.

Due to the booming applications, graph processing has attracted lots of attention from both academia and industry. Though there are lots of state-of-the-art works focused on transitional architecture, it is hard to unleash the computing efficiency of the hardware. There are few works focused on the new architecture devices, and most of the existed works are experimental works, which is hard to program, with low availability. This article attempts to propose an ideal framework for the new architecture devices, which can provide a reference for future works.

VI. CONCLUSION AND FUTURE OPPORTUNITIES

Intelligent data processing for Smart CPSs has attracted much attention from both industry and academics because of the complex dynamic interaction with their environment without any prior information. Focused on the large-scale HITL data processing challenges, this article designed a set of experiments to illustrate the performance of existed GPFs and then proposed a graph-based HITL CPS data processing framework, named Barge, which can fit for different computing devices. By using the semantic model, Barge can implement and map the same code to different computing devices without any change, which can release the programmer from the strenuous and repetitive work. Furthermore, the architecture-driven programming model can make programming logic suitable for kinds of parallel devices, such as GPU, FPGA, ASIC, and many other smart devices. In addition, the data- and topology-aware graph data partition scheme of Barge will partition the large-scale graphs by considering the data structure and also the feature of the computing device to make sure that the large-scale graph to be processed efficiently on smart devices. In further, we will design and implement a hardware compatible framework that can be mapped on to kinds of parallel devices, such as GPU, FPGA, and ASIC, without change the codes.

REFERENCES

- [1] *Fog Computing and the Internet of Things: Extend the Cloud to Where the Things Are*, Cisco, San Jose, CA, USA, 2015.
- [2] M. Saadi, M. TalhaNoor, A. Imran, W. TariqToor, S. Mumtaz, and L. Wuttisittikulij, "Iot enabled quality of experience measurement for next generation networks in smart cities," *Sustain. Cities Soc.*, vol. 60, no. 9, Sep. 2020, Art. no. 102266.
- [3] X. Lin, J. Wu, S. Mumtaz, S. Garg, J. Li, and M. Guizani, "Blockchain-based on-demand computing resource trading in IoV-assisted smart city," *IEEE Trans. Emerg. Topics Comput.*, early access, Feb. 6, 2020, doi: [10.1109/TETC.2020.2971831](https://doi.org/10.1109/TETC.2020.2971831).
- [4] M. H. Cintuglu, O. A. Mohammed, K. Akkaya, and A. S. Ulugac, "A survey on smart grid cyber-physical system testbeds," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 446–464, 1st Quart., 2017.
- [5] J. Sztipanovits *et al.*, "Toward a science of cyber-physical system integration," *Proc. IEEE*, vol. 100, no. 1, pp. 29–44, Jan. 2012.
- [6] F. Hofer, "Architecture, technologies and challenges for cyber-physical systems in industry 4.0: A systematic mapping study," in *Proc. 12th ACM/IEEE Int. Symp. Empirical Softw. Eng. Meas.*, New York, NY, USA, 2018, pp. 1–10.
- [7] M. I. Ashraf, M. Bennis, C. Perfecto, and W. Saad, "Dynamic proximity-aware resource allocation in vehicle-to-vehicle (V2V) communications," in *Proc. IEEE Globecom Workshops*, Dec. 2016, pp. 1–6.
- [8] S. Mumtaz, A. Gameraio, and J. Rodriguez, "EESM for IEEE 802.16e: WiMaX," in *Proc. 7th IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, May 2008, pp. 361–366.
- [9] C.-W. Tsai, C.-F. Lai, M.-C. Chiang, and L. T. Yang, "Data mining for Internet of Things: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 77–97, Qua. 2014.
- [10] Y. Liu, X. Fang, M. Xiao, and S. Mumtaz, "Decentralized beam pair selection in multi-beam millimeter-wave networks," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2722–2737, Jun. 2018.
- [11] J. Wu, S. Guo, J. Li, and D. Zeng, "Big data meet green challenges: Big data toward green applications," *IEEE Systems J.*, vol. 10, no. 3, pp. 888–900, Sep. 2016.
- [12] J. Wu, S. Guo, H. Huang, W. Liu, and Y. Xiang, "Information and communications technologies for sustainable development goals: State-of-the-art, needs and perspectives," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2389–2406, 3rd Quart., 2018.
- [13] K. Zhang, L. Lan, J. T. Kwok, S. Vucetic, and B. Parvin, "Scaling up graph-based semisupervised learning via prototype vector machines," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 3, pp. 444–457, Mar. 2015.
- [14] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, Jan. 2016.
- [15] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, doi: [10.1109/TKDE.2020.2981333](https://doi.org/10.1109/TKDE.2020.2981333).
- [16] Y. Lou, M. Uddin, N. Brown, and M. Cafarella, "Knowledge graph programming with a human-in-the-loop: Preliminary results," in *Proc. Workshop Hum. Loop Data Anal.*, New York, NY, USA, 2019, pp. 1–7.
- [17] A. Holzinger *et al.*, "Interactive machine learning: Experimental evidence for the human in the algorithmic loop," *Int. J. Speech Technol.*, vol. 49, no. 7, pp. 2401–2414, Jul. 2019.
- [18] A. Doan, "Human-in-the-loop data analysis: A personal perspective," in *Proc. Workshop Human Loop Data Analy.*, New York, NY, USA, 2018, pp. 1–5.
- [19] Y. Wang, A. Davidson, Y. Pan, Y. Wu, A. Riffel, and J. D. Owens, "Gunrock: A high-performance graph processing library on the GPU," in *Proc. 20th ACM SIGPLAN Symp. Princ. Pract. Parallel Program.*, New York, NY, USA, 2015, pp. 1–12.
- [20] H. Wang, L. Geng, R. Lee, K. Hou, Y. Zhang, and X. Zhang, "SEP-graph: Finding shortest execution paths for graph processing under a hybrid framework on GPU," in *Proc. 24th Symp. Princ. Pract. Parallel Program.*, New York, NY, USA, Feb. 2019, p. 38.
- [21] A. Gharaibeh, L. Beltr ao Costa, E. Santos-Neto, and M. Ripeanu, "A yoke of oxen and a thousand chickens for heavy lifting graph processing," in *Proc. 21st Int. Conf. Parallel architectures compilation Techn.*, New York, NY, USA, 2012, pp. 345–354.
- [22] F. Khorasani, K. Vora, R. Gupta, and L. N. Bhuyan, "CuSha: Vertex-centric graph processing on GPUs," in *Proc. 23rd Int. Symp. High-Perform. Parallel Distrib. Comput.*, New York, NY, USA, 2014, p. 239–252, p. 239.
- [23] A. H. Nodehi Sabet, J. Qiu, and Z. Zhao, "Tigr: Transforming irregular graphs for GPU-friendly graph processing," in *Proc. Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, New York, NY, USA, 2018, p. 622–636.
- [24] K. Meng, J. Li, G. Tan, and N. Sun, "A pattern based algorithmic autotuner for graph processing on GPUs," in *Proc. 24th Symp. Princ. Pract. Parallel Program.*, New York, NY, USA, Feb. 2019, p. 201–213.
- [25] D. B. West, *Introduction to Graph Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.
- [26] J. Leskovec. (2009). *Stanford Network Analysis Project*. <http://snap.stanford.edu/index.html>
- [27] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.
- [28] Y. Virkar and A. Clauset, "Power-law distributions in Binned empirical data," *Ann. Appl. Statist.*, vol. 8, no. 1, pp. 89–119, Mar. 2014.
- [29] S. Kanev *et al.*, "Profiling a warehouse-scale computer," *SIGARCH Comput. Archit. News*, vol. 43, no. 3S, p. 158–169, Jun. 2015.
- [30] J. Zhong and B. He, "Medusa: Simplified graph processing on GPUs," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 6, pp. 1543–1552, Jun. 2014.
- [31] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin, "Powergraph: Distributed graph-parallel computation on natural graphs," in *Proc. 10th USENIX Conf. Operating Syst. Design Implement.*, 2012, p. 17–30.
- [32] X. Zhu, W. Han, and W. Chen, "Gridgraph: Large-scale graph processing on a single machine using 2-level hierarchical partitioning," in *Proc. 2015 USENIX Conf. Usenix Annu. Tech. Conf.*, 2015, p. 375–386.

- [33] A. Kyrola, G. Blelloch, and C. Guestrin, "Graphchi: Large-scale graph computation on just a PC," in *Proc. 10th USENIX Conf. Oper. Syst. Design Implement.*, 2012, p. 31–46.
- [34] L. Fangling *et al.*, "Edge cluster based large graph partitioning and iterative processing in BSP," *J. Comput. Res. Develop.*, vol. 52, no. 4, p. 960, 2015.
- [35] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica, "Graphx: Graph processing in a distributed dataflow framework," in *Proc. 11th USENIX Conf. Oper. Syst. Des. Implement.*, 2014, p. 599–613.
- [36] M. Zaharia *et al.*, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proc. 9th USENIX Conf. Netw. Syst. Design Implement.*, 2012, p. 2.
- [37] A. Roy, L. Bindschaedler, J. Malicevic, and W. Zwaenepoel, "Chaos: Scale-out graph processing from secondary storage," in *Proc. 25th Symp. Operating Syst. Princ.*, New York, NY, USA, 2015, pp. 410–412.
- [38] G. Malewicz *et al.*, "Pregel: A system for large-scale graph processing," in *Proc. Int. Conf. Manage. Data*, New York, NY, USA, 2010, p. 6.
- [39] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein, "Graphlab: A new framework for parallel machine learning," in *Proc. Conf. Uncertainty Artif. Intell.*, Arlington, VA, USA, 2010, p. 340–349.
- [40] R. Chen, J. Shi, Y. Chen, B. Zang, H. Guan, and H. Chen, "Powerlyra: Differentiated graph computation and partitioning on skewed graphs," *ACM Trans. Parallel Comput.*, vol. 5, no. 3, p. 15, Jan. 2019.
- [41] R. R. McCune, T. Weninger, and G. Madey, "Thinking like a vertex: A survey of vertex-centric frameworks for large-scale distributed graph processing," *ACM Comput. Surv.*, vol. 48, no. 2, p. 12, Oct. 2015.
- [42] A. Roy, I. Mihailovic, and W. Zwaenepoel, "X-stream: Edge-centric graph processing using streaming partitions," in *Proc. ACM Symp. Operating Syst. Princ.*, New York, NY, USA, 2013, pp. 472–488.
- [43] P. Yuan, W. Zhang, C. Xie, H. Jin, L. Liu, and K. Lee, "Fast iterative graph computation: A path centric approach," in *Proc. Int. Conf. for High Perform. Comput., Netw., Storage Anal.*, Nov. 2014, p. 401.
- [44] X. Zhu, W. Chen, W. Zheng, and X. Ma, "Gemini: A computation-centric distributed graph processing system," in *Proc. 12th USENIX Conf. Operating Syst. Des. Implement.*, 2016, p. 301–316.
- [45] R. Chen, X. Ding, P. Wang, H. Chen, B. Zang, and H. Guan, "Computation and communication efficient graph processing with distributed immutable view," in *Proc. 23rd Int. Symp. High-Perform. Parallel Distrib. Comput.*, New York, NY, USA, 2014, pp. 215–226.
- [46] S. Seo, E. J. Yoon, J. Kim, S. Jin, J.-S. Kim, and S. Maeng, "HAMA: An efficient matrix computation with the MapReduce framework," in *Proc. IEEE 2nd Int. Conf. Cloud Comput. Technol. Sci.*, Nov. 2010, pp. 721–726.
- [47] S. Salihoglu and J. Widom, "GPS: A graph processing system," in *Proc. 25th Int. Conf. Scientific Stat. Database Manage.*, New York, NY, USA, 2013, pp. 1–4.
- [48] Z. Xiang, L. Bo, S. Haichuan, and X. Weidong, "A revised BSP-based massive graph computation model," *Chin. J. Comput.*, vol. 40, no. 1, pp. 223–235, Jan. 2017.
- [49] X. Shi *et al.*, "Frog: Asynchronous graph processing on GPU with hybrid coloring model," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 29–42, Jan. 2018.
- [50] G. Dai, Y. Chi, Y. Wang, and H. Yang, "FPGP: Graph processing framework on FPGA a case study of breadth-first search," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays*, New York, NY, USA, 2016, p. 10–105.
- [51] T. Oguntebi and K. Olukotun, "GraphOps: A dataflow library for graph analytics acceleration," in *Proc. ACM/SIGDA Int. Symp. Field-Program. Gate Arrays - FPGA*, New York, NY, USA, 2016, p. 111–117.2016, p. 111.
- [52] T. J. Ham, L. Wu, N. Sundaram, N. Satish, and M. Martonosi, "Graphicionado: A high-performance and energy-efficient accelerator for graph analytics," in *Proc. 49th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO)*, Oct. 2016, pp. 1–4.
- [53] M. M. Ozdal *et al.*, "Energy efficient architecture for graph analytics accelerators," in *Proc. ACM/IEEE 43rd Annu. Int. Symp. Comput. Archit. (ISCA)*, Jun. 2016, p. 166–177.

[Home](#) > [Multimedia Tools and Applications](#) > Article

1194: Secured and Efficient Convergence of Artificial Intelligence and Internet of Things

[Published: 23 January 2021](#)

Learning based MIMO communications with imperfect channel state information for Internet of Things

[Dan Deng](#), [Xingwang Li](#)  & [Varun G. Menon](#)

[Multimedia Tools and Applications](#) **80**, 31265–31276 (2021)

261 Accesses | **1** Citations | [Metrics](#)

Abstract

Imperfect channel state information (CSI) may seriously worsen the system performance for classical MIMO communications. In order to overcome the impacts of imperfect CSI for Internet of things, we propose a deep convolutional neural network (DCNN) based MIMO detection algorithm, where the DCNN is trained offline and works online to refine the imperfect CSI and improve the bit error rate of the wireless systems. Two types of learning based detectors, i.e., with or without accurate CSI, are proposed in this paper to reduce the detrimental effects of imperfect CSI. The

of Higher Education Institutions in Henan Province Grant 20A510007, the Natural Science Foundation of Shaanxi Province under Grant 2020JQ-844, the Fundamental Research Funds for the Universities of Henan Province under Grant NSFRF180309, the Key Research and Development Program of Shanxi under Grant 201903D121117.

Author information

Authors and Affiliations

**School of Information Engineering,
Guangzhou Panyu Polytechnic, Guangzhou,
511483, China**

Dan Deng

**School of Physics and Electronic
Information Engineering, Henan
Polytechnic University, Jiaozuo, 454150,
China**

Xingwang Li

**Department of Computer Science and
Engineering, SCMS School of Engineering
and Technology, Ernakulam, 683576, India**

Varun G. Menon

Corresponding author

Correspondence to [Xingwang Li](#).

Additional information

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and

institutional affiliations.

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Deng, D., Li, X. & Menon, V.G. Learning based MIMO communications with imperfect channel state information for Internet of Things. *Multimed Tools Appl* **80**, 31265–31276 (2021). <https://doi.org/10.1007/s11042-020-10387-6>

Received	Revised	Accepted
21 June 2020	23 October 2020	22 December 2020

Published	Issue Date
23 January 2021	August 2021

DOI

<https://doi.org/10.1007/s11042-020-10387-6>

Keywords

Imperfect CSI **MIMO communications**

Deep learning **Detection**

INTERNATIONAL RESEARCH JOURNAL OF SCIENCE ENGINEERING AND TECHNOLOGY



ISSN 2454-3195

An Internationally Indexed Peer Reviewed & Refereed Journal

WWW.RJSET.COM
www.isarasolutions.com

Published by iSaRa Solutions

IoT based Power Analyzer for an Automated Home

Susmi Jacob, Shilpa P C, and Binu John

Susmi Jacob is with Department of Computer Science, SCMS school of Engineering and Technology, Kerala

Email: susmijacob@scmsgroup.org

Email: shilpape@scmsgroup.org Email: binujohn@scmsgroup.org

Abstract— As technology enhances, houses have become smarter and energy-efficient. Traditional switches are gradually being replaced with automated centralized switches with remotecontrols in modern homes. The inmates find it difficult to activate the traditional wall switches located throughout the residence when they are needed. Smartphones give a modern option for remote-controlled home automation. The main goal of this study is to design and construct a low-cost Internet Of Things (IoT) enabled energy monitoring system that can be benefited in different applications such as energy management in smart automated homes as well as electricity billing systems. We develop a power analyzer using an Arduino board and Current Transformer (CT) sensor, as well as gadgets that can be controlled remotely using an Android OS smartphone. At the beneficiary end, a Bluetooth module is connected to an Arduinoboard, while at the transmitter end, a GUI application running on a PDA sends ON/OFF bearings to the authority where homeequipment is installed. Through this technique, the pieces of equipment are turned ON/OFF by tapping the correct spots on the GUI. Additionally, we recorded the daily and monthly readings of power consumed to monitor the power consumption in a month and to give necessary warnings and suggestions to the consumer.

I. INTRODUCTION

Energy conservation is an urgent requirement of this era. The idea of resourceful equipment in assorted areas such as air conditioning, refrigeration, lighting, etc will lead to energy-efficient utilization of household devices. Energy auditing is an inevitable mechanism for analyzing the systematic energy utilization of equipment and devices. It provides a better way to control the use of electricity in case of excess usage and also helps the user to fix the inaccuracy in the electricity bill which may show excess amount sometimes [1].

Domestic electricity bill which presents surplus amount that causes disapproval for the users. By using a smart power analyzer system user monitors the energy utilization details at the equipment level and governs it rather than calculating the fixed monthly expenditure. This may also aid the user to restore the normal

appliances with energy-efficient and smart devices[2]. Critically, the checking power system can caution the user on startling overabundance utilization brought about by the improper working, absence of timely maintenance, and so forth. Further, energy management in the proper way leads to the better utilization of the resources, and thereby we can reduce the cost.

Thus the wastage of energy can be reduced to meet future needs by protecting the precious resource. Similarly, the cost estimation can be predicted for each industrial unit or home for better utilization of the energy. Moreover, this comparison and the analysis of cost estimation for each industrial unit or home will reduce the production cost which will result in the tremendous profit of the industry.

The importance of saving electricity attributes to the fact that electricity is generated from natural resources, which are limited and reducing as time goes. The unsustainable use of natural resources not only affects the balance of nature but also makes the planet completely unfit to live. Saving electricity can reduce its production, thereby it reduces the manpower and cost of production. Similarly by producing electricity from other means like coal accelerates pollution. Thus pollution can then be controlled by limiting the production of electricity by consuming electricity efficiently and effectively to save the planet.

A. Motivation

Energy consumed by each industrial unit or home can be measured and analyzed based on time stamps. Thus, the energy used by different industrial units or home can be compared and analyze which consume more energy. Hence,

the industrial units or homes which possess more energy consumption can be taken care and inference can be made that whether the industrial unit or home are actually needed this much energy or whether this consumption is due to any faults in machines. Also, it can control electronic devices from anywhere at any time using the internet or within a Wi-Fi connection. Thus wastage of electricity gets reduced by controlling devices at the correct time. The electricity meter stationed in the individual constructions displays the energy utilized by the buildings. There is an urgent necessity for a novel power analyzing system.

The wastage of energy can be reduced to save the future by protecting their precious resource. By reducing the wastage of electricity, the production of electricity can be decreased. Electricity is produced from natural resources such as water etc. So by limiting the production of electricity will lead to less exploitation of these natural resources. Similarly, electricity can be produced from other means like coal which increases pollution. Thus, pollution, manpower, and cost can be reduced by limiting the production of electricity which can be achieved by saving and using electricity in efficient and effective ways to save the planet.

B. Contribution

Internet of Things (IoT) is relied upon to achieve an enormous measure of progress in the field of pervasive computing. IoT-based energy management framework can contribute a lot to the preservation of energy. It can control different electric devices in the home from anywhere at any time using an internet connection or Wi-Fi. Similarly, the cost estimation can be predicted for each industrial unit or home. And this comparison and analysis of cost estimation from each industrial unit or home will reduce the production cost, hence will result in tremendous profit to the industry.

An IoT Based Power Analyzer is a general-purpose real-time mobile application. It is used to measure the usage of power consumed by each industrial unit or home. It also estimates the cost of each industrial unit or home.

We propose the implementation of a smart plug, an energy observation system that provides real-time information on device-level energy consumption. An Arduino microcontroller board, an ENC28J60 LAN module, and a current electrical device detecting element are used in the suggested device. The present sensor technology employed is non-invasive. The device's computer software is written in Android and the statistics are stored in a server. The end product is a smart plug that uses the Arduino-android platform to monitor a distant device.

The next half essentially carries out the style of an IoT sensible Home System (IoTSHS) which gives the remote control to sensible home appliances via android mobile phones, similarly like PC/Laptop. The controller accustomed style the IoTSHS is Arduino Uno micro-controller. A temperature detector is provided to analyse the surrounding temperature and alert the users if the regulation of the fan speed is required. The designed IoTSHS will edge the total elements within the community by facilitating technologically improved remote dominance for the sensible home.

The organization of the paper is as follows. Section II summarises the literature survey.

II. RELATED WORK

We could find several substantiating facts which are beneficial for our research work. Literature survey have been done in the area of power analyzer as well as home automation.

A. Power Analyzer

A survey done on already existing literature on energy consumption and analysing reveals that a tremendous change have been reported in the implementation [5]. For the necessities of acknowledging energy-sparing and outflow decrease for the smart network, this paper presents an observing foundation of the energy, the board framework which is coordinated with flexible energy. It governs, reads, and evaluate the processed readings, of energy creation, energy transportation, and energy utilization for the smart system. The design of the community energy system is split into the acquisition layer, transmission layer, and management layer. The

information correspondence between information securing devices and server embraces Ethernet and TCP/IP convention. The information correspondence between field instruments and information procurement gadgets receives the RS-485 interface and Modbus correspondence convention.

The advancement of an observing platform for a smart network can improve the administration level as well as diminish the energy utilization of the network energy framework. A few studies done on the possibilities of better energy utilization proposes various methods [6]. The users are going to be alerted when the electricity usage in their home exceeds the limits to avoid the wastage of energy consumption.

Various studies done on the data acquisition and control of energy utilization efficient method. Smart Home Energy Management Systems Based on Non-Intrusive Load Monitoring [7] proposed a unique system of good domestic energy management systems incorporating each approaches, in order that correct energy utilization watching and assuming communication with the smart home device at the same time achieved. The good parts directly management the appliances, whereas the essential controller coordinates the info assortment. The key point is that the competence of mechanically aligning the appliances to their corresponding sockets, reducing the need for manual initial setup. We assure that our smart framework, if sophisticated widely, will profit not solely separate households by reducing current bills.

A Server Agent is the term given to the innovative microcontroller used in this project. While the initial PC-based server is offline, this agent collects information from buyers. Once all knowledge area units have dropped, the Server Agent can turn down the PC-based server again. This procedure minimises the amount of energy used [4].

We conducted several research on An Efficient Home Energy Management Solution based on Automatic Meter Reading, with a specific focus on household energy consumption, and proposed a simple and

effective system for reducing power waste in a home. They designed a home energy management system (HEMS) that uses a simple energy management mechanism to make it easier to implement. It also makes use of AMR (Automatic Meter Reading) network-based power line communication (PLC). They installed HEMS in real customers' homes and validated the results, resulting in a significant energy conservation outcome that is vital for power reduction.[9].

GAPMR stands for "automated power metre reading system using GSM network." It is a system that consists of GSM digital power metres installed in the client unit and an electricity e-billing system at the energy supplier's end. The GSM digital electric metre (GPM) could be a single part IEC61036, a standard digital kWh electric metre with incorporated GSM electronic equipment that uses the GSM network to report power usage readings. The readings are send back to the energy supplier as short messaging system (SMS) via a wireless manager[8]. On the service provider's side, there's a degree e-billing system that's used to monitor all SMS metre readings, encrypt the charge value, update the information, and send charge notifications to its various customers by SMS, email, web portal, and letters. The effectiveness and efficiency of automatic meter reading

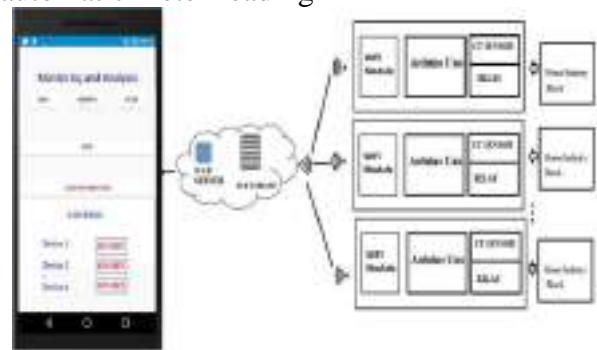


Fig. 1: Proposed system architecture

was demonstrated by implementing a GAPMR system. Also the charge and notification through the employment of GSM network can also be monitor by this system.

B. Home Automation

The research into the implementation of IoT for condition monitoring in homes shows that condition monitoring and energy management in the home may be done in an inexpensive, flexible, and cost-effective manner. The designed system's major tasks include remote control and management of home devices such as electrical lamps, heaters, and so on, as well as unassertive surveillance of domestic usage and supplying closed intelligence to reduce energy consumption using IoT technology.[3]. This will assist and schedule the individual's operation time according to the energy demand.

The majority of current sensible Web connection is provided by TV set-top boxes, allowing users to install and run additional advanced applications or plugins/add-ons for a certain platform. It means that a sensible—a wise TV set-top box is a good option for acting as a hub that integrates a variety of smart home solutions. This study presents a framework for managing household appliances that is enabled by a smart TV set-top box.

Many buttons (often dozens) are designed on the remote controller in home areas as the quality of devices/appliances improves, yet many of them are rarely used. A user is also perplexed by the controller, despite the fact that he or she only wants to do a simple task. This confusion in addition ends up in a far better likelihood of mal-functions. Additionally to the current a typical ways in which of communication between remote controllers and connected devices, like ventricose language (XML) messages, unit generally bandwidth-consumptive. The asymmetrical feature of Point-n-Press provides for simple and intuitive management by informing the target device and displaying the target's management interface on the remote controller's screen. Exclusively sensible pops that unit relevant to the present context unit by using the state dependencies of home device/equipment actions. Two real prototypes are being used to test the feasibility of the proposed theme. According to the findings, Point-n-Press could be a useful and relevant management theme for IoT-based smart homes.

III. PROPOSED SYSTEM

An IoT Based Power Analyzer with home automation is a real time IoT based mobile application which analyze the energy consumed by each industrial unit. It also predicts the cost estimation of each industrial unit or home which incorporates with online data storage. It clearly specifies the average energy used per day, per month and per year along with cost estimation. Energy consumed by different industrial units or home can be measured and analyzed based on time stamps. Thus, the energy used by different industrial units or home can be compare and analyze which consume more energy. So the industrial unit or home which possess more energy consumption can be taken care and inference can be made that whether the industrial units or home are actually needed this much energy or these much consumption

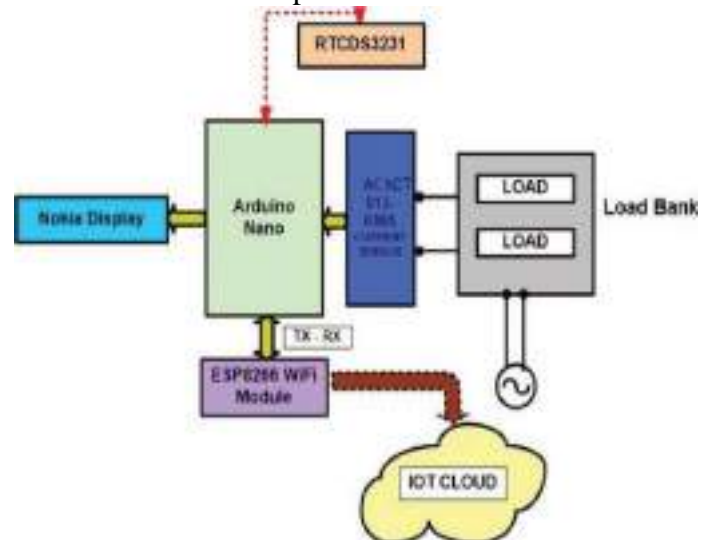


Fig. 2: Power Analyzer Architecture

is due to any faults in machines. By identifying faulty devices before their life expired, the production cost can be reduced to some extent because the faulty devices can be repaired before it get fully damaged.

The second section focuses on the design of an IoT sensible House System (IoTSHS), which might provide access to a smart home via a mobile device, similar to a PC or laptop. The Arduino Uno microcontroller is the standard controller for the IoTSHS. A temperature detector

is included to indicate the temperature in the area and inform the user if the fan speed needs to be adjusted. By giving improved remote dominance for the sensible house, the developed IoTSHS will edge the total elements within society.

A. Product Perspective

This app can control electronic equipment and to measure the usage or consumption of electricity for each device. Cost of consumption per devices can also available in this app. Thus the electricity usage can be analyzed and provides an efficient way electricity usage. Customer can analysis on a real time based usage of various electrical device and control consumption of energy through the app. The cost of this system is very less as compared to existing system in market. The feature that makes the application unique is that no other application has the facilities to measure power consumption of each device.

IV. SYSTEM ARCHITECTURE

The Current sensor which is a vital part of our system is attached to the phase wires of industrial units or home. Then it senses the current usage by industrial/home and give to arduino, where current sensors are attached to the arduino with help of extra circuit which contain 3.5mm audio connectors. This circuit helps to limit the voltage coming from industrial units/home because arduino have only 5v capacity. Current sensors give the analog value of current to arduino which convert to digital value within arduino (in build Analog to Digital Convertor). From this arduino values are passed to database through Wi-Fi module. Then data processes in database which provides data to the user through

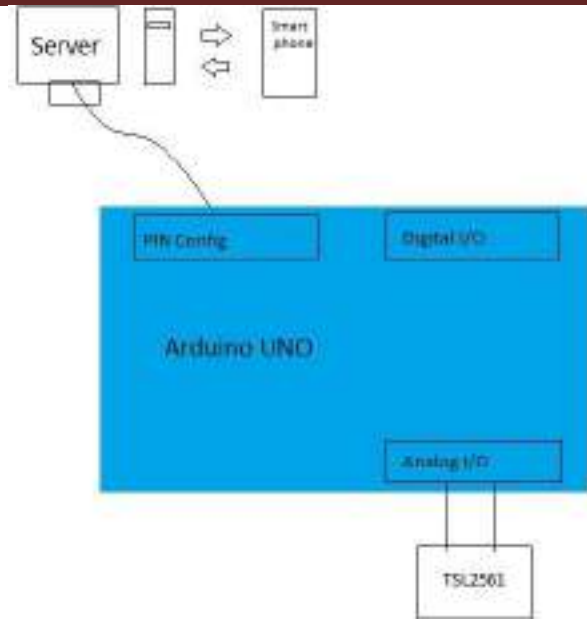


Fig. 3: Sensor Module outline

android application. The continuous monitoring values are given to user. Along with that the daily, monthly and yearly usage analyses are also given. The prediction of electricity bill given to user provides the energy consumption and cost estimation of industrial units [10].

1) *Power Analyzer Module:* The block diagram to show

the operation of the power analyzer is shown in figure 2. The current sensor SCT013030 is the main part of the circuit. The electricity measures are detected in real time and passed to the server through ESP 8266 WiFi module connected to the Arduino Nano. The SCT013 030 split core current sensor is capable of detecting a maximum current of 30A and provides a peak output voltage of 1V peak for that current. The output voltage thus generated is then transmitted to the Arduino Nano microcontroller via the input of the analog-to-digital converter (ADC). This voltage waveform is shifted up to 2.5 VDC. The rms value of the output signal is calculated and the power is calculated in the program.

2) *Home Automation Architecture:* Figure 4 depicts the

basic diagram of the proposed IoT Smart Home (IoT SHS) system. This is a low-cost, easily manageable, and profitable product. By providing

superior remote control for home appliances, it addresses the entire company's segment.

To control devices remotely, it comes standard with WiFi. Lights, fans, and sockets are usually found in every room of any house where our items can be installed. This product does not affect the room's electrical distribution wiring; everything remains the same except for the relays, which are wired in series with the switch or socket in the distribution box.

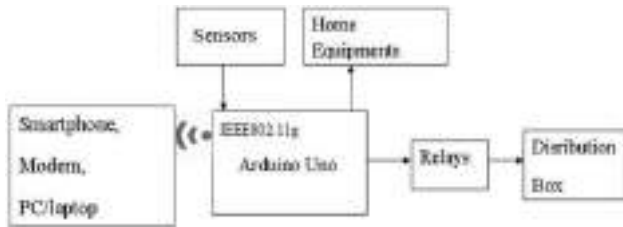


Fig. 4: Home Automation Architecture

The controller is a WiFi-based microcontroller (Arduino UNO) that serves as the system's brain and controls all of the other components. The ambient temperature is indicated through a temperature sensor.

3) *Sensor Module*: The TSL2561 is a cheap, but sophisticated, light sensor. To better predict the human eye's response, the TSL2561 integrates infrared and visible light sensors. This is contrast with simpler sensors, such as photoelectric sensors and photodiodes. The TSL2561 can measure both very small and very large amounts of light because it is a built-in sensor (it absorbs light for a pre calculated amount of time). The block diagram is shown in figure 3.

Circuit Diagram : The CT sensor cannot be directly connected to the Arduino since the high current from the

sensor can damage the Arduino, So we connect firstly the burden resistor of low resistance of about 33 is connected represented by R_b in the figure. Then the voltage divider bias consisting of R_1 and R_2 and additionally a bypass capacitor C_1 of 470F to block the digital signal, there by getting only the analog current signal to the Arduino. The analog pins of the arduino is indicated by notation $A_0 - A_5$ we have connected the CT sensor to pin A_0 . Then by clamping the CT sensor on a live wire gives the current output.

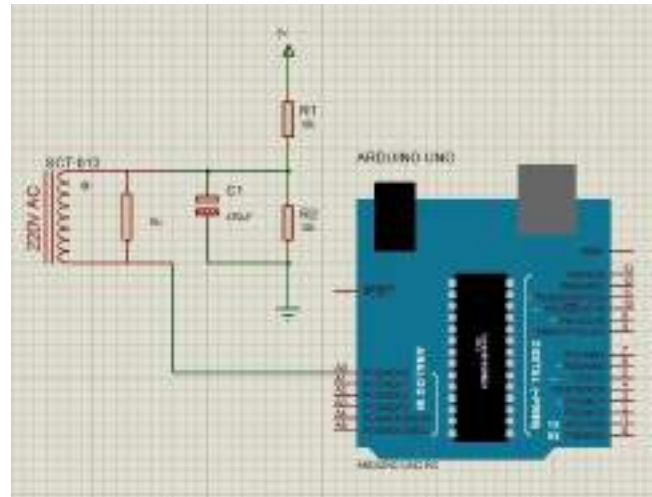


Fig. 5: (CT sensor interfaced with arduino)

V. EXPERIMENTS AND DISCUSSION

Initially, login to the Digital space using username and password and check the username and password is valid by comparing in database. If it is valid it enter to the next module else unsuccessful login. There are different phases for the app where we can do the control of our home appliances automatically through the application.

Now we measure the usage statistics of the home that is read through the CT sensor. We can analyse the power usage in various manner according to our requirements. The app will provide the provision to calculate the power consumption in daily, monthly and year wise. Also have the facility to monitor the live power consumption. The app interface for the power analysis is shown in figure 6.



Fig. 6: Application Interface



Fig. 7: Daily usage analysis



Fig. 8: Monthly and yearly usage analysis

For calculating the daily consumption we can choose the option in the app and it will give the consumption as a graph which is shown in the figure 7. By taking the power consumption in

this manner we would be able to control the usage.

We have also done the analysis of monthly and yearly power consumption rates and plotted as a graph. It is very convenient to track the miss usage of power. The results from the analysis is shown in the figure 8. The next part in the experiment was the cost estimation. It will give the summary of average power usage in a Day/Month/Year base. Also we have done the average cost estimation for a day, Month and Year.

CONCLUSION

IoT based energy the board framework can contribute a ton into preservation of energy. It can control different electric devices in home from anywhere at any time using internet connection or Wi-Fi. Similarly the cost estimation can be predicted for each industrial unit or home. We proposed an IoT Based Power Analyzer is a general purpose real time mobile application. It estimates the cost of each industrial unit or home. We propose creating a smart plug, which is an energy observation system that provides real-time information on energy use at the device level. An Arduino microcontroller board, an ENC28J60 LAN module, and a current electrical device sensing element are used in the suggested device. The end product is a smart plug that uses the Arduino-android platform to monitor a remote device. The second section focuses on the design of an IoT sensible House System (IoTSHS), which might provide access to a smart home via a mobile device, similar to a PC or laptop. The Arduino Uno microcontroller is the standard controller for the IoTSHS. By connecting sample appliances and successfully controlling them from a wireless mobile device, the home automation system has been experimentally proved to perform satisfactorily.

REFERENCES

- [1] Aldabbagh, Ghadah, Raneen Alzafarani, and Ghadi Ahmad. "Energy Efficient IoT Home Monitoring and Automation System (EE-HMA)."
- [2] Kodali, Ravi Kishore, and Subbachary Yerroju. "Energy

- efficient home automation using IoT." In 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), pp. 151-154. IEEE, 2018.
- [3] Jabbar, Waheb A., Tee Kok Kian, Roshahliza M. Ramli, Siti Nabila Zubir, Nurthaqifah SM Zamrizaman, Mohammed Balfaqih, Vladimir Shepelev, and Soltan Alharbi. "Design and fabrication of smart home with Internet of Things enabled automation system." *IEEE Access* 7 (2019): 144059-144074.
- [4] Gray, Chrispin, Robert Ayre, Kerry Hinton, and Leith Campbell. "'smart' is not free: Energy consumption of consumer home automation systems." *IEEE Transactions on Consumer Electronics* 66, no. 1 (2019):87-95.
- [5] Kitayama, Ryosuke, Takashi Takenaka, Masao Yanagisawa, and Nozomu Togawa. "Scalable and small-sized power analyzer design with signal-averaging noise reduction for low-power IoT devices." In 2016 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 978-981. IEEE, 2016.
- [6] Gao, Xiaofei, and Qin Zhou. "A low consumption DSP based power analyzer." In The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), pp. 164-168. IEEE, 2014.
- [7] Hui, Li, Wang Gui-rong, Wei Jian-ping, and Duan Peiyong. "Monitoring platform of energy management system for smart community." In 2017 29th Chinese Control And Decision Conference (CCDC), pp.1832-1836. IEEE, 2017.
- [8] Chaudhari, Sneha, Purvang Rathod, Ashfaque Shaikh, Darshan Vora, and Jignasha Ahir. "Smart energy meter using Arduino and GSM." In 2017 International Conference on Trends in Electronics and Informatics(ICEI), pp. 598-601. IEEE, 2017.
- [9] Al-Ali, Abdul-Rahman, Imran A. Zualkernan, Mohammed Rashid, Ragini Gupta, and Mazin AliKarar. "A smart home energy management system using IoT and big data analytics approach." *IEEE Transactions on Consumer Electronics* 63, no. 4 (2017): 426-434.
- [10] Hlaing, Win, Somchai Thepphaeng, Varunyou Nontaboot, Natthanan Tangsunantham, Tanayoot Sangsuwan, and Chaiyod Pira. "Implementation of WiFi-based single phase smart meter for Internet of Things (IoT)." In 2017 International Electrical Engineering Congress (iEECON), pp. 1-4. IEEE, 2017.

Hardware Impaired Ambient Backscatter NOMA Systems: Reliability and Security

Xingwang Li¹, Senior Member, IEEE, Mengle Zhao², Student Member, IEEE,
 Ming Zeng³, Member, IEEE, Shahid Mumtaz⁴, Senior Member, IEEE,
 Varun G. Menon⁵, Senior Member, IEEE, Zhiguo Ding⁶, Fellow, IEEE,
 and Octavia A. Dobre⁷, Fellow, IEEE

Abstract—Non-orthogonal multiple access (NOMA) and ambient backscatter communication have been envisioned as two promising technologies for the Internet-of-things due to their high spectral efficiency and energy efficiency. Motivated by this fact, we consider an ambient backscatter NOMA system in the presence of a malicious eavesdropper. Under the realistic assumptions of residual hardware impairments (RHIs), channel estimation errors (CEEs) and imperfect successive interference cancellation (ipSIC), we investigate the physical layer security (PLS) of the ambient backscatter NOMA systems with emphasis on reliability and security. In order to further improve the security of the considered system, an artificial noise scheme is proposed where the radio frequency (RF) source acts as a jammer that transmits interference signals to the legitimate receivers and eavesdropper. On this basis, the analytical expressions for the outage probability (OP) and the intercept probability (IP) are derived. To gain more insights, the asymptotic analysis and corresponding diversity orders for the OP in the high signal-to-noise ratio (SNR) regime are carried out, and the asymptotic behaviors of the IP in the high main-to-eavesdropper ratio (MER) region are explored as well. Finally, the correctness of the theoretical analysis is verified by the Monte Carlo simulation

results. These results show that compared with the non-ideal conditions, the reliability of the considered system is high under ideal conditions, but the security is low.

Index Terms—Ambient backscatter, artificial noise, channel estimation errors, Internet-of-Things, imperfect successive interference cancellation, NOMA, physical layer security, residual hardware impairments.

I. INTRODUCTION

A LARGE number of intelligent devices will be supported for the wireless networks with Internet-of-things (IoT) and massive machine-type communication [1], [2]. To this end, non-orthogonal multiple access (NOMA) has been identified as a promising solution to serve massive connections due to high spectral efficiency and low latency [3], [4].¹ The distinguishing feature of NOMA is that a plurality of users are allowed to share the same time/frequency/code resources by power multiplexing through superposition coding [5]. At the receiver, the signals can be extracted with the aid of successive interference cancellation (SIC) [6]. From the perspective of coverage, NOMA can enhance the performance of the cell edge users by allocating more power to them [7].

On a parallel avenue, backscatter communication has emerged as a promising paradigm for the green sustainable IoT applications due to its ultralow-power and low cost [8]. A well-known backscatter communication application for the IoT is radio frequency identification (RFID) that consists of one reader and one tag. More exactly, the tag modulates and reflects the incident signal from the source through a mismatched antenna impedance to passively transmit information, and the reader performs demodulation after receiving the reflected signal [9]. However, the traditional backscatter communication technology is limited by the power consumption resulting from the active transmission [10]. To tackle this limitation, the work in [11] proposed ambient backscatter prototypes. This technology utilizes environmental wireless signals (e.g., digital TV broadcasting or cellular signals) to collect energy and transmit information through battery-free tags.

Ambient backscatter technology has drawn great attention from both academia and industry [12]–[16]. A framework for evaluating the ultimate achievable rates of point-to-point networks with ambient backscatter devices was proposed

¹Generally, NOMA can be divided into code-domain NOMA and power-domain NOMA. In this article, we use NOMA to refer to the power-domain NOMA.

Manuscript received July 30, 2020; revised November 26, 2020; accepted January 4, 2021. Date of publication January 11, 2021; date of current version April 16, 2021. The work of Xingwang Li was supported by the Key Scientific Research Projects of Higher Education Institutions in Henan Province under Grant 20A510007, the Outstanding Youth Science Foundation of Henan Polytechnic University under Grant J2019-4, the Natural Science Foundation of China under Grant 61901367 and 62001320; The work of Octavia A. Dobre was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), through its Discovery program. The associate editor coordinating the review of this article and approving it for publication was D. B. Da Costa. (Corresponding author: Xingwang Li.)

Xingwang Li and Mengle Zhao are with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China (e-mail: lixingwangbupt@gmail.com; zhaomenglephu@163.com).

Ming Zeng is with the Department of Electrical Engineering and Computer Engineering, Université Laval, Quebec, QC G1V 0A6, Canada (e-mail: ming.zeng@gel.ulaval.ca).

Shahid Mumtaz is with the Institute of Telecommunications, 3810078 Aveiro, Portugal, and also with the ARIES Research Center, Universidad Antonio de Nebrija, E-28040 Madrid, Spain (e-mail: smumtaz@av.it.pt).

Varun G. Menon is with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India (e-mail: varunmenon@ieee.org).

Zhiguo Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, U.K. (e-mail: zhiguo.ding@manchester.ac.uk).

Octavia A. Dobre is with Memorial University, St. John's, NL A1B 3X9, Canada (e-mail: odobre@mun.ca).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCOMM.2021.3050503>.

Digital Object Identifier 10.1109/TCOMM.2021.3050503

in [12], where the impact of the backscatter transmission on the performance of the legacy systems was considered. In [13], the authors analyzed the outage performance of the ambient backscatter communication systems with a pair of passive tag-reader by deriving the exact and asymptotic expressions for the outage probability (OP). Guo *et al.* in [14] exploited the NOMA technology to support massive tag connections. According to the unique characteristics of the cooperative ambient backscatter systems, the authors of [15] proposed three symbiotic transmission schemes, where the relationships between the primary and backscatter transmissions were commensal, parasitic, and competitive. The authors in [16] investigated the effects of co-channel interference and the energy harvesting (EH) on the achievable OP of the ambient backscatter communication systems with multiple backscatter links. In addition, the ambient backscatter technology has been widely used in smart phones, agriculture and other sectors [17], [18]. The authors in [17] developed a new type of environmental leaf sensor tag in agricultural applications using the ambient backscatter technology. In [19], the authors investigated a hybrid device-to-device (D2D) communication paradigm through integrating the ambient backscatter technology into wireless powered communication network (WPCNs) to improve the performance of current WPCNs, as well as the performance of the hybrid D2D communication system; it was demonstrated that the ambient backscatter technology can be well combined with current or future wireless communications to improve network performance.

Another well-known fact is that the transmission of wireless signals is vulnerable to fronted threats due to the broadcast nature of wireless communication environments. The traditional key encryption technologies has high computation complexity, and thus, are not suitable for small-volume backscatter devices with limited storage and computing power [20]. As a result, they may not be applied for solving the security communication problem of the ambient backscatter NOMA systems [21].

As an alternative, physical layer security (PLS) has been proposed as a promising mechanism to enhance the security of wireless communication systems from an information theoretic perspective [22], [23]. By exploiting the inherent random characteristics of wireless channels, PLS can achieve secure communication for wireless communication networks without being eavesdropped by illegal eavesdroppers, which has sparked a great deal of research interests, e.g., see [24]–[28] and the references therein. In [24], the secrecy outage performance of a multi-relay NOMA network was investigated, where three relay selection schemes were proposed. With the emphasis on the cognitive radio networks (CRNs), the authors of [25] evaluated the reliability-security tradeoff by deriving the connection outage probability and the secrecy outage probability for the cooperative NOMA aided CRNs. Additionally, the secrecy rate was studied under the traditional backscatter communications systems in [26], where the reader and eavesdropper were equipped with multiple antennas. To enhance the security of the ambient backscatter communication systems, an optimal tag selection scheme for the multi-tag ambient backscatter systems was designed in [27]. By the virtue

of artificial noise, an enhanced PLS scheme for multi-tag ambient backscatter system was designed, in which the bit error rate and secrecy rate were investigated in [28]. Moreover, the authors of [29] proposed to combined multiple-input multiple-output technology with artificial noise technology to enhance the secrecy performance of NOMA systems.

Unfortunately, the common feature of the aforementioned contributions is that perfect radio frequency (RF) components are assumed, which may not be realistic in practical communication systems. In practice, all RF front-ends are vulnerable to several types of hardware impairments, such as amplifier non-linearities, in-phase/quadrature imbalance, phase noise, and quantization error [30]–[34]. These impairments can be generally eliminated by using some compensation and calibration algorithms. However, owing to estimation errors, inaccurate calibration, and time-varying hardware characteristics, there are still some residual hardware impairments (RHIs), which can be modeled as an additive distortion noise to the transmitted/received signals [35]. To this end, a great deal of works have studied the impact of RHIs on system performance [36], [37]. In [36], the authors investigated the effects of RHIs on the achievable sum rate of the unmanned aerial vehicle-aided NOMA relaying networks. Considering two types of relay selection schemes, the impact of RHIs on the multiple-relay amplify-and-forward (AF) network was studied by deriving the tight closed-form expressions for the OP [37].

Moreover, another limitation of the above research works is that perfect channel state information (CSI) is assumed available at receivers, which is not practical. In fact, it is a great challenge to obtain perfect channel knowledge due to channel estimation errors (CEEs) and feedback delay [38]. The related research works about imperfect CSI have been reported in [39]–[41]. The outage performance of the down-link cooperative NOMA systems based on wireless backhaul unreliability and imperfect CSI was studied by deriving the exact and asymptotic OP expressions at the receivers [39]. A proportional fair scheduling algorithm was proposed to achieved high throughput and fairness, which was extended to the multi-user NOMA scenarios with imperfect CSI in [40]. The authors of [41] considered a more practical scenarios, where the outage performance of the AF relay systems was analyzed in the presence of RHIs and CEEs. Therefore, it is of high practical relevance to look into the realistic scenario with imperfect CSI and RHIs.

A. Motivation and Contribution

Ambient backscatter communication has been identified as a cutting-edge technology, which can support communication using ambient RF signal without requiring active RF transmission. However, several major challenges related to interference, security and hardware impairments need to be addressed to implement such networks, which motivates this study. Specifically, the joint effects of RHIs, CEEs and imperfect SIC (ipSIC) on the secure performance of the ambient backscatter NOMA systems have not yet been well investigated. To fill this gap, this article makes an in-depth study of the joint effects of the three non-ideal factors on the

reliability and the security of the ambient backscatter NOMA systems. In order to improve the security, we propose an artificial noise scheme, where the RF source sends the signal and artificial noise simultaneously. This scheme is feasible since it is carried out without changing the original system framework [42], [43]. Specifically, the analytical expressions for the OP and the intercept probability (IP) are derived for the far reader, the near reader and the tag under ideal and non-ideal conditions, respectively. To obtain more insights, the asymptotic behaviors for the OP in the high signal-to-noise ratio (SNR) regime and the asymptotic behaviors for the IP in the high main-to-eavesdropper ratio (MER) region are explored. The essential contributions of this article are summarized as follows:

- We consider a novel secure framework for the ambient backscatter NOMA systems in the presence of RHIs, CEEs, and ipSIC. To improve secure performance, an artificial noise scheme is designed.
- We derive the analytical expressions for the OP and the IP the far reader, the near reader and the tag under ideal and non-ideal conditions to evaluate the reliability and the security. The results show that a smaller power coefficient of artificial noise or a larger interfering factor of readers can enhance the impact of artificial noise on balancing the reliability-security trade-off.
- In order to obtain deeper insights, we carry out the asymptotic analysis for the OP in the high SNR region as well as the diversity orders under ideal and non-ideal conditions. Moreover, the asymptotic behaviors of the IP in the high MER regime are explored by introducing the MER. The obtained results indicate that there are error floors for the OP due to the CEEs and the reflection coefficient.

B. Organization and Notations

The remainder of this article is organised as follows. In Section II, we introduce the ambient backscatter NOMA model. In Section III, the reliability is investigated by deriving the analytical and asymptotic expressions for the OP, while the expressions of IP are derived to analyze the security. In Section IV, some numerical results are provided to validate the correctness of the theoretical analysis. Section V concludes this article and summarizes key findings.

We use $E\{\cdot\}$ to denote the expectation operation. A complex Gaussian random variable with mean μ and variance σ^2 reads as $\mathcal{CN}\{\mu, \sigma^2\}$. $\Pr\{\cdot\}$ denotes the probability and $Kv(\cdot)$ represents the v -th order modified Bessel function of the second kind, while $n!$ denotes the factorial operation. Finally, $f_X(\cdot)$ and $F_X(\cdot)$ are the probability density function (PDF) and the cumulative distribution function (CDF) of a random variable, respectively.

II. SYSTEM MODEL

As illustrated in Fig. 1, we consider a downlink ambient backscatter NOMA system, which consists of one ambient RF source (S), one tag (T), two readers (R_f , R_n) and one eavesdropper (E). In this study, S transmits the signal to

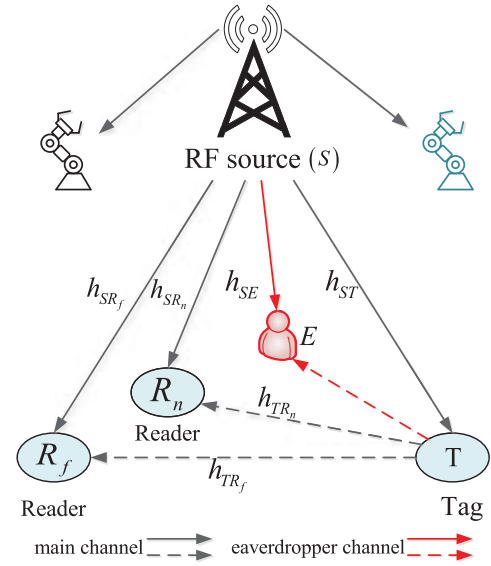


Fig. 1. Ambient backscatter NOMA system model.

readers and tag in the same time/frequency resource block. Meanwhile, T , acting as the backscatter device, can transmit its own information to the readers by reflecting the signals from S , whereas E can intercept the signal intended for readers. We consider the following assumptions: i) All the nodes are equipped with a single antenna; ii) For convenience, RHIs exist at S , readers and E but not at the tag; iii) All links are subject to Rayleigh fading.

Under practical considerations, the perfect CSI may be unavailable due to some CEEs. The common way to obtain CSI is channel estimation. For this purpose, by adopting linear minimum mean square error (MMSE), the channel can be modeled as $h_{AB} = \hat{h}_{AB} + e_{AB}$ [44], $AB = \{Si, ST, Ti\}$, $i \in (R_f, R_n, E)$, where \hat{h}_{AB} is the estimated channel of h_{AB} , and $e_{AB} \sim \mathcal{CN}(0, \sigma_{e_{AB}}^2)$ denotes the corresponding channel estimation errors which can be modeled as a Gaussian random variable, where the variance of CEE $\sigma_{e_{AB}}^2$ indicates the quality of CSI².

To improve the security communication of ambient backscatter NOMA systems, we consider injecting artificial noise $z(t)$ with $E(|z(t)|^2) = 1$ at S . Due to the CEEs, the artificial noise will cause interference to the readers and eavesdropper. Then, the superposition message at S can be written as

$$x_s = \sqrt{a_1 P_s} x_1 + \sqrt{a_2 P_s} x_2 + \sqrt{P_J} z(t), \quad (1)$$

where P_s is the transmit power for the desired signals at S ; a_1 and a_2 are the power allocation coefficients for the near reader and the far reader with $a_1 + a_2 = 1$ and $a_1 < a_2$, respectively; x_1 and x_2 are the corresponding transmitted signals of R_n and R_f with $E(|x_1|^2) = E(|x_2|^2) = 1$; P_J is

²In reality, $\sigma_{e_{AB}}^2$ is a function of the average SNR, e.g., $\sigma_{e_{AB}} \propto 1/(1 + \gamma)$. $AB = \{Si, ST, Ti\}$, $i \in (R_f, R_n, E)$, where γ represents the transmit SNR at S . For the convenience of mathematical calculations, we have used this simplified but fairly realistic channel estimation error model that is widely used in academia [45].

the transmitted power of the artificial noise with $P_J = \varphi_J P_S$, with $\varphi_J \in (0, 1]$ as the power coefficient of artificial noise.

Then, T backscatters the S signal to R_f , R_n and E with its own signal $c(t)$, with $E(|c(t)|^2) = 1$ [46]. Therefore, R_f and R_n receive the signals from S and the backscattered from T ; E can intercept the signals from S and the backscattered from T . Considering the RHIs and CEEs, the received signals at i ($i \in \{R_f, R_n, E\}$) can be expressed as

$$y_i = \beta h_{Ti} h_{ST} (x_s c(t) + \eta_{Si}) + h_{Si} (x_s + \eta_{Si}) + n_i, \quad (2)$$

where β is a complex reflection coefficient used to normalize $c(t)$; $n_i \sim \mathcal{CN}(0, N_0)$ is the complex additive white Gaussian noise (AWGN); $\eta_{Si} \sim \mathcal{CN}(0, \kappa_{Si}^2 P_S)$ can be modelled by a Gaussian random variable. This has been supported and validated by theoretical investigations and measurements [47],³ κ_{Si} denotes the level of hardware impairment at transceivers, which can be measured in practice based on the error vector magnitude (EVM) [49]; h_{Si} , h_{Ti} and h_{ST} are the channel coefficients $S \rightarrow i$, $T \rightarrow i$ and $S \rightarrow T$, respectively.

According to the NOMA protocol, R_f can decode the signals x_2 , and R_n and E can decode the signals x_2 , x_1 and $c(t)$ in turn with the aid of SIC. In addition, the readers can only eliminate part of the interference due to the presence of CEEs. Then, the received signal-to-interference-plus-noise ratio (SINR) of i ($i \in \{R_n, R_f, E\}$) can be given as⁴

$$\gamma_i^{x_2} = \frac{|\hat{h}_{Si}|^2 a_2 \gamma}{\gamma \left[|\hat{h}_{ST}|^2 \left(B_i |\hat{h}_{Ti}|^2 + M_i \right) + C_i |\hat{h}_{Ti}|^2 + Q_i |\hat{h}_{Si}|^2 + \psi_i \right] + 1}, \quad (3)$$

$$\gamma_i^{x_1} = \frac{|\hat{h}_{Si}|^2 a_1 \gamma}{\gamma \left[|\hat{h}_{ST}|^2 \left(B_i |\hat{h}_{Ti}|^2 + M_i \right) + C_i |\hat{h}_{Ti}|^2 + O_i |\hat{h}_{Si}|^2 + \psi_i \right] + 1}, \quad (4)$$

$$\gamma_i^{c(t)} = \frac{\beta^2 |\hat{h}_{Ti}|^2 |\hat{h}_{ST}|^2 \gamma}{\gamma \left[|\hat{h}_{ST}|^2 \left(m_i |\hat{h}_{Ti}|^2 + M_i \right) + C_i |\hat{h}_{Ti}|^2 + \xi_i |\hat{h}_{Si}|^2 + \psi_i \right] + 1}, \quad (5)$$

where $\gamma = P_S/N_0$ represents the transmit SNR at S ; ε is the parameter of ipSIC; $B_{R_f} = \beta^2 (1 + \varpi \varphi_J + \kappa_{S R_f}^2)$, $C_{R_f} = B_{R_f} \sigma_{e_{ST}}^2$, $M_{R_f} = B_{R_f} \sigma_{e_{TR_f}}^2$, $Q_{R_f} = a_1 + \varpi \varphi_J + \kappa_{S R_f}^2$, $\psi_{R_f} = B_{R_f} \sigma_{e_{TR_f}}^2 \sigma_{e_{ST}}^2 + \sigma_{e_{S R_f}}^2 (1 + \varpi \varphi_J + \kappa_{S R_f}^2)$; ϖ is the interference factor, reflecting the degree of interference of the artificial noise to the readers, with $0 \leq \varpi \leq 1$; $B_{R_n} = \beta^2 (1 + \varpi \varphi_J + \kappa_{S R_n}^2)$, $C_{R_n} = B_{R_n} \sigma_{e_{ST}}^2$, $M_{R_n} = B_{R_n} \sigma_{e_{TR_n}}^2$, $Q_{R_n} = a_1 + \varpi \varphi_J + \kappa_{S R_n}^2$, $\psi_{R_n} = B_{R_n} \sigma_{e_{TR_n}}^2$

³It is worth noting that when the compensation algorithms are applied to mitigate hardware impairments [48], the Gaussian model is particularly applicable to residual distortion.

⁴It should be pointed out that R_f only needs to decode its own signal x_2 , that is, the SINR of R_f is $\gamma_{R_f}^{x_2}$.

$$\sigma_{e_{ST}}^2 + \sigma_{e_{S R_n}}^2 (1 + \varpi \varphi_J + \kappa_{S R_n}^2), O_{R_n} = \varepsilon a_2 + \varpi \varphi_J + \kappa_{S R_n}^2, m_{R_n} = \beta^2 (\kappa_{S R_n}^2 + \varpi \varphi_J), \xi_{R_n} = \varepsilon + \varpi \varphi_J + \kappa_{S R_n}^2; B_E = \beta^2 (1 + \varphi_J + \kappa_{S E}^2), C_E = B_E \sigma_{e_{ST}}^2, M_E = B_E \sigma_{e_{TE}}^2, Q_E = a_1 + \varphi_J + \kappa_{S E}^2, \psi_E = B_E \sigma_{e_{TE}}^2 \sigma_{e_{ST}}^2 + \sigma_{e_{SE}}^2 (1 + \varphi_J + \kappa_{S E}^2), O_E = \varepsilon a_2 + \varphi_J + \kappa_{S E}^2, m_E = \beta^2 (\kappa_{S E}^2 + \varphi_J), \xi_E = \varepsilon + \varphi_J + \kappa_{S E}^2.$$

III. PERFORMANCE ANALYSIS

In this section, we investigate the reliability and security of the ambient backscatter NOMA systems in term of OP and IP. In addition, the asymptotic OP and diversity orders in the high SNR regions are examined, as well as the asymptotic IP in the high MER regime. In order to facilitate comparison, we discuss the ideal and non-ideal situations in this section.

A. Ideal RF ($\kappa = 0$, $\sigma_e^2 = 0$)

1) OP Analysis

OP for R_f : The outage event occurs at R_f when R_f cannot successfully decode x_2 . Thus, the OP at R_f can be expressed as

$$P_{out}^{R_f} = 1 - \Pr \left(\gamma_{R_f}^{x_2} > \gamma_{th2}^{R_f} \right), \quad (6)$$

where $\gamma_{th2}^{R_f}$ is the target rate of R_f .

Theorem 1: For Rayleigh fading channels, the analytical expression for the OP of the far reader under ideal conditions can be obtained as

$$P_{out}^{R_f, id} = 1 + \Delta_1^{R_f} e^{-\frac{\Delta_1^{R_f} \gamma_{th2}^{R_f}}{\lambda_{S R_f} \gamma (a_2 - Q_{R_f} \gamma_{th2}^{R_f})}} \text{Ei} \left(-\Delta_1^{R_f} \right), \quad (7)$$

where $\Delta_1^i = \frac{\lambda_{S i} (a_2 - Q_i \gamma_{th2}^i)}{\lambda_{S T} \lambda_{T i} B_i \gamma_{th2}^i}$, ($i \in \{R_n, R_f, E\}$), and $\text{Ei}(p)$ is the exponential integral function [50] expressed by

$$\text{Ei}(p) = \frac{(-p)^{i-1}}{(i-1)!} [-\ln p + \psi(i)] - \sum_{m=0}^{\infty} \frac{(-p)^m}{(m-i+1)m!}, \quad (8)$$

with

$$\begin{cases} \psi(1) = -v \\ \psi(i) = -v + \sum_{m=1}^{i-1} \frac{1}{m} \quad i > 1, \end{cases} \quad (9)$$

where $v \approx 0.577$ is the Euler constant.

Proof: See Appendix A. ■

Corollary 1: At high SNRs, the asymptotic expression for the OP of R_f of the ambient backscatter NOMA systems under ideal conditions is given as

$$P_{out, \infty}^{R_f, id} = 1 + \Delta_1^{R_f} e^{-\Delta_1^{R_f}} \text{Ei} \left(-\Delta_1^{R_f} \right). \quad (10)$$

OP for R_n : To successfully decode x_1 at R_n , two conditions are needed to be met simultaneously: 1) R_n can successfully decode x_2 ; 2) R_n can successfully decode its own information x_1 . Therefore, the OP of R_n can be expressed as

$$P_{out}^{R_n} = 1 - \Pr \left(\gamma_{R_n}^{x_2} > \gamma_{th2}^{R_n}, \gamma_{R_n}^{x_1} > \gamma_{th1}^{R_n} \right), \quad (11)$$

where $\gamma_{th1}^{R_n}$ is the target rate of R_n .

Theorem 2: For Rayleigh fading channels, the analytical expression for the OP of the near reader under ideal conditions can be obtained as

$$P_{out}^{R_n, id} = 1 + \frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}} e^{-\left(\frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma} + \frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}}\right)} \times \text{Ei}\left(-\frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}}\right), \quad (12)$$

where $\varsigma_{R_n} = \max\left\{\frac{\gamma_{th1}^{R_n}}{a_1 - O_{R_n} \gamma_{th1}^{R_n}}, \frac{\gamma_{th2}^{R_n}}{a_2 - Q_{R_n} \gamma_{th2}^{R_n}}\right\}$.

Proof: By substituting (3) and (4) into (11), we can obtain the result of (12) after some mathematical manipulations as in the proof of **Theorem 1**. ■

Corollary 2: At high SNRs, the asymptotic expression for the OP of R_n of the ambient backscatter NOMA systems is given as

$$P_{out, \infty}^{R_n, id} = 1 + \frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}} e^{-\frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}}} \times \text{Ei}\left(-\frac{\lambda_{SR_n}}{\lambda_{ST} \varsigma_{R_n} \lambda_{TR_n} B_{R_n}}\right). \quad (13)$$

OP for T: The T signals can be successfully decoded when x_2 and x_1 are perfectly decoded at R_n . Thus, the OP of BD can be expressed as

$$P_{out}^T = 1 - \text{Pr}\left(\gamma_{R_n}^{x_2} > \gamma_{th2}^{R_n}, \gamma_{R_n}^{x_1} > \gamma_{th1}^{R_n}, \gamma_{R_n}^{c(t)} > \gamma_{thc}^{R_n}\right), \quad (14)$$

where $\gamma_{thc}^{R_n}$ is the target rate for R_n decoding tag signals.

Theorem 3: For Rayleigh fading channels, the analytical expression for the OP of T under ideal conditions in (15), as shown at the bottom of the page.

In (15), $\vartheta_k = \cos[(2k-1)\pi/(2N)]$, N is an accuracy-complexity trade-off parameter, $\Delta_5^i = \beta^2 - m_i \gamma_{thc}^i$ ($i \in \{R_n, R_f, E\}$), $\Delta_9 = \frac{\lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \varsigma_{R_n} B_{R_n}}$, $\Delta_{10} = \frac{(\vartheta_k+1)\gamma_{thc}^{R_n}}{2\lambda_{TR_n} \lambda_{ST} \gamma \Delta_5^{R_n}}$, $\Delta_{11} = \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{TR_n} \lambda_{ST} \Delta_5^{R_n}}$, and $\Delta_{12} = \frac{(\vartheta_k+1)\gamma_{thc}^{R_n}}{2\lambda_{TR_n} \lambda_{ST} \gamma \Delta_5^{R_n}}$.

Proof: See Appendix B. ■

Corollary 3: At high SNRs, the asymptotic expression for the OP of T under ideal conditions of the ambient backscatter NOMA systems can be expressed as

$$P_{out, \infty}^{T, id} = 1 + \Delta_9 e^{\Delta_9} \text{Ei}(-\Delta_9) - \Delta_{11} e^{\Delta_{11}} \text{Ei}(-\Delta_{11}). \quad (16)$$

3) IP Analysis

User i ($i \in \{R_f, R_n, T\}$) will be intercepted if E can successfully wiretap j 's signal, i.e., $\gamma_E^p > \gamma_{thj}^E$, $p \in \{x_2, x_1, c(t)\}$, $j \in (2, 1, c)$. Thus, the IP of i by E can be expressed as

$$P_{int}^i = \text{Pr}\left(\gamma_E^p > \gamma_{thj}^E\right), \quad (17)$$

where γ_{thj}^E is the secrecy SNR threshold of i .

Theorem 4: The analytical expressions for the IP of the far reader and the near reader under ideal conditions can be respectively obtained as

$$P_{int}^{R_f, id} = -\Delta_1^E e^{\Delta_1^E - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_2 - Q_E \gamma_{th2}^E)}} \text{Ei}(-\Delta_1^E), \quad (18)$$

$$P_{int}^{R_n, id} = -\Delta_{16}^E e^{\Delta_{16}^E - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_1 - O_E \gamma_{th2}^E)}} \text{Ei}(-\Delta_{16}^E), \quad (19)$$

where $\Delta_{16}^E = \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}{\lambda_{ST} \lambda_{TE} B_E \gamma_{th2}^E}$.

For ideal conditions, the analytical expression for the IP of T in (20), as shown at the bottom of the page.

Proof: See Appendix C. ■

Moreover, for the further investigation of the ambient backscatter NOMA secure communication systems, we also study the asymptotic behaviors of IP in the high MER region [51]. MER is introduced to distinguish the channel state of the main link and eavesdropping link, being defined as $\lambda_{me} = \frac{\lambda_{ST}}{\lambda_{TE}}$.

Corollary 4: At high MERs, the asymptotic expression for the IP of R_f of the ambient backscatter NOMA systems under ideal conditions is given as

$$P_{int, \infty}^{R_f, id} = -\Delta_1'^E e^{-\frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_2 - Q_E \gamma_{th2}^E)}} (1 + b_1') \text{Ei}(-b_1'), \quad (21)$$

where $\Delta_1'^E = \frac{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E)}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}$, and $b_1' = \frac{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}$.

Proof: The proof follows by taking λ_{me} large in (21) and simplifying the expressions by utilizing $e^x \approx 1 + x$ if $x \rightarrow 0$. Similarly, we can also obtain (22), (38) and (39). ■

Corollary 5: At high MERs, the asymptotic expression for the IP of R_n of the ambient backscatter NOMA systems under

$$P_{out}^{T, id} = 1 + \Delta_9 e^{\Delta_9 - \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma}} \text{Ei}(-\Delta_9) + \frac{\gamma_{thc}^{R_n} \pi}{N \lambda_{TR_n} \lambda_{ST} \gamma \Delta_5^{R_n}} \sum_{k=0}^N e^{-\left(\varsigma_{R_n} B_{R_n} \Delta_{10} + \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma}\right)} K_0\left(2\sqrt{\Delta_{10}}\right) \sqrt{1 - \vartheta_k^2} - \Delta_{11} e^{\Delta_{11} + \frac{1}{\lambda_{SR_n} \gamma \xi_{R_n}}} \text{Ei}(-\Delta_{11}) - \frac{\gamma_{thc}^{R_n} \pi}{N \lambda_{TR_n} \lambda_{ST} \gamma \Delta_5^{R_n}} \sum_{k=0}^N e^{\frac{1}{\lambda_{SR_n} \gamma \xi_{R_n}} - \frac{\vartheta_k+1}{2\lambda_{SR_n} \gamma \xi_{R_n}}} K_0\left(2\sqrt{\Delta_{10}}\right) \sqrt{1 - \vartheta_k^2}. \quad (15)$$

$$P_{int}^{T, id} = 1 - \frac{\pi \gamma_{thc}^E}{N \lambda_{ST} \lambda_{TE} \gamma \Delta_5^E} \sum_{k=0}^N (\vartheta_k + 1) K_0\left((\vartheta_k + 1) \sqrt{\frac{\gamma_{thc}^E}{\lambda_{ST} \lambda_{TE} \gamma \Delta_5^E}}\right) \sqrt{1 - \vartheta_k^2} - \frac{2}{\lambda_{ST} \lambda_{TE}} e^{\frac{1}{\lambda_{SE} \xi_E \gamma}} \int_{\frac{\gamma_{thc}^E}{\Delta_5^E}}^{\infty} e^{-\frac{\Delta_5^E y}{\lambda_{SE} \xi_E \gamma_{thc}^E}} K_0\left(2\sqrt{\frac{y}{\lambda_{ST} \lambda_{TE}}}\right) dy. \quad (20)$$

ideal conditions is given as

$$P_{int,\infty}^{R_n,id} = -\Delta_{16}' e^{-\frac{\gamma_{th2}^E}{\lambda_{SE}\gamma(a_1-O_E\gamma_{th2}^E)}} (1+b_2') \text{Ei}(-b_2'), \quad (22)$$

where $\Delta_{16}' = \frac{\lambda_{SE}(a_1-O_E\gamma_{th2}^E)}{\lambda_{me}\lambda_{TE}^2 B_E\gamma_{th2}^E}$, and $b_2' = \frac{\lambda_{SE}(a_1-O_E\gamma_{th2}^E)}{\lambda_{me}\lambda_{TE}^2 B_E\gamma_{th2}^E}$.

Corollary 6: At high MERs, the asymptotic expression for the IP of T of the ambient backscatter NOMA systems under ideal conditions in (23), as shown at the bottom of the page.

Proof: The proof follows by taking λ_{me} large in (23) and simplifying the expressions by utilizing $e^{-x} \approx 1-x$ and $K_0(x) \approx -\ln(x)$ if $x \rightarrow 0$. ■

B. Non-Ideal RF ($\kappa \neq 0$, $\sigma_e^2 \neq 0$)

1) OP Analysis

OP for R_f : According to the definition of OP in (6), we can obtain Theorem 5.

Theorem 5: For Rayleigh fading channels, the analytical expression for the OP of the far reader can be obtained as

$$P_{out}^{R_f,ni} = 1 + \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f} - \frac{\gamma_{th2}^{R_f}}{\lambda_{SR_f}\gamma(a_2-Q_{R_f}\gamma_{th2}^{R_f})}} \text{Ei}(-\Delta_2^{R_f}), \quad (24)$$

where $\Delta_2^i = \left(\frac{M_i\gamma_{th2}^i}{\lambda_{S_i}(a_2-Q_i\gamma_{th2}^i)} + \frac{1}{\lambda_{ST}} \right) \frac{\lambda_{S_i}(a_2-Q_i\gamma_{th2}^i) + \lambda_{T_i}C_i\gamma_{th2}^i}{\lambda_{T_i}B_i\gamma_{th2}^i}$,

$\Delta_3^i = \frac{\psi_i\gamma_{th2}^i}{\lambda_{S_i}(a_2-Q_i\gamma_{th2}^i)}$, ($i \in \{R_n, R_f, E\}$).

Proof: See Appendix A. ■

Corollary 7: At high SNRs, the asymptotic expression for the OP of R_f of the ambient backscatter NOMA systems is given as

$$P_{out,\infty}^{R_f,ni} = 1 + \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f}} \text{Ei}(-\Delta_2^{R_f}). \quad (25)$$

OP for R_n : According to the definition of OP in (11), we can obtain Theorem 6.

Theorem 6: For Rayleigh fading channels, the analytical expression for the OP of the near reader can be obtained as

$$P_{out}^{R_n,ni} = 1 + \frac{\lambda_{SR_n}}{\lambda_{ST}\varsigma_{R_n}\lambda_{TR_n}B_{R_n}} e^{-\left(\frac{\varsigma_{R_n}}{\lambda_{SR_n}\gamma} + \Delta_4^{R_n}\right)} \text{Ei}(-\Delta_4^{R_n}), \quad (26)$$

where $\Delta_4^i = \frac{(\lambda_{ST}\varsigma_i M_i + \lambda_{S_i})(\lambda_{S_i} + \varsigma_i \lambda_{T_i} C_i)}{\lambda_{S_i} \lambda_{ST} \varsigma_i \lambda_{T_i} B_i} + \frac{\varsigma_i \psi_i}{\lambda_{S_i}}$, ($i \in \{R_n, R_f, E\}$).

Proof: By substituting (3) and (4) into (11), we can obtain the result of (26) after some mathematical manipulations, as in the proof of **Theorem 1**. ■

Corollary 8: At high SNRs, the asymptotic expression for the OP of R_n of the ambient backscatter NOMA systems is given as

$$P_{out,\infty}^{R_n,ni} = 1 + \frac{\lambda_{SR_n}}{\lambda_{ST}\varsigma_{R_n}\lambda_{TR_n}B_{R_n}} e^{-\Delta_4^{R_n}} \text{Ei}(-\Delta_4^{R_n}). \quad (27)$$

OP for T : According to the definition of OP in (14), we can obtain Theorem 7. ■

$$P_{int,\infty}^{T,id} = 1 + \frac{\pi\gamma_{thc}^E}{N\lambda_{me}\lambda_{TE}^2\gamma\Delta_5^E} \sum_{k=0}^N (\vartheta_k + 1) \ln \left(\frac{\vartheta_k + 1}{2} \sqrt{\frac{\gamma_{thc}^E}{\lambda_{me}\lambda_{TE}^2\gamma\Delta_5^E}} \right) \sqrt{1 - \vartheta_k^2} + \frac{2}{\lambda_{me}\lambda_{TE}^2} e^{\frac{1}{\lambda_{SE}\xi E\gamma}} \int_{\frac{\gamma_{thc}^E}{\Delta_5^E}}^{\infty} e^{-\frac{\Delta_5^E y}{\lambda_{SE}\xi E\gamma_{thc}^E}} \ln \left(\sqrt{\frac{y}{\lambda_{me}\lambda_{TE}^2}} \right) dy. \quad (23)$$

$$P_{out}^{T,ni} = 1 - \frac{2\lambda_{SR_n}}{\lambda_{TR_n}\lambda_{ST}\varsigma_{R_n}B_{R_n}} e^{-\left(B_5 + \frac{\lambda_{TR_n}\varsigma_{R_n}B_{R_n}\gamma_{thc} + \varsigma_{R_n}}{\lambda_{SR_n}\lambda_{TR_n}\gamma\Delta_5} + \frac{\varsigma_{R_n}}{\lambda_{SR_n}\gamma}\right)} \sum_{v=1}^{\infty} (-1)^v \frac{1}{B_4^v} \left(\frac{(B_3 + \Delta_6)}{B_1} \right)^{\frac{v}{2}} K_v \left(2\sqrt{(B_3 + \Delta_6)B_1} \right) + \frac{\lambda_{SR_n}\xi_{R_n}\gamma_{thc}^{R_n}}{\lambda_{ST}\lambda_{TR_n}\Delta_5^{R_n}} e^{A_2^{R_n}} \left(\frac{\pi}{N} \sum_{k=0}^N e^{-\left(\frac{2(A_3^{R_n} + \Delta_8^{R_n})}{A_4^{R_n}(\vartheta_k + 1)} - \frac{A_1^{R_n}A_4^{R_n}(\vartheta_k + 1)}{2}\right)} \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} - \frac{1}{\vartheta_k + 1} \right) + 2K_0 \left(2\sqrt{-A_1^{R_n}(A_3^{R_n} + \Delta_8^{R_n})} \right) \right). \quad (28)$$

$$B_2 = \frac{2\varsigma_{R_n}B_{R_n}M_{R_n}C_{R_n}(\gamma_{thc}^{R_n})^2}{\lambda_{SR_n}(\Delta_5^{R_n})^2} + \frac{(\lambda_{BD_n}\varsigma_{R_n}C_{R_n}M_{R_n} + \lambda_{SR_n}M_{R_n} + \lambda_{BR_n}\varsigma_{R_n}B_{R_n}\psi_{R_n})\gamma_{thc}^{R_n}}{\lambda_{SR_n}\lambda_{BR_n}\Delta_5^{R_n}}, \quad (29)$$

$$B_3 = \left[\lambda_{TR_n}\varsigma_{R_n}M_{R_n}C_{R_n}^2\gamma(\gamma_{thc}^{R_n})^2 (B_{R_n}\gamma_{thc}^{R_n} + \Delta_5^{R_n}) + (\lambda_{SR_n}M_{R_n} + \lambda_{TR_n}\varsigma_{R_n}B_{R_n}\psi_{R_n})C_{R_n}\gamma(\Delta_5^{R_n}\gamma_{thc}^{R_n})^2 \right] / (\Delta_5^{R_n})^2 + (\lambda_{TR_n}\varsigma_{R_n}C_{R_n} + \lambda_{SR_n})\psi_{R_n}\gamma\gamma_{thc}^{R_n}, \quad (30)$$

$$B_4 = \frac{\varsigma_{R_n}\lambda_{SR_n}\lambda_{BR_n}C_{R_n}\gamma(B_{R_n}\gamma_{thc}^{R_n} + \Delta_5^{R_n}) + \lambda_{SR_n}^2\Delta_5^{R_n}\gamma}{\varsigma_{R_n}B_{R_n}}. \quad (31)$$

Theorem 7: For Rayleigh fading channels, the analytical expression for the OP of T in (28), as shown at the bottom of the previous page..

$$\begin{aligned} \text{In (28), } \Delta_6 &= \frac{\lambda_{TR_n} \zeta_{R_n} C_{R_n} \gamma_{thc}^{R_n} (B_{R_n} \gamma_{thc}^{R_n} + \Delta_5) + \lambda_{SR_n} \gamma_{thc}^{R_n} \Delta_5^{R_n}}{\Delta_5^{R_n} \gamma}, \\ \Delta_7^i &= (\lambda_{Si} \xi_i - \lambda_{Ti} C_i) \gamma_{thc}^i, \quad \Delta_8^i = (\lambda_{Ti} C_i \gamma_{thc}^i + \Delta_7^i) \gamma, \\ A_1^i &= \frac{-1}{\lambda_{Si} \xi_i \lambda_{ST} \lambda_{TE} \Delta_5^i \gamma^2}, \quad A_2^i = -\left(\frac{C_i \gamma_{thc}^i}{\lambda_{ST} \Delta_5^i} + \frac{M_i \gamma_{thc}^i}{\lambda_{TE} \Delta_5^i} \right), \\ A_3^i &= \frac{\lambda_{Ti} M_i (C_i \gamma_{thc}^i)^2 + (\lambda_{Ti} \psi_i \Delta_5^i + \Delta_7^i M_i) C_i \gamma_{thc}^i}{\Delta_5^i} + \Delta_7^i \psi_i \gamma^2, \\ A_4^i &= \lambda_{Si}^2 \xi_i^2 \gamma_{thc}^i, \quad B_1 = \frac{\lambda_{SR_n} + \lambda_{ST} \zeta_{R_n} M_{R_n}}{\lambda_{SR_n}^2 \lambda_{ST} \lambda_{TE} R_n \gamma \Delta_5^{R_n}} + \frac{\zeta_{R_n} C_{R_n} B_{R_n} \gamma_{thc}^{R_n}}{\lambda_{SR_n}^2 \lambda_{TE} R_n (\Delta_5^{R_n})^2}, \\ (i \in \{R_n, R_f, E\}), \quad B_5 &= \frac{(\lambda_{ST} \zeta_{R_n} C_{R_n} + \lambda_{SR_n}) C_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{SR_n} \Delta_5^{R_n}} + B_2 + \frac{\zeta_{R_n} \psi_{R_n}}{\lambda_{SR_n}}, \end{aligned}$$

where B_2 , B_3 and B_4 are as shown at the bottom of the previous page.

Proof: See Appendix B. ■

Corollary 9: At high SNRs, the asymptotic expressions for the OP of T of the ambient backscatter NOMA systems in (32), as shown at the bottom of the page.

Then, in order to obtain more insights, the diversity orders for R_f , R_n and T are investigated, which can be defined as [52]:

$$d = - \lim_{\gamma \rightarrow \infty} \frac{\log(P_{out}^\infty)}{\log \gamma}. \quad (33)$$

Corollary 10: The diversity orders of R_f , R_n and T are given as:

$$d_{R_f}^{id} = d_{R_f}^{ni} = d_{R_n}^{id} = d_{R_n}^{ni} = d_T^{id} = d_T^{ni} = 0. \quad (34)$$

Remark 1: From **Corollaries 1-3** and **Corollaries 7-10**, for both ideal and non-ideal conditions we can observe that: 1) RHIs, CEEs and ipSIC have detrimental effects on the reliability of the considered systems; 2) From (3), (4) and (5), we can see that the increase of SNR leads to higher received SINR, and therefore lowers OP for R_f , R_n and T ; 3) When the transmit SNR goes to infinity, the received SINR also grows to infinity, while the asymptotic outage performance of the R_f , R_n and T becomes a constant, indicating that there are error floors for the OP; 4) From (34), it can be observed that the diversity orders of the considered system are zero due to the fixed constant for the OP in the high SNR regime.

2) **IP Analysis:** According to the definition of IP in (17), we can obtain Theorem 8.

Theorem 8: The analytical expressions for the IP of the far reader and the near reader can be respectively obtained as

$$P_{int}^{R_f, ni} = -\Delta_1^E e^{\Delta_2^E - \Delta_3^E - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_2 - Q_E \gamma_{th2}^E)}} \text{Ei}(-\Delta_2^E), \quad (35)$$

$$P_{int}^{R_n, ni} = -\Delta_{16} e^{\Delta_{17} - \Delta_{18} - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_1 - O_E \gamma_{th2}^E)}} \text{Ei}(-\Delta_{17}), \quad (36)$$

where $\Delta_{17} = \left(\frac{M_E \gamma_{th2}}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)} + \frac{1}{\lambda_{ST}} \right) \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{TE} B_E \gamma_{th2}^E}$,

$$\Delta_{16} = \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}{\lambda_{ST} \lambda_{TE} B_E \gamma_{th2}^E}, \quad \text{and } \Delta_{18} = \frac{\psi_E \gamma_{th2}^E}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}.$$

For non-ideal conditions, the analytical expression for the IP of T in (37), as shown at the bottom of the page.

$$\text{In (27), } \Delta_{13} = 1 / (\lambda_{ST} \lambda_{TE} \Delta_5^E), \quad \Delta_{14} = \frac{C_E \gamma_{thc}^E (\lambda_{ST} + \lambda_{TE})}{\lambda_{TE} \lambda_{ST} \Delta_5^E}, \quad \text{and } \Delta_{15} = \frac{C_E^2 (\gamma_{thc}^E)^2}{(\Delta_5^E)^2 \lambda_{TE}} + \left(\psi_E + \frac{1}{\gamma} \right) \gamma_{thc}^E.$$

Proof: See Appendix C. ■

Corollary 11: At high MERs, the asymptotic expression for the IP of R_f of the ambient backscatter NOMA systems is given as

$$P_{int, \infty}^{R_f, ni} = -\Delta_1' e^{\Delta_2' - \Delta_3' - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_2 - Q_E \gamma_{th2}^E)}} (1 + b_1') \text{Ei}(-(\Delta_2' + b_1')), \quad (38)$$

where $\Delta_2' = \frac{M_E}{\lambda_{TE} B_E} + \frac{M_E C_E \gamma_{th2}^E}{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E) B_E}$, $\Delta_3' = \frac{\psi_E \gamma_{th2}^E}{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E)}$,

$$\text{and } b_1' = \frac{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}.$$

Corollary 12: At high MERs, the asymptotic expression for the IP of R_n of the ambient backscatter NOMA systems is given as

$$P_{int, \infty}^{R_n} = -\Delta_{16}' e^{\Delta_{17}' - \Delta_{18}' - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_1 - O_E \gamma_{th2}^E)}} \times (1 + b_2') \text{Ei}(-(\Delta_{17}' + b_2')), \quad (39)$$

where $\Delta_{17}' = \frac{M_E}{\lambda_{TE} B_E} + \frac{M_E C_E \gamma_{th2}^E}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E) B_E}$, $\Delta_{16}' = \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}$,

$$\Delta_{18}' = \frac{\psi_E \gamma_{th2}^E}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}, \quad \text{and } b_2' = \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}.$$

Corollary 13: At high MERs, the asymptotic expression for the IP of T of the ambient backscatter NOMA systems in (40), as shown at the bottom of the next page.

$$\begin{aligned} P_{out, \infty}^{T, ni} &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TE} \Delta_5^{R_n}} e^{A_2^{R_n}} \left(\frac{\pi}{N} \sum_{k=0}^N e^{-\left(\frac{2A_3^{R_n}}{A_4^{R_n} (\vartheta_k + 1)} - \frac{A_1^{R_n} A_4^{R_n} (\vartheta_k + 1)}{2} \right)} \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} + \frac{1}{\vartheta_k + 1} \right) - 2K_0 \left(2\sqrt{-A_1^{R_n} A_3^{R_n}} \right) \right) \\ &\quad + \frac{2\lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \zeta_{R_n} B_{R_n}} e^{-B_5} \sum_{v=1}^{\infty} (-1)^v \frac{1}{B_4^v} \left(\frac{B_3}{B_1} \right)^{\frac{v}{2}} K_v \left(2\sqrt{B_3 B_1} \right). \quad (32) \end{aligned}$$

$$\begin{aligned} P_{int}^{T, ni} &= \frac{\lambda_{SE} \xi_E \gamma_{thc}^E}{\lambda_{ST} \lambda_{TE} \Delta_5^E} e^{A_2^E} \left(\frac{\pi}{N} \sum_{k=0}^N e^{-\left(\frac{2(A_3^E + \Delta_8^E)}{A_4^E (\vartheta_k + 1)} - \frac{A_1^E A_4^E (\vartheta_k + 1)}{2} \right)} \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} - \frac{1}{\vartheta_k + 1} \right) + 2K_0 \left(2\sqrt{-A_1^E (A_3^E + \Delta_8^E)} \right) \right) \\ &\quad + 2\sqrt{\Delta_{15} \Delta_{13}} e^{-\Delta_{14}} K_1 \left(2\sqrt{\Delta_{13} \Delta_{15}} \right). \quad (37) \end{aligned}$$

TABLE I
TABLE OF PARAMETERS FOR NUMERICAL RESULTS

Power sharing coefficients of NOMA	$a_1 = 0.2, a_2 = 0.8$
Noise power	$N_0 = 1$
Reflection coefficient	$\beta = 0.1$
ipSIC parameter	$\varepsilon = 0.01$
Power coefficient of artificial noise	$\varphi_J = 0.1$
Interfering factor of readers	$\varpi = 0.5$
RHIs parameter	$\kappa_{SR_f} = \kappa_{SR_n} = \kappa_{SE} = \kappa = 0.1$
Channel fading parameters	$\{\lambda_{SR_f}, \lambda_{SR_n}, \lambda_{SB}, \lambda_{SE}, \lambda_{TR_f}, \lambda_{TR_n}, \lambda_{TE}\} = \{4, 6, 1, 0.5, 1, 2, 0.3\}$
CEEs parameter	$\sigma_{e_{SR_f}}^2 = \sigma_{e_{SR_n}}^2 = \sigma_{e_{SB}}^2 = \sigma_{e_{SE}}^2 = \sigma_{e_{TR_f}}^2 = \sigma_{e_{TR_n}}^2 = \sigma_{e_{TE}}^2 = \sigma_e^2 = 0.05$
Targeted data rates (OP)	$\{\gamma_{th1}^R, \gamma_{th2}^R = \gamma_{th2}^f, \gamma_{thc}^R\} = \{1.2, 1, 0.001\}$
Targeted data rates (IP)	$\{\gamma_{th1}^E, \gamma_{th2}^E, \gamma_{thc}^E\} = \{0.12, 0.3, 0.01\}$

$$\ln(40), A_1' = -\frac{1}{\lambda_{SE} \xi_E \lambda_{me} \lambda_{TE}^2 \Delta_5^E \gamma^2}, \Delta_{13}' = \frac{1}{\lambda_{me} \lambda_{TE}^2 \Delta_5^E},$$

$$K_1(2\sqrt{\Delta_{13}' \Delta_{15}}) \approx I_1(2\sqrt{\Delta_{13}' \Delta_{15}}) (\ln(\sqrt{\Delta_{13}' \Delta_{15}}) + \nu) + \frac{1}{2} (\sqrt{\Delta_{13}' \Delta_{15}})^{-1} - \frac{1}{2} \sum_{l=0}^3 \frac{(\sqrt{\Delta_{13}' \Delta_{15}})^{2l+1}}{l!(l+1)} \left(\sum_{k=1}^l \frac{1}{k} + \sum_{k=1}^{l+1} \frac{1}{k} \right).^5$$

Proof: The proof follows by taking λ_{me} large in (40) and simplifying the expressions by utilizing $e^{-x} \approx 1 - x$ and $K_0(x) \approx -\ln(x)$ if $x \rightarrow 0$. ■

Remark 2: From **Theorem 4**, **Theorem 8**, **Corollaries 4-6** and **Corollaries 11-13**, whether the ideal or non-ideal conditions, the following observations can be inferred: 1) RHIs, CEEs and ipSIC can enhance the security of the ambient backscatter NOMA systems; 2) When the reflection coefficient β increases, both $P_{int}^{R_f}$ and $P_{int}^{R_n}$ decrease, while P_{int}^T increases; 3) Increasing φ_J can reduce the IP, thereby improving the reliability-security trade-off of the considered systems; 4) as λ_{me} grows, the security for R_n and R_f is improved, while the security for T is reduced.

IV. NUMERICAL RESULTS

In this section, simulation results are provided to verify the correctness of our theoretical analysis in Section III. The results are verified by using Monte Carlo simulations with 10^6 trials. Unless otherwise stated, we set the parameters as shown in Table I is at the top of this page.

Fig. 2 plots the OP and the IP versus the transmit SNR for the far reader, the near reader and T under both ideal

⁵For large MER, in order to achieve a better approximation effect, we only need to consider the first three terms of l , i.e. $l = 1, 2, 3$.

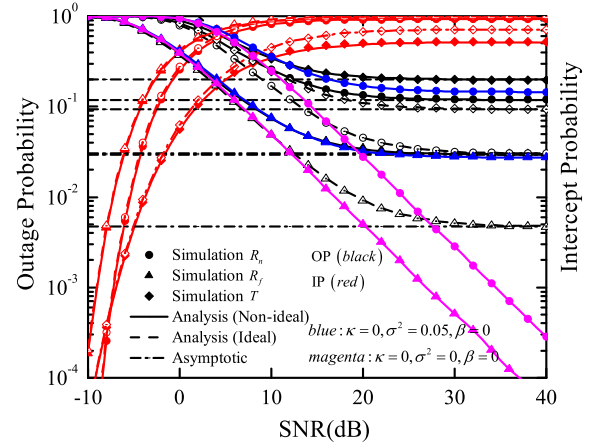


Fig. 2. OP and IP versus the transmit SNR.

and non-ideal conditions, with $\kappa = 0.1$ and $\sigma_e^2 = 0.05$. For the purpose of comparison, the considered system performance under ideal conditions is provided with $\kappa = 0$, $\sigma_e^2 = 0$, as well as with $\sigma_e^2 = 0$, $\beta = 0$ and $\kappa = 0$. It is shown that the theoretical results match well the simulations across the entire SNR region. We can also observe that the OP approaches a fixed non-negative constant due to the fixed estimation error and β in the high SNR region, which results in zero diversity order. These results verify the conclusion in **Remark 1**. Moreover, RHIs have a positive impact on IP, which reveals that the ideal communication systems are more vulnerable to be eavesdropped than the non-ideal communication systems.

$$P_{int,\infty}^{T,ni}$$

$$= -\frac{\pi \lambda_{SE} \xi_E \gamma_{thc}^E}{N \lambda_{me} \lambda_{TE}^2 \Delta_5^E} e^{-\frac{M_E \gamma_{thc}^E}{\lambda_{TE} \Delta_5^E}} \left(1 - \frac{C_E \gamma_{thc}^E}{\lambda_{me} \lambda_{TE} \Delta_5^E} \right) \sum_{k=0}^N e^{-\frac{2(A_3^E + \Delta_8^E)}{A_4^E (\vartheta_k + 1)}} \left(1 + \frac{A_1' A_4^E (\vartheta_k + 1)}{2} \right) \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} - \frac{1}{\vartheta_k + 1} \right)$$

$$+ \frac{\lambda_{SE} \xi_E \gamma_{thc}^E}{\lambda_{me} \lambda_{TE}^2 \Delta_5^E} e^{-\frac{M_E \gamma_{thc}^E}{\lambda_{TE} \Delta_5^E}} \left(1 - \frac{C_E \gamma_{thc}^E}{\lambda_{me} \lambda_{TE} \Delta_5^E} \right) \ln \left(\sqrt{-A_1^E (A_3^E + \Delta_8^E)} \right)$$

$$+ 2\sqrt{\Delta_{15} \Delta_{13}'} K_1(2\sqrt{\Delta_{13}' \Delta_{15}}) e^{-\frac{C_E \gamma_{thc}^E}{\lambda_{TE} \Delta_5^E}} \left(1 - \frac{C_E \gamma_{thc}^E}{\lambda_{me} \lambda_{TE} \Delta_5^E} \right).$$

(40)

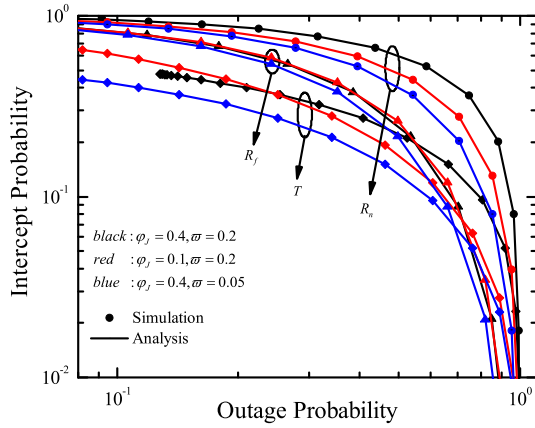
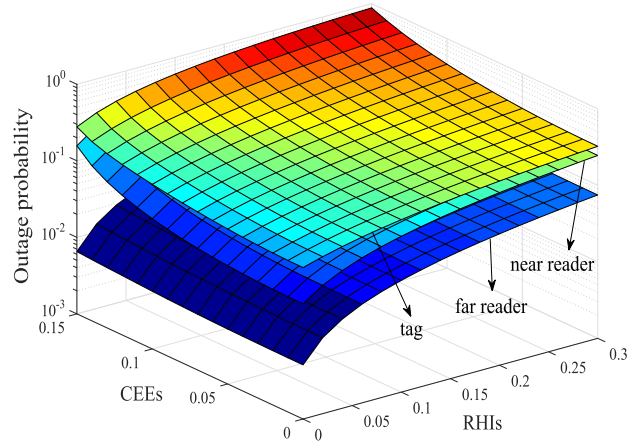


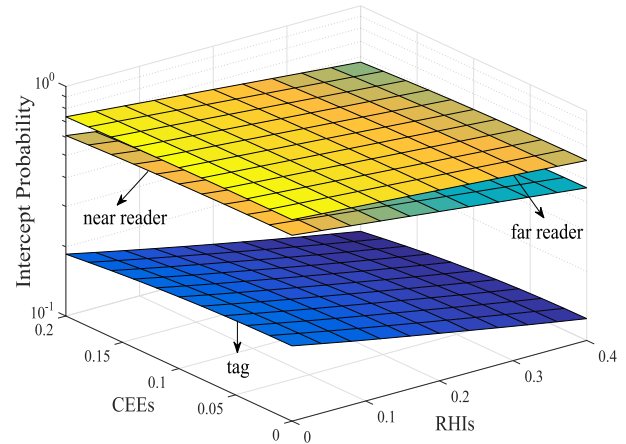
Fig. 3. IP versus OP for different power coefficient of artificial noise φ_J .



(a) OP versus RHIs and CEEs.

Finally, we can also see that there exists a trade-off between reliability and security.

Fig. 3 shows the impact of OP versus IP for different power coefficient of the artificial noise φ_J and interference factor ϖ , with $\varphi_J = \{0.1, 0.4\}$ and $\varpi = \{0.2, 0.05\}$. In this simulation, we assume $\kappa = 0$ and $\sigma_e^2 = 0$. Results show that when OP increases, IP decreases, and vice-versa. This means that there exists a trade-off between OP and IP. In addition, we can observe that as the power coefficient of the artificial noise φ_J becomes smaller, the reliability-security trade-off of the considered system degrades significantly. This is because the interference signals at eavesdropper become more dominant, resulting in a higher IP. Similarly, the interference factor ϖ of the readers increases so as to result in a higher OP, which indicates that the reliability-security trade-off degrades obviously. It is noted that the IP of T is the smallest, implying that T has the most secure performance. Therefore, in order to improve the reliability-security trade-off of the considered system by artificial noise, the design with a larger power coefficient of the artificial noise and smaller interference factor of the reader is more important.



(b) IP versus RHIs and CEEs.

Fig. 4 presents the OPs and IPs versus RHIs κ and CEEs σ_e^2 . In this simulation, we set SNR = 25 dB and $\varphi_J^{R_n} = 0.05$ for the OP, while SNR = 5 dB and $\varphi_J^E = 0.2$ for the IP. According to Figs. 4 (a) and (b), as κ grows, $P_{out}^{R_f}$, $P_{out}^{R_n}$ and P_{out}^T increase, while $P_{int}^{R_f}$, $P_{int}^{R_n}$ and P_{int}^T decrease. Likewise, when increasing σ_e^2 , the OPs of R_f , R_n and T increase, whereas the corresponding IPs decrease. It means that the reliability of T is the worst, while it has better security. Moreover, for R_f , R_n and T , the fluctuation for the OP and IP of RHIs is more obvious than that of CEEs, which shows that the reliability and security of the readers are more dependent on the level of RHIs. Finally, we can also observe that as RHIs change, the OP of far reader changes drastically. In contrast, the change for OP of T is the least obvious, and this happens because T can eliminate part of interference caused by the far and near readers.

Fig. 5 illustrates the OP and IP versus the transmit SNR for different ε and β , respectively. In this simulation, we set: $\varepsilon = \{0, 0.05\}$, $\beta = \{0.2, 0.12\}$ for OP; $\varepsilon = \{0, 0.3\}$,

$\beta = \{0.1, 0.3\}$ for IP. As can be seen in Fig. 5 (a), error floors for the OP occur in the high SNR regime. OP decreases as the transmit SNR increases, which is determined by the values of ε and β . More specifically, under perfect SIC ($\varepsilon = 0$), the outage behaviors of R_f , R_n and T improve remarkably when β increases; Similarly, for a fixed β , the increase of ε also leads to lower reliability of R_n and T . By comparing Fig. 5 (a) with Fig. 5 (b), we can observe that ε and β have opposite effects on IP for the far reader, near reader, and T , while β has identical effects on T , i.e., as ε increases, the corresponding IPs at the near reader and T decrease. Additionally, the increase of β reduces the security of T , but enhances the security of both the far and near readers. It is worth noting that OPs of R_f and R_n are more sensitive to β , which is due to the increase of interference from the backscatter link. For IP, T is more sensitive to β . This happens because when β increases, E is more likely to eavesdrop the information of $c(t)$ successfully.

Fig. 6 presents the IP versus the MER for R_f , R_n , and T under ideal conditions with $\kappa = 0$, $\sigma_e^2 = 0$, as well as non-ideal conditions with $\kappa = 0.1$, $\sigma_e^2 = 0.05$. In this simulation, we set

Fig. 4. OP and IP versus RHIs and CEEs.

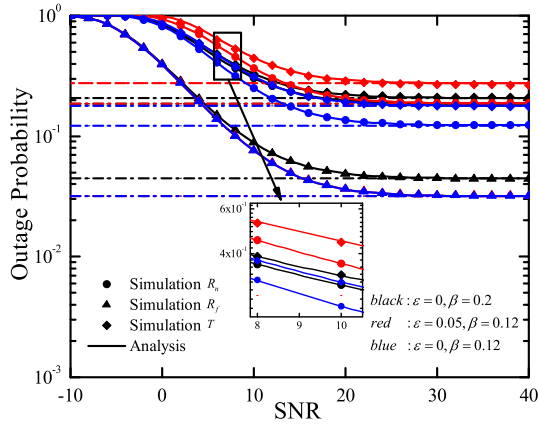
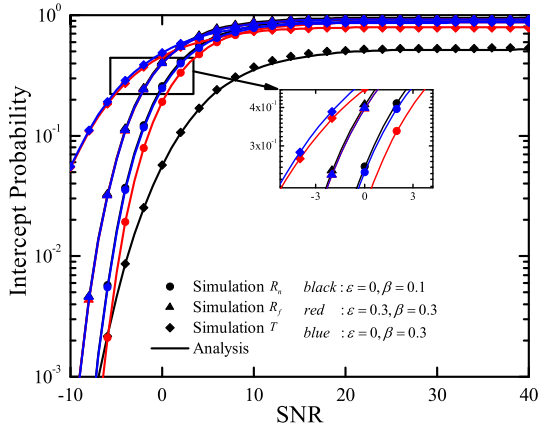
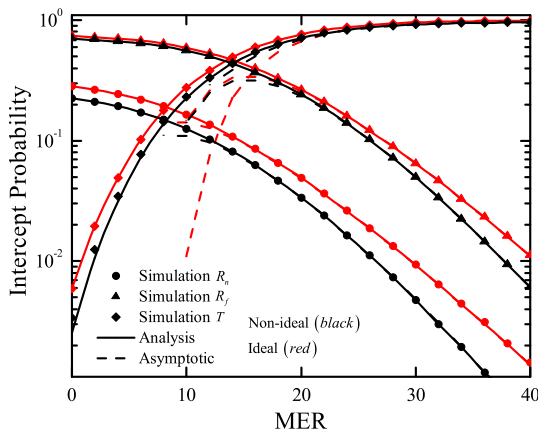
(a) OP versus the transmit SNR for different ε and β .(b) IP versus the transmit SNR for different ε and β .Fig. 5. OP and IP versus the transmit SNR for different ε and β .

Fig. 6. IP versus MER for ideal and non-ideal conditions.

SNR = 5 dB, $\lambda_{TE} = 2$, and $\{\gamma_{th1}^E, \gamma_{th2}^E, \gamma_{thc}^E\} = \{0.3, 0.3, 1\}$. From Fig. 6, we can observe that the asymptotic results are strict approximation of the IP in the high MER regime and the RHIs can enhance the security of R_f , R_n , and T . In addition, the IP of R_f is much larger than that of R_n when R_f and

R_n have the same target rate, which is due to the fact that R_f allocates more power. Therefore, considering the small power allocation coefficients a_1 and high target rate γ_{th1}^E of the R_n , it is difficult for the information of R_n to be eavesdropped by E . Finally, we can also observe that as MER grows, the security for R_n and R_f is improved, while the security for T is reduced.

V. CONCLUSION

In this article, we investigate the joint impacts of RHIs, CEEs and ipSIC on the reliability and the security of the ambient backscatter NOMA systems in terms of OP and IP. To improve the security performance, an artificial noise scheme was proposed, where the RF source simultaneously sends the signal and artificial noise to the readers and tag. The analytical expressions for the OP and the IP were derived. Furthermore, the asymptotic OP in the high SNR regime and the asymptotic IP in the high MER region were analyzed. Numerical results showed that: 1) RHIs, CEEs and ipSIC have negative effects on the OP but positive effects on the IP; 2) Compared with CEEs, RHIs have a more serious impact on the reliability and security of the considered system; 3) There exists a trade-off between reliability and security, and this trade-off can be optimized by reducing the power coefficient of the artificial noise or increasing the interfering factor of readers; 4) There are error floors for the OP due to the CEEs and the reflection coefficient; 5) As MER becomes large, the security for R_n and R_f improves, while the security for T reduces. In addition, the increase of β reduces the reliability but enhances the security for the far reader and near reader. Finally, we can conclude that the optimal reliability-security trade-off performance can be achieved by adjusting the power coefficient of the artificial noise and interference factor of the reader, which further drives the applicability of ambient backscatter communications in the IoT networks.

APPENDIX A: PROOF OF THEOREM 1

Substituting (3) into (6), the OP of R_f can be expressed as

$$P_{out}^{R_f} = 1 - \underbrace{\text{Pr}(\gamma_{R_f}^{x_2} > \gamma_{th2}^{R_f})}_{I_1}, \quad (\text{A.1})$$

where I_1 is calculated as follows:

$$\begin{aligned} I_1 &= \text{Pr}(\gamma_{R_f}^{x_2} > \gamma_{th2}^{R_f}) \\ &= \int_{\alpha_1}^{\infty} \frac{1}{\lambda_{SR_f}} e^{-\frac{x}{\lambda_{SD_f}}} \frac{1}{\lambda_{TR_f}} e^{-\frac{y}{\lambda_{TR_f}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dx dy dz \\ &\stackrel{u=z+\alpha}{=} \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f} - \frac{\gamma_{th2}^{R_f}}{\lambda_{SR_f} \gamma(a_2 - Q_{R_f} \gamma_{th2}^{R_f})}} \int_{\alpha}^{\infty} e^{-\alpha_3 u} \frac{1}{u} du \\ &\stackrel{l_1}{=} 1 + \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f} - \frac{\gamma_{th2}^{R_f}}{\lambda_{SR_f} \gamma(a_2 - Q_{R_f} \gamma_{th2}^{R_f})}} \text{Ei}(-\Delta_2^{R_f}), \end{aligned} \quad (\text{A.2})$$

where $\alpha_1 = \frac{(B_{R_f} z + C_{R_f}) \gamma \gamma_{th2}^{R_f} y + M_{R_f} \gamma \gamma_{th2}^{R_f} z + (\psi_{R_f} + 1) \gamma \gamma_{th2}^{R_f}}{(a_2 - Q_{R_f} \gamma_{th2}) \gamma}$, $\alpha_2 = \frac{\lambda_{SR_f} (a_2 - Q_{R_f} \gamma_{th2}) + \lambda_{TR_f} C_{R_f} \gamma_{th2}^{R_f}}{\lambda_{TR_f} B_{R_f} \gamma_{th2}^{R_f}}$, $\alpha_3 = \frac{M_{R_f} \gamma_{th2}^{R_f}}{\lambda_{SR_f} (a_2 - Q_{R_f} \gamma_{th2})} + \frac{1}{\lambda_{ST}}$, and the step l_1 is obtained by utilizing [53, eq. (3.352)]. Finally, substituting (A.2) into (A.1), we can obtain (24); then, substituting $\kappa = 0$ and $\sigma_e^2 = 0$ into (24), we can obtain (7).

Similarly, substituting (3) and (4) into (11), the (12) and (26) can be obtained.

APPENDIX B: PROOF OF THEOREM 3

Substituting (3), (4) and (5) into (14), the OP of T can be expressed as

$$P_{out}^T = 1 - \underbrace{\text{Pr} \left(\gamma_{R_n}^{x_2} > \gamma_{th2}^{R_n}, \gamma_{R_n}^{x_1} > \gamma_{th1}^{R_n}, \gamma_{R_n}^{c(t)} > \gamma_{thc}^{R_n} \right)}_{I_2}, \quad (\text{B.1})$$

• Non-ideal conditions

For non-ideal conditions, I_2 is calculated as (B.2), as shown at the bottom of the page.

By using some mathematical manipulations, we can obtain

$$\begin{aligned} I_{21} &= \int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} \alpha_5 e^{-\frac{\varsigma_{R_n} (M_{R_n} z + \psi_{R_n} + \frac{1}{\gamma})}{\lambda_{SR_n}} - \alpha_4} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dz \\ &= \int_0^{\infty} \frac{\lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \varsigma_{R_n} B_{R_n}} e^{-\alpha_6} \frac{1}{u + B_4} e^{-(B_1 u + \frac{B_3 + \Delta_6}{u})} du \\ &= \frac{\lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \varsigma_{TR_n}} e^{-B_6} \sum_{v=1}^{\infty} (-1)^v \frac{1}{B_4^v} \int_0^{\infty} u^{v-1} e^{-(B_1 u + \frac{B_3 + \Delta_6}{u})} du \\ &\stackrel{l_2}{=} \frac{2 \lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \varsigma_{R_n} B_{R_n}} e^{-\alpha_6} \sum_{v=1}^{\infty} (-1)^v \frac{1}{B_4^v} \\ &\quad \times \left(\frac{(B_3 + \Delta_6)}{B_1} \right)^{\frac{v}{2}} K_v \left(2 \sqrt{(B_3 + \Delta_6) B_1} \right), \quad (\text{B.3}) \end{aligned}$$

where $u = \lambda_{SR_n} \lambda_{TR_n} \Delta_5^{R_n} z - \lambda_{SR_n} \lambda_{TR_n} C_{R_n} \gamma_{thc}^{R_n}$, $\alpha_4 = \frac{[\lambda_{TR_n} \varsigma_{R_n} (B_{R_n} z + C_{R_n}) + \lambda_{SR_n} (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n} + M_{R_n} \gamma_{thc}^{R_n} z]}{\lambda_{SR_n} \lambda_{TR_n} (\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n})}$, $\alpha_5 = \frac{\lambda_{SR_n}}{\lambda_{TR_n} \varsigma_{R_n} (B_{R_n} z + C_{R_n}) + \lambda_{SR_n}}$, $\alpha_6 = B_5 + \frac{\lambda_{TR_n} \varsigma_{R_n} B_{R_n} \gamma_{thc}^{R_n} + \frac{\varsigma_{R_n}}{\lambda_{SR_n}}}{\lambda_{SR_n} \lambda_{TR_n} \gamma_{thc}^{R_n}}$, and l_2 is obtained by utilizing [15, eq. (3.471)].

$$\begin{aligned} I_{22} &= \int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} e^{\frac{\psi_{R_n} + \frac{1}{\gamma}}{\lambda_{SR_n} \varsigma_{R_n}} + \left(\frac{M_{R_n}}{\lambda_{SR_n} \varsigma_{R_n}} - \frac{1}{\lambda_{ST}} \right) z - \alpha_7} \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} (\lambda_{TR_n} \Delta_5^{R_n} z + \Delta_7^{R_n})} dz \\ &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TR_n} \Delta_5^{R_n}} e^{A_2} \int_0^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u + A_4^{R_n}} du \\ &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TR_n} \Delta_5^{R_n}} e^{A_2} \left[\underbrace{\int_0^{A_4^{R_n}} e^{-\left((-A_5) u + \frac{A_3}{u} \right)} \frac{1}{u + A_4^{R_n}} du}_{l_3} \right. \\ &\quad \left. + \underbrace{\int_{A_4^{R_n}}^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u + A_4^{R_n}} du}_{l_4} \right], \quad (\text{B.4}) \end{aligned}$$

where $u = \lambda_{SR_n} \xi_{R_n} \lambda_{TR_n} \Delta_5^{R_n} z - \lambda_{SR_n} \xi_{R_n} \lambda_{TR_n} C_{R_n} \gamma_{thc}^{R_n}$,

$$\alpha_7 = \frac{(\lambda_{TR_n} \Delta_5^{R_n} z + \Delta_7^{R_n}) (\psi_{R_n} + 1/\gamma + M_{R_n} z)}{\lambda_{SR_n} \xi_{R_n} \lambda_{TR_n} [\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}]},$$

and l_3 can be approximated by the Gaussian-Chebyshev quadrature [54], i.e., $l_3 \approx$

$$\frac{\pi}{N} \sum_{k=0}^N \frac{1}{(\vartheta_k + 3)} e^{-\left(\frac{2(A_3^{R_n} + \Delta_8^{R_n})}{A_4^{R_n} (\vartheta_k + 1)} - \frac{A_1^{R_n} A_4^{R_n} (\vartheta_k + 1)}{2} \right)} \sqrt{1 - \vartheta_k^2}.$$

Next, due to $A_4^{R_n} \leq 1$, l_4 can be expressed as

$$\begin{aligned} l_4 &\approx \int_{A_4^{R_n}}^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u} du \\ &= \int_0^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u} du \end{aligned}$$

$$\begin{aligned} I_2 &= \text{Pr} \left(\varsigma_{R_n} \gamma \left[(B_{R_n} |\hat{h}_{ST}|^2 + C_{R_n}) |\hat{h}_{TR_n}|^2 + M_{R_n} |\hat{h}_{ST}|^2 + \psi_{R_n} + \frac{1}{\gamma} \right] < |\hat{h}_{SR_n}|^2 < \frac{(\Delta_5^{R_n} |\hat{h}_{ST}|^2 - C_{R_n} \gamma_{thc}^{R_n}) |\hat{h}_{TR_n}|^2 - M_{R_n} |\hat{h}_{ST}|^2 \gamma_{thc}^{R_n} - (N_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\xi_{R_n} \gamma_{thc}^{R_n}} \right) \\ &= \int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} \int_{\frac{M_{R_n} \gamma_{thc}^{R_n} z + (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}}}^{\infty} \int_{\varsigma_{R_n} [(B_{R_n} z + C_{R_n}) y + M_{R_n} z + \psi_{R_n} + \frac{1}{\gamma}]}^{\frac{(\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}) y - M_{R_n} \gamma_{thc}^{R_n} z - (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\xi_{R_n} \gamma_{thc}^{R_n}}} \frac{1}{\lambda_{SR_n}} e^{-\frac{x}{\lambda_{SR_n}}} \frac{1}{\lambda_{TR_n}} e^{-\frac{y}{\lambda_{TR_n}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dx dy dz \\ &= \underbrace{\int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} \int_{\frac{M_{R_n} \gamma_{thc}^{R_n} z + (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}}}^{\infty} e^{-\frac{1}{\lambda_{SR_n}} \varsigma_{R_n} [(B_{R_n} z + C_{R_n}) y + M_{R_n} z + \psi_{R_n} + \frac{1}{\gamma}]} \frac{1}{\lambda_{TR_n}} e^{-\frac{y}{\lambda_{TR_n}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dy dz}_{I_{21}} \\ &\quad - \underbrace{\int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} \int_{\frac{M_{R_n} \gamma_{thc}^{R_n} z + (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}}}^{\infty} e^{-\frac{(\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n}) y - M_{R_n} \gamma_{thc}^{R_n} z - (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n}}{\lambda_{SR_n} \xi_{D_n} \gamma_{thc}^{R_n}}} \frac{1}{\lambda_{TR_n}} e^{-\frac{y}{\lambda_{TR_n}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dy dz}_{I_{22}}. \quad (\text{B.2}) \end{aligned}$$

$$\begin{aligned}
P_{out}^{T,id} &= 1 - \int_{\frac{(\psi_{R_n}+1/\gamma)\gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} \left(e^{-\frac{\varsigma_{R_n}(B_{R_n}y+\psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n}y - (\psi_{R_n}+1/\gamma)\gamma_{thc}^{R_n}}{\lambda_{SR_n}\xi_{R_n}\gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST}\lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST}\lambda_{TR_n}}} \right) dy \\
&= 1 - \underbrace{\int_0^{\infty} \left(e^{-\frac{\varsigma_{R_n}(B_{R_n}y+\psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n}y - (\psi_{R_n}+1/\gamma)\gamma_{thc}^{R_n}}{\lambda_{SR_n}\xi_{R_n}\gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST}\lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST}\lambda_{TR_n}}} \right) dy}_{I_{31}} \\
&\quad + \underbrace{\int_0^{\frac{(\psi_{R_n}+1/\gamma)\gamma_{thc}^{R_n}}{\Delta_5^{R_n}}} \left(e^{-\frac{\varsigma_{R_n}(B_{R_n}y+\psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n}y - (\psi_{R_n}+1/\gamma)\gamma_{thc}^{R_n}}{\lambda_{SR_n}\xi_{R_n}\gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST}\lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST}\lambda_{TR_n}}} \right) dy}_{I_{32}}. \quad (B.6)
\end{aligned}$$

$$\begin{aligned}
& - \int_0^{A_4^{R_n}} e^{-\left((-A_1^{R_n})u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u} du \\
&= 2K_0 \left(2\sqrt{-A_1^{R_n} \left(A_3^{R_n} + \Delta_8^{R_n} \right)} \right) \\
&\quad - \frac{\pi}{N} \sum_{k=0}^N \frac{1}{\vartheta_{k+1}} e^{-\left(\frac{2(A_3^{R_n} + \Delta_8^{R_n})}{A_4^{R_n}(\vartheta_{k+1})} - \frac{A_1^{R_n} A_4^{R_n}(\vartheta_{k+1})}{2} \right)} \sqrt{1 - \vartheta_k^2}. \quad (B.5)
\end{aligned}$$

$$\begin{aligned}
& \times \int_{\frac{M_{E^*} \gamma_{thc}^{E^*} z + (\psi_{E^*} + 1/\gamma) \gamma_{thc}^{E^*}}{\Delta_5^{E^*} z - C_E \gamma_{thc}^{E^*}}}^{\infty} e^{-\frac{(\Delta_5^{E^*} z - C_E \gamma_{thc}^{E^*}) y - M_{E^*} \gamma_{thc}^{E^*} z - (\psi_{E^*} + 1/\gamma) \gamma_{thc}^{E^*}}{\lambda_{SE} \xi_{E^*} \gamma_{thc}^{E^*}}} \\
& \times \frac{1}{\lambda_{TE}} e^{-\frac{y}{\lambda_{TE}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dy dz. \quad (C.1)
\end{aligned}$$

Similar to the derivation process of I_{22} , after some mathematical manipulations, $P_{int}^{T,ni}$ can be obtained.

• Ideal conditions

Substituting $\kappa = 0$ and $\sigma_e^2 = 0$ into (5), $C_E = M_E = 0$. Then, the IP of T at the ideal conditions is given by

$$\begin{aligned}
P_{int}^{T,id} &= \int_0^{\infty} \int_{\frac{\gamma_{thc}^{E^*}}{\gamma_{\Delta_5^{E^*}}}}^{\infty} \left(1 - e^{-\frac{\Delta_5^{E^*} \gamma y - \gamma_{thc}^{E^*}}{\lambda_{SE} \xi_{E^*} \gamma_{thc}^{E^*}}} \right) \\
&\quad \frac{2}{\lambda_{TE} \lambda_{ST}} K_0 \left(2\sqrt{\frac{y}{\lambda_{TE} \lambda_{ST}}} \right) dy, \quad (C.2)
\end{aligned}$$

After some mathematical manipulations, we can obtain $P_{int}^{T,id}$.

By substituting l_3 and (B.5) into (B.4), I_{22} can be obtained; substituting (B.3) and (B.4) into (B.2), I_2 can be derived.

• Ideal conditions

Substituting $\kappa = 0$ and $\sigma_e^2 = 0$ into (3), (4) and (5), $C_{R_f} = M_{R_f} = C_{R_n} = M_{R_n} = 0$. Then, the OP of T under ideal conditions is given (B.6), as shown at the top of the page.

In (B.6), I_{31} can be obtained by utilizing [53, eq. (6.611)], I_{32} can be approximated by the Gaussian-Chebyshev quadrature [54]. Thus, I_{31} and I_{32} can be expressed as

$$I_{31} = \Delta_{11} e^{\Delta_{11} + \frac{1}{\lambda_{SR_n} \gamma_{\xi_{R_n}}}} \text{Ei}(-\Delta_{11}) - \Delta_{9} e^{\Delta_{9} - \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma}} \text{Ei}(-\Delta_{9}), \quad (B.7)$$

$$\begin{aligned}
I_{32} &= \frac{\gamma_{thc}^{R_n} \pi}{N \lambda_{TR_n} \lambda_{ST} \gamma \Delta_5^{R_n}} \sum_{k=0}^N K_0 \left(2\sqrt{\Delta_{10}} \right) \sqrt{1 - \vartheta_k^2} \\
&\quad \times \left[e^{-\left(\varsigma_{R_n} B_{R_n} \Delta_{10} + \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma} \right)} - e^{-\frac{1}{\lambda_{SR_n} \gamma_{\xi_{R_n}}} - \frac{\vartheta_{k+1}}{2 \lambda_{SR_n} \gamma_{\xi_{R_n}}}} \right]. \quad (B.8)
\end{aligned}$$

Similarly, substituting (B.7) and (B.8) into (B.6), we can obtain $P_{out}^{T,id}$.

APPENDIX C: PROOF OF THEOREM 4

According to I_1 , we can obtain $P_{int}^{R_f}$ and $P_{int}^{R_n}$. Substituting (5) into (24), the IP of T can be expressed as

• Non-ideal conditions

$$\begin{aligned}
P_{int}^{T,ni} &= \int_{\frac{C_E \gamma_{thc}^{E^*}}{\Delta_5^{E^*}}}^{\infty} \int_{\frac{M_{E^*} \gamma_{thc}^{E^*} z + (\psi_{E^*} + 1/\gamma) \gamma_{thc}^{E^*}}{\Delta_5^{E^*} z - C_E \gamma_{thc}^{E^*}}}^{\infty} \frac{1}{\lambda_{TE} \lambda_{ST}} e^{-\left(\frac{y}{\lambda_{TE}} + \frac{z}{\lambda_{ST}} \right)} dy dz \\
&\quad - \int_{\frac{C_E \gamma_{thc}^{E^*}}{\Delta_5^{E^*}}}^{\infty}
\end{aligned}$$

REFERENCES

- [1] X. Liu, H. Ding, and S. Hu, "Uplink resource allocation for NOMA-based hybrid spectrum access in 6G-enabled cognitive Internet of Things," *IEEE Internet Things J.*, early access, Jul. 3, 2020, doi: 10.1109/JIOT.2020.3007017.
- [2] S. Jacob *et al.*, "A novel spectrum sharing scheme using dynamic long short-term memory with CP-OFDMA in 5G networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 3, pp. 926–934, Sep. 2020.
- [3] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.
- [4] M. Zeng, G. I. Tsiropoulos, O. A. Dobre, and M. H. Ahmed, "Power allocation for cognitive radio networks employing non-orthogonal multiple access," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–5.
- [5] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.
- [6] X. Li, J. Li, Y. Liu, Z. Ding, and A. Nallanathan, "Residual transceiver hardware impairments on cooperative NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 680–695, Jan. 2020.
- [7] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.
- [8] X. Lu, D. Niyato, H. Jiang, D. I. Kim, Y. Xiao, and Z. Han, "Ambient backscatter assisted wireless powered communications," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 170–177, Apr. 2018.
- [9] B. Lyu, Z. Yang, H. Guo, F. Tian, and G. Gui, "Relay cooperation enhanced backscatter communication for Internet-of-Things," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2860–2871, Apr. 2019.

- [10] X. Lu, D. Niyato, H. Jiang, E. Hossain, and P. Wang, "Ambient backscatter-assisted wireless-powered relaying," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 4, pp. 1087–1105, Dec. 2019.
- [11] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, "Ambient backscatter: Wireless communication out of thin air," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 39–50, Sep. 2013.
- [12] D. Darsena, G. Gelli, and F. Verde, "Modeling and performance analysis of wireless networks with ambient backscatter devices," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1797–1814, Apr. 2017.
- [13] W. Zhao, G. Wang, S. Atapattu, C. Tellambura, and H. Guan, "Outage analysis of ambient backscatter communication systems," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1736–1739, Aug. 2018.
- [14] J. Guo, X. Zhou, S. Durrani, and H. Yanikomeroglu, "Design of non-orthogonal multiple access enhanced backscatter communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6837–6852, Oct. 2018.
- [15] H. Guo, Y.-C. Liang, R. Long, and Q. Zhang, "Cooperative ambient backscatter system: A symbiotic radio paradigm for passive IoT," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1191–1194, Aug. 2019.
- [16] Y. Ye, L. Shi, X. Chu, and G. Lu, "On the outage performance of ambient backscatter communications," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7265–7278, Aug. 2020.
- [17] S. N. Daskalakis, A. Georgiadis, G. Goussetis, and M. M. Tentzeris, "Low cost ambient backscatter for agricultural applications," in *Proc. IEEE-APS Top. Conf. Antennas Propag. Wireless Commun. (APWC)*, Sep. 2019, p. 201.
- [18] S. Wei, J. Wang, and Z. Zhao, "Poster abstract: LocTag: Passive WiFi tag for robust indoor localization via smartphones," in *Proc. IEEE INFOCOM-IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Jul. 2020, pp. 1342–1343.
- [19] X. Lu, H. Jiang, D. Niyato, D. I. Kim, and Z. Han, "Wireless-powered device-to-device communications with ambient backscattering: Performance modeling and analysis," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1528–1544, Mar. 2018.
- [20] X. Li, H. Mengyan, Y. Liu, V. G. Menon, A. Paul, and Z. Ding, "I/Q imbalance aware nonlinear wireless-powered relaying of B5G networks: Security and reliability analysis," *IEEE Trans. Netw. Sci. Eng.*, early access, Sep. 3, 2020, doi: [10.1109/TNSE.2020.3020950](https://doi.org/10.1109/TNSE.2020.3020950).
- [21] M. Abbasi, A. Shokrollahi, M. R. Khosravi, and V. G. Menon, "High-performance flow classification using hybrid clusters in software defined mobile edge computing," *Comput. Commun.*, vol. 160, pp. 643–660, Jul. 2020, doi: [10.1016/j.comcom.2020.07.002](https://doi.org/10.1016/j.comcom.2020.07.002).
- [22] A. D. Wyner, "The wire-tap channel," *Bell Syst. Tech. J.*, vol. 54, no. 8, pp. 1355–1387, Oct. 1975.
- [23] N.-P. Nguyen, M. Zeng, O. A. Dobre, and H. V. Poor, "Securing massive MIMO-NOMA networks with ZF beamforming and artificial noise," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.
- [24] H. Lei *et al.*, "Secrecy outage analysis for cooperative NOMA systems with relay selection schemes," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6282–6298, Sep. 2019.
- [25] B. Li, X. Qi, K. Huang, Z. Fei, F. Zhou, and R. Q. Hu, "Security-reliability tradeoff analysis for cooperative NOMA in cognitive radio networks," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 83–96, Jan. 2019.
- [26] Q. Yang, H.-M. Wang, Q. Yin, and A. L. Swindlehurst, "Exploiting randomized continuous wave in secure backscatter communications," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3389–3403, Apr. 2020.
- [27] Y. Zhang, F. Gao, L. Fan, X. Lei, and G. K. Karagiannidis, "Secure communications for multi-tag backscatter systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1146–1149, Aug. 2019.
- [28] J. Y. Han, M. J. Kim, J. Kim, and S. M. Kim, "Physical layer security in multi-tag ambient backscatter communications—Jamming vs. Cooperation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.
- [29] M. Zeng, N.-P. Nguyen, O. A. Dobre, and H. V. Poor, "Securing downlink massive MIMO-NOMA networks with artificial noise," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 685–699, Jun. 2019.
- [30] E. Balti and M. Guizani, "Impact of non-linear high-power amplifiers on cooperative relaying systems," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4163–4175, Oct. 2017.
- [31] A.-A.-A. Boulogeorgos, V. M. Kapinas, G. K. Karagiannidis, and R. Schober, "I/Q-imbalance self-interference coordination," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4157–4170, Jun. 2016.
- [32] X. Li, M. Zhao, Y. Liu, L. Li, Z. Ding, and A. Nallanathan, "Secrecy analysis of ambient backscatter NOMA systems under I/Q imbalance," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12286–12290, Oct. 2020.
- [33] A.-A.-A. Boulogeorgos, N. D. Chatzidiamantis, and G. K. Karagiannidis, "Non-orthogonal multiple access in the presence of phase noise," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1133–1137, May 2020.
- [34] T. Schenk, *RF Imperfections in High-Rate Wireless Systems: Impact and Digital Compensation*. New York, NY, USA: Springer-Verlag, 2008.
- [35] E. Bjornson, J. Hoydis, M. Kountouris, and M. Debbah, "Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.
- [36] X. Li, Q. Wang, Y. Liu, T. A. Tsiftsis, Z. Ding, and A. Nallanathan, "UAV-aided multi-way NOMA networks with residual hardware impairments," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1538–1542, Sep. 2020.
- [37] P. K. Sharma and P. K. Upadhyay, "Cognitive relaying with transceiver hardware impairments under interference constraints," *IEEE Commun. Lett.*, vol. 20, no. 4, pp. 820–823, Apr. 2016.
- [38] J. Cui, Z. Ding, and P. Fan, "Outage probability constrained MIMO-NOMA designs under imperfect CSI," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8239–8255, Dec. 2018.
- [39] S. Lee, T. Q. Duong, and R. Woods, "Impact of wireless backhaul unreliability and imperfect channel estimation on opportunistic NOMA," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10822–10833, Nov. 2019.
- [40] J. He, Z. Tang, Z. Tang, H. Chen, and C. Ling, "Design and optimization of scheduling and non-orthogonal multiple access algorithms with imperfect channel state information," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10800–10814, Nov. 2018.
- [41] A. K. Mishra, D. Mallick, and P. Singh, "Combined effect of RF impairment and CEE on the performance of dual-hop fixed-gain AF relaying," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1725–1728, Sep. 2016.
- [42] X. Li, M. Huang, J. Li, Q. Yu, K. Rabie, and C. C. Cavalcante, "Secure analysis of multi-antenna cooperative networks with residual transceiver HIs and CEEs," *IET Commun.*, vol. 13, no. 17, pp. 2649–2659, Oct. 2019.
- [43] X. Ding, T. Song, Y. Zou, X. Chen, and L. Hanzo, "Security-reliability tradeoff analysis of artificial noise aided two-way opportunistic relay selection," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 3930–3941, May 2017.
- [44] T. Yoo and A. Goldsmith, "Capacity and power allocation for fading MIMO channels with channel estimation error," *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2203–2214, May 2006.
- [45] O. S. Badarneh and R. Mesleh, "A comprehensive framework for quadrature spatial modulation in generalized fading scenarios," *IEEE Trans. Commun.*, vol. 64, no. 7, pp. 2961–2970, Jul. 2016.
- [46] M. T. Mamaghani, A. Kuhestani, and K.-K. Wong, "Secure two-way transmission via wireless-powered untrusted relay and external jammer," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8451–8465, Sep. 2018.
- [47] E. Bjornson, M. Matthaiou, and M. Debbah, "A new look at dual-hop relaying: Performance limits with hardware impairments," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4512–4525, Nov. 2013.
- [48] C. Studer, M. Wenk, and A. Burg, "MIMO transmission with residual transmit-RF impairments," in *Proc. Int. ITG Workshop Smart Antennas (WSA)*, Feb. 2010, pp. 189–196.
- [49] S. Stefania, B. Matthew, and T. Issam, *LTE—The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. New York, NY, USA: Wiley, 2011.
- [50] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, 10th ed. New York, NY, USA: Academic, 1972.
- [51] X. Li *et al.*, "Physical layer security of cooperative NOMA for IoT networks under I/Q imbalance," *IEEE Access*, vol. 8, pp. 51189–51199, Mar. 2020.
- [52] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H. V. Poor, *MIMO Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2007.
- [53] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*. New York, NY, USA: Academic, 2007.
- [54] F. B. Hildebrand, *Introduction to Numerical Analysis*. New York, NY, USA: Dover, 1987.



Xingwang Li (Senior Member, IEEE) received the M.Sc. degree from the University of Electronic Science and Technology of China in 2010 and the Ph.D. degree from the Beijing University of Posts and Telecommunications in 2015. From 2010 to 2012, he was with Comba Telecom, Ltd., Guangzhou, China, as an Engineer. From 2017 to 2018, he was a Visiting Scholar with Queen's University Belfast, Belfast, U.K., for one year. He is currently an Associate Professor with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, China. His research interests include MIMO communication, cooperative communication, hardware constrained communication, NOMA, physical layer security, UAV communication, and the Internet of Things. He has served as many TPC members, such as the IEEE Globecom, IEEE WCNC, IEEE VTC, and IEEE ICC. He has also served as the Co-Chair for the IEEE/IET CSNDSP 2020 of the Green Communications and Networks Track. He also serves as an Editor on the Editorial Board for IEEE ACCESS, *Computer Communications*, *Physical Communication*, the *KSII Transaction on Internet and Systems*, and *IET Quantum Communication*. He is also the Lead Guest Editor of the Special Issue on *UAV-enabled 5G/6G networks: Emerging Trends and Challenges of Physical Communication*, the Special Issue on *Recent Advances in Physical Layer Technologies for the 5G-Enabled Internet of Things of Wireless Communications and Mobile Computing*, and the Special Issue on *Recent Advances in Multiple Access for 5G-enabled IoT of Security and Communication Networks*.



Mengle Zhao (Student Member, IEEE) received the B.Sc. degree in electronic information engineering from the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, China, in 2018, where she is currently pursuing the M.Sc. degree in communication and information systems with the School of Physics and Electronic Information Engineering. Her current research interests include non-orthogonal multiple access, physical layer security, in-phase and quadrature-phase imbalance, cooperative communication, and backscatter device.



Ming Zeng (Member, IEEE) received the B.E. and master's degrees from the Beijing University of Post and Telecommunications, China, in 2013 and 2016, respectively, and the Ph.D. degree in telecommunications engineering from the Memorial University of Newfoundland, Canada, in 2020. He is currently an Assistant Professor with the Department of Electrical Engineering and Computer Engineering, Université Laval, Canada. He has authored or coauthored more than 45 articles and conferences in first-tier IEEE journals and proceedings. His research interests include resource allocation for beyond 5G systems and machine learning empowered optical communications. He serves as an Associate Editor for IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. His work has been cited over 1050 times per Google Scholar.



Shahid Mumtaz (Senior Member, IEEE) is an IET Fellow, an IEEE ComSoc and ACM Distinguished speaker. He was a recipient of the IEEE ComSoc Young Researcher Award in 2020. He is the Founder and the EiC of the *IET Journal of Quantum communication*. He is the Vice Chair of the Europe/Africa Region-IEEE ComSoc: Green Communications and Computing Society and the IEEE standard on P1932.1: Standard for Licensed Unlicensed Spectrum Interoperability in Wireless Mobile Networks.

He is currently a Senior 5G Consultant with Huawei, Sweden. He is serving as a scientific expert and an evaluator for various research funding agencies. He has authored four technical books, 12 book chapters, over 250 technical articles (over 150 journals/Transactions, over 80 conference). Most of his publication is in the field of wireless communication. He was a recipient of the Alain Bensoussan fellowship in 2012. He was also a recipient of the NSFC Researcher Fund for Young Scientist in 2017 from China. He received the IEEE Best Paper Award in the area of mobile communications.



Varun G. Menon (Senior Member, IEEE) received the Ph.D. degree in computer science and engineering from Sathyabama University, India, in 2017. He is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. He has authored or coauthored more than 50 research articles in peer-reviewed and highly indexed international journals and conferences. His research interests include Internet of Things, fog computing and networking, underwater acoustic sensor networks, scientometrics, educational psychology, ad-hoc networks, wireless communication, opportunistic routing, and wireless sensor networks. He is an Editorial Board Member of the IEEE Future Directions. He is a Distinguished Speaker of the ACM. He has served over 20 conferences, such as the IEEE ICC, ICCCN 2020, IEEE COINS 2020, SigTelCom, ICACCI, and ICDMAI in leadership capacities, including the program co-chair, the track chair, and the session chair, and a technical program committee member. He is currently a Guest Editor of IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE SENSORS JOURNAL, and IEEE INTERNET OF THINGS JOURNAL. He is an Associate Editor of *IET Quantum Communications*.



Zhiguo Ding (Fellow, IEEE) received the B.Eng. degree in electrical engineering from the Beijing University of Posts and Telecommunications in 2000 and the Ph.D. degree in electrical engineering from the Imperial College London in 2005. From 2005 to 2018, he was with Queen's University Belfast, the Imperial College London, Newcastle University, and Lancaster University. From 2012 to 2020, he was an Academic Visitor with Princeton University. Since 2018, he has been with The University of Manchester as a Professor

in communications. His research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He received the Best Paper Award from the IET ICWMC-2009 and IEEE WCSP-2014, the EU Marie Curie Fellowship 2012–2014, the Top IEEE TVT Editor 2017, the IEEE Heinrich Hertz Award 2018, the IEEE Jack Neubauer Memorial Award 2018, the IEEE Best Signal Processing Letter Award 2018, and the Web of Science Highly Cited Researcher 2019. He was an Editor of IEEE WIRELESS COMMUNICATION LETTERS and IEEE COMMUNICATION LETTERS from 2013 to 2016. He is serving as an Area Editor for IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY and an Editor for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and *Journal of Wireless Communications and Mobile Computing*.



Octavia A. Dobre (Fellow, IEEE) received the Dipl.-Ing. and Ph.D. degrees from the Polytechnic Institute of Bucharest, Romania, in 1991 and 2000, respectively. From 2002 to 2005, she was with the New Jersey Institute of Technology, USA. In 2005, she joined Memorial University of Newfoundland, Canada, where she is currently a Professor and the Research Chair. She was a Visiting Professor with the Massachusetts Institute of Technology, USA, and the Université de Bretagne Occidentale, France. Her research interests encompass various wireless

technologies, such as non-orthogonal multiple access and full duplex, optical and underwater communications, and machine learning for communications. She has (co-)authored over 300 refereed articles in these areas.

Dr. Dobre also served as the general chair, the technical program co-chair, the tutorial co-chair, and the technical co-chair of symposia at numerous conferences. She was the Editor-in-Chief (EiC) of IEEE COMMUNICATIONS LETTERS, a senior editor, an editor, and a guest editor for various prestigious journals and magazines. She serves as the EiC of IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.

Dr. Dobre is a fellow of the Engineering Institute of Canada. She was a Royal Society Scholar, a Fulbright Scholar, and a Distinguished Lecturer of the IEEE Communications Society. She received the Best Paper Awards from various conferences, including the IEEE ICC, IEEE Globecom, IEEE WCNC, and IEEE PIMRC.



Physical Communication

Volume 46, June 2021, 101315

Full length article

Hardware impaired modify-and-forward relaying with relay selection: Reliability and security

Hongxing Peng^a , Hongyan Qi^a , Xingwang Li^{a, b} , Yuan Ding^c , Jun Wu^a  , Varun G. Menon^d ^a School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo 454000, China^b Henan Chuitian Technology Company Ltd., Hebi, China^c School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, Scotland, UK^d Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam, India

Received 10 November 2020, Revised 12 February 2021, Accepted 26 February 2021, Available online 4 March 2021, Version of Record 10 March 2021.

Show less  Share  Cite<https://doi.org/10.1016/j.phycom.2021.101315> [Get rights and content](#) 

Abstract

In this paper, we consider the physical layer security of cooperative multiple relays networks, where the source tries to communicate the destination via modify-and-forward (MF) relaying in the presence of eavesdropper. More practical, transceiver residual hardware impairments (TRHIs) and channel estimation errors (CEEs) are taken into account. To improve secure performance and energy efficiency, the $K - th$ best relay is selected since the best relay is not available due to some schedule and/or other reasons. More specifically, we investigate the reliability and security by invoking the outage probability(OP) and intercept probability(IP). To obtain more useful insights, the asymptotic behaviors for the OP are examined in the high signal-to-noise ratio (SNR) regime, followed by the diversity orders. The numeric results show that: (1) The secure performance is improved by employing MF compared with decode-and-forward (DF); (2) The reliability increases as the total number of relays increases; (3) There is an error floor for the outage probability due to the CEEs.

Introduction

With the development of Internet-of-Things (IoT) and Mobile Internets (MIN), the future beyond fifth generation (B5G) mobile communication networks will meet the demands of massive connections and ultra-reliable low-latency communications (URLLC) [1], [2], [3]. In order to achieve the above demands, secure communication has been identified as a crucial guarantee for the future wireless networks. Traditionally, secure communication is ensured by using encryption algorithms at the transmitter and decryption at the receiver. This not only imposes extra computational overhead and system complexity but also insecurity with the rapid development of computer

technology. In light of this fact, Physical Layer Security (PLS) has been proposed as an effective way to ensure security of wireless communication network, which has sparked a great deal of interests from academia and industry[4], [5].

PLS, originally proposed by Wyner[6], investigated the reliable communication from the point of information theory, which has sparked a great deal of research interests[7], [8], [9], [10], [11], [12], [13], [14]. In[7], authors derived analytical expressions for non-zero secrecy capacity and the secrecy outage probability of single-input single-output (SISO) systems over Rician/Nakagami-m fading channels. Authors in[8] focused on the PLS of single-input multiple-output (SIMO) systems, and a media-based modulation scheme was proposed. Extending multiple distributed antenna arrays, Forssell et al. proposed a new physical layer authentication approach of SIMO systems[9]. In[10], the opportunistic access point selection was used to discuss the outage performance for mobile edge computing (MEC) network, in which employed selection combining (SC) and switch-and-stay combining (SSC) two protocols. Regarding to multiple-input multiple-output (MIMO) cognitive wiretap system, Lei et al. has studied the secrecy outage probability performance of optimal antenna selection and suboptimal antenna selection schemes over Nakagami-m fading channel[11]. For MIMO system with unknown noise statistics, the authors developed a generalized maximum likelihood (ML) estimation to detect signals in[12]. For improving physical layer security, Yan et al. considered a multi-input multi-output cognitive radio (MIMO-CR) system and derived the secrecy outage probability analysis by proposed optimal antenna selection (OAS) and suboptimal antenna selection (SAS) schemes[13]. Employing the large scale antenna array can improve spectral efficiency and enhance wireless security, in[14] a large scale MIMO was introduced into the physical layer, with the purpose of tracking with the short range interception problem, the secrecy performance of amplify-and-forward (AF) and DF was analyzed. Considering hardware impairments at transceivers, authors in[15] investigated the reliability and security of ambient backscatter NOMA networks.

Cooperative relaying is an effective way to provide diversity gain and enhance edge coverage. Thus, cooperative communication has received extensive research in wireless networks. In the[16], the performance of a multi-carrier cooperative underwater acoustic communication (UWAC), in which fixed features in the underwater channel, has been analyzed. Cao et al. introduced the cooperative relay technique into conventional underlay/overlay D2D communications, where proposed adaptive mode selection and spectrum allocation schemes to ensure better performance of the cellular and D2D users[17]. With the advantage of improving network capacity, a Capacity-Optimized Cooperative topology control scheme, in which including the upper layer network capacity and the physical layer cooperative communications, has been proposed[18]. The security of cooperative communication networks has also attracted many researchers[19], [20], [21]. In a dual-hop cooperative AF relaying network, the expressions in terms of the secrecy outage probability and ergodic secrecy capacity have been derived, for the consideration, an effective secrecy diversity order has also been investigated[19]. Though small cell networks can meet the data traffic demands, it is constrained when converting between base stations. Based on this situation, the achievable sum rate, symbol error rate and outage probability in a cooperative transmission mechanism, have been explored by combining Rician/Gamma fading channels with zero-forcing receivers[20]. In the presence of an eavesdropper and co-channel interference, Vahidian et al. considered two opportunistic relay selection techniques to achieve physical layer security, where the first scheme was that the selected relay minimized the leakage information at the eavesdropper node, the second scheme was that the selected relay maximized achievable capacity of the destination node[21]. For cache-aided multi-relay networks, Xia et al. discussed secrecy outage performance in[22]. Though the multi-relay cooperative network can reduce the network complexity and improves the spatial diversity of the network, it does not make full use of the frequency band. Relay selection has been considered as an effective scheme to use frequency and ensure the secrecy and protect the source message in cooperative relay communication, which appears in rich literature[23], [24], [25]. In order to improve the PLS of cooperative wireless networks and prevent eavesdropping attacks, two protocols, where called AF and DF, were studied. Considering the existence of eavesdropping, the intercept probability expressions and the diversity order performance of relay selection was derived and evaluated, where using asymptotic intercept probability analysis[23]. Since the opportunistic relay selection has limits in the confidentiality, two scheme, where the one assumed that the eavesdropping CSI can be known at any time and the achievable secure rate can be maximized and the other one assumed a general understanding of the eavesdropper channel and was suitable for practical application, were proposed in cooperative networks[24]. Ikki et al. in[25] investigated the performance of the best-relay selection scheme in the cooperative networks, where the selected best relay needed to achieve the maximum SNR at the destination node, and also derived the expressions of the outage probability and average channel capacity. Fan et al. in[26] discussed the outage performance and optimized the cache placement with multiple amplify-

and-forward relay networks, which applied the best relay. However, the best relay may not be available. The authors in [27] explored the OP and the throughput by employing a relay selection scheme, where the HIs and interference were considered. For enhancing work efficiency, Bao et al. adopted three opportunity relay selection schemes to analyze the PLS performance in [28].

In practice, radio frequency (RF) frond-ends are limited by some imperfections, such as residual hardware impairments (RHIs) [29], [30], phase noise [31], [32], non-linear power amplify [33], [34] and in-phase/quadrature phase (I/Q) imbalance [35]. For terrestrial relays that are interfered by co-channel interference (CCI), Guo et al. in [29] investigated outage probability (OP) and throughput performance of the considered system under HIs, where a partial relay selection scheme was used. Considering the impact of RHIs, the authors analyzed the achievable sum-rate of the unmanned aerial vehicle (UAV)-aided non-orthogonal multiple access (NOMA) multi-way in [30]. In [31], the authors focused on the analysis of average symbol error rate (ASER) by different fading scenarios, where random phase noise was considered. In [32], the authors proposed a physical layer authentication scheme of MIMO system by jointly utilizing channel and phase noise, analyzed the security, covertness, robustness of the proposed scheme, and estimated the channel gain and phase noise. Considering the high-power amplifier (HPA) non-linear, Balti et al. analyzed the outage probability (OP), the bit error rate, and the capacity of the cooperative relaying systems, in which the opportunistic relay selection with outdated CSI was used to select the best relay [33]. Taking the effect of the HPA, Belkacem et al. discussed the OP and ergodic sum rate in NOMA systems, and further explored the asymptotic OP in the high SNR region [34]. In this respect, Zhang et al. in [36] proposed four linear precoding techniques to mitigate I/Q imbalance of down-link massive MIMO systems, namely widely linear zero-forcing, widely linear matched filter, widely linear minimum mean-squared error and widely linear block-diagonalization. The security and reliability of the ambient backscatter NOMA system were studied by deriving analytical expressions for the outage probability and the intercept probability [35]. In addition, it is impossible to obtain perfect channel state information (CSI) due to channel estimation errors (CEEs) [5], [37]. In [37], authors analyzed the security-reliability tradeoff of multiple DF relays networks, where the CEEs was taken into account. Li et al. in [5] investigated PLS of wireless-powered decode-and-forward (DF) multi-relay networks by joint considering non-linear energy harvesters, I/Q imbalance and CEEs.

To further improve the system secure performance, a MF protocol was originally proposed by Kim in [38], where relay first decodes the received information and then forwards the modified information to the receiver. The secure performance can be achieved that the secret can only be shared between relay and destination via unique CSI. However, eavesdropper cannot decode information since the CSI of between relay and destination is not know in the eavesdropper. On this basis, the authors have investigated the PLS of MF cooperative communications [39], [40], [41]. Utilizing the principle of physical-layer-network coding, a novel secure physical layer network coding MF (SPMF) was proposed in cooperative relay network in [39], without CEEs. Compared with [39], Vien et al. in [40] discussed the analytical expressions for the secrecy outage probability of SPMF networks by considering both direct transmission or relaying transmission scenarios. The authors focused on the secure performance analysis of MF multi-relay and multi-eavesdropper networks, where three relay selection criteria are considered according to the level of channel knowledge acquisition in [41], however, the RHIs was not considered.

The above studies on MF protocol security performance are based on ideal conditions, however, in real communication systems, this becomes impractical. Motivated by this, we focuses on the reliability and security performance of cooperative multi-relay networks, where the $K - th$ best relay is selected to communicate with destination by using MF protocol. In practice, RHIs and CEEs are considered. In this study, we assume that all nodes are equipped with single antenna and all links experience Rayleigh fading and path loss. Specifically, we derive the theoretical analytical expressions of outage probability and intercept probability. To get more insights, we also study the asymptotic expressions and the diversity order of the outage probability. Some research involved non-ideal HIs and imperfect CSI on DF relaying networks in [42], [43], [44]. Guo et al. in [42] evaluated the effect of HIs on DF multiple relaying networks, adopting switch-and-examine combining with post-selection (SECps) scheduling scheme. The authors discussed the OP with HIs in the DF terrestrial relays, where used a multi-relay selection (MRS) and single-relay selection (SRS) schemes in [43]. In [44], taking the HIs and CEEs two factors, the reliability performance for a cognitive satellite-terrestrial relay network (CSTRN) was investigated, and the half-duplex decode-and-forward (DF) mode was adopted. For the purpose of comparison, the results of DF protocol are provided. The main contributions of this paper are as follows:

- Different from the most existing works, considering RHIs and CEEs, we propose a K-th best relay selection scheme. This happens that the best relay is not available or the best relay is scheduled. Moreover, the MF protocol is considered by decoding the original information and forwarding the modified information the destination in the presence of eavesdropper.
- We investigate the reliability of the considered cooperative MF multi-relay networks by deriving the theoretical analytical expression for the outage probability. For the purpose of comparison, we consider both ideal conditions and non-ideal conditions.
- We investigate the security of the considered cooperative MF multi-relay networks by deriving the theoretical analytical expression for the intercept probability. For the purpose of comparison, the results of the considered systems with DF protocol are taken into account.
- We further study the asymptotic condition and the diversity order of the outage probability in the high signal-to-noise ratio (SNR) regime. It illustrates that outage probability has error floor at high SNRs in the presence of CEEs. It also indicates there is a tradeoff between the outage probability and the intercept probability in the presence of CEEs, RHIs. This means that the optimal can be obtained by carefully selecting parameter values.

The remainder of this paper is organized as follows. In Section2, we present the system model of the considered networks. In Section3, we investigate the security and reliability by deriving the intercept probability and the outage probability both non-ideal conditions and ideal conditions. In Section4, we analyze and discuss the asymptotic behavior and diversity order of the outage probability under high SNRs. The numerical results are given in Section5. Finally, the conclusions are drawn in Section6.

Section snippets

System model and statistical characteristics

We consider a cooperative MF relaying network as shown in Fig. 1, which consists of one source S , one legitimate destination D , one illegitimate eavesdropper E , and N relays R_n , $n=\{1, 2, \dots, N\}$. We assume that all nodes are equipped a single antenna, and the direct link between S and D is absent due to the heavy blockage[45]. For convenience, we also assume that channel coefficients about S to R_n , S to E , R_n to E , R_n to D are all marked as h_i , $i \in (SR_n, R_nD, R_nE, SE)$.

In practice, owing to CEEs, it is ...

Reliability and security analysis

In this section, we study the reliability and security of the considered system in terms of the outage probability and the intercept probability, and the asymptotic analysis and the diversity orders are carried out. For comparison, the results of DF protocol are also presented in this section....

Asymptotic analysis and diversity order

To obtain useful insights, we investigate the asymptotic analysis and the diversity order of the OP...

Numerical results and discussion

In this section, we present the analytical and simulation results to verify our analysis in Sections3 Reliability and security analysis, 4 Asymptotic analysis and diversity order. In all evaluations, unless otherwise explicitly specified, we assume that the parameters of those results are set as follows: $\sigma_{eji}^2 = \sigma_e^2$, $\alpha = 3$. Moreover, Monte Carlo simulations have been conducted with 10^4 channels trials.

Fig.2 plots the OP and IP versus the average transmit SNR under the ideal and non-ideal...

Conclusion

In this paper, we consider the reliability and security of multi-relay networks by presenting a new MF protocol, where the two factors of RHIs and CEEs are taken into account. Specifically, the exact expressions of the OP and IP have been derived. Numerical results reveal that: (i) the MF is effective for system security compared with the DF; (ii) the K_{th} ($K_{th} > 1$) best relay selection schemes can solve the best relay unavailable. (iii) RHIs and CEEs have detrimental impact on reliability; and...

CRedit authorship contribution statement

Hongxing Peng: Conceptualization, Methodology, Supervision. **Hongyan Qi:** Investigation, Software, Data curation, Writing - original draft, Address comments. **Xingwang Li:** Investigation, Software, Data curation, Writing - original draft, Supervision, Writing - reviewing and editing, Address comments. **Yuan Ding:** Conceptualization, Methodology, Writing - reviewing and editing, Address comments. **Jun Wu:** Conceptualization, Methodology, Writing - reviewing and editing. **Varun G. Menon:** Visualization,...

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper....

Acknowledgments

This work was support in party by Henan Scientific and Technological Research Project under grant 212102210557, in part by Key Scientific Research Projects of Higher Education Institutions in Henan Province under grant 20A510007, and in part by the National Natural Science Foundation of China under grant 61601414....

Hongxing Peng received his B.Sc. and M.Sc. degrees in industrial automation from and control theory and control engineering from Henan Polytechnic University, China in 1999 and 2003, respectively. He then received Ph.D. degree in detection technology and automatic equipment from Beijing Institute of Technology in 2009. He is currently an Associated Professor with the School of Physics and Electronic Information Engineering. His research interests include wireless communication and...

[Special issue articles](#) [Recommended articles](#)

References (52)

WuY. *et al.*

Massive access for future wireless communications systems

IEEE Wireless Commun. (2020)

SuttonG.J. *et al.*

Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives

IEEE Commun. Surveys Tuts. (2019)

Li Xingwang, Wang Qunshu, Liu Meng, Li Jingjing, Peng Hongxing, PiranMd Jalil, Li Lihua, Cooperative wireless-powered...

CaoK.

Improving physical layer security of uplink NOMA via energy harvesting jammers

IEEE Trans. Inf. Forensics Secur. (2021)

Li Xingwang, Huang Mengyan, Liu Yuanwei, MenonVarun G., Paul Anand, Ding Zhiguo, I/Q imbalance aware nonlinear...

WynerA.D.

The wire-tap channel

Bell Labs Tech. J. (1975)

S. Iwata, T. Ohtsuki, P.Y. Kam, Performance analysis of physical layer security over Rician/Nakagami-m fading channels,...

T. Mao, Z. Wang, Physical-layer security enhancement for SIMO-MBM systems, in: Proc. IEEE Global Commun. Conf....

H. Forssell, R. Thobaben, J. Gross, Performance analysis of distributed SIMO physical layer authentication, in: Proc...

Xiaj.

Opportunistic access point selection for mobile edge computing networks

IEEE Trans. Wireless Commun. (2021)



View more references

Cited by (2)

[Research on Physical Layer Security of Cooperative NOMA System Based on MF Protocol](#) ↗

2023, Dianzi Yu Xinxu Xuebao/Journal of Electronics and Information Technology

[Secrecy sum-rate based illegitimate relay selection](#) ↗

2023, Australian Journal of Electrical and Electronics Engineering



Hongxing Peng received his B.Sc. and M.Sc. degrees in industrial automation from and control theory and control engineering from Henan Polytechnic University, China in 1999 and 2003, respectively. He then received Ph.D. degree in detection technology and automatic equipment from Beijing Institute of Technology in 2009. He is currently an Associated Professor with the School of Physics and Electronic Information Engineering. His research interests include wireless communication and Internet-of-thing (IoT).



Hongyan Qi (S'19) received the B.Sc. degree in communication and information systems with the School of Physics and Electronic Information Engineering, Henan Polytechnic University in 2016. She is currently pursuing the M.Sc degree in communication and information systems with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo China. She current research interests include cooperative communication, simultaneous wireless information and power transfer, hardware-constrained communication.



Xingwang Li (S'12-M'15-SM'20) received the M.Sc. and Ph.D. degrees from University of Electronic Science and Technology of China and Beijing University of Posts and Telecommunications in 2010 and 2015. From 2010 to 2012, he worked at Comba Telecom Ltd. in Guangzhou China, as an engineer. He spent one year from 2017 to 2018 as a visiting scholar at Queen's University Belfast, Belfast, UK. He is also a visiting scholar at State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications from 2016 to 2018. He is currently an Associated Professor with the School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo China. His research interests include MIMO communication, cooperative communication, hardware constrained communication, non-orthogonal multiple access, physical layer security, unmanned aerial vehicles, and the Internet of Things. He has served as many TPC members, such as the IEEE GLOBECOM, IEEE WCNC, IEEE VTC, IEEE ICC and so on. He has also served as the Co-Chair for the IEEE/IET CSNDSP 2020 of the Green Communications and Networks Track. He also serves as an Editor on the Editorial Board for IEEE ACCESS, COMPUTER COMMUNICATIONS, PHYSICAL COMMUNICATION, KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS and IET QUANTUM COMMUNICATION. He is also the Lead Guest Editor for the Special Issue on UAV-enabled B5G/6G networks: Emerging Trends and Challenges of PHYSICAL COMMUNICATION, Special Issue on Recent Advances in Physical Layer Technologies for the 5G-Enabled Internet of Things of WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, and Special Issue on Recent Advances in Multiple Access for 5G-enabled IoT of SECURITY AND COMMUNICATION NETWORKS.



Yuan Ding (M'19) received his Bachelor's degree from Beihang University (BUAA), Beijing, China, in 2004, received his Master's degree from Tsinghua University, Beijing, China, in 2007, and received his Ph.D. degree from Queen's University of Belfast, Belfast, UK, in 2014, all in Electronic Engineering. He was a radio frequency (RF) Engineer in Motorola R&D Centre (Beijing, China) from 2007 to 2009, before joining Freescale Semiconductor Inc. (Beijing, China) as an RF Field Application Engineer, responsible for high power base-station amplifier design, from 2009 to 2011. He is now an Assistant Professor at the Institute of Sensors, Signals and Systems (ISSS) in Heriot-Watt University, Edinburgh, UK. His research interests are in antenna array, physical layer security, and 5G related areas. Dr. Ding was the recipient of the IET Best Student Paper Award at LAPC 2013 and the recipient of the Young Scientists Awards in General Assembly and Scientific Symposium (GASS), 2014 XXXIst URSI.



Jun Wu received the M.Sc. and Ph.D. degrees from Henan Polytechnic University, China in 2006 and 2020. He is currently an Associated Professor with the School of Physics and Electronic Information Engineering. His research interests include wireless communication and Internet-of-thing (IoT).



Varun G Menon is currently an Associate Professor in Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. He is a Senior Member of IEEE and a Distinguished Speaker of ACM Distinguished Speaker. Dr. Varun G Menon is currently a Guest Editor for IEEE Transactions on Industrial Informatics, IEEE Sensors Journal, IEEE Internet of Things Magazine and Journal of Supercomputing. He is an Associate Editor of IET Quantum Communications and also an Editorial Board Member of IEEE Future Directions: Technology Policy and Ethics. His research interests include Internet of Things, Fog Computing and Networking, Underwater Acoustic Sensor Networks, Cyber Psychology, Hijacked Journals, Ad-Hoc Networks, Wireless Sensor Networks.

[View full text](#)



All content on this site: Copyright © 2024 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the Creative Commons licensing terms apply.



[\[BACK\]](#)

Computers, Materials & Continua
DOI:10.32604/cmc.2021.015426
Article



Energy-Efficient Transmission Range Optimization Model for WSN-Based Internet of Things

Md. Jalil Piran¹, Sandeep Verma², Varun G. Menon³ and Doug Young Suh^{4,*}

¹Department of Computer Science and Engineering, Sejong University, Seoul, Korea

²Department of Electronics and Communication Engineering, D.B.R.A. National Institute of Technology, Jalandhar, India

³SCMS School of Engineering and Technology, Ernakulam, India

⁴Department of Electronics Engineer, Kyung Hee University, Yongin, Korea

*Corresponding Author: Doug Young Suh. Email: suh@khu.ac.kr

Received: 02 November 2020; Accepted: 10 December 2020

Abstract: With the explosive advancements in wireless communications and digital electronics, some tiny devices, sensors, became a part of our daily life in numerous fields. Wireless sensor networks (WSNs) is composed of tiny sensor devices. WSNs have emerged as a key technology enabling the realization of the Internet of Things (IoT). In particular, the sensor-based revolution of WSN-based IoT has led to considerable technological growth in nearly all circles of our life such as smart cities, smart homes, smart healthcare, security applications, environmental monitoring, etc. However, the limitations of energy, communication range, and computational resources are bottlenecks to the widespread applications of this technology. In order to tackle these issues, in this paper, we propose an Energy-efficient Transmission Range Optimized Model for IoT (ETROMI), which can optimize the transmission range of the sensor nodes to curb the hot-spot problem occurring in multi-hop communication. In particular, we maximize the transmission range by employing linear programming to alleviate the sensor nodes' energy consumption and considerably enhance the network longevity compared to that achievable using state-of-the-art algorithms. Through extensive simulation results, we demonstrate the superiority of the proposed model. ETROMI is expected to be extensively used for various smart city, smart home, and smart healthcare applications in which the transmission range of the sensor nodes is a key concern.

Keywords: Internet of Things; wireless sensor networks; routing; transmission range optimization; energy-efficiency; hot-spot problem; linear programming

1 Introduction

1.1 Background and Problem Statement

Data-driven wireless sensor networks (WSNs) are widely applied to enhance the Internet of Things (IoT) in terms of the data throughput, energy efficiency, and self-management [1]. WSN-based IoTs are composed of wireless sensor nodes, which realize data collection and communication [2,3]. In this framework, the sensor nodes are deployed in the physical environment to sense the phenomena and report their readings in a distributed manner to the sinks [4]. However, the sensor nodes exhibit certain limitations in terms of energy, computation resources, and communication range [5,6].

When a WSN-based IoT is deployed over a large application area, the nodes perform multihop communication due to the limited transmission range, and direct data transmission cannot be realized. Furthermore, it has been reported that a larger number of relay nodes on the path of data delivery to the sink corresponds to a higher probability of these nodes closer to the sink suffering from hot-spot

problem [7]. In such a scenario, the number of intermediate nodes should be reduced to decrease the emergence of a no-connection zone for distantly located nodes.

Moreover, the battery of the sensor nodes may not be able to be changed or recharged. Therefore, it is necessary to ensure efficient power consumption in a WSN-based IoT [8]. Furthermore, transmitting one kilobyte of data corresponds to the processing of three million instructions [9]. Therefore, data transmission in the WSNs should be minimized with regard to the distance between any two entities among sensor nodes, cluster heads (CHs), or sinks [10].

One solution is to maximize the transmission range between nodes. The key concept of transmission range maximization is that if a sensor initiates a data packet transmission to a sink located 1000 m away, the least number of relay sensors should be selected to forward the packet. The communication range of sensor nodes depends on their transmission power and the volume of the packet to be transmitted. Transmission over long distances requires a higher energy [11,12]. Therefore, it is necessary to determine the maximum possible distance (transmission range) to which the sensor nodes can transmit the data packets.

Many researchers have attempted to reduce the energy consumption by avoiding the hot-spot problem [13]. In particular, Verma et al. [14] proposed the multiple sink-based genetic algorithm-based optimized clustering (MS-GAOC) approach, in which four data collection sinks were incorporated outside the network. However, the cost of using four sinks may be prohibitive in various applications.

Moreover, researchers generally apply the corona-based model to avoid hot-spot problems. A survey of the various corona-based approaches has been presented in an existing study [15]. Nevertheless, even corona-based methods are not sufficiently reliable in mitigating the hot-spot problem. In fact, the literature review indicates that the concept of transmission range adjustment for the sensor nodes, to realize direct data transfer to the sink or transfer with the least possible number of intermediate nodes, has not been extensively investigated.

1.2 Motivation

The review pertaining to the mitigation of hot-spot problems indicated that the optimization-based approach can provide a balanced solution to specific problems. Therefore, in this work, we used linear programming (LP) to compute the maximum data transmission range [16,17]. In particular, LP exhibits remarkable exploration and exploitation capabilities, enabling fast convergence to the optimal solution. Moreover, LP is highly computationally efficient [16].

1.3 Our Contributions

In the context of the aforementioned problems, the key contributions of this work are as follow:

- a) We propose an energy-efficient transmission range optimized model for IoT (ETROMI) to optimize the transmission range of the sensor nodes to reduce the hot-spot problem in WSN-based IoT.
- b) The mathematical model and formulation using LP is presented.
- c) The simplex method is used to solve the defined problem.
- d) The proposed model's performance of the proposed model is analyzed in terms of various aspects, and the optimal solution is identified.

1.4 Paper Organization

The remaining paper is structured as follows. Section 2 presents the background of transmission range adjustment algorithms and describes the existing work pertaining to the hot-spot problem in WSNs. Section 3 describes the system model and explains the LP formulation. Section 4 describes the performance evaluation of ETROMI, which is used to compute the maximized distance corresponding to the transmission range of a node. The concluding remarks, along with the limitations and scope for future work, are presented in Section 5.

2 Related Work

In this section, we discuss the existing work focused on addressing the hot-spot problem through various state-of-the-art techniques and on realizing the transmission range adjustment of a sensor node.

2.1 Approaches to Solve the Hot-Spot Problem

In applications involving an extremely large network area, the sensor nodes inevitably perform multi-hop communication [18]. In this process, a hot-spot is created at the nodes located nearest to the sink. Several researchers have addressed this concern through various topology-based methods. Moreover, the many-to-one approach (many sensor nodes corresponding to one sink) has been widely implemented through corona-based structures [13]. Many researchers use the term “energy-hole,” which is equivalent in meaning to a hot-spot.

Elkamel et al. [19] proposed an unequal clustering method to overcome the hot-spot problem by placing the small and large clusters nearer to and farther from the sink, respectively. However, the proposed technique failed to eliminate the hot-spot problem, and the network’s energy consumption was high. Verma et al. [7,14] implemented multiple data sinks in a given network to mitigate the hot-spot problem. In their former and latter studies, the authors used the conventional approach and the genetic algorithm, respectively. However, the network incurred a higher financial cost owing to the use of multiple data sinks. The authors in [20] proposed a virtual-force-based energy-hole mitigation strategy to ensure sensor nodes’ uniform distribution. Moreover, the network was composed of various annuli, and virtual gravity was used to optimize the sensor node positions in each annulus. However, due to the multi-hop communication, the number of overheads in each annulus was extremely high, which increased the energy consumption in the network. Sharmin et al. [21] proposed a strategy in which the network was partitioned into several wedges, and residual energy was considered to combine the various wedges. The head node was selected based on the distance between the innermost corona and node. However, the inefficient selection of the head node led to the mediocre performance of this strategy.

In addition to the static network scenario, certain researchers introduced sink mobility to curb the hot-spot problem. Sahoo et al. [22] proposed a particle-swarm-optimization-based energy-efficient clustering and sink mobility (PSO-ECSM) technique, in which the sink mobility was used to alleviate the hot-spot problem. However, the mobility scenario was not efficiently utilized, and the slow convergence of the PSO degraded the performance of the proposed scheme. Furthermore, Kaur et al. [23] introduced dual sink mobility outside the network to target unattended applications. Although the authors implemented the PSO-based sink mobility, the use of the dual sink introduced overheads in the network, which increased the energy consumption. In addition, the data delivery was required to be synchronized when using the two sinks in the network. Certain other researchers also employed the sink mobility scenario to alleviate the hot-spot problem. However, it was observed that the use of sink mobility limited the applicability of the approaches in various real-time scenarios.

2.2 Transmission Range Adjustment Algorithms

In addition to the network topological changes associated with the introduction of the corona-based model, the characteristics of sensor nodes have been examined. The focus of the present study is to optimize the transmission range. Although certain researchers have attempted to adjust the transmission range to alleviate the hot-spot problem, the proposed approaches suffer from the inherent problems, which limit their relevance.

In an existing strategy [24] pertaining to the transmission range adjustment, the network was divided into various concentric sets termed as coronas. Every corona was assigned a transmission range level. Furthermore, the authors presented an ant colony optimization (ACO)-based transmission range adjustment strategy [24] to prolong the network lifetime. Liu [25] considered the energy consumption balancing (ECB) and energy consumption minimization (ECM) techniques to avoid the occurrence of energy holes. The authors exploited the short-trip moving scheme for the ACO, which helped in decreasing the complexity and in the amelioration of convergence speed. Furthermore, the authors considered a reference transmission distance to implement the ECB and ECM techniques. Xin et al. [26] were the first to attempt to solve the many-to-one data transmission problem, particularly in strip-based WSNs. The authors adjusted the transmission range based on the computation of the accurate distance. The objective was to prolong the network lifetime. However, the proposed algorithm was applicable only for strip-based WSNs, for example, railway track, bridge, and tunnel systems.

In summary, only a few studies have been focused on addressing the hot-spot problem through transmission range adjustment, and this approach exhibits considerable scope for improvement. Furthermore, the use of LP for energy-hole mitigation in the transmission range adjustment context is yet to be explored. Therefore, we implement these aspects in our proposed strategy.

3 System Model

In this section, we describe the network assumptions and the system model.

3.1 Network Assumptions

The following network assumptions are considered to implement ETROMI.

- a) The WSN is composed of one sink and several sensor nodes that collect data and transfer them to the sink.
- b) Each sensor has a unique ID.
- c) There is no dispute for medium access, and thus, proportional fair channel access is available to all the sensors.
- d) The minimum cost forwarding approach is employed as the multi-hop-routing protocol.
- e) The sensor nodes are homogeneous, i.e., all the nodes have the same configuration in terms of energy, computational resources, transmission range, etc.
- f) The entire network is static, including the CHs and the sink.
- g) The entire network has ideal conditions in terms of security, physical medium factors, reflection, refraction, splitting of signals, and presence of other obstacles.

3.2 Fundamental Principle of ETROMI

Assume a WSN with N sensor nodes, in which one of the sensor nodes initiates a data packet with the intent to transmit it to the sink (final receiver base station). In the conventional clustering method, a CH collects the data from the sensors node in the corresponding cluster and forwards the data toward the sink via the other CHs. However, this approach is not efficient because the CHs suffer from battery limitations, even more than the other data collecting elements, but must be involved in all transmissions.

In contrast, the lifetime of the WSNs depends on the remaining energy of the members, i.e., the sensor nodes. Therefore, the number of forwarding nodes must be minimized. In this study, we assume that instead of always selecting the CHs to receive and forward the data packets, the sensors select the farthest sensor node in their transmission range. In other words, the sensor nodes increase their transmission power to transmit a data packet over a longer distance. In this configuration, the number of nodes that are involved in a transmission are minimized, which can considerably improve the energy efficiency. In Fig. 1, the red dotted line represents the routing procedure in a clustering-based method.

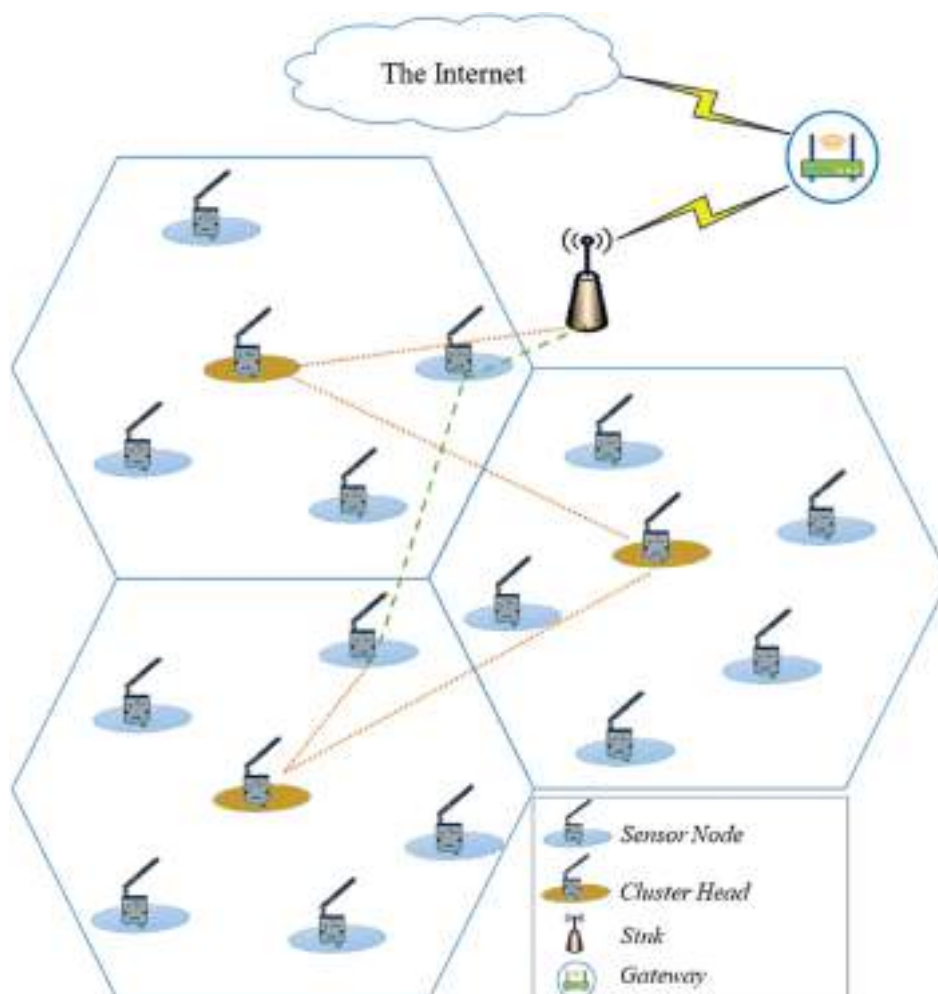


Figure 1: Routing procedure in WSNs

In this approach, the nodes send their packets to their corresponding CH, which then forwards the packet to the next CH and so on. Finally, the closest CH delivers the packet to the sink. In contrast, in the approach represented by the green dashed line, the node that initiates the packet sends the packet to the farthest node, and the receiver node follows the same principle and send the packet to the farthest node in its transmission range. Consequently, the number of nodes involved in the transmission procedure is less than that in the clustering-based method.

Consider a network involving 100 sensor nodes. Sensor 1 initiates a data packet and wants to send it to node 100. As mentioned earlier, in the WSNs, the topology is multi-hop. In other words, node 1 sends its packet to its neighbor, which receives the packet and forwards it to the neighboring nodes, excluding the node that the packet was received from. This process continues until node 100 receives the packet. The problem then is to determine the number of nodes in the transmission process that receive and forward the packet. This number of the relay nodes should be minimized to abate the energy consumption and, in turn, prolongs the network lifetime.

One solution is to increase the transmission range of the nodes involved in the transmission process from the source to the destination. In this case, a node selects a neighboring node, which is far from it, but in its range, i.e., on the edge of its transmission range, and the number of intermediate nodes is decreased. To this end, we consider the energy consumption accounted to transmission and reception of data packets and also the magnitude of data packets. The total required energy can be expressed as

$$E_t = E^T + P_i \times d_{ij} + W \times d^m + \sum_{k=1}^{n-1} D_k \tag{11}$$

The list of main symbols used in this paper are listed in [Tab. 1](#).

Table 1: List of symbols

Symbol	Definition
D_i	The distance between node i to the sink
E^I	The initial power
E_0	Total energy consumption of a link
E^{rx}	The receiving power
E^{tx}	Transmission power
P_i	Packet volume
d_{ij}	The distance between node i and j
N	The number of nodes

3.3 LP in ETROMI

Consider a WSN-based IoT represented by graph $G = (V, E)$, in which $V = v_1, v_2, v_3, \dots, v_n$ is the set of sensor nodes, and $E = e_1, e_2, e_3, \dots, e_n$ is the set of direct wireless links between the nodes, such that $E \subseteq V \times V$. Link (i, j) exists if and only if $j \in L_i$, where L_i is the set of all nodes that can be reached by sensor i directly with a certain transmission power level. Furthermore, each sensor i has the initial power E^I . The transmission energy consumed by node i to send a data packet to the neighboring sensor j is $E_{ij}^{tx} = \{e_{1j}^{tx}, e_{2j}^{tx}, e_{3j}^{tx}, \dots, e_{nj}^{tx}\}$; $E_{ij}^{rx} = \{e_{1j}^{rx}, e_{2j}^{rx}, e_{3j}^{rx}, \dots, e_{nj}^{rx}\}$ is the energy required for a node to receive a packet from node i ; and $P = \{p_1, p_2, p_3, \dots, p_n\}$ is the set of the packet volumes.

The objective function is to maximize the transmission distance with respect to the packet volume and the transmission and receiving energies, that is

$$\max \sum_{i=1}^n D_i \tag{12}$$

$$\text{s.t. } \sum_{i=1}^n D_i + \sum_{i=1}^n A_i \times E_i \tag{13}$$

$$\sum_{i=1}^n E_i \leq E^I \tag{14}$$

$$\sum_{i=1}^n P_i \leq P \tag{15}$$

$$E^I + E^T = E^I \tag{16}$$

Constraint (3) specifies that the total number of packets received or transmitted to/from node i must be less than a threshold for a specific time slot. Constraints (4) and (5) control the maximum transmission and receiving energy consumptions, respectively. Constraint (6) ensures that the total consumed energy for transmission and receiving by node i , is not less than the initial energy of the node.

4 Performance Evaluation

To evaluate the performance of the technique, we consider that a data packet is to be sent from a sensor node to the sink via two intermediate sensor nodes. Therefore, four sensor nodes are involved in the process: one sender, one sink, and two relay nodes. The size of each packet is 50 bits, and the transmission and receiving power are 100 and 70 W, respectively.

4.1 Linear Problem

The linear problem can be expressed as

$$\begin{aligned}
 \max \quad & x_1 + x_2 + x_3 & (1) \\
 \text{s.t.} \quad & x_1 + x_2 + x_3 \leq 100 & (2) \\
 & -2x_1 + 2x_2 - 2x_3 \leq 100 & (3) \\
 & 4x_1 - 2x_2 + x_3 \leq 70 & (4) \\
 & x_1, x_2, x_3 \geq 0 & (5)
 \end{aligned}$$

By adding slack variables to constraints, the primal problem in standard format is represented as follows:

$$\begin{aligned}
 \max \quad & x_1 + x_2 + x_3 + 0s_1 + 0s_2 + 0s_3 & (1) \\
 \text{s.t.} \quad & x_1 + x_2 + x_3 + s_1 = 100 & (2) \\
 & -2x_1 + 2x_2 - 2x_3 + s_2 = 100 & (3) \\
 & 4x_1 - 2x_2 + x_3 + s_3 = 70 & (4) \\
 & x_1, x_2, x_3, s_1, s_2, s_3 \geq 0 & (5)
 \end{aligned}$$

4.2 Simplex Method

We use the simplex method to solve the problem. The simplex method is used to solve LP models by using slack variables, tableaus, and pivot variables to determine the optimal solution of an optimization problem [27]. To solve the optimization problem, the following steps are performed:

- a) Obtain the standard form,
- b) Introduce slack variables,
- c) Create the tableau,
- d) Identify the pivot variables,
- e) Create a new tableau,
- f) Check for optimality,
- g) Identify the optimal values.

The procedure starts with an initialization phase, followed by several iterations to determine the optimal solution.

The initialization step for our optimization problem is presented in Tab. 2. After the first step, x_6 is the leaving variable, x_1 is the entering variable, and 4 is the pivot element.

Table 2: Stating section

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	RHS	Θ
$0x_4$	-1	-2	4	1	0	0	50	-
$0x_5$	-2	5	-2	0	1	0	100	-
$0x_6$	4	-2	-1	0	0	1	70	70/4
$C_j - Z_j$	1	1	1	0	0	0	0	

Subsequently, we apply the first iteration, as indicated in Tab. 3. Upon completing this iteration, the leaving variable is x_5 , the entering variable is x_2 , and the pivot element is 4.

Table 3: Iteration I

$0x_4$	0	-5/2	15/4	1	0	1/4	270/4	-
$0x_5$	0	4	-5/2	0	1	1/2	135	135/4
$1x_1$	1	-1/2	-1/4	0	0	1/4	70/4	
$C_j - Z_j$	0	3/2	5/4	0	0	-1/4	70/4	

We continue by applying the second iteration as indicated in Tab. 4, which results in a leaving variable x_4 , an entering variable x_3 , and a pivot element, 35/16.

Table 4: Iteration II

$0x_4$	0	0	35/16	1	5/8	9/16	1215/8	486/7
$1x_2$	0	1	-5/8	0	1/4	1/8	135/4	-
$1x_1$	1	0	-9/16	0	1/8	5/16	275/8	-
$C_j - Z_j$	0	0	35/16	0	-3/8	-7/16	545/8	

We proceed to the third iteration, in which all $C_j - Z_j$ values are zero or negative; therefore, the simplex method is terminated at this step, as indicated in Tab. 5. The optimal solution for the defined problem is presented in Tab. 6.

Table 5: Iteration III

$1x_3$	0	0	1	16/35	2/7	9/35	486/7	
$1x_2$	0	1	0	2/7	3/7	2/7	540/7	
$1x_1$	1	0	0	9/35	2/7	16/35	514/7	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

Table 6: The optimal solution

Z	220
x_1	514/7
x_2	540/7
x_3	486/7

4.3 Duality

The duality refers to a specific relationship between an LP problem and another problem, both of which involve the same original data, albeit located differently [28]. The former and latter problems are referred to as the primal and dual problems, respectively. The feasible regions, optimal solutions, and optimal values of these problems must be strongly correlated. The duality and optimality conditions obtained from these aspects are a basis for the LP theory. Once either of the primal or dual problems is solved, both the problems can be solved owing to duality. To convert the primal problem to a dual problem, the following steps are performed:

- a) If the primal problem corresponds to “Maximize,” the dual problem corresponds to “Minimize.”
- b) The number of variables in the dual problem is equal to the number of constraints in the primal problem.
- c) The number of constraints in the dual problem, is equal to the number of variables in the primal problem.
- d) The coefficients of the objective function in the dual problem, are equal to the right-hand side (RHS) values in the primal problem.
- e) The RHS values in the dual problem are equal to the coefficients of the objective function in the primal problem.
- f) The coefficient variables in the constraints of the dual problem correspond to the transpose matrix of the coefficient variables in the primal problem.
- g) “ \leq ” constraints in the primal problem are “ \geq ” constraints in the dual problem, and vice versa.
- h) The variables in the dual problem are denoted as “y”.
- i) The objective function is denoted as “w”. The primal problem is as follows:

Our primal problem is as below:

$$\begin{aligned} \text{max } w &= 50x_1 + 100x_2 + 70x_3 && \text{---(1)} \\ \text{s.t. } x_1 + 2x_2 + 4x_3 &\leq 100 && \text{---(2)} \\ -2x_1 + 5x_2 - 2x_3 &\leq 100 && \text{---(3)} \\ 4x_1 - 2x_2 - x_3 &\leq 70 && \text{---(4)} \\ x_1, x_2, x_3 &\geq 0 && \text{---(5)} \end{aligned}$$

Coefficient matrix of basic variables in objective function is $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$, Coefficient matrix of

basic variables in constraints is $\begin{bmatrix} -1 & -2 & 4 \\ -2 & 5 & -2 \\ 4 & -2 & -1 \end{bmatrix}$, and RHS matrix is $RHS = \begin{bmatrix} 50 \\ 100 \\ 70 \end{bmatrix}$. Based on

the aforementioned steps, our dual problem will be as bellow:

$$\begin{aligned} \text{min } w &= 100y_1 + 100y_2 + 70y_3 && \text{---(1)} \\ \text{s.t. } y_1 + 2y_2 + 4y_3 &\geq 50 && \text{---(2)} \\ -2y_1 + 5y_2 - 2y_3 &\geq 100 && \text{---(3)} \\ 4y_1 - 2y_2 - y_3 &\geq 70 && \text{---(4)} \\ y_1, y_2, y_3 &\geq 0 && \text{---(5)} \end{aligned}$$

By adding surplus variables, the dual problem is as follows:

$$\begin{aligned} \text{min } w &= 100y_1 + 100y_2 + 70y_3 + 0s_1 + 0s_2 + 0s_3 && \text{---(1)} \\ \text{s.t. } y_1 + 2y_2 + 4y_3 - s_1 &= 50 && \text{---(2)} \\ -2y_1 + 5y_2 - 2y_3 - s_2 &= 100 && \text{---(3)} \\ 4y_1 - 2y_2 - y_3 - s_3 &= 70 && \text{---(4)} \\ y_1, y_2, y_3, s_1, s_2, s_3 &\geq 0 && \text{---(5)} \end{aligned}$$

$$4x_1 - 2x_2 - x_3 - a_1 = 1 \tag{10}$$

$$x_1, x_2, x_3, a_1, a_2, a_3 \in \mathbb{R} \tag{11}$$

As indicated in the dual problem, no identity matrix exists for the coefficients of the variables in the constraints; therefore, artificial variables must be introduced. In this case, the dual problem in the standard format is:

$$\text{max } -50y_1 + 100y_2 + 70y_3 + 0y_4 + 0y_5 + 0y_6 - M a_1 - M a_2 - M a_3 \tag{12}$$

$$a_1 - y_1 - 2y_2 + 4y_3 + a_4 = 1 \tag{13}$$

$$-2y_1 + 5y_2 - 2y_3 + a_5 = 1 \tag{14}$$

$$4x_1 - 2x_2 - x_3 + a_6 = 1 \tag{15}$$

$$x_1, x_2, x_3, a_4, a_5, a_6 \in \mathbb{R} \tag{16}$$

By adding artificial variables, an identity matrix can be generated, and the simplex method can be implemented.

As indicated in Tab. 7, after completing the initialization section, the leaving variable is a_3 , the entering variable is x_1 , and the pivot element is 4. Subsequently, we implement the first iteration, as indicated in Tab. 8. Upon completing iteration I, the leaving variable is a_2 , the entering variable is x_2 , and our pivot element is 4. We then proceed to the second iteration, as indicated in Tab. 9. After the second iteration, the leaving variable is a_1 , the entering variable is x_3 , and the pivot element is 35/16. We attempt to determine the optimal solution by using the two-phase simplex method. All the artificial variables are removed, and the problem can be solved through the other variables. We then apply the third iteration in two phases, as indicated in Tabs. 10 and 11.

Table 7: Starting section

Min	50	100	70	0	0	0	-M	-M	-M	RHS	Θ
	y_1	y_2	y_3	y_4	y_5	y_6	a_1	a_2	a_3		
-Ma ₁	-1	-2	4	-1	0	0	1	0	0	1	-
-Ma ₂	-2	5	-2	0	-1	0	0	1	0	1	-
-Ma ₃	4	-2	-1	0	0	-1	0	0	1	1	1/4
$C_j - W_j$	-1	-1	-1	1	1	1	0	0	0	3	

Table 8: Iteration I

-Ma ₁	0	-9/4	15/4	-1	0	-1/4	1	0	1/4	5/4	-
-Ma ₂	0	4	-9/4	0	-1	-1/2	0	1	1/2	3/2	3/8
50y ₁	1	-1/2	-1/4	0	0	-1/4	0	0	1/4	1/4	-
$C_j - W_j$	0	-3/2	-5/4	1	1	3/4	0	0	1/4	11/4	

Table 9: Iteration II

-Ma ₁	0	0	35/16	-1	-5/8	-9/16	1	5/8	9/16	35/16	1
100y ₂	0	1	-5/8	0	-1/4	-1/8	0	1/4	1/8	3/8	-
50y ₁	1	0	-9/16	0	-1/8	-5/16	0	1/8	5/16	7/16	-
$C_j - W_j$	0	0	-35/16	1	5/8	9/16	0	3/8	7/16	35/16	

Table 10: Phase I, Iteration III

70y ₃	0	0	1	-16/35	-2/7	-9/35	16/35	2/7	9/35	1
100y ₂	0	1	0	-2/7	-3/7	-2/7	2/7	3/7	2/7	1
50y ₁	1	0	0	-9/35	-2/7	-16/35	9/35	2/7	16/35	1
$C_j - W_j$	0	0	0	0	0	0	1	1	1	0

Table 11: Phase II, Iteration III

$70y_3$	0	0	1	$-16/35$	$-2/7$	$-9/35$	1
$100y_2$	0	1	0	$-2/7$	$-3/7$	$-2/7$	1
$1y_1$	1	0	0	$-9/35$	$-2/7$	$-16/35$	1
$C_j - W_j$	0	0	0	$514/7$	$540/7$	$486/7$	220

Finally, it is observed that the primal solution, presented in Tab. 12, is equal to the dual solution, presented in Tab. 13, that is $Z^* = W^*$.

Table 12: Primal optimal Solution

Z	220
x_1	$514/7$
x_2	$540/7$
x_3	$486/7$

Table 13: Dual optimal solution

W	220
y_1	1
y_2	1
y_3	1

4.4 Sensitivity Analysis

Sensitivity analysis is aimed at examining the influence of changes in the variables, such as the RHS, coefficients of the objective function, and constraints, on the solution. We start with Tab. 14 and make the some changes as explained in the next subsection.

Table 14: Simplex optimum tableau

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	RHS
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$486/7$
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$540/7$
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$514/7$
$C_j - Z_j$	0	0	0	-1	-1	-1	220

4.4.1 Change in the Objective Function Coefficient for Non-Basic Variables

In the last iteration, no non-basic variables of the objective function exist. Therefore, if one of the coefficients is changed, the optimal solution is not influenced.

4.4.2 Change in the RHS Value

Suppose the intention is to change the first RHS to b_1 ; then we have $\begin{bmatrix} 50 \\ 100 \\ 70 \end{bmatrix}$. To calculate new RHS;

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} b_1 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} 16b_1 + 1638 \\ 2b_1 + 240 \\ 9b_1 + 2129 \end{bmatrix} \quad (27)$$

$$\begin{cases} 16b_1 + 1638 \geq 0 \Rightarrow b_1 \geq -102 \\ 2b_1 + 240 \geq 0 \Rightarrow b_1 \geq -120 \\ 9b_1 + 2129 \geq 0 \Rightarrow b_1 \geq -236 \end{cases} \quad (28)$$

Because, $b_1 \geq -815/8$. We suppose a b_1 value beyond the specified range; as an example -110 .

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} -110 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} -278 \\ 220 \\ 226 \end{bmatrix} \quad (29)$$

Now, we continue the tableau with new RHS values;

As indicated in Tab. 15, the primal solution is not feasible; therefore, we attempt to find the optimal solution through the dual problem.

Table 15: New RHS values

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	RHS	Θ
$1x_3$	0	0	1	16/35	2/7	9/35	278/7	
$1x_2$	0	1	0	2/7	3/7	2/7	220/7	
$1x_1$	1	0	0	9/35	2/7	16/35	226/7	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

The initialization step, as the first iteration, is presented in Tab. 16. After completing the initialization step, the leaving variable is x_1 , the entering variable is x_5 , and the pivot element is $3/7$.

Table 16: Initialization step

$1x_3$	0	0	1	16/35	2/7	9/35	278/7
$1x_2$	0	1	0	2/7	3/7	2/7	220/7
$1x_1$	1	0	0	9/35	2/7	16/35	226/7
$C_j - Z_j$	0	0	0	-1	-1	-1	220
Θ	-	0	-	-7/2	-7/3	-7/2	

Subsequently, we implement the second iteration, as indicated in Tab. 17. After finishing the second iteration, it is noted that the primal is feasible; the leaving variable is x_5 , the entering variable is x_2 , and the pivot element is $7/3$.

Table 17: Iteration II

$1x_3$	0	-2/3	1	28/105	0	7/105	394/21	-
$0x_5$	0	7/3	0	2/3	1	2/3	220/3	220/7
$1x_1$	1	-2/3	0	7/105	0	28/105	34/3	-
$C_j - Z_j$	0	7	0	-1/3	0	-1/3	30.1	

We then implement the third iteration as indicated in Tab. 18. All the values for $C_j - Z_j$ are zero or negative; therefore, the process is terminated at this step. The optimal solution is as presented in Tab. 19.

Table 18: Iteration III

$1x_3$	0	0	1	48/105	6/21	37/105	278/7
$1x_2$	0	1	0	2/7	3/7	2/7	220/7
$1x_1$	1	0	0	27/105	6/21	48/105	226/7
$C_j - Z_j$	0	0	0	-1	-1	-115/105	724/7

Table 19: Optimal solution

Z	724/7
x_1	226/7
x_2	220/7
x_3	278/7

4.4.3 Change in the Objective Function Coefficient for the Basic Variable

We consider the case in which the coefficient of x_1 changes. Suppose the coefficient of x_1 is c_1 .

Any change in the coefficient of the basic variables of the objective function affects the value of $C_j - Z_j$.

$$\text{For } x_3 = C_j - Z_j = 0 - \left[\left(1 + \frac{36}{105}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{9}{105}\right) \right] = -\frac{26}{15} - \frac{26}{15} \tag{39}$$

$$\text{For } x_2 = C_j - Z_j = 0 - \left[\left(1 + \frac{2}{7}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{2}{7}\right) \right] = -\frac{26}{7} - \frac{2}{7} \tag{40}$$

$$\text{For } x_1 = C_j - Z_j = 0 - \left[\left(1 + \frac{27}{105}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{18}{105}\right) \right] = -\frac{18c_1}{15} - \frac{18}{15} \tag{41}$$

If $C_j - Z_j \leq 0$ then the present solution remains optimal solution;

$$\frac{-26}{15} - \frac{26}{15} \geq 0 \text{ for } c_1 \geq -\frac{26}{9} \tag{42}$$

$$\frac{-26}{7} - \frac{2}{7} \geq 0 \text{ for } c_1 \geq -\frac{2}{7} \tag{43}$$

$$\frac{-18c_1}{15} - \frac{18}{15} \geq 0 \text{ for } c_1 \leq -\frac{18}{18} \tag{44}$$

In this case, the range of c_1 is greater than $-26/9$. Thus, we assign c_1 beyond this range, for example $c_1 = -4$, and implement the first iteration, as indicated in Tab. 20.

Table 20: Iteration I

Maximize	-4	1	1	0	0	0	RHS	Θ
	x_1	x_2	x_3	x_4	x_5	x_6		
$1x_3$	0	0	1	16/35	2/7	9/35	486/7	270
$1x_2$	0	1	0	2/7	3/7	2/7	540/7	270
$-4x_1$	1	0	0	9/35	2/7	16/35	514/7	1285/8
$C_j - Z_j$	0	0	0	2/7	3/7	9/7	-1030/7	

Upon completing the first iteration, the leaving variable is x_3 , the entering variable is x_6 , and the pivot element is 9/35.

The second iteration is presented in Tab. 21. All the values for $C_j - Z_j$ are zero or negative; therefore, the process is terminated at this step. We conclude that the optimal solution is as follows: $x_1 = -50, x_2 = 0, x_3 = 270$ and $z = 470$.

Table 21: Iteration II

$1x_3$	0	0	$35/9$	$16/9$	$10/9$	1	270
$1x_2$	0	1	$-10/9$	$-2/9$	$1/9$	0	0
$-4x_1$	1	0	$-16/9$	$-7/9$	$-2/9$	0	-50
$C_j - Z_j$	0	0	$-80/9$	$-42/9$	$-19/9$	-1	470

4.4.4 Change in the Constraint Coefficient Corresponding to Non-basic Variables

In the last iteration, no non-basic variable of the objective function exists. Therefore, if one of the coefficients is changed, the optimal solution is not influenced.

4.4.5 Addition of a New Variable

Consider a new variable x_7 with coefficient $c_7 = 12$ and $P_7 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}$, then;

$$B^{-1}P_7 = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 212/35 \\ 12/7 \\ 61/35 \end{bmatrix} \quad (40)$$

In this case, we perform three iterations as indicated in Tabs. 22–24.

Table 22: Iteration I

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	RHS	Θ
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$486/7$	
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$540/7$	
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$514/7$	
$C_j - Z_j$	0	0	0	-1	-1	-1	220	

Table 23: Iteration II

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	12	RHS	Θ
$1x_3$	0	0	1	$16/35$	$2/7$	$9/35$	$212/35$	$486/7$	$104/9$
$1x_2$	0	1	0	$2/7$	$3/7$	$2/7$	$12/7$	$540/7$	45
$1x_1$	1	0	0	$9/35$	$2/7$	$16/35$	$61/35$	$514/7$	$2570/61$
$C_j - Z_j$	0	0	0	-1	-1	-1	$137/105$		

Table 24: Iteration III

Maximize	$\frac{1}{x_1}$	$\frac{1}{x_2}$	$\frac{1}{x_3}$	$\frac{0}{x_4}$	$\frac{0}{x_5}$	$\frac{0}{x_6}$	12	RHS	Θ
$12x_7$	0	0	$3/18$	$8/9$	$15/9$	$27/18$	1	$104/18$	
$1x_2$	1	0	$37/127$	$68/319$	$65/319$	$67/319$	0	$36/18$	
$1x_1$	0	1	$-18/63$	$-78/63$	$-153/63$	$-144/63$	0	$235/18$	
$C_j - Z_j$	0	0	-1	$-87/9$	$-160/9$	$-143/9$	0	115	

Upon completing the first iteration, the leaving variable is x_3 , the entering variable is x_7 , and the pivot element is $212/35$.

The third iteration is presented in Tab. 12, in which all the values for $C_j - Z_j$ are zero or negative and; therefore, the program is terminated at this step, and the optimal solution is as indicated in Tab. 25.

Table 25: Optimal solution

Z	115
x_1	$235/18$
x_2	$36/18$
x_7	$104/18$

4.4.6 Addition of a New Constraint

To examine the influence of the addition of a new constraint to the problem, we consider $x_3 \leq 40$:

As indicated in Tab. 26, the optimal solution is as follows: $x_1 = 514/7$, $x_2 = 540/7$, $x_3 = 486/7$, and $Z = 20$.

Table 26: Additional constraint

Maximize	1	1	1	0	0	0	0	RHS
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	
$1x_3$	0	0	1	16/35	2/7	9/35	0	486/7
$1x_2$	0	1	0	2/7	3/7	2/7	0	540/7
$1x_1$	1	0	0	9/35	2/7	16/35	0	514/7
$0x_7$	0	0	1	0	0	0	1	40
$C_j - Z_j$	0	0	0	-1	-1	-1	0	220

5 Conclusion and Future Direction

The transmission range of a sensor node defines whether the communication mode is single-hop or multi-hop. In this paper, we proposed the use of ETROMI, which can determine the maximum distance to which a sensor node can transmit data with the least possible number of relay nodes. We presented an LP-based analytical model to determine the transmission range of the sensor node. Moreover, we explained the mathematical model associated with the ETROMI to reduce the energy consumption of WSN-based IoT. A key concern about the ETROMI is that it considers the ideal conditions involving no obstacles between the sensor nodes and the sink. Therefore, the model performance is specific to the circumstances. Furthermore, the network is assumed to be homogeneous, whereas homogeneity does not exist in an actual network due to the different factors associated with network deployment. In future work, we aim to extend our work to address the aforementioned scenarios.

Funding Statement: This research was supported by Korea Electric Power Corporation (Grant Number: R18XA02).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. S. Kumar and V. K. Chaurasiya. (2018). "A strategy for elimination of data redundancy in Internet of Things (IoT) based wireless sensor network (WSN)," *IEEE Systems Journal*, vol. 13, pp. 1650–1657.
2. P. Swarna, P. Maddikunta, M. Parimala, S. Koppu, T. Gadekallu et al. (2020). , "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," *Computer Communications*, vol. 160, pp. 139–149.
3. R. Vinayakumar, M. Alazab, S. Srinivasan, Q. Pham, S. Padannayil et al. (2020). , "A visualized bot net detection system based deep learning for the Internet of Things networks of smart cities," *IEEE*

Transactions on Industry Applications, vol. 56, no. 4, pp. 4436–4456.

4. M. Piran, Y. Cho, J. Yun and D. Y. Suh. (2014). “Cognitive radio-based vehicular ad hoc and sensor networks (CR-VASNET),” *International Journal of Distributed Sensor Networks*, vol. 2014, pp. 1–11.
5. T. M. Behera, S. K. Mohapatra, U. C. Samal and M. S. Khan. (2019). “Hybrid heterogeneous routing scheme for improved network performance in WSNs for animal tracking,” *Internet of Things*, vol. 6, pp. 1–9.
6. T. M. Behera, S. K. Mohapatra, U. C. Samal, M. S. Khan, M. Daneshmand et al. (2019). , “Residual energy-based cluster-head selection in WSNs for IoT application,” *IEEE Internet of Things Journal*, vol. 6, pp. 5132–5139.
7. S. Verma, N. Sood and A. K. Sharma. (2019). “A novelistic approach for energy efficient routing using single and multiple data sinks in heterogeneous wireless sensor network,” *Peer-to-Peer Networking and Applications*, vol. 12, pp. 1110–1136.
8. Y. Liu, C. Yang, L. Jiang, S. Xie and Y. Zhang. (2019). “Intelligent edge computing for IoT-based energy management in smart cities,” *IEEE Network*, vol. 33, pp. 111–117.
9. D. K. Gupta. (2013). “A review on wireless sensor networks,” *Network and Complex Systems*, vol. 3, no. 1, pp. 18–23.
10. L. Krishnasamy, R. K. Dhanaraj, G. D. Ganesh, G. Reddy, M. K. Aboudaif et al. (2020). , “A heuristic angular clustering framework for secured statistical data aggregation in sensor networks,” *Sensors*, vol. 20, pp. 1–15.
11. S. Bhattacharya, P. Maddikunta, S. Somayaji, K. Lakshmana, R. Kaluri et al. (2020). , “Load balancing of energy cloud using wind driven and firefly algorithms in internet of everything,” *Journal of Parallel and Distributed Computing*, vol. 142, pp. 16–26.
12. C. Iwendi, P. K. Maddikunta, T. R. Gadekallu, K. Lakshmana, A. K. Bashir et al. (2020). , “A metaheuristic optimization approach for energy efficiency in the IoT networks,” *Software: Practice and Experience*, vol. 22, no. 6, pp. 1–14.
13. H. Asharioun, H. Asadollahi, T. C. Wan and N. Gharaei. (2015). “A survey on analytical modeling and mitigation techniques for the energy hole problem in corona-based wireless sensor network,” *Wireless Personal Communications*, vol. 81, pp. 161–187.
14. S. Verma, N. Sood and A. K. Sharma. (2019). “Genetic algorithm-based optimized cluster head selection for single and multiple data sinks in heterogeneous wireless sensor network,” *Applied Soft Computing*, vol. 85, pp. 1–21.
15. A. U. Rahman, A. Alharby, H. Hasbullah and K. Almuzaini. (2016). “Corona based deployment strategies in wireless sensor network: A survey,” *Journal of Network and Computer Applications*, vol. 64, pp. 176–193.
16. D. Bertsimas and J. N. Tsitsiklis. (1997). *Introduction to Linear Optimization*, vol. 6. Belmont, MA: Athena Scientific.
17. V. Tabus, D. Moltchanov, Y. Koucheryavy, I. Tabus and J. Astola. (2015). “Energy efficient wireless sensor networks using linear-programming optimization of the communication schedule,” *Journal of Communications and Networks*, vol. 17, pp. 184–197.
18. V. Sandeep, N. Sood and A. K. Sharma. (2019). “QoS provisioning-based routing protocols using multiple data sink in IoT-based WSN,” *Modern Physics Letters*, vol. 34, pp. 1–36.
19. R. Elkamel, A. Messouadi and A. Cherif. (2019). “Extending the lifetime of wireless sensor networks through mitigating the hot spot problem,” *Journal of Parallel and Distributed Computing*, vol. 133, pp. 159–169.
20. C. Sha, C. Ren, R. Malekian, M. Wu, H. Huang et al. (2019). , “A type of virtual force-based energy-hole mitigation strategy for sensor networks,” *IEEE Sensors Journal*, vol. 20, pp. 1105–1119.
21. N. Sharmin, A. Karmaker, W. Lambert, M. Alam and M. Shawkat. (2020). “Minimizing the energy hole problem in wireless sensor networks: A wedge merging approach,” *Sensors*, vol. 20, pp. 1–25.
22. B. Sahoo, T. Amgoth and H. Pandey. (2020). “Particle swarm optimization based energy efficient clustering and sink mobility in heterogeneous wireless sensor network,” *Ad Hoc Networks*, vol. 106, pp. 1–21.
23. S. Kaur and V. Grewal. (2020). “A novel approach for particle swarm optimization-based clustering with dual sink mobility in wireless sensor network,” *International Journal of Communication Systems*, vol. 33, no. 16, pp. 1–
24. M. Liu and C. Song. (2012). “Ant-based transmission range assignment scheme for energy hole problem in wireless sensor networks,” *International Journal of Distributed Sensor Networks*, vol. 8, pp. 1–12.

25. X. Liu. (2016). "A novel transmission range adjustment strategy for energy hole avoiding in wireless sensor networks," *Journal of Network and Computer Applications*, vol. 67, pp. 43–52.
26. H. Xin and X. Liu. (2017). "Energy-balanced transmission with accurate distances for strip-based wireless sensor networks," *IEEE Access*, vol. 5, pp. 16193–16204.
27. V. Zhadan. (2019). "Two-phase simplex method for linear semidefinite optimization," *Optimization Letters*, vol. 13, pp. 1969–1984.
28. S. Nasser and D. Darvishi. (2018). "Duality results on grey linear programming problems," *The Journal of Grey System*, vol. 30, pp. 127–142.



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Detection and robustness evaluation of android malware classifiers

M. L. Anupama¹ · P. Vinod² · Corrado Aaron Visaggio³ · M. A. Arya¹ · Josna Philomina¹ · Rincy Raphael⁴ · Anson Pinhero¹ · K. S. Ajith¹ · P. Mathiyalagan⁴

Received: 7 August 2020 / Accepted: 31 May 2021 / Published online: 26 June 2021
© The Author(s), under exclusive licence to Springer-Verlag France SAS, part of Springer Nature 2021

Abstract

Android malware attacks are tremendously increasing, and evasion techniques become more and more effective. For this reason, it is necessary to continuously improve the detection performances. With this paper, we wish to pursue this purpose with two contributions. On one hand, we aim at evaluating how improving machine learning-based malware detectors, and on the other hand, we investigate to which extent adversarial attacks can deteriorate the performances of the classifiers. Analysis of malware samples is performed using static and dynamic analysis. This paper proposes a framework for integrating both static and dynamic features trained on machine learning methods and deep neural network. On employing machine learning algorithms, we obtain an accuracy of 97.59% with static features using SVM, and 95.64% is reached with dynamic features using Random forest. Additionally, a 100% accuracy was obtained with CART and SVM using hybrid attributes (on combining relevant static and dynamic features). Further, using deep neural network models, experimental results showed an accuracy of 99.28% using static features, 94.61% using dynamic attributes, and 99.59% by combining both static and dynamic features (also known as multi-modal attributes). Besides, we evaluated the robustness of classifiers against evasion and poisoning attack. In particular comprehensive analysis was performed using permission, APIs, app components and system calls (especially n -grams of system calls). We noticed that the performances of the classifiers significantly dropped while simulating evasion attack using static features, and in some cases 100% of adversarial examples were wrongly labelled by the classification models. Additionally, we show that models trained using dynamic features are also vulnerable to attack, however they exhibit more resilience than a classifier built on static features.

Keywords Static features · Dynamic features · Hybrid features · Fisher score · Adversarial examples · Attack models

1 Introduction

Malicious code is a software intentionally written for bypassing security controls and performing unauthorized actions that are not allowed to the attacker and can cause a damage to the victim. The techniques for analyzing malicious code can be divided into static analysis and dynamic analysis. Static analysis techniques scan the source code and don't require

✉ Corrado Aaron Visaggio
visaggio@unisannio.it

M. L. Anupama
anupama.ml@scmsgroup.org

P. Vinod
vinod.p@cusat.ac.in

M. A. Arya
aryanand54@gmail.com

Josna Philomina
josnaphilomina@scmsgroup.org

Rincy Raphael
rincyraphael2019@srec.ac.in

K. S. Ajith
ajithks273@gmail.com

P. Mathiyalagan
mathiyalagan.p@srec.ac.in

¹ Present Address: Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Cochin, India

² Department of Computer Applications, Cochin University of Science and Technology, Cochin, India

³ Department of Engineering, University of Sannio, Benevento, Italy

⁴ Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, Affiliated by Anna University, Chennai, India

the execution of the programs to be examined. Thus, the study can be conducted without compromising the systems. Static analysis gained wider acceptance amongst the analysts as it is quick and harmless, even though encryption, obfuscation and the use of runtime libraries obstruct the static analysis. Dynamic analysis, on the contrary, aims to uncover the runtime behaviour of the application by executing the application on the real device or in a sandbox environment [8]. Dynamic analysis is not limited by code obfuscation and can provide details about the malware behavior.

By combining both static and dynamic analysis, it is possible to leverage the advantages of both approaches: malware scanners that use both the types of analysis are generally known as hybrid malware detectors. Static analysis is conducted by extracting structural features from the file, while dynamic analysis uses features that require the execution of the app, like system calls, network traces, and control flow graphs.

Despite the large literature investigating the advantages and limitations of using machine learning for detecting malware, further studies are necessary for consolidating the body of knowledge on this topic and removing all the uncertainties research pointed out so far, for different reasons. Recent works collect evidence that anti-malware tools are diminishing their ability to recognize malware, due mainly to the rapid increment of variants [20,40,50]. Spatial and temporal bias can make untrustworthy some results, since training or testing sets are not completely representative of the malware (and goodware) population [34]. Adversarial attacks could easily deteriorate the robustness of machine and deep learning based classifiers [10,19], while there is not a complete convergence about which are the best machine learning algorithms for malware detection [16,44]. For this reason with this paper we aim at providing a two-fold contribution to the state of the art: adding further evidence about the performance of machine and deep learning algorithms in detecting malware, and studying to which extent adversarial samples may alter the effectiveness of classifiers.

More in detail, in this work we explore the usage of the Fisher score [52] to select the most relevant attributes for the classifiers. The features obtained are used to build diverse classifiers using Logistic Regression, Classification and Regression Trees, Random Forest and Support Vector Machine algorithms. A comprehensive analysis of the machine learning models is conducted to identify the optimal classification model that can be deployed for detecting unseen or future samples. Finally, we realized three attack models which leverage adversarial examples and evaluate how the classifiers performances degrade. We observed that a minor perturbation of attributes significantly dropped the detection rate, and all the modified malware samples (tainted/adversarial examples) bypassed the detection.

Finally, the main contributions of this research work are as follows:

- We implement a feature selection algorithm based on Fisher score for ranking attributes, and show that classifiers trained on the relevant attributes selected in this way can improve the detection rate.
- We create multi-modal features (hybrid features) classifiers and obtain an accuracy of 100% with CART, SVM, and an accuracy of 99.59% with deep neural network.
- We realize three attack models based on hamming distance, k-means and app's components for creating adversarial samples. These specimens are created by inserting permissions and app's components into malicious apps. We observed that classifiers' performance dropped drastically. In particular, Hamming distance based attack increases the average False Positive Rate of machine learning classifiers and deep neural network by 55.86% and 45.94% respectively. All the adversarial samples developed using k-means clustering are successfully evaded ($FNR = 100\%$). Finally, 90.13% and 100% tainted applications created by injecting especially crafted app's components deceived classifiers based on machine learning approaches and deep neural network.

The paper is organized as follows. Section 2 discusses the related work. In Sect. 3, proposed methodology is presented. The adversarial attacks are introduced in Sect. 4 while the attacks are elaborated in Sect. 5. The experiments and obtained results are given in Sect. 6. Evaluation on obfuscated samples are discussed in Sect. 7. Finally, the concluding remarks and direction for future work is given in Sect. 8.

2 Related work

This section discusses existing malware detection and classification models based on both machine learning and deep learning. Patel *et al.* [33] proposed a hybrid android malware detection system. It extracts both permission and behaviour-based features. Then, performed feature selection using information gain. Finally, rule generation module classifies applications as benign or malicious. In [46], authors have mentioned another hybrid malware detector that uses SVM classifier to classify app as benign or malware. It detects zero-day malware with a true positive rate of 98.76%. Damodaran *et al.* [16] conducted a comparative analysis on malware detection system employing static, dynamic, and hybrid analysis. They found that behavioural data produce an highest AUC of 0.98 using Hidden Markov Models (HMMs) trained on 785 samples. In [47], authors initially utilize APIMonitor to obtain static features from apps. Then, it involves the usage of APE_BOX to obtain dynamic features. Finally, they

apply SVM for classification. MADAM [38] demonstrated how KNN classifier can achieve 96.9% detection rate.

Significant Permission Identification, SigPID [27] is another malware detection system that uses a three-layered pruning by mining the permission data to identify the most significant permissions that result in differentiating benign and malicious apps. It then uses machine-learning classifiers (SVM and decision tree) for classification and achieved over 90% of detection accuracy. In [15], authors initially disassemble applications by using Androguard to obtain the frequency of API calls used by the application. Finally, it is observed that a particular set of APIs is more frequent in malicious apps. It can detect malicious apps with 96.69% accuracy and 95.25% detection rate, by using SVM.

Crowdroid [11] is an Android malware detector which uses dynamic analysis and then employs two-means clustering algorithm for classifying benign and malicious apps. In [18], authors have presented an Android Malware Detection system which extracts system calls by executing the applications in a sandbox environment. They implemented their approach in MALINE tool and can detect malware with low rates of false positives by employing machine learning algorithms. Afonso *et al.* [2] propose another android malware detection system that uses dynamic features such as API calls and system call traces along with machine learning to identify malware with high detection rate.

Authors in [21] present a machine-learning-based Android malware detection and family identification approach, RevealDroid, that aims at reducing the sets of features used in the classifiers. This approach leverages categorized Android API usage, reflection-based features, and features from native binaries of apps. Besides accuracy and efficiency, authors evaluate also obfuscation resilience using a large dataset of more than 54,000 malicious and benign apps. The experimental results show an accuracy of 98.

Tam *et al.* [43] propose a mechanism for reconstructing behaviors of Android malware by observing and dissecting system calls. This mechanism allows CopperDroid to obtain events of interest, especially intra- and inter-process communications. This makes CopperDroid agnostic to the underlying invocation methods. Experimental results showed that CopperDroid discloses additional behaviors on more than 60% of the analyzed dataset.

In [4] authors analyze the permissions used by an application that requires during installation. It uses clustering and classification techniques and also allows user to identify malicious applications installed on the phone and also provides a provision to remove them. The drawback of this system is that if a new unknown family of a malware is supposed to be detected then a new cluster has to be created considering the same family's permission. CSCdroid [51] builds a Markov chain by using system calls. Then, it constructs the target feature vector from the probability matrix.

Finally, it uses the Support Vector Machine classifier to detect malware, achieving an F1-score of 98.11% and a true positive rate of 97.23%.

Kimet *al.* [25] propose an Android malware detection method, that uses opcode features, API features, strings, permissions, app's components, and environmental features, to generate a multimodal malware detection model. With these static features, they trained their initial networks. Later, they trained the final network, with initial network results. The model produces an accuracy of 98%. Paper [41] proposes a malware detection model-based on RNN and CNN. It involves the usage of the static feature opcode. Finally authors conclude that their accuracy exceeds 92%, for even small training datasets. Malware Classification using SimHash and CNN, MCSC [32] is a model leveraging opcode sequences as static features, that combine malware visualization, and deep learning techniques, resulting in a classification accuracy of 99.26%.

In [39] authors propose a deep neural network-based malware detector using static features. It consists of three components, the first component extracts features, the second component is a DNN classifier, and the final component is a score calibrator which translates the output of a NN to a score. Achieved 95% detection rate, at 0.1% false-positive rate (FPR). MalDozer [24] is another highly accurate malware detection model that relies on deep learning techniques and raw sequences of API method calls. Deep android malware detection [29] is another model developed based on the static analysis of the raw opcode sequence from a disassembled program. Features indicative of malware are automatically learned by the network from the raw opcode sequence thus removing the need for hand-engineered malware features. This model has proposed a much simpler training pipeline.

A comprehensive analysis and comparison of deep neural networks(DNNs) and various classical machine learning algorithms for static malware detection are discussed in [45]. The authors have concluded that DNNs perform comparably well and are well suited to address the problem of malware detection using static PE features. A malware classification method using Visualization and deep learning is mentioned in [26]. It requires no expert domain knowledge. Initially, the files are visualized as grayscale images then experimented on deep learning architectures involving different combinations of Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). A deep learning approach that amends the convolutional deep learning models to use the support vector machine is presented in [3]. The authors have finally concluded that, among their three models, the model with 5 layers has the best accuracy compared to those with 2 and 3 layers.

A CNN based windows malware detector that uses API calls and their corresponding category as dynamic features that finally resulted in the achievement of 97.97% accuracy

for the N-grams counselled by the Relief Feature Selection Technique is described in [36]. In [28], the authors have designed a method based on a convolutional neural network applied to the system calls occurrences through dynamic analysis. They obtained an accuracy ranging between 0.85 and 0.95.

In [48], authors have presented a method based on back-propagation neural network to detect malware. It builds Markov chains from system call sequences and then applies the back-propagation neural network to detect malware. They experimented on a dataset of 1,189 benign ones, and 1,227 malicious applications and obtained an F1-score of 0.983.

KuafuDet [14] is a two-phase detection system, where features are extracted in the first phase and the second phase is an online detection phase. Camouflaged malicious applications which is a form of adversarial examples are developed, and similarity-based filtering is used to identify false negatives.

Xu et al. [49] applied genetic programming for evading PDF malware classifiers. It uses the probabilities assigned by the classifiers to estimate the fitness of variants. PDFrate and Hidost were the two PDF classifiers used for the evaluation. Authors reported 500 evasive variants created in 6 days. The evaluation of adversarial attacks were performed on Android malware detectors. Authors in [13] proposed a system called DroidEye which extracts features from Android apps and represents each observation as a binary vector. Further, they evaluated the attack on standard classifiers used for identifying malware. To improve the robustness of these classifiers, they transformed the binary vectors as continuous probabilities. Experiments were performed on samples collected from Comodo Cloud Security Center, and reported that DroidEye improved the security of system without effecting the detection performance. Adversarial crafting attacks on neural network were experimented in [23]. The attack was demonstrated on malware detection system trained on DREBIN dataset, where each application was represented as a binary vector. They reported a classification accuracy of 97% with FNR value of 7.6%. The trained model was subjected with adversarial examples generated by modifying AndroidManifest.xml and achieved a misclassification rate of 85%. In addition, they hardened neural network using adversarial training and defensive distillation, and reported that the later approach reduced the misclassification rates. Comprehensive experiments considering permissions [12] were performed for binary classification (malware vs benign) and multi class classification. Their study demonstrated that carefully selecting permissions can lead to accurate detection and classification. Further, to evaluate robustness of permission-based detection, top benign permissions were added to the malicious applications. They showed that a small number of requested benign permissions decreases ANN performance. However, ANN recovers on larger permission

request, indicating identical performance as observed with unmodified malware applications.

Demontis et al. [17] developed an adversary-aware machine learning detector against evasion attacks. Authors propose a secure learning solution which is able to retain computational efficiency and scalability on large datasets. The method outperforms state of the art classification algorithms, without loss of accuracy when there aren't well-crafted attacks.

Pierazzi et al. [35] propose a formalization of problem-space attacks. They uncover new relationships between feature space and problem space, providing necessary and sufficient conditions for the existence of problem-space attacks. This work shows that adversarial malware can be produced automatically.

In our work, we build machine learning and deep learning models using static, dynamic and hybrid techniques. We found that DNN obtained better performance using hybrid features. Further, we conducted comprehensive analysis on adversarial attacks by proposing three approaches for creating adversarial examples, and conclude that malware classifiers can be easily defeated by introducing tiny perturbations.

The Table 13 summarizes the main contributions of each analyzed work.

3 Methodology

This section describes the methods used by the hybrid malware detector that we will study.

3.1 Static analysis

An Android application explicitly requires the user to approve the necessary permissions during the installation. As a consequence, the collection of permissions can reflect the application behaviour. Standard and non-standard permissions may be extracted from AndroidManifest.xml file using Android Asset Packaging Tool (aapt) command. We developed a parser to read the manifest file to extract `<user-permission>` and the `<permission>` tag containing the permission name. Besides, our parser captures the application components: activities, services, broadcast receivers, and content providers are obtained by decompiling the given app using APKTool [6].

3.2 Dynamic analysis

We use dynamic analysis for capturing the sequence of system calls while the application executes and interacts with the operating system. Given an Android application, the procedure for extracting system calls is shown in Fig. 1. Initially the application is installed in Nexus_5_API_22 Android emula-

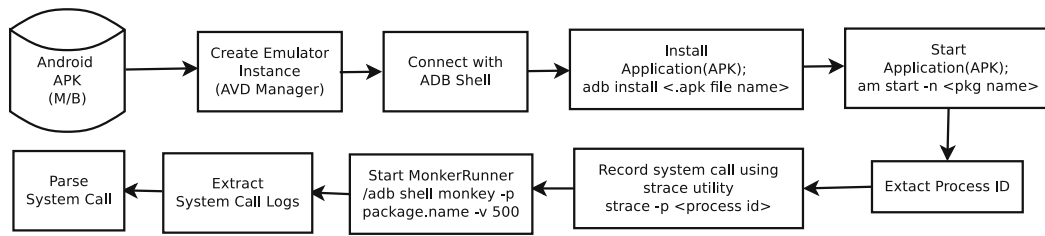


Fig. 1 System call extraction

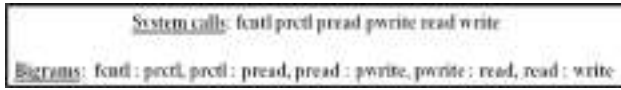


Fig. 2 Sample set of system calls and bigrams

tor using the ADB install command. The Monkey tool [30] is used for interacting with the application and generating system calls. Monkey tool is configured to give automatic user inputs (events) which are: making a call, sending SMS, changing geo location, updating the battery charging status, incoming call 200 times in a minute. Then system calls are recorded using the strace utility. Once the specified events are completed, the application is uninstalled with the ADB uninstall command, and the emulator is set into a clean state for the next app installation. We consider system call names and ignored the parameters of call. In order to avoid the presence of rare system calls in the feature space, we collected five execution traces for each applications. We noticed a longer call trace in case of benign application compared to malicious apps. Also, we noticed top 10 frequently invoked system calls in malicious applications were brk, bind, fchown32, sendto, gettimeofday, epoll_wait, getuid, getpid, clock_gettime and mprotect.

3.3 System call bigram generation

Bigrams are generated from the obtained system calls in a separate text document for each application. Figure 2 shows a sample set of features (system calls) and their corresponding bigrams. The detailed architecture of the proposed Hybrid Malware Detector is shown in Fig. 3.

3.4 Fisher score algorithm

In order to select the most relevant features, an algorithm was developed implementing the Fisher-score.

The algorithm takes as input a set of system calls. Initially, the mean for benign samples is computed, then for malware samples; the variance for benign and malware samples is obtained. The Fisher score is computed for benign and for malware samples. Finally, the Fisher scores obtained are

Algorithm 1 Fisher score algorithm

Input: $F = \{f_1, f_2, f_3 \dots f_m\}$ where f_i represents a feature

Output: $D = \{f_1, f_2, f_3 \dots f_k\}$ where $k \ll m$

- 1: Start
- 2: for $i = 1$ to m do
- 3: $\mu_B = n_m(\mu_{f_i}^B - \mu_{f_i})^2$ ▷ Mean
- 4: end for
- 5: for $i = 1$ to m do
- 6: $\mu_M = n_m(\mu_{f_i}^M - \mu_{f_i})^2$ ▷ Mean
- 7: end for
- 8: for $i = 1$ to m do
- 9: for $j = 1$ to n do
- 10: $\sigma_B = (f_{ji} - \mu_{f_i}^B)^2$ ▷ Variance
- 11: end for
- 12: end for
- 13: for $i = 1$ to m do
- 14: for $j = 1$ to n do
- 15: $\sigma_M = (f_{ji} - \mu_{f_i}^M)^2$ ▷ Variance
- 16: end for
- 17: end for
- 18: $F(f_i)_b = \frac{\mu_B}{\sigma_B}$ ▷ Fisher score
- 19: $F(f_i)_m = \frac{\mu_M}{\sigma_M}$ ▷ Fisher score
- 20: $F(f_i)_{bm} = \frac{\mu_B + \mu_M}{\sigma_B + \sigma_M}$ ▷ Fisher score
- 21: Sort the fisher scores obtained in descending order.
- 22: Stop

sorted in descending order. The steps involved are shown in Algorithm 1.

3.5 Features vector table generation

A sample features vector table is a dataframe consisting of a collection of features. $F_1, F_2, F_3, \dots, F_p$, represent 'p' features (permissions, system calls or app's component). $S_1, S_2, S_3, \dots, S_q$ represent 'q' samples. Class labels in the last column are represented as either '0' or '1'. '0' denotes a benign app while '1' denotes a malware. The values in the table denoted by $v_{11}, v_{12}, \dots, v_{qp}$ refer to the occurrence of a particular feature in a sample. In the case of static features, the occurrence of an attribute is represented by '1' while the absence of an attribute is represented by '0'. While in case of dynamic features and app's components, the elements of vectors are the number of times the p^{th} system call or the app's component was invoked by the q^{th} sample.

In the case of hybrid analysis, the features vector tables produced by both static and dynamic analysis are combined.

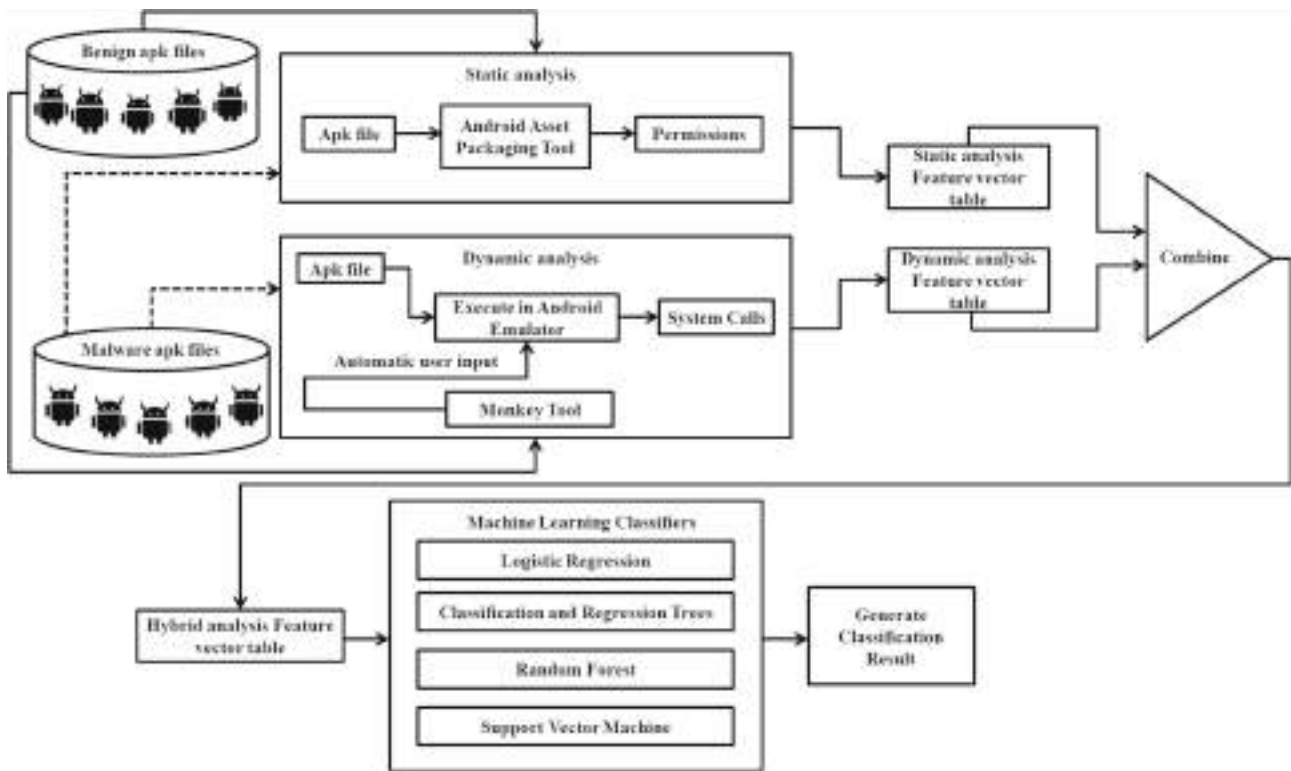


Fig. 3 The architecture of the proposed Hybrid Malware Detector

$F_1, F_2, F_3, \dots, F_p$ represent the relevant attributes obtained after the features selection phase.

3.6 Machine learning unit

The training set is given to the classifier in the form of features vector table. Test data are supplied to it, thus the trained model assigns class labels to each sample in the test set. Here machine learning is run on features obtained with the static analysis (considering the permissions as the features), dynamic analysis (in this case the features are the system call bigrams), and hybrid analysis (permissions and system call bigram are used jointly). Each of these features vector tables is given as input to machine learning classifier and the performances of the different machine learning classifiers are compared. Also the features vector table generated after features selection based on Fisher score is given as input to machine learning and the obtained performances are compared.

3.6.1 Training and testing

There are different techniques for training and testing. One is train-test split and the other is cross validation. In train test split, data are loaded in, then are split into training and test sets. The model is finally fitted to the training data. The

predictions are based on the input training data while are tested on the test data. With cross validation the dataset is split into k subsets: $k - 1$ of these subsets are used for the training while the last subset is hold for test. For our experiment k is fixed to 10.

3.6.2 Classifiers

Classification is a supervised learning approach, i.e. each sample of the training set is explicitly assigned to a category identified by a label. A classifier is an assumption or a function with discrete values that is used to assign class labels to input test samples. The machine learning classifiers in the proposed system used: the Logistic Regression (LR), Classification and Regression Trees (CART), Random Forest (RF), and Support Vector Machine (SVM). The features vector table is the input to the machine learning unit, which then generates the trained model, used to assign class labels to the samples of the test dataset.

4 Adversarial attacks on classifier

In the previous experiments, we discussed feature engineering for developing a classification model to accurately detect malware and benign apps. In this section, we discuss how

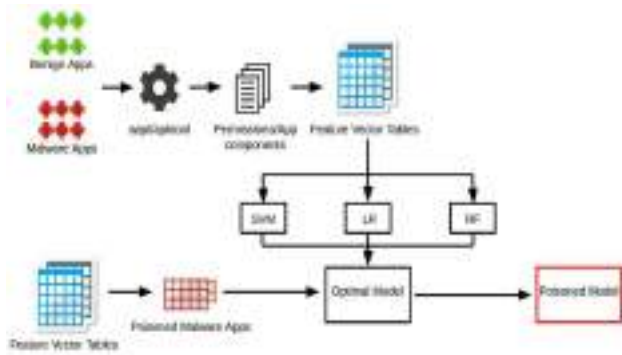


Fig. 4 Adversarial attack

adversarial attacks degrade the robustness of machine learning classifiers, thus we proposed three attack models. In the first phase, we develop the models for classification. Next we perform a poisoning attack on the optimal model. Figure 4 depicts the architecture of the proposed method. The dataset consists of benign applications and malware. Features such as permissions and app components are extracted using `apktool` and `apktool`. Using the extracted features, Feature Vector Tables(FVT) are created. The FVTs are given as input to the machine learning classifiers and DNN for training. In the next phase, the attack is launched on the classifiers. For the attack, 10% of total malware apps are chosen randomly as the test set. Hamming distance and KMeans clustering techniques are used for injecting additional permissions to malicious seed samples. App components are inserted by adding a perturbation in the FVT of app components. Attacks are explained in Sect. 5. The adversarial malware samples are presented to the trained model for predicting the modified applications. Further, we compute the performance of DNN when supplying adversarial samples. The classification accuracy, F1-score, precision and recall of the classifiers are evaluated before and after the attacks. We found that the classification accuracy of the classification model dropped to 40% and 10% for permissions and app components respectively.

4.1 Feature extraction

After data collection, features extraction is performed. In this approach, static features such as permissions and app components are extracted.

For extracting permissions, Android Asset Packaging Tool (AAPT) utility is used, which helps us to view, create and update zipped packages. To extract app’s components, applications are disassembled using `apktool`. `apktool` is an utility for reverse engineering Android applications resources(APK).

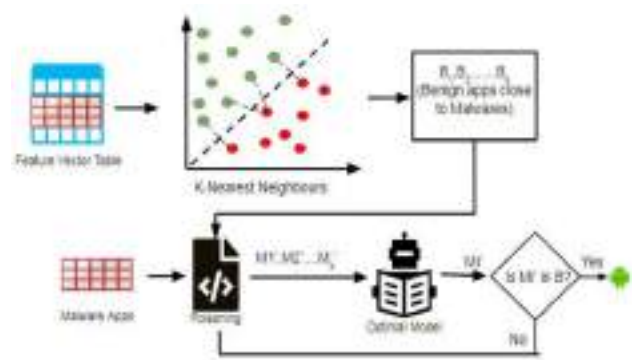


Fig. 5 Evasion Attack based on Hamming Distance

5 Evasion attack

Evasion attack is the process of injecting certain perturbations at test time to increase the error rate of the machine learning classifiers. Initially, classifiers say H is trained using dataset $D = (X_i, y_i)_{i=1}^n$, where $X_i \in [1, 0]^d$ is a d dimensional feature vector for permissions and $X_i \in [integer]^4$ is a four-dimensional feature vector since there are four app components. $y_i \in [1, 0]$ are the class labels where $i \in [1, \dots, n]$. When the dataset is given to the classifiers as input, it performs a classification and response y is generated by $s.t.H(X) = y$. The goal of the is to add a small perturbation to feature vectors of X , $H(X + \mu) = H(X^*)$ such that $H(X^*) = y'$ and $y' \neq y$. For permissions the perturbation $\mu \in [1, 0]$ and for app components, the perturbation is X_{ij_avg} or X_{ij_max} , where X_{ij_avg} is the average of an app component values in the dataset D and X_{ij_max} is the maximum of an app component values in the dataset D .

Three types of attacks are proposed in this study using (a) hamming distance (b) K-means and (c) statistical methods. In the attack scenario, an adversary will add extra attributes to each malicious samples in the test set, until the classifiers wrongly labels suspicious files as legitimate. For the interest of deceiving classifiers, discriminant attributes characteristic of legitimates apps are inserted in the malware applications. In this context by discriminant attributes, we refer to subset of prominent features in one class but at the same time this set is rarely used in alternate class or vice-versa. This will result the decision boundaries of the target classes to overlap thereby increase misclassification.

5.1 Attack using hamming distance

The Hamming distance-based attack is performed using permissions. The attack model is shown in Fig. 5. A set of malware sample is randomly chosen as a test set. In the next step, the Hamming distance between a malwares in the test set and all benign samples are calculated.

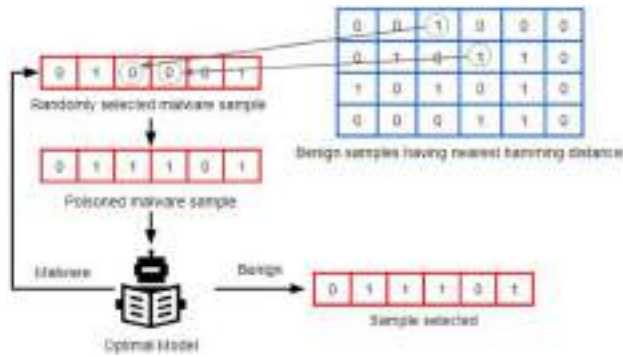


Fig. 6 Evasion attack an example

For example, let the feature vector of malware sample be $M = 1011011001$ and that of benign be $B = 0100110011$. The Hamming distance between M and B is $d(1011011001, 0100110011)$, i.e.

$$1011011001 \oplus 0100110011 = 111111010$$

$$d(1011011001, 0100110011) = 7$$

The benign samples are arranged in ascending order of the distance with the malware seed sample. 0.5% of legitimate files that are close to the malwares are selected. Finally, the attack is performed on selected malware having feature vectors nearly identical to the legitimate app vectors. As the comparison performed over the entire feature space is computationally expensive. Hence, we randomly choose features, and if an attribute is present in benign (logic 1) and absent in malware(logic 0), then that feature is added to the malware sample. Figure 6 shows the addition of permissions to a malware sample.

Steps for adding features to the malware sample are:

- Select a malware sample from the test set.
- 0.5% of the nearest benign samples are shortlisted after calculating the Hamming distance.
- Perform XOR operation between the malware sample and the first benign sample in the shortlist.
- Randomly select an index where XOR gives a logic 1 as output.
- If the selected index has a logic 1 in a benign sample and logic 0 in the malware sample, then add a 1 to the corresponding index in the malware sample to get a new sample.
- The new sample is given to the optimal model for classification.
- If all of the three classifiers in the model predict the new sample as a benign one, then malware is selected and continue the iteration. Otherwise, randomly choose an alternate index, and compare its value in both malware and benign samples.

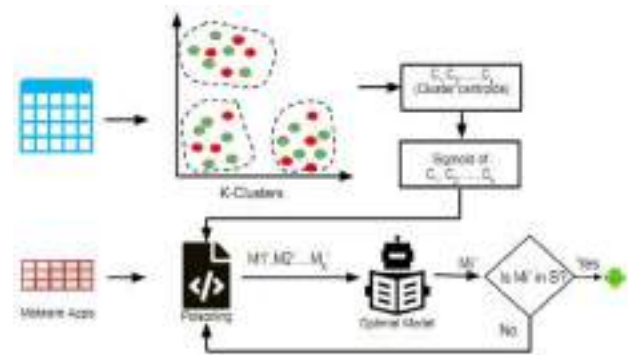


Fig. 7 Poisoning Attack Using KMeans Clustering

- These modified samples are presented to DNN for prediction, finally, the performance of DNN is recorded.

In the algorithm 2, lines 5 to 13 show step for calculating Hamming distance, which are stored in a two-dimensional array A of n rows and 2 columns, where n equals the number of benign samples. The elements of the first column indicate a benign vector and the second column is the Hamming distance to the malware sample. In line 14, values are sorted in ascending order to obtain the legitimate files close to the malware sample. The XOR operation in line 20 is computed to restrict unnecessary comparisons in future. The aim is to obtain the index of a feature that is present in a benign but absent in malware samples.

5.2 Evasion attack using KMeans clustering

In this approach, we cluster benign applications using K-Means clustering. The groups or clusters are formed by representing each legitimate application as a vector of permissions. The attack model is presented in Figure 7 and steps involved are described in algorithm 3. Further, the process of creating adversarial examples using K-Means is discussed below:

1. Randomly choose k centroids.
2. Calculate the Euclidean distance of malware seed sample to the centroids.
3. Assign each seed to the closest centroid and update the centroids by finding the mean value of all the data points in the cluster. This way we cluster all seed examples to the clusters which have similarity based on explicit permission declaration.
4. Compute XOR operation of each seed sample with the centroid vector.
5. Randomly choose an index, if the selected index has the value 1 in the centroid vector and 0 in the malicious seed, modify the vector of the malicious seed sample. This cor-

Algorithm 2 Evasion Attack using permissions (Hamming Distance)

```

Input: Dataset  $D$ , Testset  $T$ , Classifiers  $H$ , Number of benign samples to be shortlisted  $\beta$ , perturbation limit  $\delta$ 
Output: Evaded Samples
1:  $i \leftarrow 0$  ▷ iteration counter
2: repeat
3:    $x \leftarrow T[i]$  ▷ initialize  $i^{th}$  malware sample vector from T to x
4:    $j \leftarrow 0$  ▷ iteration counter
5:   repeat
6:      $b \leftarrow D[j, 1 : m]$  ▷ initialize  $j^{th}$  benign sample vector from D to b
7:     if  $b[m]=0$  then ▷  $m^{th}$  column represents the class label of a vector
8:        $h \leftarrow \text{hamming\_distance}(x, b)$ 
9:       if  $h \neq 0$  then
10:         $A[j][2] \leftarrow h$  ▷ A is a 2 dimensional array where,  $1^{st}$  column has benign samples  $2^{nd}$  column has the
           distance to x
11:        end if
12:      end if
13:    until  $j \leq |D|$  and
14:    sort A in ascending order of distances
15:     $l \leftarrow A[1 : \beta]$  ▷ l is the 2- dimensional array of benign samples with the shortest distance to malware x
16:     $j \leftarrow 0$ 
17:    repeat
18:       $c \leftarrow 0$  ▷ count of perturbation added
19:       $b \leftarrow l[j]$  ▷ benign vector in A
20:       $a \leftarrow b \text{ XOR } x$ 
21:      select a random number  $\gamma$  s.t.a[ $\gamma$ ] = 1
22:      if  $b[\gamma]=1$  and  $x[\gamma]=0$  then
23:         $x[\gamma] \leftarrow 1$  ▷ adding perturbation
24:         $c \leftarrow c + 1$ 
25:       $P \leftarrow H\_predict(x)$  ▷ testing classifier with evaded sample
26:      if  $p=0$  then ▷ classifier predict it as benign
27:         $i = i + 1$ 
28:        goto 2
29:      else
30:        if  $c < \delta$  then
31:          goto 21
32:    until  $j \leq |l|$ 
33: until  $i \leq |T|$ 

```

responds to the addition of permissions in the malware apk.

6. The new sample with injected permissions are presented to all the classification models. If the models wrongly predict the tainted sample as benign, we select such adversarial samples to perform evasion against the deep neural network.
7. However, if the classification model labels modify samples as malicious, we repeat the process by selecting randomly index of the seed vector. This process is continued until a minimum fraction of permissions is injected into the malicious samples.

5.3 Evasion attack using app’s components

App’s components are the basic building blocks of an Android application. The four main app components are Activity, Services, Provider and Receiver. Activities are used for user interaction, Services are an entry point for keeping an app running in the background, the

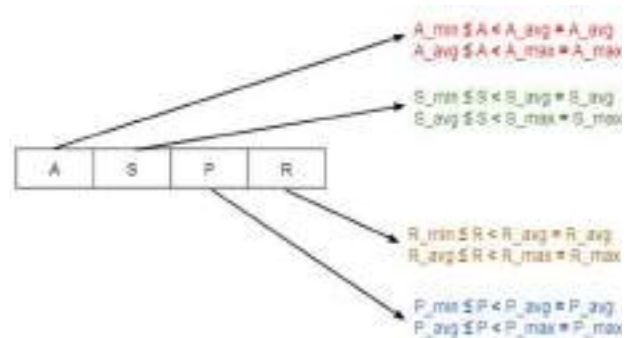


Fig. 8 Modification of app components

Table 1 Statistics of Application components for the legitimate apps

Metrics	Activity	Services	Provider	Receivers
Minimum	0	0	0	0
Average	57	43	24	18
Maximum	130	112	79	53

Table 2 Comparison between different machine learning classifiers on static, dynamic and hybrid analysis

Method to detect malware	Technique to evaluate predictive models	Classifier	A (%)	F1(%)	P (%)	R (%)
Static analysis	K-fold	LR	96.89	95.61	96.88	94.39
		CART	97.26	96.17	96.45	95.85
		RF	96.41	94.77	99.24	90.70
		SVM	97.59	96.60	97.85	95.40
	Train-test split	LR	96.57	95.25	97.37	93.22
		CART	96.57	95.32	96.05	94.59
		RF	95.67	93.81	99.31	88.88
		SVM	97.10	96.00	97.94	94.14
Dynamic analysis	K-fold	LR	93.00	90.46	88.66	92.37
		CART	93.71	91.52	89.27	93.37
		RF	95.64	94.07	92.21	96.05
		SVM	93.41	90.52	93.67	87.73
	Train-test split	LR	92.77	90.43	88.55	92.39
		CART	93.56	91.47	89.57	93.46
		RF	95.47	93.99	92.17	95.89
		SVM	93.53	90.95	94.21	87.90
Hybrid analysis	K-fold	LR	93.80	91.50	90.20	92.80
		CART	100	100	100	100
		RF	98.54	97.98	98.06	97.90
		SVM	100	100	100	100
	Train-test split	LR	93.19	90.89	89.88	91.93
		CART	100	100	100	100
		RF	98.03	97.31	97.99	96.65
		SVM	100	100	100	100

Receiver helps in delivering events outside the app environment and the Provider manages the shared set of app data. The `AndroidManifest.xml` file contain following tags: `<activity>`, `<services>`, `<provider>` and `<receiver>`. To create samples that can evade classifiers, we count the occurrence of app components defined in the legitimate applications. Figure 8 shows the approach of perturbing malicious apk. In Fig. 8 A_{min} , A_{avg} and A_{max} denote the minimum, average and maximum occurrence of activities in all the benign samples. Similarly S_{min} , S_{avg} and S_{max} is the minimum, average and maximum number of services in the manifest file, R_{min} , R_{avg} , R_{max} denote receiver and P_{min} , P_{avg} and P_{max} is the estimate of providers declared in goodwill. A , S , R and P are the estimates of activity, services, receiver and provider in a seed malware sample. The number of injected components in a malware seed is either average or a maximum number of specific component appearing in benign applications.

We consider a malware seed sample with 20 activities, 50 services, 35 provider, and 2 receivers respectively. The statistics of the app components in the set of benign applications are shown in Table 1.

Using the approach detailed in Fig. 8, the app components of the malware seed sample are modified. The first feature value $A = 20$ is in the range $A_{min} < 20 < A_{avg}$, hence the activity (A) in the seed example is updated to $A_{avg} = 57$. The count of services in the seed is altered to $S = S_{max} = 112$, as S is in the range $S_{avg} < 50 < S_{max}$. Similarly the old value of $P = 35$ is updated to P_{max} , as P is in the range $P_{avg} < 35 < P_{max}$, likewise R is modified to $R_{avg} = 18$. Finally, the seed malware application is augmented with 57 activities, 112 services, 24 providers, and 18 receivers. If the modified app is wrongly labelled by the classification models, then a set of such samples have the potential to deceive detection. Otherwise, we increment the count of each component by a value of 3 until the modified app is miss-classified by the classification models.

6 Experimental evaluation

The study consists of two experiments. The purpose of the first experiment was to compare the performances of classifiers trained with features obtained with static, dynamic, and

Table 3 Comparison of the results obtained for static, dynamic, and hybrid analysis based on Deep Learning

Method to detect malware	A (%)	P (%)	R (%)	F1(%)
Static analysis	99.28	98.99	99.08	99.04
Dynamic analysis	94.61	90.54	95.51	92.96
Hybrid analysis	99.59	99.63	99.27	99.45

hybrid analysis. The second experiment aims at evaluating how the performances of classifiers degrade when subjected to the adversarial examples.

6.1 Dataset and experimental setting

For the first experiment we consider, 5,694 benign applications, and 3,197 malware applications. The benign applications were downloaded from the Android App store “9apps”. The Drebin dataset [7] is considered as the malware dataset as it is widely used for experiments and testing of malware classifiers and detectors. Subsequently, in the second experiment for evaluating the robustness of the machine learning and deep learning models, we augmented both malware and benign dataset retaining apks from the first experiment. A total of 11, 447 applications comprising 6,072 benign apks (from 18 different categories) and 5,375 malware apks were collected. Employing VirusTotal¹ we accepted as benign those apps that were labelled as goodware by the majority of antivirus offered by VirusTotal.

All our experiments were conducted on a system with an i7 processor, 8GB RAM, 256 SSD and, 1TB HDD, running the 64-bit Ubuntu operating system. The software requirements were Android Studio and, Anaconda. Anaconda Python distribution was used to execute machine learning in Python language with the help of libraries Scikit-learn, Keras, Matplotlib. Classifiers used in this study are logistics Regression, Random Forest, Support Vector Machine and Deep Neural Network. Hyperparameters for classifiers are tuned using a random search method.

6.2 Evaluation metrics

The metrics used for evaluating the performance of the classifiers are accuracy, the F1, precision and recall. Malware classified as malware represents the True Positive (TP), malware classified as benign represents False Negative (FN), benign app classified as malware represents False Positive (FP) and benign application classified as benign app represents True Negative (TN). Accuracy, precision, recall and

$F1$ are defined with the following equations.

$$Accuracy(A) = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precision(P) = \frac{TP}{TP + FP} \quad (2)$$

$$Recall(R) = \frac{TP}{TP + FN} \quad (3)$$

$$False\ Positive\ Rate(FPR) = \frac{FP}{FP + TN} \quad (4)$$

$$F1\ score(F1) = 2 * \left(\frac{P \times R}{P + R} \right) \quad (5)$$

6.3 Results of experiment-I

Static Analysis :In static analysis, the attribute length in the experiments carried out is 3,360. The highest accuracy and $F1$ was observed for the SVM classifier, even if the best precision is obtained by RF in k-fold and in train-test split, while recall is better for CART classifier in both k-fold and train-test split.

Dynamic Analysis :In dynamic analysis, the attribute length is 2,425. It is observed that RF produced the highest accuracy, $F1$ and recall compared to LR, CART and SVM classifier, even if the precision is greater for SVM classifier in k-fold and 2.04% in train-test split.

Hybrid Analysis :In hybrid analysis, the feature length is 5,785. It is observed that CART and SVM classifier obtained the highest accuracy, $F1$, precision and recall: we can conclude that the hybrid features provide the highest performances.

Using Fischer score prominent attributes were selected to obtain variable feature vector comprising of 10%, 20%, 30%, 40% and 50% of original feature space (which is 3,360, 2,425 and 5,785, respectively as discussed above). Table 2 reports comparison between different machine learning classifiers on static, dynamic and hybrid analysis. Specifically we report the results of k-fold cross-validation and train-test split approach to evaluate predictive models.

6.3.1 Performance of deep neural network

The conventional machine learning algorithms accurately detect unknown samples if specialised feature engineering methods are put in place for extracting attributes representative of target classes. Thus, it demands discovering attribute selection methods that can capture the behaviour of the applications capable of categorizing samples into a specific class. Usually extracting a subset of features from a feature space by applying diverse feature selection approaches is time-consuming. Even if a set of significant attributes are derived, the next challenge is the adoption of a suit-

¹ <https://www.virustotal.com/gui/>.

Algorithm 3 Evasion attack on permission:K-Means clustering**Input:** Dataset D , Test set T , Classifiers H , Number of clusters ρ , the threshold for sigmoid function f - \mathfrak{S} , perturbation limit δ **Output:** Evaded Samples

```

1: procedure K-Means clustering (Dataset  $D$ )
2:   initially choose  $\rho$  data points from  $D$  as centroids
3:   (re)assign each vector in  $D$  to the cluster to which it is closer relying on the mean value of the object in the cluster
4:   update the cluster means
5:   centres  $\leftarrow$  cluster_means
6: end procedure
7:  $i \leftarrow 0$  ▷ iteration counter
8: repeat
9:    $c \leftarrow$  centers[ $i$ ] ▷ initialize  $i^{th}$  centroid vector from centers to  $c$ 
10:   $j \leftarrow 0$ 
11:  repeat
12:     $s \leftarrow f[c[j]]$  ▷  $f$  is the sigmoid function applied on each value in  $j^{th}$  centroid
13:    if  $s > T$ 
14:       $c[j]=1$ 
15:    else
16:       $c[j]=0$ 
17:       $j = j + 1$ 
18:    until  $j \leq |c|$ 
19:     $i = i + 1$ 
20: until  $i < \rho$ 
21:  $i \leftarrow 0$  ▷ iteration counter
22: repeat
23:   $x \leftarrow T[i]$  ▷ initialize  $i^{th}$  malware sample vector from  $T$  to  $x$ 
24:   $j \leftarrow 0$  ▷ iteration counter
25:  repeat
26:     $c \leftarrow$  centers[ $j, 1 : m$ ] ▷ initialize  $j^{th}$  centroid  $v$ 
27:     $a \leftarrow c$  XOR  $x$ 
28:    select a random number  $\gamma$  s.t.  $a[\gamma] = 1$ 
29:    if  $b[\gamma] = 1$  and  $x[\gamma] = 0$  then
30:       $x[\gamma] \leftarrow 1$  ▷ adding perturbation
31:       $c = c + 1$ 
32:       $P \leftarrow H\_predict(x)$  ▷ testing classifier with evaded sample
33:      if  $p = 0$  then ▷ classifier predict it as benign
34:         $j = j + 1$ 
35:        goto 24
36:      else
37:        if  $c < \delta$  then ▷ check if number of perturbations added is beyond the limit
38:          goto 27
39:        end if
40:      end if
41:       $j = j + 1$ 
42:    until  $j \leq |l|$ 
43:   $i = i + 1$ 
44: until  $i \leq |T|$ 

```

able approach for representing applications, in particular feature vector representation. Both the aforesaid techniques, i.e., feature engineering and attribute representation require domain-specific knowledge. The dark side of such a proposal for security systems is the threat of adversarial attacks affecting the integrity and availability of such malware scanners.

To overcome the limitations posed by conventional machine learning algorithms, deep learning neural network models are used as an extension in this study. The primary objective is to improve the detection of malicious apks without the need of implementing feature selection and representation. Thus, we developed three DNN models for predicting samples by using attributes such as (a) permissions (b) sys-

tem calls and (c) a combination of permissions and system calls. Further, before deploying the classification models for predicting apks, hyper-parameters were tuned. In particular, we investigated fixing the best optimizer from a collection of optimizers (rmsprop, adam) and initializers from a collection of initializers (glorot_uniform, uniform). Additionally, we tuned drop-out rate, epochs and batch size. Further, speeding the search of optimal hyper-parameters GridSearchCV approach was adopted. The number of epochs, batch size, and the dropout rate is different in all three models. A small description of these parameters and their values are discussed below.

The dataset has to be propagated forward and backwards through the neural network and this denotes one epoch. But it is too large to pass the entire dataset in one epoch. So it is divided into smaller batches. In the initial static analysis model, the number of epochs is 50 and its batch size is set to 500. In the dynamic analysis model, the number of epochs is raised to 250 and its batch size is reduced to 200. In the hybrid analysis model, the number of epochs is 150 and its batch size is set to 300.

Dropout is a technique used to reduce overfitting, which randomly ignores some layer's output. In the static analysis model, its rate is 0.0, which denotes no outputs from the layer. For both dynamic and hybrid analysis models, it is 0.4. That is, 40% of the neurons in the neural networks are ignored.

Table 3 reports the results obtained for static, dynamic, and hybrid analysis based on deep learning. The static analysis model using deep learning has the highest accuracy, precision, recall, and $F1$ compared to the highest performance static analysis SVM model based on machine learning. That is, accuracy, precision, recall, $F1$ is increased by 1.69%, 1.14%, 3.68% and 2.44%. The machine learning-based RF model has a better performance compared to the deep learning-based model for dynamic analysis. That is, accuracy is greater by 1.03%, precision is greater by 1.67%, recall is greater by 0.54%, and $F1$ is greater by 1.11%.

Finally, in hybrid analysis, the machine learning-based CART and SVM models exhibit higher accuracy, precision, recall, and $F1$ compared to the deep learning-based model. That is, accuracy is higher by 0.41%, precision is higher by 0.37%, recall is higher by 0.73%, and $F1$ is higher by 0.55%. However, comparing the results of static, dynamic, and hybrid models using deep learning, the hybrid model has the highest performance. This again shows that hybrid models can exhibit better results than standalone static and dynamic models.

6.3.2 Comparative analysis

The proposed system that uses multi-modal features, i.e. hybrid features is compared with the following solutions developed on the same dataset

Surendran et al. [42] proposed GSDroid, which leverages graphs for representing system calls sequence extracted from applications in lower-dimensional space. Experiments were conducted on 2,500 malware and benign samples. Malware applications included 1,250 apps from Drebin and the same number of goodware downloaded from Google Playstore. GSDroid reported 99.0% accuracy and $F1$. Bernardi et al. [9] adopted an approach based on model checking for detecting Android malware on 1,200 apk's from Drebin dataset. They created a system calls execution fingerprint (SEF); the obtained SEFs were given as an input to the classifier, reporting 0.94 as True Positive Rate. Finally, SAMADroid [8] is

a 3-level malware detection system that operates on a local host and remote server. Random forest model trained on static features resulted in 99.07% accuracy. However, through our solution based on hybrid features, the accuracy of DNN and SVM is 99.59% and 100% respectively which is far better than the solutions discussed above.

6.3.3 Execution time

The time for detecting samples in our system can be measured based on the time consumed in each module. Here, we discuss the time expended for extracting system calls. Each application is executed for 60 seconds in an emulator, with 200 random events generated by Android Monkey. Overall an average of 92 seconds is required for the entire operation, which comprises booting a clean virtual machine, installing the app, generating the system call logs, copying logs to the host and finally reloading fresh VM. After extracting features, we created a data structure known as the feature vector table (FVT), which is a collection of the feature vectors. We represent the feature space as a binary tree that requires $O(\log n)$. FVT is presented to the classification algorithms for building classifiers. Finally, training Random forest, SVM, CART, LR and DNN requires 5,296 ms, 4,750 ms, 4,076 ms, 899 ms and 6,322 ms respectively.

6.4 Experiment-II: performance of classifiers on adversarial examples

In the following section, we discuss the performance of classifiers presented with adversarial samples. These evasive applications are developed by injecting additional permissions and app components. Additionally, we report the attributes responsible for transforming malware apk's to legitimate applications.

6.4.1 Adversarial applications developed with similarity measure

Table 4 shows the performance of different classification models. It can be seen that $F1$ for predicting applications in the test set is in the range of 0.964-0.970. We randomly selected 537 malicious applications from the test set and determined the similarity with legitimate applications. Extra permissions absent in malware samples but present in the benign dataset were added to these malicious applications. After submitting such tainted (adversarial) applications, the average detection rate and false-positive rate of classifiers obtained are 44.13% and 55.86% respectively. Overall 300 tainted malware samples were created from 537 malware seed samples by merely altering permissions identical to 0.5% of benign applications.

Table 4 Performance of classifier on Adversarial Examples developed using Hamming Distance

Training Set Classifiers	A	F1	P	R
<i>Before Attack</i>				
LR	0.964	0.958	0.975	0.943
RF	0.964	0.958	0.987	0.930
SVM	0.965	0.960	0.975	0.946
<i>Test Set</i>				
LR	0.937	0.967	1.0	0.937
RF	0.931	0.964	1.0	0.931
SVM	0.942	0.970	1.0	0.942
FNR	TPR	#Evaded sample	Mean attributes	Standard
<i>After Attack</i>				
55.86%	44.13%	300	altered 7.02	deviation 6.108

Table 5 Permission-based attack on Deep Neural Network, adversarial examples have high similarity (Hamming distance) with the legitimate applications

Dropout	A(%)	F1(%)	P(%)	R(%)
<i>Before Attack</i>				
0.6	98.38	98.25	99.08	97.32
FNR(%)	A(%)	F1(%)	P(%)	R(%)
<i>After Attack</i>				
45.94	51.62	68.02	1.0	51.62

Similarly we simulated an identical permission-based attack on a deep neural network. In this way, the statistics of permissions in adversarial samples should be close to legitimate applications. The results in Table 5 show a decrease in F1 (68.02%) after the attack, consequently an increase in 45.94% of False Negative Rate is obtained. Overall, 300 malware samples in the test set evaded the detection by merely changing 38 permissions in the malicious applications.

The distribution of evaded malware samples is shown in Fig. 9. It is seen that 50.27% malicious samples (270 nos.) can bypass DNN by solely changing 1 to 5 permissions, 27.5% adversarial samples evade detection by altering 6 to 10 features. As opposed to this, 2 to 4 samples require the addition of 20 permissions to escape detection.

In Fig. 10, we show permissions that are frequently inserted majorly in adversarial samples. In particular, we show the top 25 permissions injected in malware applications through which they escape detection.

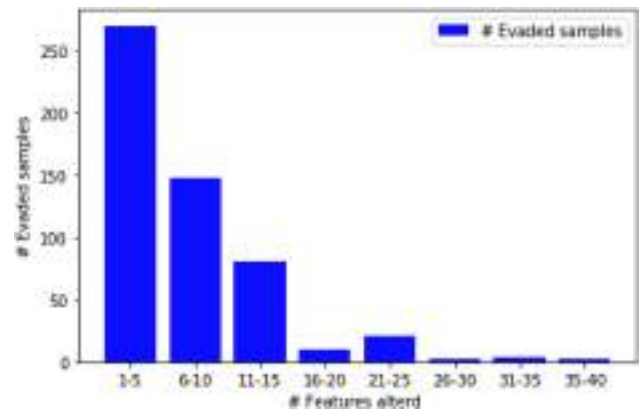


Fig. 9 Number of evaded samples vs number of permissions inserted

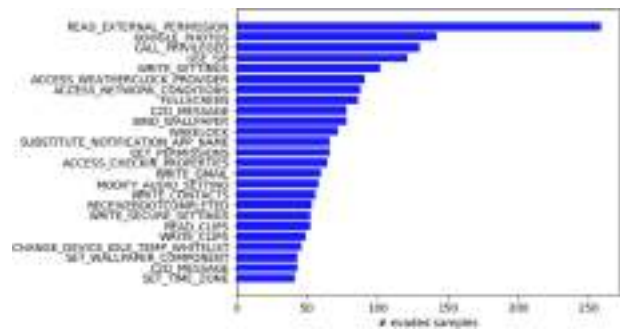


Fig. 10 Inserted permissions in adversarial samples

6.4.2 Adversarial applications generated by estimating the similarity with clusters of goodwill

In the previous scenario, the similarity of malware applications reserved for generating adversarial samples (A) is computed with all benign applications (B) which were not part of the training set. The overall computational cost of

estimating similarity using Hamming distance (discussed in Sect. 6.4.1) is $O(\mathcal{A} \times \mathcal{B})$. In this experiment, using the K-Means clustering approach, we create ρ clusters of benign samples (\mathcal{B}). Now the distance of each malware sample in \mathcal{A} is computed with benign applications in ρ centroids, hence the complexity is $O(\mathcal{A} \times \rho)$ which is less than $O(\mathcal{A} \times \mathcal{B})$. The centroids are real-valued vectors. As the feature vectors are in binary form, the centroids are converted to binary-valued vectors using a sigmoid function. The threshold τ for the sigmoid function is considered. If a value of the sigmoid function is greater than τ , then the number is mapped to 1 otherwise is retained as 0. For example, let us consider centroid of a cluster as [3.21, 5.13, 0.77, 6.71, 2.54, 1.78, 7.89, 4.62], and the threshold is assumed as 0.5. Thus, centroid is transformed to binary vector as [0, 1, 0, 1, 0, 0, 1, 0]. In this work, the τ is in the range of 0.5 to 0.6 obtained in increments of 0.02. Experiments are conducted for different cluster size, i.e., $k = 3, 5$ and 10, shown in Table 6.

From Table 6, 100% evasion of adversarial samples are obtained at threshold of 0.5 for $k = 3, 5$ and 10. The average number of permissions inserted for cluster size $k = 3$ is higher compared to $k = 5$ and 10. Further, we observe as threshold increases the average percentage of evasions decreases.

6.4.3 Evaluation of poisoning attack on app components

In this scenario we randomly chose 537 malware applications from the test set and injected different components. The results obtained is shown in Table 7 and Table 8. The highest $F1$ and accuracy is obtained with Random forest, all other classifiers report poor accuracy. One of the fundamental reason is the lack of attributes to separate applications of target classes. Generally, DNN needs a large number of features to extract relevant attributes to perform precise prediction of the presented samples. Thus, we see that the highest accuracy of 78.1% is obtained with a deep neural network which justifies the lack of attributes for classification. Also, we observe that merely increasing the number of app’s components in the malicious application can easily deceive machine learning and deep learning classifier. In particular, the increase in the frequency of a particular component changes the direction of classification and the learned hypothesis function cannot appropriately predict the new applications.

6.4.4 Attacks using system calls

In this section, we create adversarial examples (AE) using system calls to launch evasion attack (where the attacker aims to affect the target model) and poisoning attack (adversary has the access to training data, to influence model performance). We simulate attacks on a set of machine learning and deep learning models. For deceiving models, partic-

Table 6 Adversarial examples created using k-means clustering

Threshold	Avg. Attributes altered (%)	Evasion (%)	FNR (%)	TPR (%)
<i>No. of Cluster (k = 3)</i>				
0.5	55.1	100	100	0
0.52	0.95	87.15	87.15	12.84
0.54	1.14	77.74	79.08	20.91
0.56	1.33	84.91	96.64	3.35
0.58	1.38	74.23	85.72	14.27
0.6	1	38.36	38.91	61
<hr/>				
Threshold	Avg. Attributes altered (%)	Evasion (%)	FNR (%)	TPR (%)
<i>No. of Cluster (k = 5)</i>				
0.5	12	100	100	0
0.52	0.84	90.8	98.92	9.66
0.54	0.84	73.03	76.05	33.81
0.56	0.84	73.7	78.69	21.3
0.58	1.38	45.47	48	14.27
0.6	0.707	54.45	67.03	32.9
<hr/>				
Threshold	Avg. Attributes altered (%)	Evasion (%)	FNR (%)	TPR (%)
<i>No. of Cluster (k = 10)</i>				
0.5	6.47	1	1	0
0.52	0.89	75.34	90.33	9.66
0.54	0.92	59.01	66.18	33.81
0.56	0.51	40.94	46.03	53.96
0.58	0.95	37.7	43.66	56.33
0.6	1	48.41	64.67	35.32

Table 7 Performance of classifier on evasive malware variants injected with app components

Training Phase Classifier	A (%)	F1(%)	P(%)	R(%)
<i>Before Attack</i>				
LR	75.86	76.14	66.14	89.7
RF	86.42	84.76	81.78	87.96
SVM	81.97	78.16	81.44	75.13
<i>Testing Phase</i>				
LR	89	94.18	100	89
RF	88.1	93.67	100	88.1
SVM	74.05	85.09	100	74.05
<i>After Attack</i>				
FNR	TPR	Evaded sample		
90.13%	98%	484		

Table 8 Performance of DNN on evasive malware variants injected with app components

Drop out	A	F1	P	R
<i>Before Attack</i>				
0.5	78.21%	79.98%	68.75%	95.6%
<i>After Attack</i>				
100%	0%	0%	0%	0%

ularly SVM, Random forest, dense neural networks and 1D-Convolutional Neural Network (1D-CNN). The detailed configuration deep neural network (DNN) and 1D-CNN is presented in Table 9. We assume that the attacker has partial knowledge about the system, in this context the classification algorithms. However, the attacker has access to alternate malware dataset from public repositories. With these capabilities, the adversary is capable of deriving discriminant features and use a subset of attributes to create evasive malware variants. In particular for simulating this form of attack, discriminant attributes from the training set are obtained employing *SelectKBest (SK)* and *Recursive Feature Elimination(RFE)* methods from `sklearn.feature_selection` module. Moreover, for each app, n -gram profiles are created, then each file is represented as uni-gram and bi-grams of system calls. n -grams have been extensively studied in malware detection [37] [1], and have proven to efficiently identify malicious samples from a collection of large examples consisting of both malware and goodware. Figure 11 provides the difference in the distribution of n -grams (system call grams) in malware and benign applications.

Before applying attribute selection methods, we trimmed the feature space by eliminating n -grams with a score less than or equal to 0.0001. Later, features are further synthesized using *SelectKBest* and *Recursive Feature Selection*. In the case of uni-gram 96 system calls are reduced to 83, and finally, 56 uni-grams are extracted through feature selection methods. Likewise, out of 2,364 bi-grams, 166 call grams are chosen using the threshold and finally, 83 call sequences are obtained with attribute selection methods.

We performed the prediction on 10% of randomly selected malware samples (T) excluded from the training set by appending discriminant system calls. We set the maximum attack iteration (I_{max}) to 30%, which means discriminant system calls are repeated at the end of each sample $\tau \in T$ which satisfies the condition that $|\tau| + gram \leq I_{max}$. To evaluate the efficacy of the evasion attack we measured the amount of system call gram added to each file τ : the percentage of calls appended to the file is in the range of 5%-30%, while the inserted ones in increments of 5%.

(A) Evasion attacks using system call

We performed the experiments with 247 randomly selected malware samples as the test set (10% of applications). Figure 12 provides the results attained by progressively appending system calls to the samples in the test set. Before the attack, the F1-measure of uni-gram models (SVM-SK, SVM-RFE, RF-SK AND RF-RFE) are 0.952, 0.950, 0.981 and 0.99 respectively. A significant drop in F1 is observed for each model (refer Fig. 12a) by adding 5% of system calls to each file in the test set. Overall, F1 of the model after the attack is observed between 0.10 to 0.15.

While in the case of bi-gram model, F1 score for the above mentioned classifiers are in range of 0.961 to 0.988 (also shown as 0% in Fig. 12b). We see a marginal drop in F1 for RF-RFE model and a maximum overall drop of 1.6% after the attack. Notably, adding call sequences to uni-gram models is effective compared to bi-gram ML models. We also observe that RF-RFE model trained on RFE features can withstand an evasion attack. RFE being a wrapper-type feature selection algorithm utilizes a classification algorithm to measure the importance of attributes. As the stability of RFE depends primarily on the wrapper(classification algorithm), thus relatively improved outcome is obtained with Random Forest (RF). The superior performance of Random Forest is attributed to the fact that the relevant attributes are filtered by bootstrapping the samples and features. In this way, several decision trees are created which contribute to computing the model performance.

Figure 12(c) present the results of Deep neural network (DNN) and 1D-CNN on evasion attack. For DNN F1 drops from 0.967 to 0.375 and 0.562 respectively adding extra 5% system calls in each malware samples in the test set. The classifier performance is severely affected by increasing the number of system calls being added to files. Here, we observe that a significant misclassification is obtained, however, the rate of misclassification for bi-gram models are comparably less than models trained on uni-grams. Additionally, we evaluated the robustness of 1D-CNN; results are shown in Fig. 12d. The evaluation was conducted on variable stride length which can be considered as n -grams. Before the attack, the F1 scores on distinct strides are 0.9788, 0.981 and 0.9815 respectively. However, after the evasion attack malware samples were wrongly labelled as legitimate, thus the drop in F1 by padding 5% discriminant system calls to each file are 5.88%, 2.06% and 2.75% respectively. On comparing individual models, it can be seen that the 1D-CNN offer higher resistance to evasion attacks. 1D-CNN can derive robust features without the use of a complex feature engineering process, and have a computational complexity of $O(K.N)$, where K is the kernel and N is the size of the input.

(B) Poisoning attack using system call

In the following paragraphs, we discuss the evaluation of the poisoning attack. We simulate the behaviour of an adversary

Table 9 Configuration of DNN and 1D-CNN

Model	Input	Layers	Hyperparameters
DNN (Uni-gram)	96	Layer - 1 (Hidden) Dense(128) + Dropout(0.1) + BatchNormalization Layer - 2 (Hidden) Dense(256) + Dropout(0.2) + BatchNormalization Layer - 3 (Hidden) Dense(512) + Dropout(0.3) + BatchNormalization	Learning rate = 0.0001 Epochs = 100, Batch size = 16 Optimizer = Adam Hidden layer activation = Relu Output layer activation = sigmoid
DNN (Bi-gram)	2364	Layer - 1 (Hidden) Dense(64) + Dropout(0.1) + BatchNormalization Layer - 2 (Hidden) Dense(32) + Dropout(0.2) + BatchNormalization Layer - 3 (Hidden) Dense(16) + Dropout(0.3) + BatchNormalization	Learning rate = 0.0001 Epochs = 50, Batch size = 16 Optimizer = Adam Hidden layer activation = Relu Output layer activation = sigmoid
1D-CNN (Stride 1 -3)	101681	Layer - 1 (Embedding) Embedding(32) Layer - 2 (Hidden) Conv1D(128) Layer - 3 (Hidden) MaxPooling1D Layer - 5 (Hidden) Conv1D(256) Layer - 6 (Hidden) MaxPooling1D Layer - 7 (Hidden) Conv1D(512) Layer - 8 (Hidden) MaxPooling1D Layer - 9 (Hidden) Dense(10)	Learning rate = 0.0001 Epochs = 30, Batch size = 8 Optimizer = Adam Kernel size = 3 Hidden layer activation = Relu Output layer activation = Sigmoid

who manipulates a subset of malware files in the training set by appending a set of selected system call sequence (extracted using feature selection methods). The overall objective is to maximize the classifier confidence in labelling malicious file as legitimate, or in other words, increase the probability of tainted samples classified as benign. An alternate scenario of poisoning attack is the label flipping attack, here the adversary deliberately swaps the original label of a sample with the target class label. In our study we focused on developing poisoned samples by adding extraneous system call to selected malware seed samples. Figure 13 presents the results of poisoning attack.

Practical use case of poisoning attack in malware detection domain is crowd-sourcing the malware apps for labelling and generating its signatures. Under such circumstances, a dishonest user can manipulate the samples or intentionally modify the label. However, the attack can be defeated in the presence of a large number of legitimate users, where the

class label of a suspect file is decided relying on majority voting. Mimicking such a scenario we intended to poison a very small fraction of malwares in the training set. Figure 13(a) provides the outcome of ML classifiers on padding uni-grams. We observe here that a small fraction of samples in the test set is misclassified. The overall drop in average F1 for the RF-RFE and RF-SK is 0.068%, 0.25% respectively. Likewise, in the case of SVM-SK and SVM-RFE the average drop in F1 are 3.32% and 1.596%. We can conclude that Random forest models are highly resistant to adversarial attack, specifically, the performance of RFE trained models show improved results with respect to the models trained on *SelectKbest* attributes.

Similar trends in the results are obtained for bi-gram models (refer Fig. 13b. For SVM-SK classifier the difference in F1 falls in the range of 0.004 to 0.006 compared with the model in the absence of a poisoning attack, where the F1 is 0.963. In the case of SVM-RFE the average change in F1 for the entire

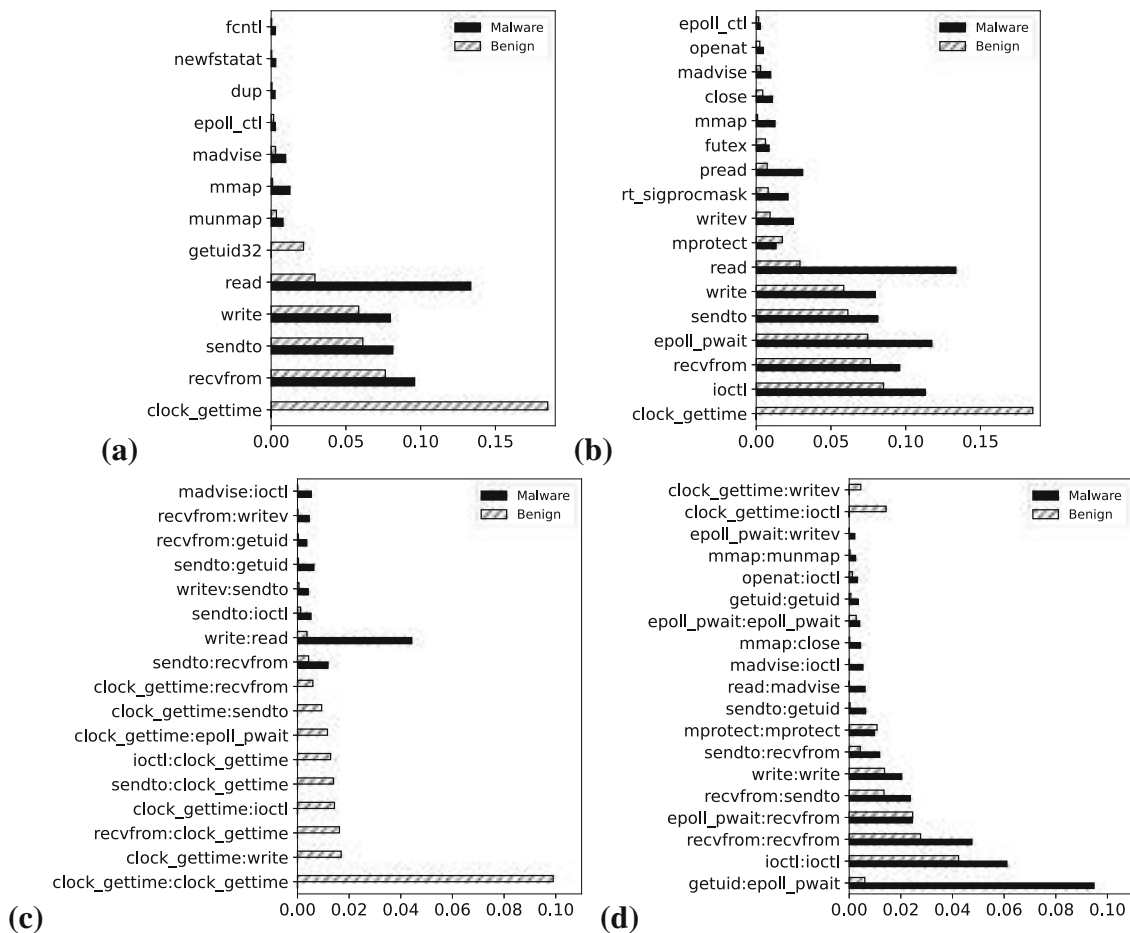


Fig. 11 System call grams **a** uni-gram SelectKBest **b** uni-grams RFE **c** bi-gram SelectKBest and **d** bi-gram RFE

range of padded system calls (i.e., 5% to 30%) is 0.575% with the standard deviation of 0.0006. A very small spread in the F1 values indicates the ineffectiveness of poisoning attacks. Identical observations can be made for Random forest models (RF-RFE and RF-SK), where the spread of F1 across a different range of padding is 0.00035 and 0.00031 respectively.

Figure 13(c) and (d) show the performance of DNN and 1D-CNN. It is evident from these figures that the attack is not severe, and a marginal drop is observed when malware samples are padded with system calls in a larger amount. However, a clear trend is not noticed in the case of deep learning models. Training set with tainted samples in certain cases also improves the classifier results. On investigating the confusion matrix we found that for larger padding size malware samples that were previously misclassified were now precisely detected by DNN. It is intuitive that malicious data points statistically closer to the legitimate files are now accurately detected.

7 Evaluation on obfuscated samples

Software developers obfuscate the source code of applications to avoid manual analysis and violations of intellectual property. Instead, malware writers use obfuscation to keep new variants of original malicious applications being detected. A vast majority of malware variants have less than 2% difference in code [22]. Anti-malware products employing pattern matching techniques fail to detect obfuscated files. By forcing an application to execute in an emulated environment, and monitoring system call invocation, obfuscated samples are identified. To generate obfuscated malware variants, we make use of an open-source obfuscator known as Obfuscapk [5]. Obfuscapk supports obfuscation techniques like trivial, renaming, encryption, code reorder and reflection. As the first step, we looked at detecting obfuscated samples in the dataset. In this step, we represented system call invocation of a file as a system call co-occurrence matrix of size $m \times m$, where m is the number of unique calls. Each element in the matrix corresponds to the occurrence of a pair of calls. The call frequencies are normalized and mapped to pixels

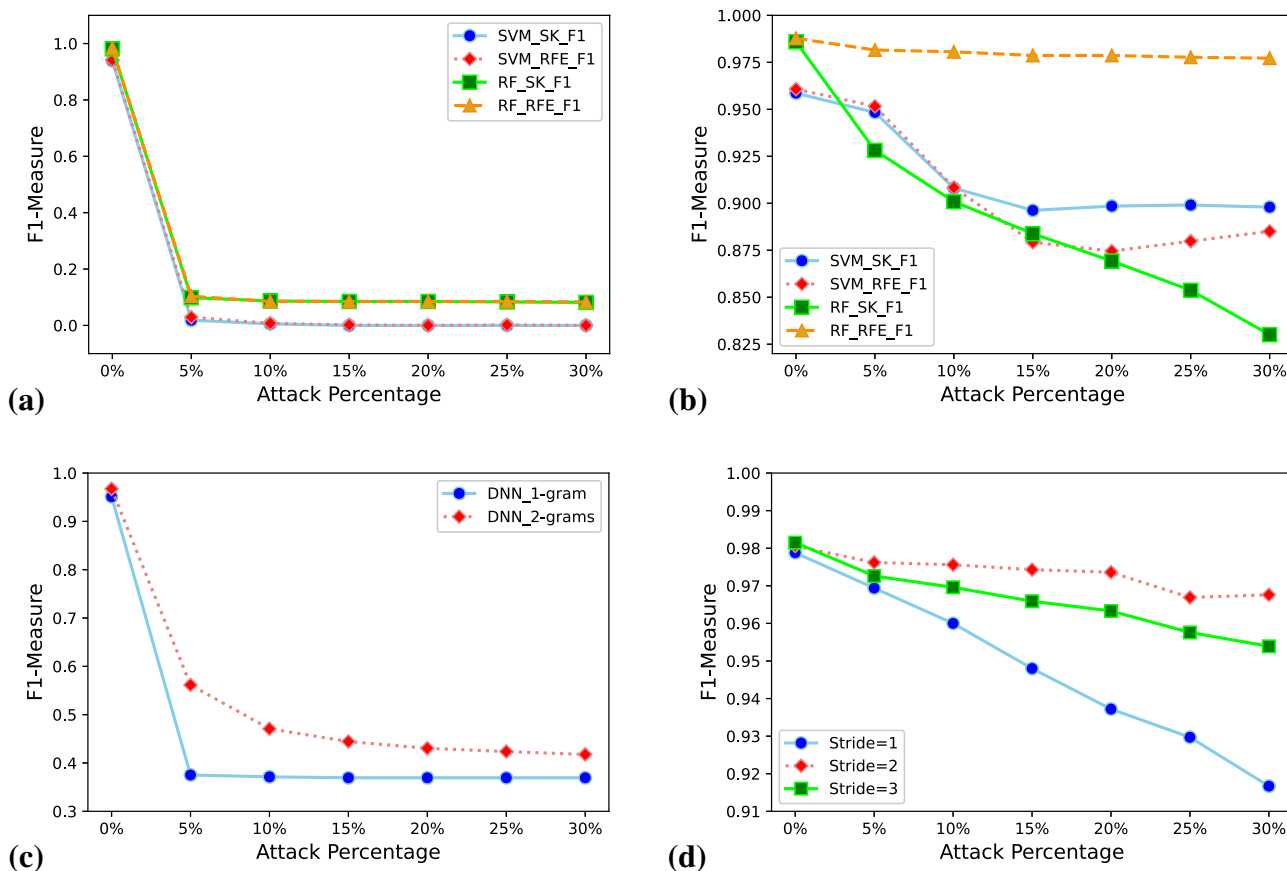


Fig. 12 Evasion attack using a uni-gram SelectKBest b bi-gram RFE c Deep neural network and d 1D-CNN

Table 10 Evaluation of obfuscated malware using system call images

Approach	Accuracy	Precision	Recall	F1	Time
Train-Test	0.966	0.936	0.980	0.957	27sec
CV	0.960	0.928	0.972	0.949	33 sec

by multiplying the normalized values with 255. Finally, the system call images corresponding to malware and benign set are used for training the 2D-CNN model for prediction. We chose CNN for developing the model as it extracts relevant patterns in images even if they are not fixed. To be precise, CNN is spatially invariant to patches of a given image. This is fundamental to code obfuscation where the blocks of code in the program are randomly rearranged by the obfuscator using branch instructions. Table 10 shows the identification of obfuscated malware using *train_test_split* and stratified ten-fold cross-validation (CV) approach.

We can observe that the highest F1 obtained by transforming apps into a system call co-occurrence matrix is 0.959. Analysis of co-occurrence matrix revealed the presence of a large number of contiguous blocks of black regions indicating the existence of zeros in this matrix. To improve the

detection, we addressed the problem by transforming malware as gray-scale images, similar to the approach in [31]. In this context, we map raw bytes of .dex files to pixels and apply image processing techniques. Initially, we investigated training ML models on images, especially on image textures extracted using a bank of Gabor filters formed by varying the kernel size, standard deviation, angle, wavelength and aspect ratio. As the feature extraction and training was computationally expensive, we considered employing 2D-CNN, which extracts features without manual intervention from raw malware binaries. For retaining the semantic information of an image, pairwise probability of bytes(pixels) were estimated. Subsequently, the probabilities are transformed into pixel values between 0-255. As a consequence, each apk is converted to a fixed size image (256×256). We train tuned Convolutional Neural Network (CNN) (learning rate = 0.0001, momentum = 0.9, epoch = 100 and batch size = 32) on the generated images of malware and benign samples. The topology of the network is presented in Table 11.

Malware samples used in the previous experiments (refer Section 6.1) [7] are obfuscated, and the performance of the CNN model is estimated under four scenario (a) malware (\mathcal{M}) vs benign (\mathcal{B}) (b) benign (\mathcal{B}) vs obfuscated malware

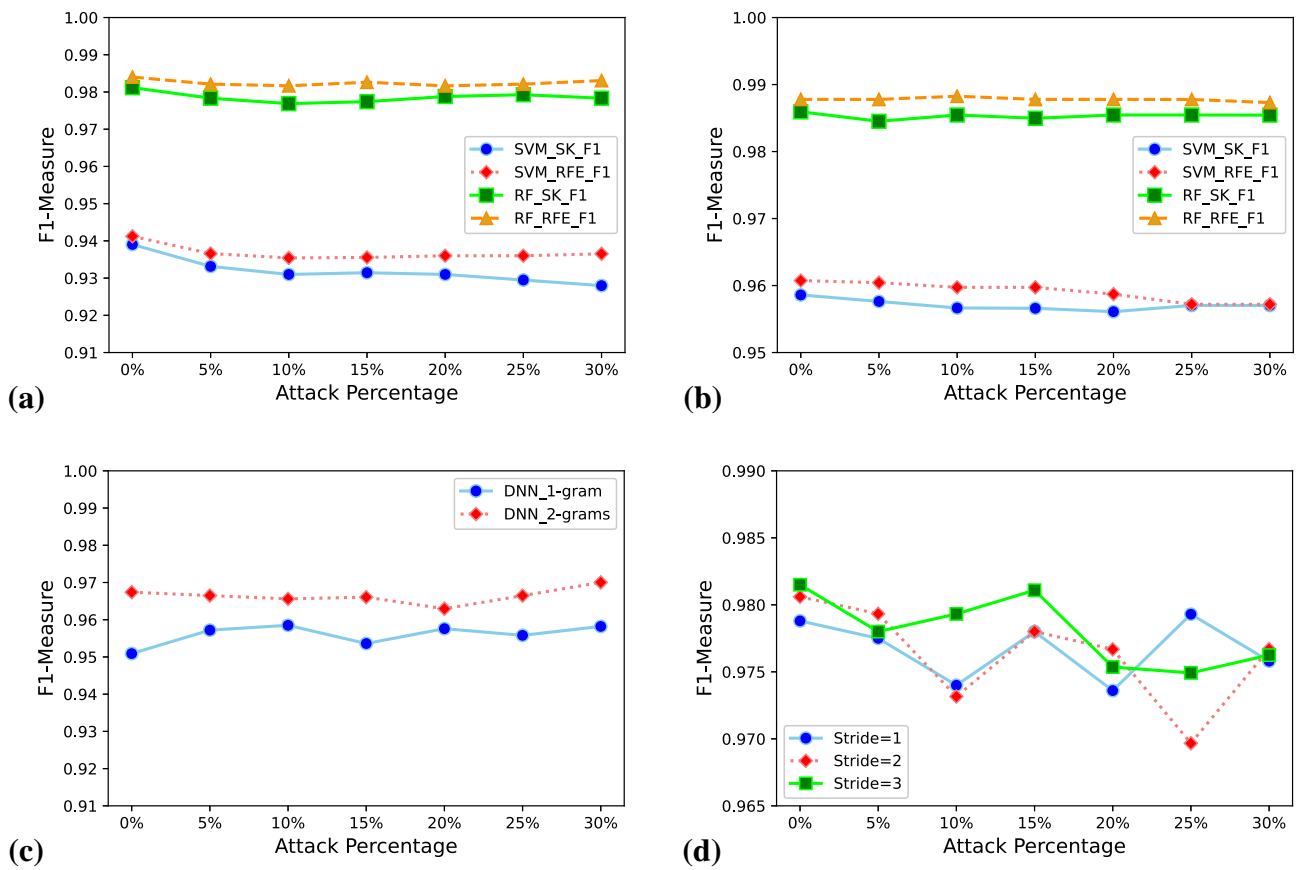


Fig. 13 Poisoning attack employing a uni-gram SelectKBest b bi-gram RFE c Deep neural network and d 1D-CNN

Table 11 Architecture of CNN

Layers	Filter size	Input Shape	Output Shape	Activation
Conv-1	64(3*3)	(64,64,1)	(none,62,62,64)	ReLU
MaxPooling-1	(2*2)	(none,62,62,64)	(none,31,31,64)	-
Conv-2	64(3*3)	(none,31,31,64)	(none,29,29,64)	ReLU
MaxPooling-2	(2*2)	(none,29,29,64)	(none,14,14,64)	-
Dense-1	(none,128)			ReLU
Dense-2(binary)	(none,1)			Sigmoid
Dense-2(categorical)	(none,14)			Softmax

(\mathcal{M}^\perp) (c) malware(\mathcal{M}) vs obfuscated malware(\mathcal{M}^\perp) and (d) malware family class (\mathcal{FC}). Figure 14 shows the classification of obfuscated malware family classification. Through this experiment we conclude that CNN accurately labels each sample in the test set to the appropriate obfuscation class. Table 12 presents the results obtained using 2D-CNN.

8 Discussion

In this study, we show that machine learning classifiers are vulnerable to adversarial attack. ML-based Malware detectors trained on static features such as permissions, APIs

and applications components can be easily attacked by carefully generating perturbed apps having statistical similarity with legitimate apps. Generally, the vector corresponding to an application is represented with boolean values. Iterative addition of features (permission, hardware feature and intents, etc) generates evasive applications with minimal effort without compromising app functionality. In this context, an attacker must modify selected attributes with a value 0 to 1. Further, changing minimum subset of attributes will force linear classifier such as logistic regression, SVM (linear kernel) to misclassify files in the test set. However, significant attempts are required to bypass the classifier trained with the sequence of system calls, as values of features are continu-

Table 12 Performance of CNN using different proportion of training and test set

Data Split	\mathcal{M} vs \mathcal{B}		\mathcal{B} vs \mathcal{M}^\perp		\mathcal{M} vs \mathcal{M}^\perp		$\mathcal{F}\mathcal{C}$	
	A	F1	A	F1	A	F1	A	F1
70:30	0.996	0.995	0.987	0.989	0.997	0.996	0.997	0.997
80:20	0.994	0.994	0.995	0.996	0.998	0.998	0.996	0.996
90:10	0.995	0.995	0.995	0.990	0.999	0.999	0.997	0.996

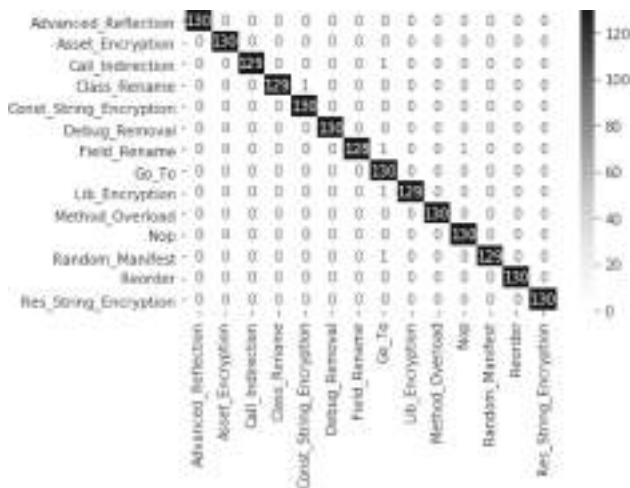


Fig. 14 Classification of obfuscated malware variants

ous. This require padding of larger amount of discriminant calls sequence to each malware sample. Intuitively it means that the modified applications will spend large execution time compared to its normal functionality. It is worth mentioning that such suspicious apps will be easily detected by monitoring the power consumption and heat dissipated of the smart device. Further, if we think in the context of designing intelligent anti-malware systems, adversarial samples generated by augmenting large number of call sequences would deliberately force the application execute longer on the device. Thus, anti-viruses making use of simple heuristic such as the utilization of memory (virtual memory, cursor, dalvik), CPU usage, number of processes created, etc, would identify such applications.

Poisoning attack using static features can be easily simulated, but considerable efforts are needed for injecting dynamic features. Especially in all cases, we observed that Random forest and non-linear classifiers such as DNN and 1D-CNN are difficult to be attacked. Besides CNN shows a good detection rate in identifying modified malicious samples and obfuscated samples, as its convolution operation is capable of identifying repeated patterns in different regions of files, be it a chunk of system call sequence or byte stream. Another important observation emerged from our experiments is that the knowledge of the feature set plays a very significant role in creating adversarial samples. Randomly

selecting attributes and injecting them into applications does not create a successful attack.

An attack can be practically demonstrated by modifying the decompiled source of a malicious app. Top-weighted features comprising permissions, APIs and app components can be inserted into the decompiled code. By progressively adding features in the AndroidManifest.xml and rebuilding it, and later resigning the app creates a modified version with extraneous attributes. In our approach feature addition is considered for maintaining functionality of the application. Although, in the case of APIs, we can shield the call to specific API by substituting the characters by applying mono-alphabetic substitution (identical to additive cipher). Here our implication is to replace a character with a new character based on the specific substitution key. This will generate an encoded representation of the API. Logically, creating a modified version of encoded API in this way resembles the creation of an obfuscated application. To maintain the functionality a decoder module can be plugged in the app, which regenerates the API call name at runtime. Further the original API is invoked through Java reflection. However, an evasion attack created by the above-mentioned strategy using API modification would fail while performing dynamic analysis, as the classifier designed on dynamic attributes can identify the call to decoded APIs during runtime. We left the implementation as an open research problem, which we plan to address in our future work.

Relying on the lessons learnt by conducting our experiments, in future we plan to propose countermeasures for evasion attack. Following are our proposal:

- Address N class problem as $N + 1$ class problem. This means we must develop a proactive system wherein the designers of the anti-malware system must simulate the behaviour of an adversary. By doing this, a large collection of adversarial samples can be approximated. A set of created samples can be used to augment the training set. In other words, classifiers are trained using malware, benign and adversarial examples.
- Development of ensembles of classifiers randomly trained on subset of attributes that periodically are modified during the re-training process. As the knowledge of features is critical for crafting attacks, it will hinder attack tactics as an adversary is unaware of classifier revision and

Table 13 Resume of the Related Work

Paper	Contributions
Patel and Buddadev [33]	Hybrid Android malware detection Permissions and behaviour-based features Rule generation
Wang et al. [46]	Hybrid malware detector Detection of zero days
Damodaran et al. [16]	Comparative analysis on malware detection system Static, dynamic, and hybrid analysis
Wu and Hung [47]	Static and dynamic features
Saracino et al. [38]	Experiment on KNN classifier
Li et al. [27]	Malware detection by mining permission SVM and decision trees for classification
Chuang and Wang [15]	Classification with frequency of API calls
Burguera et al. [11]	Dynamic analysis of Android apps Two means clustering algorithm
Dimjašević et al. [18]	Detection of Android malware through system calls
Afonso et al. [2]	Detection of Android malware API calls and system call traces
Garcia et al. [21]	Detection of Android malware Categorized Android API usage, reflection-based features, and Features from native binaries of apps
Tam et al. [43]	Reconstructing behaviors of Android malware Observing system calls
Almin and Chatterjee [4]	Analysis of permissions Clustering and classification techniques
Kim et al. [25]	Android malware detection Opcode features, API features, strings, permissions, app's Components, and environmental features
Sun and Qian [41]	Malware detection model-based on RNN and CNN
Ni et al. [32]	Opcode sequences, malware visualization, and deep learning
Saxe and Berlin [39]	Deep neural network Static features
Karbab et al. [24]	Deep learning techniques Raw sequences of API method calls
McLaughlin et al. [29]	Static analysis Raw opcode sequence from a disassembled program
Vinayakumar and Soman [45]	Comparison of deep neural networks(DNNs) andMachine learning algorithms for static malware detection
Le et al. [26]	Malware classification method using Visualization and deep learning
Agarap and Pepito [3]	Convolutional deep learning models
SI and CD [36]	CNN based windows malware detectorAPI calls
Martinelli et al. [28]	Convolutional neural network System calls
Xiao et al. [48]	Backpropagation neural network
Chen et al. [14]	Two-phase detection system
Xu et al. [49]	Genetic programming
Chen et al. [13]	Evading PDF malware classifiers
Grosse et al. [23]	Evaluation of standard classifiers
Chavan et al. [12]	Adversarial crafting attacks on neural network
Demontis et al. [17]	Experiments on permissions Binary and multiclass classification
Pierazzi et al. [35]	Adversary-aware machine learning detector
	Formalization of problem-space attacks Relationships between feature space and problem space

the features used to model the classifiers. Notably, the conclusion for assigning the labels for a sample under consideration could be based on *OR* operations, which means that if anyone among the pool of classifiers labels the sample as malware and all the others as legitimate, the target class label will be concluded as malware.

- Building classifier using a set of attributes that are difficult to be modified. This would restrict the attack surface as a modification to the aforementioned feature would affect the functionality of the program.

9 Conclusion and future work

In this paper, we present a study on malware detectors based on machine and deep learning classifiers, consisting of two experiments. In the first experiment, we propose a hybrid approach for malware detection, that lets us conclude that hybrid analysis increases the performance of classifiers concerning the independent features. The results show that with static features the SVM algorithm produces the best outcomes, and this corroborates the evidence provided by the literature. With regards to the dynamic analysis, the RF algo-

rithm showed better results, while the highest performances with the hybrid approach were obtained with CART and SVM algorithms. We extended our study by investigating the performances of the deep neural network, which also show that the hybrid features produced improved results.

In addition, we examined how evasion and poisoning attacks deteriorate the robustness of the classifiers. We showed that the evasion attack severely affects classifier performance with static features, however, evasive examples created using system calls (dynamic analysis) adversely affected the classifier outcome. We show a large collection of adversarial examples which are able to prevent from the detection. Concerning the classifiers, we observed that Random Forest and CNN offer a good resistance to adversarial attacks.

In the future, we will evaluate the performances of diverse deep learning models using multiple datasets. Additionally, we would like to test the reliability of classification systems on adversarial attacks trained on malware images techniques. In particular, we would like to explore how neurons in each layer participate in the feature extractor process.

References

1. Abou-Assaleh, T., Cercone, N., Keselj, V., Sweidan, R.: N-gram-based detection of new malicious code. In: Proceedings of the 28th Annual International Computer Software and Applications Conference, 2004. COMPSAC 2004., vol. 2, pp. 41–42. IEEE (2004)
2. Afonso, V.M., de Amorim, M.F., Grégio, A.R.A., Junquera, G.B., de Geus, P.: Identifying Android malware using dynamically obtained features. *J. Comput. Virol. Hacking Techn.* **11**(1), 9–17 (2015)
3. Agarap, A.F.: Towards building an intelligent anti-malware system: a deep learning approach using support vector machine (svm) for malware classification. arXiv preprint [arXiv:1801.00318](https://arxiv.org/abs/1801.00318) (2017)
4. Almin, S.B., Chatterjee, M.: A novel approach to detect android malware. *Procedia Comput. Sci.* **45**, 407–417 (2015)
5. Aonzo, S., Georgiu, G.C., Verderame, L., Merlo, A.: Obfuscapck: an open-source black-box obfuscation tool for Android apps. *SoftwareX* **11**, 100403 (2020)
6. APKTool: <https://ibotpeaches.github.io/Apktool/install/>
7. Arp, D., Spreitzenbarth, M., Hubner, M., Gascon, H., Rieck, K., Siemens, C.E.R.T.: Drebin: effective and explainable detection of android malware in your pocket. In: Ndss, vol. **14**, pp. 23–26. (2014)
8. Arshad, S., Shah, M.A., Wahid, A., Mehmood, A., Song, H., Hongnian, Y.: Samadroid: a novel 3-level hybrid malware detection model for android operating system. *IEEE Access* **6**, 4321–4339 (2018)
9. Bernardi, M.L., Cimitile, M., Distanto, D., Martinelli, F., Mercaldo, F.: Dynamic malware detection and phylogeny analysis using process mining. *Int. J. Inf. Secur.* **18**(3), 257–284 (2019)
10. Biggio, B., Fabio, R.: Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit.* **84**, 317–331 (2018)
11. Burguera, I., Zurutuza, U., Nadjm-Tehrani, S.: Crowddroid: behavior-based malware detection system for android. In: Proceedings of the 1st ACM Workshop on Security and Privacy in Smartphones and Mobile Devices, pp. 15–26. (2011)
12. Chavan, N., Di Troia, F., Stamp, M.: A comparative analysis of android malware. arXiv preprint [arXiv:1904.00735](https://arxiv.org/abs/1904.00735) (2019)
13. Chen, L., Hou, S., Ye, Y., Xu, S.: Droideye: fortifying security of learning-based classifier against adversarial android malware attacks. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 782–789. IEEE (2018)
14. Chen, S., Xue, M., Fan, L., Hao, S., Xu, L., Zhu, H., Li, B.: Automated poisoning attacks and defenses in malware detection systems: an adversarial machine learning approach. *Comput. Secur.* **73**, 326–344 (2018)
15. Chuang, H.Y., Wang, S.D.: Machine learning based hybrid behavior models for Android malware analysis. In: 2015 IEEE International Conference on Software Quality, Reliability and Security, pp. 201–206. IEEE (2015)
16. Damodaran, A., Di Troia, F., Visaggio, C.A., Austin, T.H., Stamp, M.: A comparison of static, dynamic, and hybrid analysis for malware detection. *J. Comput. Virol. Hacking Techn.* **13**(1), 1–12 (2017)
17. Demonits, A., Melis, M., Biggio, B., Maiorca, D.A., Rieck, K., Corona, I., Giacinto, G., Roli, F.: Yes, machine learning can be more secure! a case study on android malware detection. In: IEEE Transactions on Dependable and Secure Computing, vol.16, pp. 711–723. IEEE (2019)
18. Dimjašević, M., Atzeni, S., Ugrina, I., Rakamarić, Z.: Evaluation of android malware detection based on system calls. In: Proceedings of the 2016 ACM on International Workshop on Security And Privacy Analytics, pp. 1–8. (2016)
19. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9185–9193. (2018)
20. Gandotra, E., Bansal, D., Sofat, S.: Malware analysis and classification: a survey. *J. Inf. Secur.* **2014** (2014)
21. Garcia, J., Hammad, M., Malek, S.: Lightweight, obfuscation-resilient detection and family identification of android malware. *ACM Trans. Softw. Eng. Methodol. (TOSEM)* **26**(3), 1–29 (2018)
22. Greengard, S.: Cybersecurity gets smart. *Commun. ACM* **59**(5), 29–31 (2016)
23. Grosse, K., Papernot, N., Manoharan, P., Backes, M.I., McDaniel, P.: Adversarial examples for malware detection. In: European Symposium on Research in Computer Security, pp. 62–79. Springer, Cham (2017)
24. Karbab, E.B., Debbabi, M., Derhab, A., Mouheb, D.: MalDozer: automatic framework for android malware detection using deep learning. *Digit. Investig.* **24**, S48–S59 (2018)
25. Kim, T.G., Kang, B.J., Rho, M., Sezer, S., Im, E.G.: A multimodal deep learning method for android malware detection using various features. *IEEE Trans. Inf. Forensics Secur.* **14**(3), 773–788 (2018)
26. Le, Q., Boydell, O., Namee, B.M., Scanlon, M.: Deep learning at the shallow end: malware classification for non-domain experts. *Digit. Investig.* **26**, S118–S126 (2018)
27. Li, J., Sun, L., Yan, Q., Li, Z., Srisa-An, W., Ye, H.: Significant permission identification for machine-learning-based android malware detection. *IEEE Trans. Ind. Inf.* **14**(7), 3216–3225 (2018)
28. Martinelli, F., Marulli, F., Mercaldo, F.: Evaluating convolutional neural network for effective mobile malware detection. *Procedia Comput. Sci.* **112**, 2372–2381 (2017)
29. McLaughlin, N., del Rincon, J.M., Kang, B.J., Yerima, S., Miller, S., Sakir, S., et al.: Deep android malware detection. In: Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy, pp. 301–308. (2017)
30. MonkeyRunner: <https://developer.android.com/studio/test/monkey>
31. Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S.: Malware images: visualization and automatic classification. In: Proceedings

Blockchain-Based Secure Healthcare Application for Diabetic-Cardio Disease Prediction in Fog Computing

P. G. SHYNU¹, (Member, IEEE), VARUN G. MENON², (Senior Member, IEEE),
R. LAKSHMANA KUMAR³, (Member, IEEE), SEIFEDINE KADRY⁴, (Senior Member, IEEE),
AND YUNYOUNG NAM⁵, (Member, IEEE)

¹School of Information Technology and Engineering, Vellore Institute of Technology, Vellore 632014, India

²Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India

³Head- Centre of Excellence for Artificial Intelligence and Machine Learning, Hindusthan College of Engineering and Technology, Coimbatore 641050, India

⁴Faculty of Applied Computing and Technology, Noroff University College, 4608 Kristiansand, Norway

⁵Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538, South Korea

Corresponding author: Yunyoung Nam (ynam@sch.ac.kr)

This work was supported in part by the Korea Institute for Advancement of Technology (KIAT) Grant by the Korean Government through Ministry of Trade Industry and Energy (MOTIE) (The competency development program for industry specialist) under Grant P0012724, and in part by the Soonchunhyang University Research Fund.

ABSTRACT Fog computing is a modern computing model which offers geographically dispersed end-users with the latency-aware and highly scalable services. It is comparatively safer than cloud computing, due to information being rapidly stored and evaluated closer to data sources on local fog nodes. The advent of Blockchain (BC) technology has become a remarkable, most revolutionary, and growing development in recent years. BC's open platform stresses data protection and anonymity. It also guarantees data is protected and valid through the consensus process. BC is mainly used in money-related exchanges; now it will be used in many domains, including healthcare; This paper proposes efficient Blockchain-based secure healthcare services for disease prediction in fog computing. Diabetes and cardio diseases are considered for prediction. Initially, the patient health information is collected from Fog Nodes and stored on a Blockchain. The novel rule-based clustering algorithm is initially applied to cluster the patient health records. Finally, diabetic and cardio diseases are predicted using feature selection based adaptive neuro-fuzzy inference system (FS-ANFIS). To evaluate the performance of the proposed work, an extensive experiment and analysis were conducted on data from the real world healthcare. Purity and NMI metrics are used to analyze the performance of the rule based clustering and the accuracy is used for prediction performance. The experimental results show that the proposed work efficiently predicts the disease. The proposed work reaches more than 81% of prediction accuracy compared to the other neural network algorithms.

INDEX TERMS Fog computing, blockchain, clustering, classification, fuzzy, disease prediction.

I. INTRODUCTION

Enduring technical advancements provide significant opportunities for biomedical innovation and cost savings, but also pose an obstacle for the integration of emerging technology into medical treatment [1]. A considerable volume of work is primarily focusing on smart healthcare to address conventional healthcare limitations and satisfy rising expectations for premium healthcare. Smart healthcare could be designed and developed as a range of devices, tools, software, facilities, and organizations with conventional healthcare, biosensors,

The associate editor coordinating the review of this manuscript and approving it for publication was Amir Masoud Rahmani¹.

connected apps, and smart emergency service systems [2]. The cornerstone of intelligent healthcare is IoT end nodes that include a wide range of medical equipment and applications that link to healthcare through the Internet. Fog computing is an extension of cloud computing which can process and archive vast quantities of data that IoT devices produce near their origins.

A. MOTIVATION

Fog computing is considered to be one of the key technologies that contribute greatly to promoting IoT healthcare and surveillance applications as these systems are latency-sensitive and real-time tracking, data processing, and

decision making are critical criteria in healthcare applications such as servicing the elderly by home nursing, heart care, diabetes and some other diseases. Health data is an important topic because it includes essential, confidential knowledge. With fog computing, the aim that patients take care of their own health data locally is realized. Those safety data are housed in fog nodes such as smart phones or smart vehicles [13]. Fog computing provides tremendous advantages for fog-based application which is prone to delay. Hong *et al.*, [12] introduced Mobile Fog, which is a globally dispersed and latency-sensitive programming paradigm for Internet applications. A variety of studies has looked into the use of fog in health care. This motivates to develop the fog based health care prediction.

B. PROBLEM STATEMENT

Health care contributors are widely implicit in producing large volumes of information in a variety of formats, together with records, economic papers, clinical test findings, imaging tests, and vital sign assessments, etc., [10]. The comprehensive database created in care environments is expanding rapidly, with healthcare information struggling from numerous problems, with data access, and how information can be obtained beyond the healthcare ability. Blockchain provides the ability to enhance the data's authentication and legitimacy. It also helps to disseminate data inside the network or services. Such apps affect the cost, quality of data, and importance of providing health care within the system. Blockchain is a transparent, decentralized network without the middleman [14]. Blockchain healthcare networks do not need several verification rates which have access to data for anyone who is part of the infrastructure of blockchain. Data is rendered available to consumers and is transparent. Such innovations will continue to overcome the numerous problems facing the healthcare domain today.

Disease prediction is one of the main real-world problems in healthcare domain. Many classification algorithms [31], [33] are used to predicts the diseases accurately. Artificial neural network (ANN) is one of the classification algorithms. ANN is a massively computational parallel model with self-adaptive and self-learning capabilities, because of its large parallel structure; it takes more time to predict the outcome. ANN is not appropriate for dealing with such issues, such as ambiguous and imprecise data for which problems of uncertainty may occur at any point of the process of classification.

Fuzzy logic is used to resolve this issue in order to translate the numeric input features into their corresponding linguistic terminology. Based on linguistic properties such as low, medium and high, each input function is transformed into its corresponding membership values in this fuzzification process. Similarly, from the input features, all linguistic characteristics are extracted. By deciding the membership value in different linguistic terms, fuzzy logic is also sufficient to deal with the ambiguity problem. Adaptive Neuro Fuzzy

Inference System (ANFIS) is a hybrid model which adopts the characteristics of ANN and fuzzy logic.

C. CONTRIBUTIONS

The objective of this paper is to develop the disease prediction model using feature selection and ANFIS. Feature selection is the one of the pre-processing technique which reduces the size of the dimensionality of the dataset. This paper use Cronbach's alpha [41] for optimal feature selection.

The significant findings in this paper as follows:

- A semi-centralized Blockchain-based digital healthcare network for the protection and sharing of patient data is introduced to ensure safe and effective data storage and data sharing.
- The rule-based clustering algorithm is used to group the diabetic and cardio disease patient records.
- After this clustering, diabetic and cardio disease is predicted using Feature selection based ANFIS.
- Finally, the model is created to evaluate the performance of the proposed work in terms of various metrics.

D. PAPER ORGANIZATION

The remaining of the paper is organized as follows: The background of fog computing and blockchain explained in section II and Section III describes the reviews of the related work. Section IV explains the system and data model. The proposed methodology is defined in Section V. The experimental results are analyzed in Section VI and finally, Section VII, concludes the paper.

Fog computing is considered to be one of the key technologies that contribute greatly to promoting IoT healthcare and surveillance applications as these systems are latency-sensitive and real-time tracking, data processing, and decision making are critical criteria in healthcare applications such as servicing the elderly by home nursing, heart care, diabetes and some other diseases. Health data is an important topic because it includes essential, confidential knowledge. With fog computing, the aim that patients take care of their own health data locally is realized. Those safety data are housed in fog nodes such as smartphones or smart vehicles [13]. Fog computing provides tremendous advantages for fog-based application which is prone to delay. Hong *et al.*, [12] introduced Mobile Fog, which is a globally dispersed and latency-sensitive programming paradigm for Internet applications. A variety of studies has looked into the use of fog in health care.

Health care contributors are widely implicit in producing large volumes of information in a variety of formats, together with records, economic papers, clinical test findings, imaging tests, and vital sign assessments, etc., [10]. The comprehensive database created in care environments is expanding rapidly, with healthcare information struggling from numerous problems, with data access, and how information can be obtained beyond the healthcare ability. Blockchain provides the ability to enhance the data's authentication and legitimacy. It also helps to disseminate data inside the

network or services. Such apps affect the cost, quality of data, and importance of providing health care within the system. Blockchain is a transparent, decentralized network without the middleman [14]. Blockchain healthcare networks do not need several verification rates which have access to data for anyone who is part of the infrastructure of blockchain. Data is rendered available to consumers and is transparent. Such innovations will continue to overcome the numerous problems facing the healthcare domain today.

Disease prediction is one of the main real-world problems in healthcare domain. Many classification algorithms [31], [33] are used to predict the diseases accurately. Artificial neural network (ANN) is one of the classification algorithms. ANN is a massively computational parallel model with self-adaptive and self-learning capabilities, because of its large parallel structure; it takes more time to predict the outcome. ANN is not appropriate for dealing with such issues, such as ambiguous and imprecise data for which problems of uncertainty may occur at any point of the process of classification.

Fuzzy logic is used to resolve this issue in order to translate the numeric input features into their corresponding linguistic terminology. Based on linguistic properties such as low, medium and high, each input function is transformed into its corresponding membership values in this fuzzification process. Similarly, from the input features, all linguistic characteristics are extracted. By deciding the membership value in different linguistic terms, fuzzy logic is also sufficient to deal with the ambiguity problem. Adaptive Neuro Fuzzy Inference System (ANFIS) is a hybrid model which adopts the characteristics of ANN and fuzzy logic.

The objective of this paper is to develop the disease prediction model using feature selection and ANFIS. Feature selection is the one of the pre-processing technique which reduces the size of the dimensionality of the dataset. This paper use Cronbach's alpha [41] for optimal feature selection.

The significant findings in this paper as follows:

- A semi-centralized Blockchain-based digital healthcare network for the protection and sharing of patient data is introduced to ensure safe and effective data storage and data sharing.
- The rule-based clustering algorithm is used to group the diabetic and cardio disease patient records.
- After this clustering, diabetic and cardio disease is predicted using Feature selection based ANFIS.
- Finally, the model is created to evaluate the performance of the proposed work in terms of various metrics.

The remaining of the paper is organized as follows: The background of fog computing and blockchain explained in section II and Section III describes the reviews of the related work. Section IV explains the system and data model. The proposed methodology is defined in Section V. The experimental results are analyzed in Section VI and finally, Section VII, concludes the paper.

II. BACKGROUND

A. FOG COMPUTING

It is a distributed computing framework that expands the network's cloud infrastructure to the edge. It supports the operation and configuration of data center and end-user processing, networking, and storage facilities. Fog computing generally comprises specifications of the software that operates between sensors and the cloud, i.e., smart access points, routers or advanced fog devices, in both the cloud and edge applications. Fog computing embraces agility, computational power, networking protocols, the flexibility of the interface, cloud convergence, and disseminated data analytics to meet requirements of applications requiring short latency with large and compact geographic delivery [3].

Cisco initially coined the word fog computing [6]. Open Fog Consortium [7] describes fog computing as: 'a horizontal system-level architecture which distributes computation, storing, controlling and networking tools and services everywhere in the Cloud to Things spectrum.' The author in [8] defined as, "A situation in which a vast amount of heterogeneous, omnipresent and autonomous computers interacts and theoretically collaborate and with the network to execute storage and processing activities without third-party intervention. These activities may be to support simple network operations or new technologies and applications operating in a sandboxed environment".

The structure of fog computing is shown in Fig. 1. The cloud layer, which is the cornerstone of fog computing, conducts data virtualization, analysis, deep learning, and in the proxies of the fog layer updates laws and patterns. The proxy server acts as a web service and is more manageable. A centralized data collection enables creditworthiness and convenient data access through storing power within a cloud. A data store situated in the center of the fog computing system can be reached from both the computer layer and the fog layer [4].

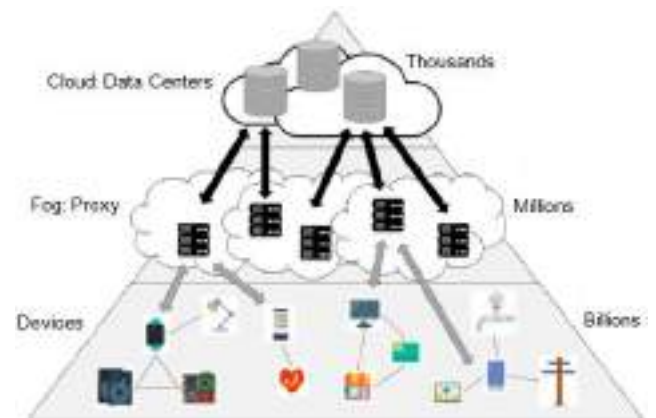


FIGURE 1. Fog computing structure [4].

Fog computing's characteristics include location recognition and low latency, spatial reach, scalability, accessibility support, real-time communications, convergence, interoperability, web analytics support, and cloud interplay. Reduced

network load, automatic connectivity assistance, context awareness, no single fault point, improved market resilience, low latency, local and large-scale delivery, reduced running costs, versatility and heterogeneity are the benefits of fog computing [5].

The Fog computing network has a wide range of applications. Fig 2 shows the applications supported by fog computing.

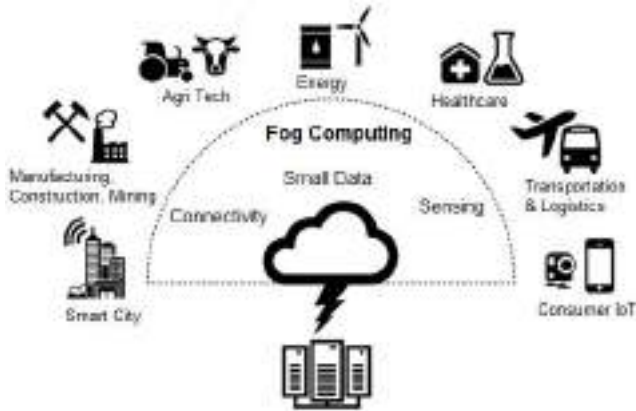


FIGURE 2. Fog computing applications.

In this paper, fog computing is used in the healthcare domain. Fog computing is a vital aspect of healthcare. It provides responses that are crucial in healthcare monitoring and incidents in real-time. Furthermore, the integration with a vast number with healthcare systems for remote collection, distribution, and cloud retrieval of medical data involves a secure network link that is not accessible.

B. BLOCKCHAIN

Blockchain is one of the most innovative technologies and a digital wallet which retains track of transactions and events occurring across the network, and whose integrity is ensured via a peer-to-peer computing network, not by any centralized entity that might eliminate the risk of a single central point. It is composed of structured documents organized in a block structure that includes transaction batches and previous key hash. Every block is chronologically linked, and the data on the Blockchain network is unchallengeable [9].

Any users have individual access rights in a blockchain network to allow transactions that are modified throughout the framework, known as consensus protocol [10]. For inserting transactions, a blockchain uses SHA256 hash. The NSA creates that, which is 64 characters large. All transactions are registered in a blockchain network though not modifying or manipulating the public ledger; Both transfers are distributed to various users across the network to transfer and update the data; a blockchain network may be duplicated to a separate venue, for example, within the same ability or healthcare distribution network, or as part of a regional or global data exchange system.

The Blockchain’s data structure is a hierarchical set of blocks shown in fig 3. Blocks are linked in the form of a tuple, while the current block stores such values as previous block hash, previous block Blockchain address etc. in its header. Every block is composed of two components: header and body. The header contains block number, previous block hash value to preserve chain reliability, current block body hash to protect transaction data integrity, timestamp, nonce, blockchain block creator address and other requested detail. Block bodies contain one or more transactions.

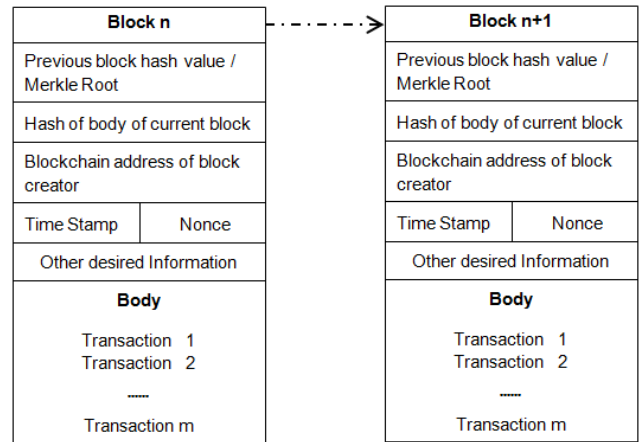


FIGURE 3. Block structure.

Decentralization, Durability, Transparency and Auditability are primary aspects of Blockchain. Public, private and consortium are kinds of Blockchain [11]. All archives are available to the public in the public Blockchain so that anyone may engage in the consensus process. Despite this, the consensus mechanism of a cooperative network will require only a collection of pre-selected nodes. As for private Blockchain, only those nodes originating from a single entity will be permitted to join the consensus process.

III. RELATED WORK

A. FOG COMPUTING IN HEALTHCARE

Health Fog framework is proposed in [15]. Fog computing is used as an intermediate layer among the cloud and end-users. Authors primarily focused on developing and addressing data protection problems in healthcare systems in a scalable way. Cloud access authentication agent is combined with Health Fog to enhance the security of the network. Besides, cryptographic features also specified to enhance Health Fog efficacy. The remote control of the patient’s healthcare in smart homes is introduced in [16] based on the principle of fog computing at the intelligent access point. For handle the patient’s real-time data at the fog layer, an event-based approach is adopted for initiating data transmission. The theory of immediate mining is used to evaluate incident difficulties by calculating the index of the temporal health of the victim.

Gia *et al.* [17] improve the health management program by leveraging the idea of fog computing at smart gateways offering specialized technologies and facilities such as distributed data processing, centralized storage and network-side monitoring. The author selects the Electrocardiogram (ECG) feature extraction as a case study. The ECG signals are analyzed with extracted features in smart gateways. Negash *et al.* [18] focus on developing an intelligent e-health interface being used in the Fog computing layer, linking a network of these gateways, both for home use and hospital use. Gateway technologies are addressed and tested when applying fog.

The idea of Fog Computing in Healthcare IoT systems is proposed in [19] via the creation of a Geo dispersed intermediate layer of information among sensor nodes and the cloud. A concept for the implementation of an intelligent e-health interface is being introduced. An IoT-based premature caution score safety screening is introduced to demonstrate the system's efficacy in a health case study. In [20], the authors propose a hierarchical computing model supported by fog for remote IoT-based patient management systems. The distributed computing system allows for the partitioning and distributing of analytics and decision-making among the fog and the cloud.

In a healthcare context, Alazeb and Panda [21] presented two separate frameworks for using fog computing. The two models are heterogeneous and homogeneous data from fog modules. They suggest a unique approach for each model to assess the harm done by malicious transactions so that actual data may be retrieved, and transactions marked for potential inquiry can be impacted. In [22], a novel architecture called Health Fog is proposed to incorporate deep learning ensemble into Edge computing devices and implemented it for real-life implementation of automated cardio disease detection. It offers healthcare as a fog service using IoT devices, handles heart patient information effectively, and comes as app requests.

B. BLOCKCHAIN IN HEALTHCARE

Healthcare information-sharing network based on Blockchain is proposed in [23]. The author uses two liberally-coupled Blockchain to manage various forms of healthcare information and also incorporates off-chain storage and on-chain authentication to meet safety and authenticity criteria. Liang *et al.*, [24] suggest a revolutionary user-centric health data exchange approach through the use of a decentralized and approved Blockchain for guarding confidentiality using the channel creation method and improve individuality protection via the blockchain-based relationship program. Evidence of validity and authentication is indefinitely recoverable from the cloud database and embedded in the blockchain network to protect the confidentiality of health records inside each document.

A secure and privacy-conserving blockchain-based PHI networking scheme was proposed in [25] for improving diagnosis in e-Health scheme. Private and consortium Blockchain is developed through the creation of their information

structures and consensus mechanisms. The private ledger manages the PHI while the ledger community keeps a database of the robust indexes of the PHI.

Griggs *et al.*, [26] propose smart, blockchain-based contracts to enable secure medical sensor research and management. The author built a network based on the Ethereum protocol using a private blockchain where the sensors connect with a mobile computer that calls smart agreement and mark logs of every activity on the Blockchain. In [27], a blockchain-based system is introduced for safe, interoperable, and proficient access by patients, clinicians, and third parties to medical data while maintaining the confidentiality of personal details of patients. Through an Ethereum-based blockchain, it makes use of smart agreement to boost access control and code obfuscation, using advanced cryptographic methods for enhanced protection.

In [28], a novel framework for the storage of medical data based on Blockchain was introduced. Users should retain valuable data in perpetuity, so where interference is alleged, the originality of the data may be checked. The author makes use of wise data management techniques and a number of cryptographic methods to protect user confidentiality. MedBlock, a blockchain-based information management program, was introduced in [29] for managing information from patients. The centralized MedBlock database in this system allows or secure entry and storage of medical information. The improved consensus process creates consensus on medical history without significant energy consumption and network congestion.

C. DISEASE PREDICTION

A novel Optimistic Unlabeled learning strategy was introduced in [30], based on clustering and 1-class classification method. This method initially clusters positive data, studies 1-class classifier models using clusters, selects negative data intersection as the Stable Negative set, and finally uses binary SVM (Support Vector Machine) classification algorithm. In [31], a scheme called ensemble classification was investigated, which is employed by combining multiple classifiers to improve the precision of weak algorithms. The author applies the algorithm for a medical dataset, demonstrating its early utility in forecasting disease.

In [32], an appropriate segmentation and classification method is presented to discern the progression of Alzheimer's disease, moderate neurological dysfunction, and common objects of control correctly. A fusion segmentation method is invented to perform segmentation using K-means clustering and graph-cutting schemes. Depending on their characteristics, the clustered regions are given labels for the classification analysis. Nilashi *et al.* [33] are developing a new knowledge-based prediction method for diseases using clustering, noise reduction, and simulation methods. Classification and Regression Trees algorithm is used to produce the knowledge-based system's fuzzy rules.

An updated variant of K-Means based on density was introduced in [34], which provides an innovative and logical

approach for choosing the early centroids. The algorithm's main concept is to pick data points that belong to dense regions and which are appropriately segregated as the initial centroids in feature space. This approach makes comparatively improved estimates of subtypes of cancer from evidence regarding gene expression. A classification algorithm for managing imbalanced datasets was introduced in [35] based on the principle of information granulation (IG). This algorithm assembles data from majority classes into granules to balance the class ratio inside the data. This algorithm first produces a collection of IGs using meta- heuristic methods and applies the data classification algorithm.

An edge-cloud-based healthcare infrastructure is proposed in [46] for real-time disease detection, monitoring, and recovery. This approach does not consider the blockchain concept. The proposed method uses blockchain for securing patient health record.

IV. SYSTEM MODEL

This section explains the proposed system model and notations used in this model. In this model, the IoT medical sensors are used to collect, patient health related data. The fog nodes collect these data and send to medical analyzer for disease analysis and prediction. Fig 4 shows the system model. It contains five entities.

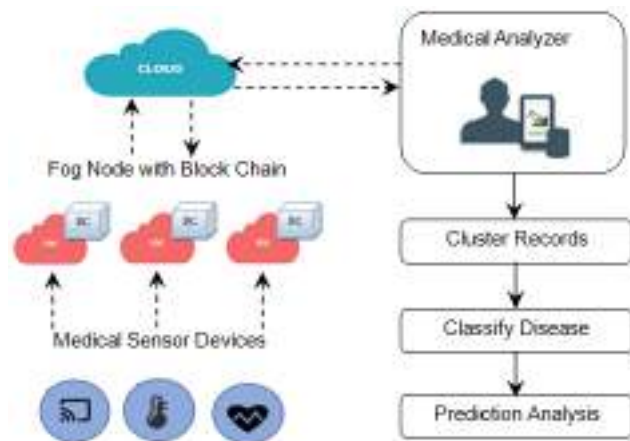


FIGURE 4. Proposed system model.

A. MEDICAL SENSOR DEVICES

Sensor devices can track human health parameters of various sorts, whether wearable systems or embedded devices. Due to their restricted computing and storage capacities, these devices collect different types of health-related data and send data that will be well managed to fog nodes.

B. FOG NODE

It is a simple platform for fog computing, which can be a network computer that manages underlying machines using processing resources, dedicated servers, or computational

servers. It collects the data from the medical sensor devices and stores into a distributed ledger called Blockchain.

C. BLOCKCHAIN

It is a cooperative network used to monitor patient health data and activity data status. No-one can access the network without authorization. This is composed of a sequence of blocks containing the previous hash block, status user health.

D. CLOUD

It is used for storage purposes. It stores encrypted patient health information, and the authenticated medical analyzer can access these encrypted data for further process.

E. MEDICAL ANALYZER

An authorized person who can access patient health information. The analyzer can group the information into two: normal patient and affected patient. The analyzer can also predict whether the patient contains diabetic or cardio diseases.

Table 1 show the notations used in clustering and classification process.

TABLE 1. Notations.

Notation	Description
D	Dataset
F_i	i^{th} feature in D
DR_i	i^{th} data record in D
A_{ij}	Attribute Value of i^{th} data record and j^{th} feature
RS	Rule Set
R_i	i^{th} rule in RS
Freq<R,C>	Frequent Rule Set (R=rule, C=Count)
R_{thr}	Rule Threshold
L+R=C	Left, Right and Class part (Rule)
cand+	Positive candidate rules
cand-	Negative candidate rules
Cl _s +	Positive clustered data
Cl _s -	Negative Clustered data
C_α	Cronbach's alpha

V. PROPOSED METHODOLOGY

This section explains the proposed Blockchain-based healthcare disease prediction with clustering and classification.

A. BLOCKCHAIN STORAGE

In the medical domain, control of access, validity, data confidentiality and integration are essential to protecting the identity of the patient and sharing data within the healthcare environment with other organizations. The traditional way to achieve control of access usually implies confidence among the data owner and the entities that store them. Such agencies are also entirely assigned servers for identifying

and implementing policies on access management. Interoperability is the capability of dissimilar information systems, software or frameworks to link data between stakeholders in a synchronized way, within and across organizational borders, to improve individual safety. The provenance of data relates to the historical record of the data and its sources, e.g., provenance in health domain data may be to provide auditability and consistency in the health record and to attain trust in the electronic health record software framework. Data integrity is the concept of data validity that concerns with the consistency required of the information. That ensures the level to which the intended data quality is achieved or surpassed decides the validity of the report [36]. Blockchain technology has several enticing features that can be used to enhance and gain a higher degree of integration, sharing of knowledge, access security, validity, and data transparency between the stakeholders listed, while trying to move towards a novel trust-building and sustaining infrastructure.

Blockchain can be described as a blockchain, capable of storing stable and permanent transactions between parties. Each block contains many elements including, user submitted valid transactions, time-stamped batches, and the previous block hash. A hash function is a function that transforms the data, it is given into a fixed-length irregular form. The timestamp reveals there must have been data at the time. The previous block hash ties the blocks together and forbids modification of any block or addition of a block between two different blocks. Blockchains allow auditing and traceability by connecting a new block to the previous one by using the latter's hash, and thereby creating a blockchain. The block transactions are generated in a Merkle tree (Fig 5) where the known root can be verified for each value of the leaf (transaction). Any non-leaf node in the Merkle tree is the hash of the values of its infant nodes. Searching for a transaction becomes really quick through using Merkle tree. Instead of checking the transactions linearly, the Merkle tree will determine more quickly whether a transaction is found in the block or not.

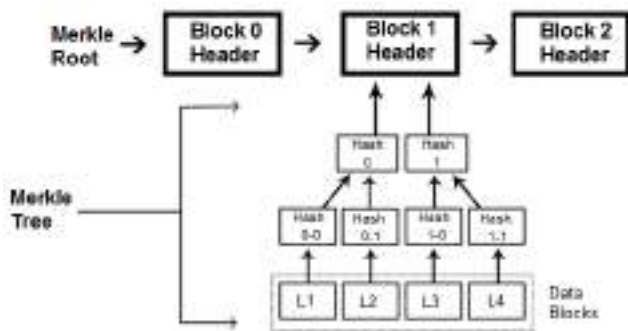


FIGURE 5. Merkle tree structure in blockchain.

This paper considers the consortium type of Blockchain, also known as semi-decentralized Blockchain. A consortium blockchain is not provided as a private blockchain to a single entity; it is conferred on a group of approved entities instead. Additionally, the blockchain consortium is

a group of predefined nodes on the network. Consortium blockchain, therefore, provides security, inherited from public Blockchain. This gives a significant degree across the network. Consortium blockchains are most commonly associated with commercial use, as a consortium of the company's works together to use blockchain technologies to boost businesses. However, this kind of Blockchain may enable specific group members to access or adopt a hybrid method of access. The root hash and its Application Program Interface (API) may be publicly accessible. External entities can, therefore, use the API to conduct several inquiries and to obtain specific information relating to the blockchain status. Table 2 shows some properties [37] of consortium blockchain.

TABLE 2. Consortium blockchain property.

Property	Value / Description
Consensus	Handled by set of nodes
Transaction Validation	Set of Authorized nodes
Transaction Reading	Any node or set of predefined node
Data Immutability	Yes
Transaction Throughput	High
Network Scalability	Low to Medium
Infrastructure	Decentralized
Features	<ul style="list-style-type: none"> ○ Applicable to tightly controlled business ○ Fee-free transaction ○ Laws on services are easier to manage ○ Effective defense from outside perturbations
Example	Hyperledger, Ethermint, Tendermint

The authorized medical analyzer collects patient information and predicts whether the patient contains diabetic or heart-related diseases.

B. DISEASE CLUSTERING

Clustering is one of the unsupervised techniques in data mining that deal with identifying groups inside a collection in unlabeled data. It is used to partition a set of data into different clusters, such that objects in the same group cluster are strongly related and distinct from objects in another cluster. Clustering technology has been widely accepted in many technologies such as pattern detection, image processing and pattern analysis of consumer transactions. It is essential during data analysis discovery and assessment, where researchers seek to find fundamental features that appear without previous knowledge of the data. However, the selection of appropriate clustering techniques and algorithms is determined by an interpretation of the data structure, the form of analysis to be carried out and the scale of the dataset.

Cluster classification in the medical domain provides a standardized, formalized approach for data discovery and identifying clinically related groupings. Efficient clustering methods are raising competition for costly health care services. It helps doctors deal with the influx of knowledge, and can assist with better facilities in strategic planning. The findings of the clustering are used to research patient independence or association and for more in-depth insight into evidence from medical surveys. All these advantages inspired the researcher to construct clustering models for grouping medical data.

Health data clustering raises a variety of new problems.

- o Information overload – Developments in medical technology combined with high processing capacities are increasing the volume of data generated and processed in the healthcare sector. Discovery of knowledge and the retrieval of information from these large databases are difficult and prohibitively costly.
- o Too many risk indicators are essential for decision-making and are heterogeneous.
- o High consumer knowledge of medical treatment and improved life expectancy creates a rising demand for better health services. Yet misdiagnosis and imprecise care strategies arise with overworked and inexperienced doctors, challenging working environments etc.
- o Choosing a suitable clustering approach and an adequate number of clusters in health care data can be challenging and often complicated.

To address this challenge, a novel rule-based clustering algorithm is proposed for the efficient cluster. This is a two-stage algorithm: in the first stage, the rules are generated based on patient information, and in the second stage, the clusters are generated based on the rules.

The pseudo-code of the rule generation algorithm has been given as follows.

This algorithm is suitable for a numerical data set. Initially, the numerical value is converted into discrete value (Low, Medium, and High) (steps 2- 12). Based on these values, the candidate rules (13-19) are generated for further process. This paper use frequency and threshold based rule generation. Based on the requirements, the candidate rules are extracted.

Consider the 15 patients fasting blood sugar level, 120, 90, 70, 45, 100, 130, 50, 35, 138, 82, 90, 50, 120, 58, 140. Table 3 shows the example.

Convert all the features values in the dataset. Count the frequencies of each record. If the record frequency is more than the R_{thr} (initially set 5 – 10 depending on the requirements), then consider the record as candidate rule. The next stage is clustering. The pseudo-code of the clustering algorithm has been given as follows.

The candidate rules are divided into three parts ($L + R = C$), i.e. left, right and a class variable. Based on the C (class variable), $cand_+$ and $cand_-$ rules are generated. Positive and negative clusters are formed based on these candidate rules if any record not matched with candidate rules then it will be considered as an outlier record.

Algorithm 1 Rule Generation

```

Input: D
Output: RS
1: RS = ∅
2: for each  $F_i \in$  Feature do
3:    $distF_i =$  get distinct value( $F_i$ )
4:   Sort( $distF_i$ )
5:   Group  $distF_i$  values into Low, Medium and High
6: end for
7: for each  $DR_i \in$  DataRecord do
8:   for each  $F_j \in$  Feature do
9:      $newA_{ij} =$  convert  $A_{ij}$  into Low, Medium, High based on Step 5
10:  end for
11: end for
12: generate newDR based on  $newA_{ij}$ 
13:  $Freq_{\langle R,C \rangle} =$  Find and Count Similar Records
14: candidate =  $Freq_{\langle R,C \rangle} \forall c > R_{thr}$ 
15: If candidate  $\neq \emptyset$ 
16:   RS = candidate
17: else
18:   RS = ∅
19: end if
    
```

TABLE 3. Data conversion example.

Steps	Value
Input Feature	120, 90, 70, 45, 100, 130, 50, 35, 138, 82, 90, 50, 120, 58, 140
Distinct Value	120, 90, 70, 45, 100, 130, 50, 35, 138, 82, 58, 140
Sort Value	35, 45, 50, 58, 70, 82, 90, 100, 120, 130, 138, 140
Group Values	(35, 45, 50, 58) = Low (70, 82, 90, 100) = Medium (120, 130, 138, 140) = High
Convert Feature	High, Medium, Medium, Low, Medium, High, Low, Low, High, Medium, Medium, Low, High, Low, High

C. DISEASE PREDICTION

Processing of medical data is a critical topic that needs to be accurate for disease prevention, diagnosis and processing. Maintaining health records has been a pivotal scientific mission. Patient data comprising of specific disease-related characteristics and symptoms will be reached with special caution to ensure professional treatment. Because the information stored in medical repository can include incomplete and redundant information, that medical data is inefficient [38]. Until implementing data mining algorithms, it is essential to contain effective data planning and reduction because this can impact the mining performance. Disease diagnosis is quicker and easier if the data is accurate, reliable and noise-free.

Selecting a feature is an effective pre-processing method in data mining designed to reduce data dimensionality.

Algorithm 2 Clustering

```

Input: D, RS
Output: Cls+, Cls-
1: cand+ = ∅, cand- = ∅
2: for each Ri ∈ RS do
3:   Split Ri into three parts (L + R = C)
4:   cand+ = L + R = C (rule with positive patients)
5:   cand- = L + R = C (rule with negative patients)
6: end for
7: for each rec ∈ newDR
8:   if (rec match with cand+) then
9:     Cls+.add(rec)
10:  else (rec match with cand-) then
11:    Cls-.add(rec)
12:  else
13:    Out.add(rec)
14:  end if
15: end for
    
```

Identifying the most severe disease-related risk factors is very important in medical diagnosis. Specific recognition of features helps delete unwanted, unnecessary features from the dataset of the disease, resulting in a simple and improved outcome. Classification and prediction is a technique of data mining that initially utilize training data to create a training model and then applies the resulting model to test data to achieve predictive results. Diverse recognition systems have been applied to disease data sets for diabetes and cardiovascular disease treatment. This paper proposes a Feature Selection and use Adaptive Neuro-Fuzzy Inference System [39], which adopts the characteristic of ANN and Fuzzy Logic for disease prediction. Fig 6 shows the prediction model workflow.

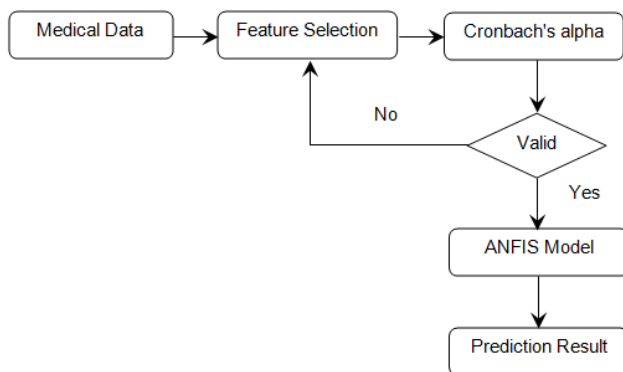


FIGURE 6. Prediction work flow.

Feature selection is a commonly used data pre-processing method in data mining that is essentially used to reduce data by removing irrelevant and redundant features from the dataset [40]. In addition, this method increases data interpretation, improves information analysis, decreases learning algorithm training times and increases prediction efficiency.

To collect more useful knowledge, different feature collection methods have been applied to the healthcare datasets. The use of feature selection methods is performed on clinical databases to predict various diseases. Different learning algorithms operate effectively and provide more reliable outcomes if there are more important and non-redundant attributes in the details. Given the vast number of redundant and unnecessary features in the medical datasets, an effective feature extraction strategy is required to mine fascinating attributes specific to the disease.

This paper proposes an optimal feature selection algorithm which uses Cronbach's alpha [41]. The Cronbach alpha measures the consistency of features in a test, i.e. the test's internal consistency. It can be measured by,

$$C\alpha = \frac{|F| \cdot CV_{avg}}{V_{avg} + (|F| - 1) \cdot CV_{avg}} \quad (1)$$

Where |F| = number of features, CV_{avg} = average of covariance, V_{avg} = average variance.

The pseudo-code of the feature selection algorithm has been given as follows.

Algorithm 3 Feature Selection

```

Input: D
Output: SF (Selected Features)
1: pc = 10, global_Cα = 0, maxIter = 100
2: for i = 1 to pc do
3:   popij = Random{0, 1}, j ∈ Fj
4:   Cαi = 0
5: end for
6: for iter = 1 to maxIter do
7:   for i = 1 to pc
8:     compute Cronbach's alpha (Ca) using (1)
9:     if (Ca > Cαi) then
10:      Cαi = Ca
11:    end if
12:  end for
13:  maxCa = max(Cαi)
14:  if (maxCa > global_Cα) then
15:    global_Cα = maxCa
16:    SF = popi(index of maxCa)
17:  end if
18:  Replace the pop which contain lowest Ca
19: end for
    
```

Randomly generate the population using the random function and assign alpha as zero (steps 2 – 5). An iterative process is used to select optimal features (6-19). The maximum iteration is set as 100. Compute Cronbach's alpha (using (1)) for each randomly generated population. Select the maximum alpha value (step 13) and population if it is more than global alpha then set selected features as population (step 16). Change the population, which contains the lowest alpha (step 17) repeat steps (6-19) until maximum iteration reached. The selected features are used in the ANFIS model to predict the disease.

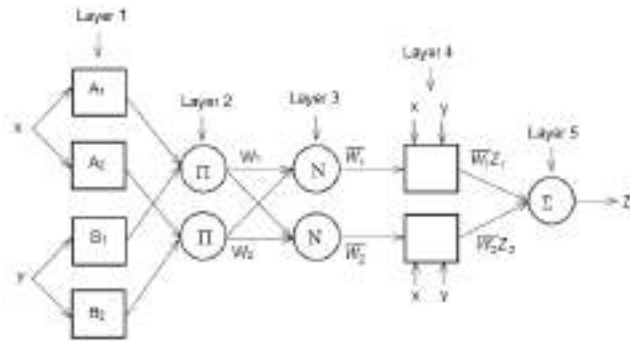


FIGURE 7. ANFIS architecture.

The ANFIS network is a neuro-fuzzy network developed by Jang in 1993 [42]. Because of ANFIS ‘adaptive property, some nodes obtain the same property, and after that, the output comes based on the constraints that belong to those nodes. For efficient optimization, two learning methods are used to adjust constraints. Of convenience, the above-suggested method should have 2-inputs and 1-output, and its rule base includes two fuzzy if-then TSK [43] fuzzy model rules. This TSK model generates fuzzy rules from the dataset input-output. If $x = A$ and $y = B$, $z = f(x, y)$. Here, $f(x, y) =$ flat function that typically denotes a polynomial.

The ANFIS architecture is depicted in fig 7. The function of each layer is defined below.

Layer 1: This layer is the membership layer which contains adaptive nodes with node functions defined as

$$L_i^1 = \mu_{A_i}(x) \quad (i = 1, 2) \tag{2}$$

$$L_i^1 = \mu_{B_{(i-2)}}(y) \quad (i = 3, 4) \tag{3}$$

where x and y denote input nodes, A and B are linguistic labels, $\mu(x)$, and $\mu(y)$ refer to membership functions.

Layer 2: This layer adopts the ‘set node’ property and each node is labeled with a ring symbol and named with multiplying the node function to act as output through input. Consider

$$L_i^2 = \omega_i = \mu_{A_i}(x) \mu_{B_i}(x) \quad (i = 1, 2) \tag{4}$$

The output ω_i represents the rules firing strength.

Layer 3: Each node in this layer is labeled with a ring symbol and called N, with the node function to regulates the firing force by measuring the proportion of the firing force of the i th node to the sum of the firing power of all laws. In fact,

$$L_i^3 = \bar{\omega}_i = \frac{\omega_i}{\sum \omega_i} = \frac{\omega_1}{\omega_1 + \omega_2}, \quad (i = 1, 2) \tag{5}$$

The outputs of that layer are called to as standardized firing ability for ease.

Layer 4: In this layer, each node is in nature, flexible, and is noticeable with a square. Node role is specified by

$$L_i^4 = \bar{\omega}_i \cdot f_i = \bar{\omega}_i (p_i x + q_i y + r_i), \quad (i = 1, 2) \tag{6}$$

where $\bar{\omega}$ is the output of layer 3 and $\{p_i, q_i, r_i\}$ is the set of parameters.

Layer 5: Each node within this layer is a constant node, and the overall result can be expressed as a linear mixture of the following parameters. Two parameter sets can be modified, $\{a_i, b_i, c_i\}$ marked as parameters of the assumption and $\{p_i, q_i, r_i\}$ marked as the subsequent parameters. The training process must harmonize the two parameters that are set to predict successful outcomes.

VI. EXPERIMENTAL RESULT

In this section, the performance of the proposed work was analyzed. The proposed work was implemented using Java (version 1.8), and the experiments are performed on an Intel(R) Pentium machine with a speed 2.13 GHz and 4.0 GB RAM using Windows 7 32-bit Operating System.

A. DATA SET

The two dataset diabetes and heart disease data set is used for the experimental result. The diabetes data set contains 768 instances, with eight numeric features. Table 4 shows the data set information.

TABLE 4. Diabetes data set information.

Feature Name	Description	Range
Pregnancies	Number of pregnancies	0 – 17
Glucose Level	Plasma glucose level	44 - 199
BP Level	Diastolic hypertension	24 - 122
Skin Thickness	The thickness of Triceps skin fold	7 – 99
Insulin	Insulin serum for 2-hours	14 – 846
BMI	Body mass index	18.2 - 67.1
Pedigree Function	A pedigree function of diabetes	0.078 – 2.42
Age	Age in Years	21 – 81
Class Label	The patient has diabetes or not	0 or 1

The heart disease data set contains 800 instances, with six numeric features and eight categorical attributes. Table 5 shows the data set information.

B. EVALUATION METRICS

This section explains the evaluation metrics for the experimental result.

1) PURITY

This measure evaluates the clustering consistency. The purity of the final clusters can be seen when opposed to the

TABLE 5. Heart disease data set details.

Feature Name	Description	Range
Age	Age in years	29 – 77
Sex	Patient Gender	0, 1
CPT	Chest pain type	1, 2, 3, 4
Trest_bps	Resting BP	94 – 200
Chol	Cholesterol in Serum	126 – 546
FBS	Fasting Blood Sugar	0, 1
RestECG	Resting Electrocardiographic	0, 1, 2
Thalach	Maximum Heart rate achieved	71 – 202
Exang	Exercise-Induced Angina	0, 1
OldPeak	ST depression induced by exercise relative to rest	0 – 6.2
Slope	Slope of the peak exercise	1, 2, 3
CA	No of major vessels	0, 1, 2, 3
Thal	Defect value	3, 6, 7
Class Label	Patient have heart disease or not	0 or 1

ground truth groups. It can be calculated as,

$$Purity = \frac{\sum_{i=1}^{|C|} n_i^d}{|C|} \tag{7}$$

where $|C|$ is the total number of clusters, n_i^d is the number of instances with the leading class label in Cluster C_i and n_i indicates the number of the instances in the cluster C_i

2) NMI (NORMALIZED MUTUAL INFORMATION)

It measures the mutual experience, followed by a normalization process, between the resulting cluster labels and ground truth labels. It can be calculated as

$$NMI = \frac{\sum_{i,j} n_{ij} \log \frac{n * n_{ij}}{n_i * n_j}}{\sqrt{(\sum_i n_i + \log \frac{n_i}{n})(\sum_j n_j + \log \frac{n_j}{n})}} \tag{8}$$

where n_{ij} is the number of instances belonging to the class i found in the cluster j and $n_i(n_j)$ is the number of instances in the cluster i (j)

3) ACCURACY

Overall prediction result

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

Where TP = true positive i.e. properly predicted disease as normal. FP = false positive i.e. wrongly predicted disease as affected TN = true negative i.e. properly predicted

disease as affected. FN = false negative i.e. wrongly predicted disease as normal.

C. EXECUTION TIME COMPARISON

This section compares the execution time of blockchain hash generation, rule generation and cluster formation for diabetic and heart disease data.

Fig. 8 shows the blockchain hash generation for diabetic and heart disease data set.

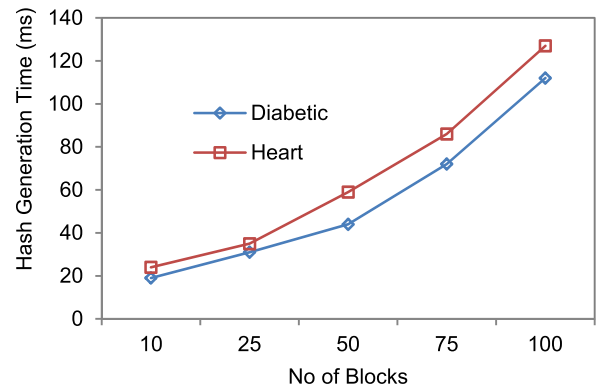


FIGURE 8. Blockchain hash generation time.

Fig. 9 shows the transaction creation time for two data set. It is the time taken to create a transaction for a given block. This paper use blockchain for secure storage purpose. The other parameters of the blockchain (latency, throughput and bandwidth) are out of scope.

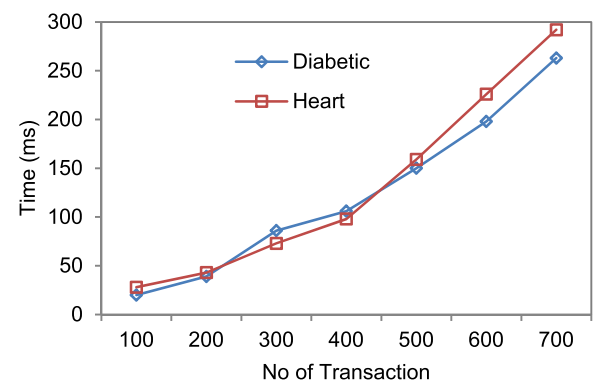


FIGURE 9. Transaction creation time.

Fig. 10 shows the execution time for rule generation and cluster formation for diabetic and heart disease data set. For two data sets, the cluster formation time is less than compared to the rule generation. The rule generation takes more time because it converts all the original data set into low, medium, high value to generate the candidate rules.

Fig. 11 shows the running time for the feature selection process. When increasing the number of iterations, the running time also increases. The proposed feature selection algorithm is compared with binary cuckoo search (BCS) [45]

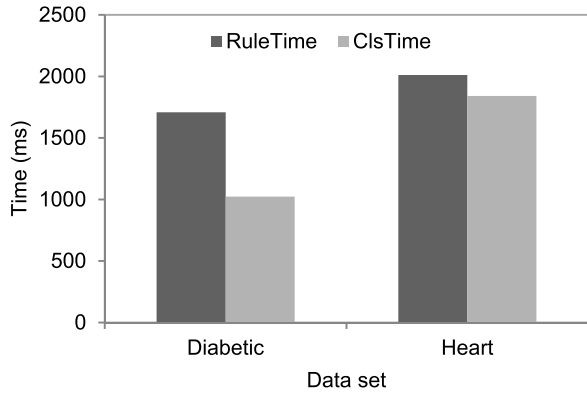


FIGURE 10. Execution time for rule and cluster formation.

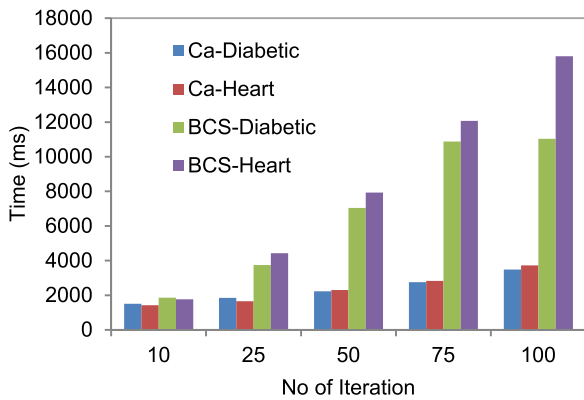


FIGURE 11. Feature selection running time.

algorithm. The BCS algorithm takes more execution time for feature selection.

D. CLUSTERING RESULT

This section explains the rule-based clustering performance result.

Fig. 12 and 13 show the rule count for diabetic and heart data. The rules are increased when the number of instances

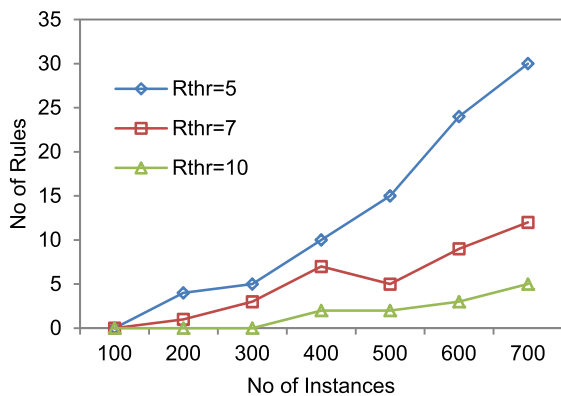


FIGURE 12. Instances vs rules for diabetic data.

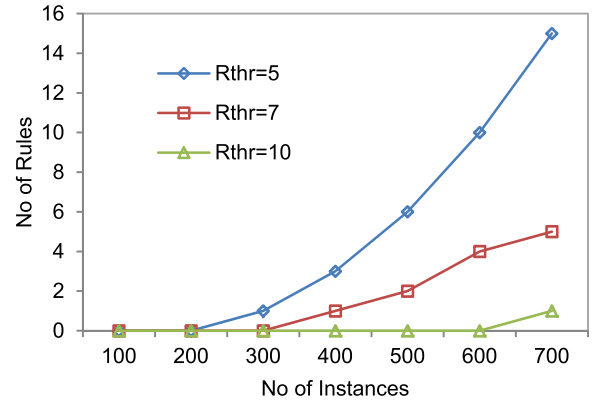


FIGURE 13. Instances vs rules for heart data.

is increased. Three threshold values (5, 7, 9) are used for experiments. More rules are generated for the threshold value $R_{thr} = 5$ for both diabetic and heart data set.

Fig 14 shows the candidate rule count with positive and negative rules for diabetic and heart disease for $R_{thr} = 5$.

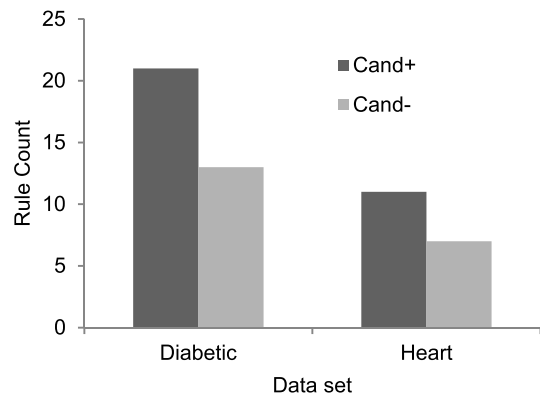


FIGURE 14. Candidate rule count $R_{thr} = 5$.

Fig 15 and 16 shows the purity and NMI result for diabetic and heart disease data set. For diabetic data set, the purity achieved 77%, and for heart disease 81%. The NMI value is more than 70% for both diabetic and heart disease data set when increasing the number of rules.

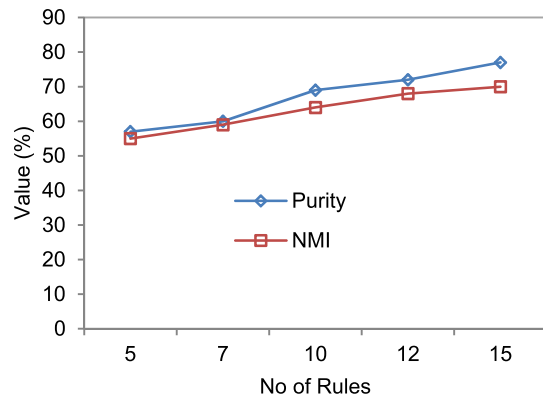


FIGURE 15. Purity and NMI for diabetic data.

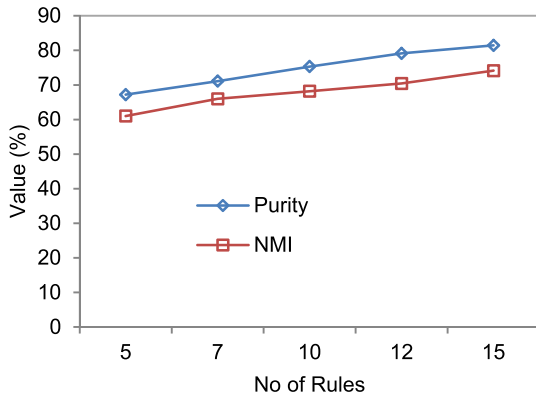


FIGURE 16. Purity and NMI for heart data.

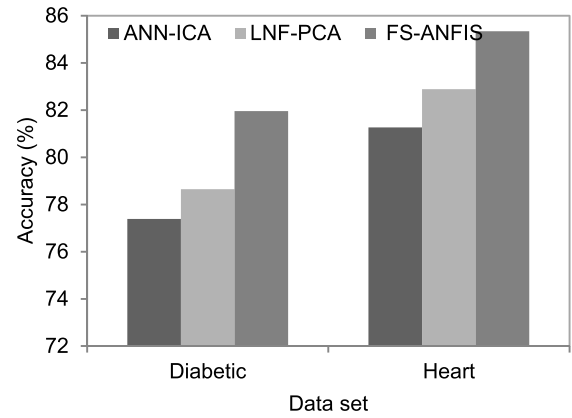


FIGURE 19. Accuracy comparison.

E. PREDICTION RESULT

This section explains the FS-ANFIS prediction performance result.

Fig. 17 shows the Cronbach’s alpha for a different population. The percentage of alpha value > 75 is acceptable consistency, and more than 90 is excellent consistency. Both the data set achieved good consistency.

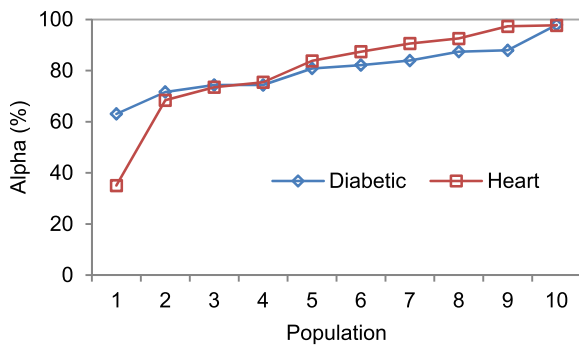


FIGURE 17. Alpha for different population.

Fig 18 shows the alpha value for 100 iterations.

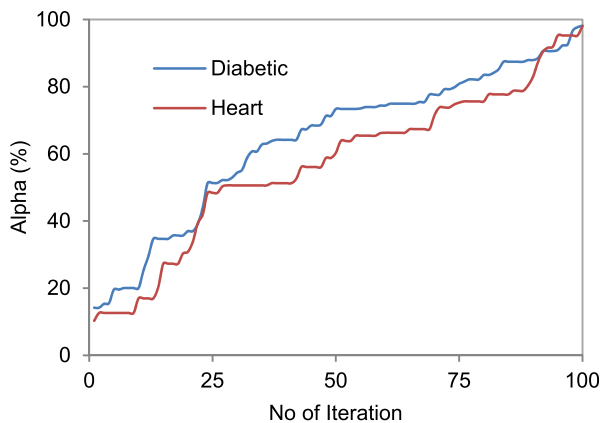


FIGURE 18. Alpha vs. no of iteration.

Fig 19 shows the accuracy comparison of 3 different algorithms. Compared to ANN-ICA (Integrated Component

Analysis) and LNF-PCA [44], the proposed algorithm obtains higher accuracy.

VII. CONCLUSION

In the current healthcare system, the use of Blockchain plays a crucial role. It can result in automated processes for collecting and verifying data, correcting and aggregating information from different resources that are indisputable, defiant to manipulation, and providing protected data, with condensed cybercrime chances and which also supports disseminated information, with system redundancy. This paper proposes efficient Blockchain-based secure healthcare services for disease prediction in fog computing. Diabetes and cardio diseases are considered for prediction. The proposed work efficiently clusters and predict the disease compared to other methods. In the future, the security and privacy for accessing patient medical data and some hybrid clustering and classification model can be added to enhance the performance of the prediction results.

REFERENCES

- [1] P. Sundaravadivel, E. Kougianos, S. P. Mohanty, and M. Ganapathiraju, "Everything you wanted to know about smart healthcare," *IEEE Consum. Electron. Mag.*, vol. 7, no. 1, pp. 18–28, Jan. 2018.
- [2] M. A. Sayeed, S. P. Mohanty, E. Kougianos, and H. P. Zaveri, "Neuro-detect: A machine learning-based fast and accurate seizure detection system in the IoMT," *IEEE Trans. Consum. Electron.*, vol. 65, no. 3, pp. 359–368, Aug. 2019.
- [3] A. V. Dastjerdi, H. Gupta, R. N. Calheiros, S. K. Ghosh, and R. Buyya, "Fog computing: Principles, architectures, and applications," 2016, *arXiv:1601.02752*. [Online]. Available: <http://arxiv.org/abs/1601.02752>
- [4] H.-J. Cha, H.-K. Yang, and Y.-J. Song, "A study on the design of fog computing architecture using sensor networks," *Sensors*, vol. 18, no. 11, p. 3633, Oct. 2018.
- [5] H. F. Atlam, R. J. Walters, and G. B. Wills, "Fog computing and the Internet of Things: A review," *Big Data Cogn. Comput.*, vol. 2, no. 10, pp. 1–18, Apr. 2018.
- [6] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in *Proc. 1st MCC Workshop Mobile Cloud Comput. - MCC*, Aug. 2012, pp. 13–15.
- [7] *Definition of Fog Computing*. Accessed: Jan. 15, 2021. [Online]. Available: <https://www.openfogconsortium.org/#definition-of-fogcomputing>
- [8] L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive definition of fog computing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 27–32, Oct. 2014.

- [9] Z. Wang, N. Luo, and P. Zhou, "GuardHealth: Blockchain empowered secure data management and graph convolutional network enabled anomaly detection in smart healthcare," *J. Parallel Distrib. Comput.*, vol. 142, pp. 1–12, Aug. 2020.
- [10] S. Tanwar, K. Parekh, and R. Evans, "Blockchain-based electronic healthcare record system for healthcare 4.0 applications," *J. Inf. Secur. Appl.*, vol. 50, Feb. 2020, Art. no. 102407.
- [11] Z. Zheng, S. Xie, H. Dai, X. Chen, and H. Wang, "An overview of blockchain technology: Architecture, consensus, and future trends," in *Proc. IEEE Int. Congr. Big Data (BigData Congress)*, Jun. 2017, pp. 557–564.
- [12] K. Hong, D. Lillethun, U. Ramachandran, B. Ottenwalder, and B. Koldehofe, "Mobile fog: A programming model for large-scale applications on the Internet of Things," in *Proc. 2nd ACM SIGCOMM Workshop Mobile Cloud Comput. - MCC*, 2013, pp. 15–20.
- [13] S. Yi, Z. Hao, Z. Qin, and Q. Li, "Fog computing: Platform and applications," in *Proc. 3rd IEEE Workshop Hot Topics Web Syst. Technol. (HotWeb)*, Nov. 2015, pp. 73–78.
- [14] N. Rifi, E. Rachkidi, N. Agoulmine, and N. C. Taher, "Towards using blockchain technology for eHealth data access management," in *Proc. 4th Int. Conf. Adv. Biomed. Eng. (ICABME)*, Oct. 2017, pp. 1–4.
- [15] M. Ahmad, M. B. Amin, S. Hussain, B. H. Kang, T. Cheong, and S. Lee, "Health fog: A novel framework for health and wellness applications," *J. Supercomput.*, vol. 72, no. 10, pp. 3677–3695, Oct. 2016.
- [16] P. Verma and S. K. Sood, "Fog assisted-IoT enabled patient health monitoring in smart homes," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1789–1796, Jun. 2018.
- [17] T. N. Gia, M. Jiang, A.-M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Fog computing in healthcare Internet of Things: A case study on ECG feature extraction," in *Proc. IEEE Int. Conf. Comput. Inf. Technol.; Ubiquitous Comput. Commun.; Dependable, Autonomic Secure Comput.; Pervas. Intell. Comput.*, Oct. 2015, pp. 1–8.
- [18] B. Negash, A. Anzanpour, I. Azimi, M. Jiang, T. Westerlund, A. M. Rahmani, P. Liljeberg, and H. Tenhunen, "Leveraging fog computing for healthcare IoT," in *Fog computing in the Internet of Things Intelligence at the edge*. Cham, Switzerland: Springer, 2017, pp. 145–169.
- [19] A. M. Rahmani, T. N. Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, and P. Liljeberg, "Exploiting smart e-health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," *Future Gener. Comput. Syst.*, vol. 78, pp. 641–658, Jan. 2018.
- [20] I. Azimi, A. Anzanpour, A. M. Rahmani, T. Pahikkala, M. Levorato, P. Liljeberg, and N. Dutt, "HiCH: Hierarchical fog-assisted computing architecture for healthcare IoT," *ACM Trans. Embedded Comput. Syst.*, vol. 16, no. 5s, pp. 1–20, Oct. 2017.
- [21] A. Alazeb and B. Panda, "Ensuring data integrity in fog computing based health-care systems," in *Proc. Int. Conf. Secur., Privacy Anonymity Comput., Commun. Storage*. Cham, Switzerland: Springer, 2019, pp. 63–77.
- [22] S. Tuli, N. Basumatary, S. S. Gill, M. Kahani, R. C. Arya, G. S. Wander, and R. Buyya, "HealthFog: An ensemble deep learning based smart healthcare system for automatic diagnosis of heart diseases in integrated IoT and fog computing environments," *Future Gener. Comput. Syst.*, vol. 104, pp. 187–200, Mar. 2020.
- [23] S. Jiang, J. Cao, H. Wu, Y. Yang, M. Ma, and J. He, "BlocHIE: A BLOCKchain-based platform for healthcare information exchange," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Jun. 2018, pp. 49–56.
- [24] X. Liang, J. Zhao, S. Shetty, J. Liu, and D. Li, "Integrating blockchain for data sharing and collaboration in mobile healthcare applications," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.
- [25] A. Zhang and X. Lin, "Towards secure and privacy-preserving data sharing in e-Health systems via consortium blockchain," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–18, Aug. 2018.
- [26] K. N. Griggs, O. Ossipova, C. P. Kohlios, A. N. Baccarini, E. A. Howson, and T. Hayajneh, "Healthcare blockchain system using smart contracts for secure automated remote patient monitoring," *J. Med. Syst.*, vol. 42, no. 7, pp. 1–7, Jul. 2018.
- [27] G. G. Dagher, J. Mohler, M. Milojkovic, and P. B. Marella, "Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology," *Sustain. Cities Soc.*, vol. 39, pp. 283–297, May 2018.
- [28] H. Li, L. Zhu, M. Shen, F. Gao, X. Tao, and S. Liu, "Blockchain-based data preservation system for medical data," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–13, Aug. 2018.
- [29] K. Fan, S. Wang, Y. Ren, H. Li, and Y. Yang, "MedBlock: Efficient and secure medical data sharing via blockchain," *J. Med. Syst.*, vol. 42, no. 8, pp. 1–11, Aug. 2018.
- [30] A. Vasighizaker and S. Jalili, "C-PUGP: A cluster-based positive unlabeled learning method for disease gene prediction and prioritization," *Comput. Biol. Chem.*, vol. 76, pp. 23–31, Oct. 2018.
- [31] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, 2019, Art. no. 100203.
- [32] P. R. Kumar, T. Arunprasad, M. P. Rajasekaran, and G. Vishnuvarathanan, "Computer-aided automated discrimination of Alzheimer's disease and its clinical progression in magnetic resonance images using hybrid clustering and game theory-based classification strategies," *Comput. Electr. Eng.*, vol. 72, pp. 283–295, Nov. 2018.
- [33] M. Nilashi, O. B. Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Comput. Chem. Eng.*, vol. 106, pp. 212–223, Nov. 2017.
- [34] N. Nidheesh, K. A. A. Nazeer, and P. M. Ameer, "An enhanced deterministic K-means clustering algorithm for cancer subtype prediction from gene expression data," *Comput. Biol. Med.*, vol. 91, pp. 213–221, Dec. 2017.
- [35] R. J. Kuo, P. Y. Su, F. E. Zulvia, and C. C. Lin, "Integrating cluster analysis with granular computing for imbalanced data classification problem—A case study on prostate cancer prognosis," *Comput. Ind. Eng.*, vol. 125, pp. 319–332, Nov. 2018.
- [36] A. Hasselgren, K. Kravevska, D. Gligoroski, S. A. Pedersen, and A. Faxvaag, "Blockchain in healthcare and health sciences—A scoping review," *Int. J. Med. Informat.*, vol. 134, Feb. 2020, Art. no. 104040.
- [37] O. Dib, K.-L. Brousmiche, A. Durand, E. Thea, and E. B. Hamida, "Consortium blockchains: Overview applications and challenges," *Int. J. Adv. Telecommun.*, vol. 11, no. 1, pp. 51–64, 2018.
- [38] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informat. J.*, vol. 19, no. 3, pp. 179–189, Nov. 2018.
- [39] E. D. ubeyli, "Adaptive neuro-fuzzy inference system for classification of ECG signals using Lyapunov exponents," *Comput. Methods Programs Biomed.*, vol. 93, no. 3, pp. 313–321, Mar. 2009.
- [40] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classif. Algor Appl.*, vol. 97, no. 7, pp. 1660–1674, Aug. 2006.
- [41] A. Christmann and S. Van Aelst, "Robust estimation of Cronbach's alpha," *J. Multivariate Anal.*, vol. 97, no. 7, pp. 1660–1674, Aug. 2006.
- [42] J.-S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 3, pp. 665–685, May/Jun. 1993.
- [43] M. Sugeno and G. T. Kang, "Structure identification of fuzzy model," *Fuzzy Sets Syst.*, vol. 28, no. 1, pp. 15–33, Oct. 1988.
- [44] H. Das, B. Naik, and H. S. Behera, "Medical disease analysis using neuro-fuzzy with feature extraction model for classification," *Informat. Med. Unlocked*, vol. 18, 2020, Art. no. 100288.
- [45] D. Rodrigues, L. A. M. Pereira, T. N. S. Almeida, J. P. Papa, A. N. Souza, C. C. O. Ramos, and X.-S. Yang, "BCS: A binary cuckoo search algorithm for feature selection," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 465–468.
- [46] R. Jayaram and S. Prabakaran, "Onboard disease prediction and rehabilitation monitoring on secure edge-cloud integrated privacy preserving healthcare system," *Egyptian Informat. J.*, to be published, doi: 10.1016/j.eij.2020.12.003.



P. G. SHYNU (Member, IEEE) received the M.E. degree in computer science and engineering from the College of Engineering, Anna University, Chennai, India, and the Ph.D. degree in Computer Science from the Vellore Institute of Technology (VIT), Vellore, India. He is currently working as an Associate Professor with the School of Information Technology and Engineering, VIT. He has published more than 30 research papers in refereed international conferences and journals.

His research interests include machine learning, cloud security and privacy, ad-hoc networks, and big data.



VARUN G. MENON (Senior Member, IEEE) is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, India. His research interests include the Internet of Things, fog computing and networking, underwater acoustic sensor networks, cyberpsychology, hijacked journals, ad-hoc networks, and wireless sensor networks. He is also a Distinguished Speaker of ACM Distinguished Speaker. He is also a Guest Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE SENSORS JOURNAL, the *IEEE Internet of Things Magazine*, and the *Journal of Supercomputing*. He is an Associate Editor of *IET Quantum Communications*. He is also an Editorial Board Member of the IEEE FUTURE DIRECTIONS: TECHNOLOGY POLICY AND ETHICS.



R. LAKSHMANA KUMAR (Member, IEEE) is currently associated with the Hindustan College of Engineering and Technology, Coimbatore, Tamil Nadu. He is also the Director-Research and Development (AI) for a Canadian-based company (ASIQC) in Vancouver region of British Columbia, Canada. He is also the Founding Member of IEEE SIG of Big Data for Cyber Security and Privacy, IEEE. He serves as a Core Member in the Editorial Advisor Board of Artificial Intelligence Group in Cambridge Scholars Publishing, U.K., *Trends in Renewable Energy Journal*, USA, *Frontiers in Communications and Networks*, -Switzerland, AI Forum (The world's leading forum for AI). He is an IEEE Brand Ambassador. He was invited as a Keynote Speaker of AVIS' 2020 (Asia Artificial Intelligence Virtual Summit 2020) which is the Asia's first biggest Virtual Summit on Artificial Intelligence held at Malaysia, in June 2020. He is a global chapter Lead of Machine Learning for Cyber Security (MLCS). He himself involves in research and expertise in AI and Blockchain technologies. He holds the certification in Data Science from John Hopkins University, USA. He also holds the Amazon Cloud Architect certification from Amazon Web Services. He is also an ACM Distinguished Speaker.



SEIFEDINE KADRY (Senior Member, IEEE) received the bachelor's degree from Lebanese University, in 1999, the M.S. degree from Reims University, France, in 2002, the EPFL (Lausanne) and Ph.D. degrees from Blaise Pascal University, France, in 2007, and the HDR degree from Rouen University, in 2017. His current research interests include data science, education using technology, system prognostics, stochastic systems, and applied mathematics. He is an ABET Program Evaluator of computing, and an ABET Program Evaluator of Engineering Tech. He is a Fellow of IET, IETE, and IACSIT. He is a Distinguished Speaker of IEEE Computer Society.




YUNYOUNG NAM (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, South Korea, in 2001, 2003, and 2007, respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with the Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia database, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.

• • •

[Home](#) > [Neural Computing and Applications](#) > Article

S.I: ML4BD_SHS | [Published: 27 May 2021](#)

An intelligent heart disease prediction system based on swarm-artificial neural network

[Sudarshan Nandy](#), [Mainak Adhikari](#), [Venki Balasubramanian](#), [Varun G. Menon](#) , [Xingwang Li](#) & [Muhammad Zakarya](#)

[Neural Computing and Applications](#) **35**, 14723–14737 (2023)

666 Accesses | **12** Citations | [Metrics](#)

Abstract

The accurate prediction of cardiovascular disease is an essential and challenging task to treat a patient efficiently before occurring a heart attack. In recent times, various intelligent healthcare frameworks have been designed with different machine learning and swarm optimization techniques for cardiovascular disease prediction. However, most of the existing strategies failed to achieve higher accuracy for cardiovascular disease prediction due to the lack of data-recognized techniques and proper prediction methodology. Motivated by the existing challenges, in this paper, we propose an intelligent healthcare framework for predicting

cardiovascular heart disease based on Swarm-Artificial Neural Network (Swarm-ANN) strategy. Initially, the proposed Swarm-ANN strategy randomly generates predefined numbers of Neural Networks (NNs) for training and evaluating the framework based on their solution consistency. Additionally, the NN populations are trained by two stages of weight changes and their weight is adjusted by a newly designed heuristic formulation. Finally, the weight of the neurons is modified by sharing the global best weight with other neurons and predicts the accuracy of cardiovascular disease. The proposed Swarm-ANN strategy achieves 95.78% accuracy while predicting the cardiovascular disease of the patients from a benchmark dataset. The simulation results exhibit that the proposed Swarm-ANN strategy outperforms the standard learning techniques in terms of various performance matrices.

This is a preview of subscription content, [access via your institution.](#)

Access options

Buy article PDF

39,95 €

Price includes VAT (India)

29. Jan MA, Khan F, Khan R, Mastorakis S, Menon VG, Watters P, Alazab M (2020) A lightweight mutual authentication and privacy-preservation scheme for intelligent wearable devices in industrial-CPS. *IEEE Trans Ind Inform* 1–11

30. Shynu PG, Menon VG, Kumar RL, Kadry S, Nam Y (2021) Blockchain-based secure healthcare application for diabetic-cardio disease prediction in fog computing. *IEEE Access* 9:45706–45720

Author information

Authors and Affiliations

Computer Science and Engineering, ASETK, Amity University Kolkata, Kolkata, India

Sudarshan Nandy

Mobile & Cloud Lab, Institute of Computer Science, University of Tartu, Tartu, Estonia

Mainak Adhikari

School of Science, Engineering and Information Technology, Federation University, Mount Helen, Australia

Venki Balasubramanian

Computer Science and Engineering, SCMS

School of Engineering and Technology,

Ernakulam, India

Varun G. Menon

**School of Physics and Electronic
Information Engineering, Henan
Polytechnic University, Jiaozuo, China**
Xingwang Li

**Department of Computer Science, Abdul
Wali Khan University, Mardan, Pakistan**
Muhammad Zakarya

Corresponding author

Correspondence to [Varun G. Menon](#).

Ethics declarations

Conflict of interest

The authors declare that there are no potential conflicts of interest in this work.

Additional information

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rights and permissions

[Reprints and Permissions](#)

About this article

Cite this article

Nandy, S., Adhikari, M., Balasubramanian, V. *et al.* An intelligent heart disease prediction system based on swarm-artificial neural network. *Neural Comput & Applic*

35, 14723–14737 (2023). <https://doi.org/10.1007/s00521-021-06124-1>

Received

21 February 2021

Accepted

11 May 2021

Published

27 May 2021

Issue Date

July 2023

DOI

<https://doi.org/10.1007/s00521-021-06124-1>

Keywords

Artificial neural network

Heuristic formulation

Swarm optimization

Back-propagation

Classification model

Heart disease prediction

A Survey of Computational Intelligence for 6G: Key Technologies, Applications and Trends

Baofeng Ji , Yanan Wang, Kang Song , Chunguo Li , Hong Wen ,
 Varun G. Menon , and Shahid Mumtaz

Abstract—The ongoing deployment of 5G network involves the Internet of Things (IoT) as a new technology for the development of mobile communication, where the Internet of Everything (IoE) as the expansion of IoT has catalyzed the explosion of data and can trigger new eras. However, the fundamental and key component of the IoE depends on the computational intelligence (CI), which may be utilized in the sixth generation mobile communication system (6G). The motivation of this article presents the 6G enabled network in box (NIB) architecture as a powerful integrated solution that can support comprehensive

Manuscript received March 30, 2020; revised August 13, 2020 and October 10, 2020; accepted January 1, 2021. Date of publication January 18, 2021; date of current version June 30, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61801170, Grant 61901241, Grant 61671144, and Grant 61902041, in part by the National Key Research and Development Plan under Grant 2018YFB0904905 and Grant 2020YFB2008400, in part by China Postdoctoral Science Foundation under Grant 2018M633351, in part by the LAGEO of Chinese Academy of Sciences under Grant LAGEO-2019-2, in part by the Program for Science and Technology Innovation Talents in the University of Henan Province under Grant 20HASTIT022, in part by the 21th Project of Xizang Cultural Inheritance and Development Collaborative Innovation Center in 2018, in part by the Natural Science Foundation of Xizang Named “Research of Key Technology of Millimeter Wave MIMO Secure Transmission with Relay Enhancement” in 2018, in part by Xizang Autonomous Region Education Science “13th Five-year Plan” Major Project for 2018 (XZJKY201803), in part by the Natural Science Foundation of Henan under Grant 202300410126, in part by Young Backbone Teachers in Henan Province under Grant 2018GGJS049, in part by Henan Province Young Talent Lift Project under Grant 2020HYTP009, and in part by Top Young Talents in Central Plains. Paper no. TII-20-1599. (Corresponding author: Baofeng Ji.)

Baofeng Ji is with the School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China, with LAGEO, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China, and also with the School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: fengbaoji@126.com).

Yanan Wang is with the School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China (e-mail: wangyn1009@163.com).

Kang Song is with the Qingdao University, Qingdao 266071, China (e-mail: sk@qdu.edu.cn).

Chunguo Li is with the Southeast University, Nanjing 210096, China (e-mail: chunguoli@seu.edu.cn).

Hong Wen is with the University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: sunlike@uestc.edu.cn).

Varun G. Menon is with the SCMS School of Engineering and Technology, Ernakulam 683576, India (e-mail: varunmenon@scmsgroup.org).

Shahid Mumtaz is with the Instituto de Telecomunicacoes, 1049-001 Lisboa, Portugal (e-mail: dr.shahid.mumtaz@ieee.org).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TII.2021.3052531>.

Digital Object Identifier 10.1109/TII.2021.3052531

network management and operations. The 6G enabled NIB can be used as an alternative method to meet the needs of next-generation mobile networks by dynamically reconfiguring the deployment of network functions, providing a high degree of flexibility for connection services in various situations. Especially the CI technology such as evolutionary computing, neural computing and fuzzy systems utilized as a part of NIB have inherent capabilities to handle various uncertainties, which have unique advantages in processing the variability and diversity of large amounts of data. Finally, CI technology for NIB, which is widely used is also introduced such as distributed computing, fog computing, and mobile edge computing in order to achieve different levels of sustainable computing infrastructure. This article discusses the key technologies, advantages, industrial scenario applications of CI technology as NIB, typical use cases and development trends based on IoE, which provides directional guidance for the development of CI technology as NIB for 6G.

Index Terms—Computational intelligence (CI), industrial Internet of Things (IIoT), Internet of Everything (IoE), mobile edge computing (MEC), network in box (NIB), sixth generation mobile communication system (6G).

I. INTRODUCTION

TODAY’S society has entered a fairly technologically intelligent society such as smart phones, smart watches, and smart wearable devices have become popular and dominant gradually in daily lives. The Internet of Computer (IoC) has been widely used since 1991 [1], which is utilized for people’s interaction for a long time. Subsequently, the mobile Internet appeared and brought about the significant convenience especially the emergence of Internet of Things (IoT) integrated physical entities with radio frequency identification, advanced sensing and so on [2], which dramatically extended the communication coverage and performance to achieve the new communication object [3].

IoT has moved toward IoE with the acceleration of the pace of intelligence, where the IoE is a completely new concept [4] and has surpassed the IoT that can be connected to the Internet to people, data, things, and network programs [5], [6]. Meanwhile, IoE is a new computational paradigm that can connect the real and virtual worlds by giving daily things to processing capabilities [7], the ultimate goal of which is to create a “better human world” and knows our preferences, desires and demands and can perform the task according to our requirements without explicit instructions [8].

With the advent of the IoE, the amount of data will become more dramatically larger. In particular, there are abundant new applications and the requirements for transmission rate and spectrum width are becoming higher gradually. The development of 6G is to improve the shortcomings of 5G and have a higher rates and lower delays. Different from 5G, 6G may build a network that can realize air, space, ground, and sea integrated communications. Therefore, 6G technology will no longer be limited to breakthroughs in simple network capacity and transmission rate in the future. Its research and development is to narrow the digital divide and promote the IoE to be truly development and maturity. Compared with previous generations, 6G will not only improve communication capabilities, but also provide a communication infrastructure that supports various services or vertical fields. Therefore, 6G enabled CI technologies as network in box (NIB) has broad application prospects in user's personalized services as well as the IoE, Industrial Internet, smart factories, and other fields. In other words, 6G can truly realize the interconnection of all things and will be dedicated to creating a fully connected communication world that integrates ground communications, satellite communications, and marine communications [9].

Although the commercialization of 5G is still in its infancy, the research of 6G has already begun impressive and the candidate technologies such as terahertz (THz) communications, artificial intelligence (AI), computational intelligence (CI) and distributed intelligent computing all can be acted as NIB to improve the system performance considerably. It is worth noting that the CI technologies as part of NIB play a vital role in 6G. CI is a calculation model and intelligent tool with high fault tolerance, which is a new stage in the development and successor of AI. In recent years, 6G-based CI technology is developing at an astonishing speed, and its scope covers all fields of engineering technology, promoting the development of the information age. Even its application research has characteristics that exceed theoretical and methodological research. Fig. 1 shows the development process for NIB from IoC and IoT to IoE and lists the comparison and application. Fig. 2 expresses the key technologies and scenes of 6G enabled NIB based on IoE.

The contributions of this article are summarized as follows:

- 1) As far as we know, this is the first work that comprehensively outlines CI as a part of NIB for 6G from different aspects and perspectives. In particular, we provide a unique perspective on why CI can play an irreplaceable role as a key technology of NIB in 6G. We gave a detailed explanation on this aspect.
- 2) In addition, we have included the industrial application of NIB in the 6G field in the article, which makes this survey increase the practical application value and significance.
- 3) Finally, we use a chart to summarize and compare the various technologies involved in CI as a part of NIB for 6G and emerging key technologies have been anticipated under the development momentum of IoE.

The rest of this article is organized as follows. Section II analyzes the technical advantages for CI as NIB in 6G. Section III elaborates the key technologies of CI. Section IV describes NIB for industrial applications. Section V outlines



Fig. 1. Development comparison of IoC, IoT, and IoE.

the application scenarios and practical examples based on IoE. Section VI elaborates the typical use cases. Section VII raises several privacy security issues. Finally, Section VIII concludes this article.

II. TECHNICAL ADVANTAGES FOR CI AS NIB IN 6G

The communication technologies are still consumer applications from the 1G to 4G era, meanwhile, the 5G and 6G can involve the industrial applications such as industrial Internet and intelligent transportation. At present, 5G is mainly based on the early infrastructure for Industry 4.0 and the large specific application of 6G can be still opened and explored in the academic and industrial community. The most important requirements of 6G networks are the ability to handle large amounts of data and the connectivity of extremely high data rates per device; therefore, the CI technologies as NIB enabled by 6G can play a significant role in the future communication systems. Certainly, several important technologies such as the THz, AI, optical wireless

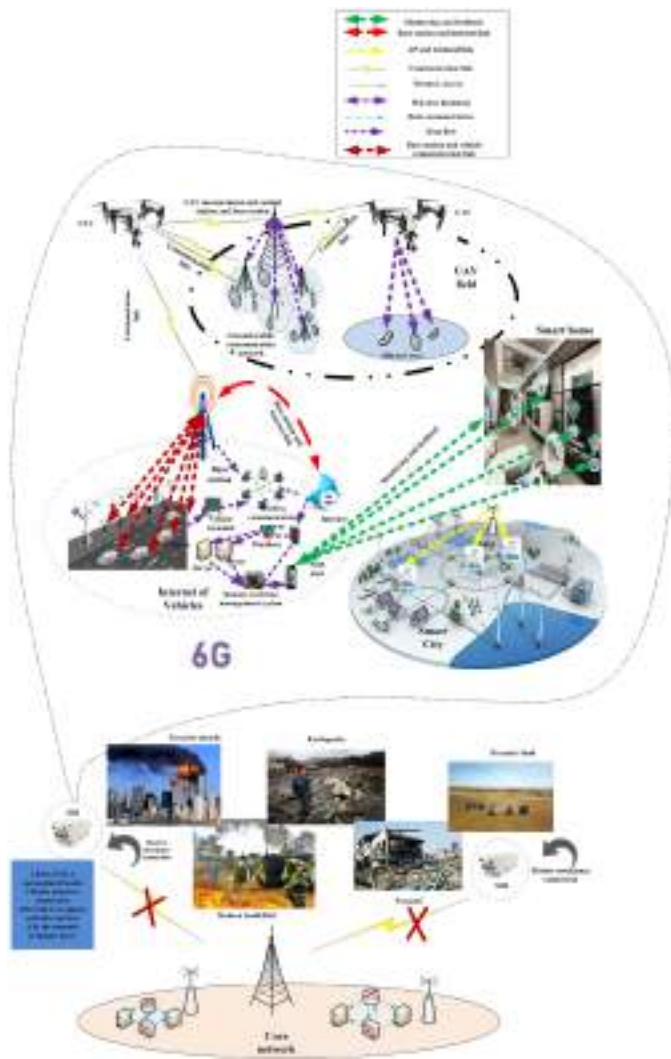


Fig. 2. Key technologies of 6G enabled CI as a part of NIB based on IoE.

communications, 3-D networks, unmanned aerial vehicles, and wireless power transmission can be also a part of the 6G system [10]–[12].

The millimeter wave band of 30G to 300 GHz has been utilized in 5G and the data speed can still provide not exceeding 100 Gbps. The THz technology adopted in 6G will be able to provide new bandwidth and allowed a large amount of data to be transmitted simultaneously. And the integration of block chain in 6G will realize the dynamic sharing of spectrum resources, the sharing of edge computing storage resources, and the sharing of distributed energy. Furthermore, the THz technology utilized in 6G network may support a variety of wireless devices to achieve real-time and remote transmission of data equivalent to the amount of human brain calculations. The THz frequency will provide a huge new bandwidth for wireless use, enabling wireless devices to remotely transmit massive amounts of computing data equivalent to the human brain in real time. For example, an unimaginable amount and type of data will be transmitted only in milliseconds. At this time,

data transmission will consume less energy and the ultrahigh gain antenna will be able to be “extremely small”. This will pave the way for smaller devices deployed in NIBs, including military-grade secure communication links that are very difficult to intercept or eavesdrop on. Some of the application scenarios may be familiar with 5G such as the remote control and so on, the difference of which is that the CI application in 6G can be dominant with AI instead of human. Therefore, the breakthrough of 6G cannot only provide fast network speed of all the data required for perception and control but also liberate a large number of heavy computational tasks from the human brain. Additionally, the submillimeter wave spectrum will be able to play an amazing role in existing technologies such as millimeter wave cameras used in dark environments, high-precision radar and terahertz-wave-based detectors for human security. Moreover, the base station of 6G may be able to access hundreds or even thousands of wireless connections at the same time and implement the compatible interaction with different transceivers such as drones, satellites, and so on to establish the integrated ground-air-space infrastructure [13]–[15]. Therefore, the CI as NIB utilized in 6G can be no longer a breakthrough in simple network capacity and transmission rate and it may pursue and achieve the ultimate goal of the IoE [16]–[18].

III. KEY TECHNOLOGIES FOR CI IN 6G

How CI technology can give full play to its technical advantages is a question worth pondering. In the 6G era, CI technology will be fully integrated into intelligent 6G network. The CI can be used to deal with the uncertainty encountered in evolutionary optimization, machine learning (ML) and data mining (DM) in the future. The CI includes neural networks, reinforcement learning (RL), evolutionary algorithms (EA), swarm intelligence (SI), fuzzy logic, artificial immune systems (AIS) and hybrid technologies such as neural fuzzy systems, fuzzy immune systems, and other types of hybrid system [19]. Therefore, this section briefly elaborates the key technologies of CI.

A. Artificial Neural Network

Artificial neural network (ANN) is a CI technology that simulates the brain processes data to deal with practical problems that need to consider multiple factors and conditions simultaneously. There are three main learning methods for artificial neuron learning.

- 1) Supervised learning [20].
- 2) Unsupervised learning.
- 3) Enhance learning.

B. Fuzzy Systems

Fuzzy system (FS) is a classic CI technology that uses fuzzy theory to solve problems in many fields. In contrast to certain logic, which can only have two possible values, fuzzy logic reasons approximately or to some extent indicates true or false. Additionally, fuzzy logic has been successfully used in control systems, power system control, and home appliance control.

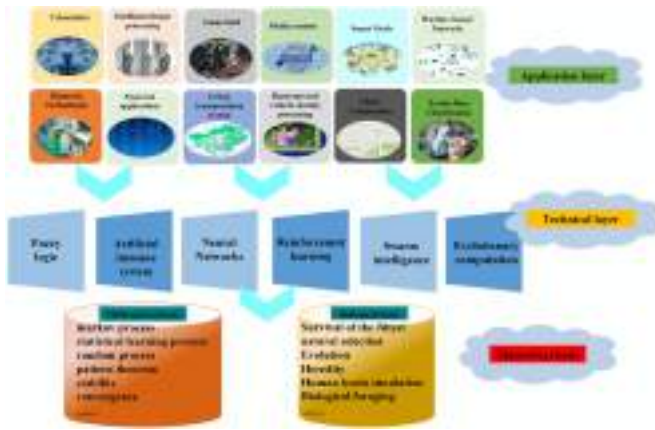


Fig. 3. CI theory, technology and application.

C. Evolutionary Computing

Evolutionary computing (EC) as a new global optimization search algorithm regardless of the function itself is continuous and general suitable for parallel processing with strong robustness for its simplicity and distinctive features such as high efficiency in the plan design control classification, clustering of time series modeling music composing and other fields has been widely applied.

D. Swarm Intelligence

SI refers to some intelligent algorithms with distributed intelligent behavior characteristics designed by birds, fish, bees, and other group behaviors [11]. The most widely accepted SI use cases are particle swarm optimization (PSO). SI algorithms have the advantages of simplicity, parallelism, and strong applicability. Therefore, it is widely used in optimization problem solving, robotics, and semiconductor manufacturing.

E. Artificial Immune System

Four decisions must be made: encoding, similarity measurement, selection and mutation in order to implement basic AIS. AIS algorithms have been successfully applied in computer security, fault detection, anomaly detection, optimization and data mining.

F. Reinforcement Learning

Traditional AI is based on ML, which is the development of technologies and algorithms that allow ML. However, the RL is a subfield of ML and is very suitable for dealing with distributed problems. It is mentioned that RL has become one of the hottest research areas in ML today with the success of Alpha Go [21].

IV. APPLICATION SCENARIOS OF CI IN 6G

In the 6G era, CI will receive widespread attention recently and becoming an important research direction of AI and computer science, which has been continuously improved with the improvement of its own performance and the expansion of its application range [22]. Fig. 3 shows the CI theory, technology, and application. The following will briefly introduce future CI applications in these areas.

A. Media Content

CI plays an important role in media content mining and processing based on big data features such as multiobjective optimization and deep learning (DL). EC such as ANN and genetic algorithm (GA) are common methods to solve complex problems. DL algorithms such as neural networks are used to detect and identify the image data.

B. Music Creation

In the field of music creation, CI technologies such as neural networks, FS and EC provide powerful tools for modeling, learning, uncertainty processing, search and optimization [23], [24]. EA is random, which makes it particularly suitable for music classification and analysis, using FS to design fitness functions to promote the imitation of phrase similarity between phrases. Use GA to generate melody motivation and use genetic algorithm to traverse the tree to construct the music structure. GA's chromosome notation can generate drum rhythms in a human-like rhythm accompaniment system. Neural networks are usually used to evaluate musical works or predict musical notes. Music systems developed using neural networks can generate music and assist in evolutionary creation [25].

C. Biometrics

CI is used in biometric systems and CI technologies such as neural networks, fuzzy logic, and EA have the characteristics of strong robustness and strong self-adaptability, which can be successfully applied to solve complex biometric recognition problems [26], [27]. In terms of face recognition and face monitoring, EA is a method to optimize the topology of neural networks and an effective face detection tool [28].

D. Finance

EC provides the possibility of trading strategies based on pattern recognition to profit from stock market transactions. Naturally inspired search technologies such as ANN can predict the direction of price changes, so neural networks are applied to exchange rate prediction [29]. Use fuzzy logic rules to design a specific fitness function in order to rank them as buying suggestions based on their fitness.

E. Intelligent Image Processing

CI can also be used in intelligent image processing such as image fusion. Combining fuzzy theory and neural network to process accurate information of noisy images and fuzzy information of noisy images. The combination of GA and neural network can improve the calculation efficiency to enhance the degree of automation of neural network modeling [30].

F. Wireless Sensor Network (WSN)

CI method is expected to produce a practical optimal/suboptimal solution to the distributed sensor scheduling problem in WSN [21]. For example, fuzzy logic is used to

determine the number of sensors and continuous PSO algorithm is used for the distributed arrangement of sensors in marine monitoring, which not only improves the network performance but also save system cost.

G. Smart Grid

CI technology can be used in smart grids. For example, critical networks based on neural network structures can overcome time-varying delays in communication channels to improve the damping performance of the power system [31]. Using adaptive design and fuzzy logic based on PSO, energy-optimized of photovoltaic systems independent of the grid can be performed.

H. Urban Traffic Control

CI technologies such as ANN, FS, and EC algorithms have flexibility, autonomy and can overcome the nonlinearity and randomness of transportation systems, so they are suitable for dynamic urban traffic control transportation systems. Traffic event detection algorithms based on fuzzy technology can have lower false alarm rates, higher detection rates, and shorter average times in order to alleviate nonperiodic congestion of expressways [32], [33]. The PSO algorithm can handle the fuzzy rules of the signal controller and it has alleviated the pressure of urban traffic to the greatest extent and reduced the waiting time of vehicles.

I. Battery Management System

CI technology can be used in designing the charge state estimator of a battery pack. Battery state of charge (SOC) is a very important parameter in the battery management system of electric vehicles or hybrid vehicles. Based on the neural network technology in CI technology, the adaptive estimator is designed to determine the SOC of the electric vehicle battery [34]. The main framework of the estimator is a three-layer feedforward neural network with four inputs and one output. The first and third layers are pure linear functions, and the middle layer is a complex neuron network structure. The hidden neuron battery pack SOC is determined by many factors and parameters, such as the discharge current, the number of ampere hours used, the average temperature of the battery module, and the module voltage. The charging state of the battery mainly depends on the current of the battery pack. In addition, the SOC estimator using the improved PSO algorithm is not only simple in structure, but also has high calculation efficiency.

J. Gaming

AI and CI algorithms are widely present in games, such as ML, RL, and GA iteration. The intelligent path search algorithm in the game mainly includes a star algorithm and GA, which is a heuristic function path calculation search algorithm. The process of path finding can be greatly reduced by designing a reasonable heuristic function in the algorithm and is widely used in game path finding [35], [36]. The use of CI technology provides an interesting alternative to scripts in most games. For example, an

evolved neural network can be used to control agent behavior instead of programming it.

K. Hyper-Spectral Remote Sensing Processing

CI theory and its algorithms have also been successfully applied in the field of hyper-spectral remote sensing processing, that is, the dimension reduction and classification of hyper-spectral remote sensing images [37], which effectively solves the problems that traditional algorithms cannot solve and has good development prospects. Generally speaking, the accuracy is guaranteed by using neural networks and transparency is achieved by using fuzzy sets.

L. Other Applications

EA has many applications in real-world parameter optimization, which is one of the most advanced methods to solve complex optimization problems today and is often used in industries such as automotive and aerospace [38], [39]. Neural network technology has the ability to continuously learn during operation in the field of automatic control [40], [41], so it can be used to detect and identify system failures and help store information for decision making. Additionally, in academia or industry, big data analysis (BDA) is becoming more and more popular and there are a large number of practical applications in IoE such as business intelligence, environmental science, and cyber security. The algorithms used in the different application areas abovementioned can be compared as shown in Fig. 4.

V. 6G-ENABLED NIB FOR INDUSTRIAL APPLICATIONS

With the rapid development of wireless transmission technology in 5G and the upcoming 6G communication system, 6G-enabled NIB has been extensively studied in academia and industry. Since one of the key features of the new generation of mobile networks is the ability to meet the needs of different vertical directions, NIB is an alternative method that can meet the needs of the next generation of mobile networks. NIB is a multigeneration 2G/3G/4G/5G/6G integrated and rapidly deployed hardware and software solution, which is a powerful and portable software and hardware integration box that integrates a core network, remote radio head and baseband unit (BBU). At the same time, NIB represents a portable and portable physical device that is flexible and can move freely or according to actual needs. The device can be used to provide connections between a group of disconnected and possibly mobile devices, and allow services such as text messages, phone calls, and Internet connections to be transferred between each other's devices. NIB equipment encapsulates part of the entire 5G or 6G mobile network, and two NIBs are connected through a standard radio interface, that is, each NIB treats the other as a preexisting legacy infrastructure component, or connects through a dedicated interface, generally providing short-term communication services. Recently, the industry has promoted the development of emergency and tactical networks, with the main purpose of increasing practicality, integrating solutions into the smallest possible physical devices. NIB also

Index	Application scenario	Key application area	Advantages	Issues
[14] [27]	Basic service	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14] [27]	Mobile CI	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Emergency service	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[27]	Disaster relief	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Intelligent manufacturing	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Wearable devices	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Smart grid	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14] [27]	Healthcare service	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Smart manufacturing	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.
[14]	Disaster relief	Establishing edge network and service coverage	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.	It has network portability and flexibility, which can be used in any scenario and can be used in any scenario.

Fig. 4. Comparison of various application scenarios.

has other features such as self-organizing functions and special services provider. Furthermore, the flexibility required for next-generation mobile networks can be achieved by including the principles of the NIB in these networks, so it can be the cornerstone of a flexible and adaptable network.

Therefore, the NIB provides services through a wireless connection and an important industrial use case for NIB is restoring basic connections in an emergency. For the case of communication infrastructure damaged and services interrupted, NIB can restore the basic communication services in the affected area in the fastest and easiest way and can quickly deploy a ready-to-use network made up of equipment that requires only minimal setup requirements. In addition, NIB is an attractive solution for handling suddenly increased traffic loads. Several NIBs can be used to offload some mobile-initiated traffic when the peak period of network usage suddenly occurs in the industry. The technology currently used in the NIB solution is mobile technology especially the combination of 6G and Wi-Fi. NIB can also be combined with microwave, Ethernet or fiber optics, Wi-Fi, telemedicine and downloading 3-D maps of buildings to improve the system and enhance the user experience in industrial applications.

NIB can also act as a traditional network and can be deployed stably to implement the wide coverage [17], [18]. NIB can provide connectivity as a stand-alone solution as well as signal connection lost. It is suitable for commercial, private, government, and military scenarios with its small, compact, and portable features. Other advantages include:

- 1) Independent, secure.
- 2) Supports up to millions of users.
- 3) No need for existing infrastructure.
- 4) Operate as a secure standalone or integrated.
- 5) Integrate 4G LTE functions into existing networks.
- 6) Can operate in any LTE band (3GPP or unlicensed).
- 7) Scalable to meet customer needs.
- 8) Suitable for air, ground, sea, disassembly and network mobile operations.

The core of the idea of combining 6G technology with NIB is to install all software and hardware modules required by the mobile network into one or several physical devices. The NIB can be deployed in a wide range of situations including extreme disasters, special rescue missions, emergency management, armed forces, peacekeeping missions and transit mobile communications networks. This node component of the radio access network (RAN) in NIB provides a seamless LTE network solution. In addition to being lighter in weight, these enhanced integration technologies translate into better quality of service and higher bit rates for packet data-intensive applications. NIB provides a rapidly deployable, high-speed 6G LTE communications network to support operations of defense, public safety and security forces. It can integrate mobile environment installations of land, air, sea, pedestrian, and unmanned systems to provide mesh communication, thereby expand system coverage. As an independent network, NIB can provide network coverage in rural and remote areas without any existing infrastructure.

VI. TYPICAL USE CASES BASED ON IOE IN 6G

In order to realize the vision of “smart connection” in the 6G era, the 6G network will be presented as a “distributed intelligent computing” network architecture. Meanwhile CI technology is also widely used in IoE applications such as fog computing, edge computing, and cloud computing to enable different levels of sustainable computing infrastructure. It can perform large-scale calculations through distributed computing resources, which enable it to solve problems that require processing very large data sets. Fig. 5 shows a simple comparison of them.

A. Mobile Edge Computing

The idea of deploying services on NIB is consistent with MEC, a technology that pushes services to the edge of the network to reduce traffic from the core network. At the same time, MEC can be defined as the implementation of edge computing, bringing computing and storage capabilities to the edge of the network within the RAN to reduce latency [42]–[45]. For example, first, MEC can support vertical segmentation services and provide emerging big data services such as video analysis to authorized third parties. Second, the MEC platform can be located at an aggregation point such as a BBU in a cloud operation deployment or it can be directly located in a mobile backhaul such as a small unit gateway. Third, for video streaming media services, MEC with edge architecture uses video analysis and video management applications to apply intelligent video acceleration solutions. Fourth, use the network as a supported

	MEC	FemtoCloud	Fog Computing
Name	Mobile Edge Cloud (MEC) and 5G	Cloudlet	Cloudlet
Implementation	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Architecture	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Advantages	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Disadvantages	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Applications	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Security	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Performance	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Scalability	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Reliability	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Flexibility	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.
Interoperability	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.	It is a collection of the virtualized and distributed computing resources located at the edge of the network.

Fig. 5. Comparison of MEC, FemtoCloud, and fog computing.

adaptive streaming media application to encapsulate multimedia content in the MEC to improve the quality of experience. Finally, use the edge as a cache to store media content and increase the life of mobile devices by forcing computational offloading [46]–[47]. In the 6G era, MEC can be widely used in various fields such as transportation systems, intelligent driving, real-time haptic control, and augmented reality.

B. Fog Computing

Fog computing, also known as fog networking, is a distributed computing infrastructure based on fog computing nodes placed on any architectural point between the terminal device and the cloud [45]. The advantages of fog computing are: first, it provides storage near the edge, which reduces the traffic load. Second, reduced data movement across the network and improved security and scalability to a certain extent. Third, reduced network bandwidth and reduced the possibility of data being attacked during transmission [48], [49]. Fog computing plays a role in advertising, entertainment and BDA as well as IoT, connected vehicles, wireless sensor and actuator networks, and cyber-physical systems [34].

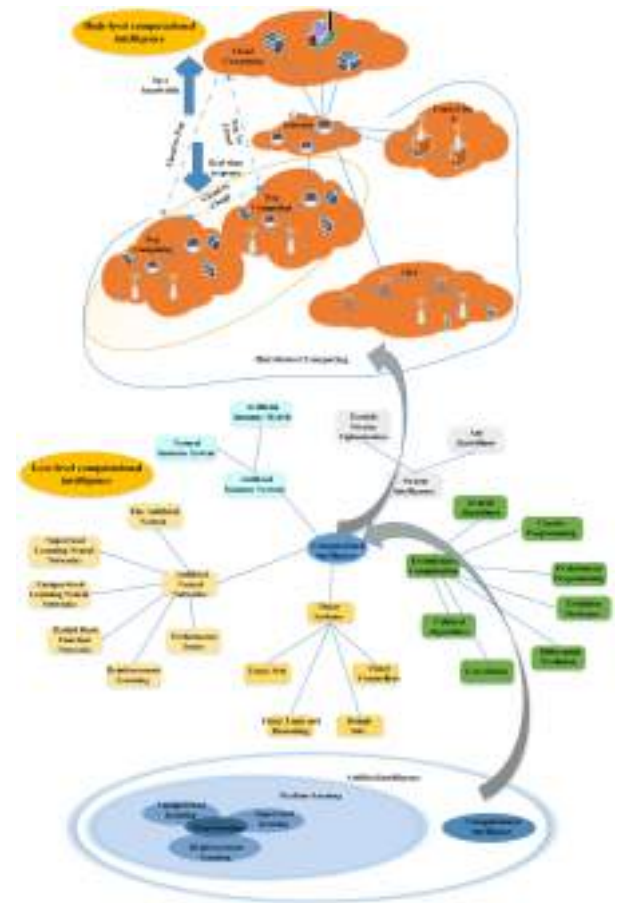


Fig. 6. Relationship for various technologies.

C. FemtoCloud

The basic idea of FemtoCloud is to be controlled by a controller to achieve the function of the cluster [43]. The advantages of FemtoCloud are: better scalability and less dependent on infrastructure. Specifically, FemtoCloud performs various tasks that reach the control device through computing services. The Femtocloud client service running on mobile devices can estimate the computing power from various mobile devices and use it with user input to determine the computing power available for sharing. Then, the Femtocloud client service shares the available information with the control device. The control device is responsible for estimating the user’s existence time and configuring the participating mobile devices to provide computing as a service cloud. However, the security of FemtoCloud may become a challenge in such application environments because of the high variability and the dynamics and instability of mobile devices. Fig. 6 shows the relationship between the various technologies.

D. Edge Cloud

Edge cloud is also a typical IoE application, which is an important area for future innovation and has many IoT application potentials. The advantages of edge cloud are: most of the data

can be processed through edge cloud or edge computing and reduce the amount of data sent to remote data centers [50], [51]. The application areas of edge cloud include smart home, smart cities, smart health, AR or VR, and machine-to-machine communication [48]. In addition, edge cloud has an absolute advantage in highly accurate 3-D indoor positioning and it saves latency and bandwidth after adopting edge cloud in terms of scalable and flexible video surveillance [52], [53].

VII. IOE IN 6G PRIVACY ISSUES AND DEVELOPMENT TRENDS

A. IoE Security

New demand of IoE emerges gradually with the rapid popularity of IoE worldwide. After integrating IoT technologies such as smart objects, BDA and communication capabilities, and the biggest problem is how to ensure security in such a large-scale scenario. The beautiful vision of 6G makes people look forward to it. But to realize these beautiful visions, we will have to face many technical needs and challenges. The huge traffic and data explosion make it more difficult to identify potential security risks in the 6G era [54]. Since the data generated by smart objects and users of the IoT can be obtained on the network, so there are three key issues for IoT devices and services to be considered: data confidentiality, privacy, and trust [55]–[58]. The goals of network security are: protect IoT devices and services that are accessed from inside and outside the device without authorization. Protect services, hardware resources, information and data in conversion and storage.

B. Cybersecurity Issues in Specific Areas

Additionally, network security issues in specific areas also deserve attention with the start of the 5G era and the arrival of the 6G era ten years later. IoE brings changes in the urban infrastructure and makes smart cities possible. The city's pipeline network, electricity, energy, transportation, and other infrastructures have countless sensors and cameras for monitoring and they will be intelligently controlled through the network. However, it also exposes risks to the hacker's vision, once criminals have the viewing authority of the camera, they illegally obtain the information they want through the camera such as a banknote transporter. Cyber security technology which is based on the key core technology of the IoE is the same as AI, big data, and internet of vehicle. Moreover, cyber security is no longer just information security.

C. Outlook

The emerging key technologies accelerate the iterative update of the IoE, which is relying on big data resources to reshape application scenarios such as transportation, medical care, and social governance which change all aspects of urban life [59]. 5G, 6G, IoE, distributed AI and other technologies will be deeply combined with the acceleration of the pace of IoE intelligence in the future. The rise of a variety of intelligent new technologies and mature commercialization are crucial to the development of the IoT toward the era of the IoE such as AI, blockchain, cloud

computing, big data, smart home, edge computing, IoT, 5G, 6G and so on [60]–[62]. In the future, the intelligent technology combined with IoE will continue to heat up our smart lives and we can more easily manage data and control our equipment in more directions.

VIII. CONCLUSION

With the development of wireless technology, 5G would not be able to fully meet the growing demand for wireless communications in 2030. Therefore, 6G would need to be rolled out. The 6G was still in the research stage. The application of 6G technology and NIB in industry would be a new research area. In addition, 6G with CI technology could help us process a large amount of data in the IoE field. This article analyzed 6G technical advantages, described 6G enabled NIB for industrial applications, introduced the basis of CI key technologies thoroughly and summarized relevant application of CI in different scenarios based on IoE, which could use IoE-based distributed computation such as MEC and fog computing for typical use cases. As one of the research directions of 6G technology, distributed intelligent computing had laid a certain foundation for the development of communication technology. Additionally, several privacy issues and challenges were also elaborated in this article.

REFERENCES

- [1] A. Ghosh, D. Chakraborty, and A. Law, "Artificial intelligence in Internet of Things," *CAAI Trans. Intell. Technol.*, vol. 3, no. 4, pp. 208–218, 2018.
- [2] R. Khan *et al.*, "Future Internet: the Internet of Things architecture, possible applications and key challenges," in *Proc. 10th Int. Conf. Front. Inf. Technol.*, 2017, pp. 257–260.
- [3] S. Charmonman and P. Mongkhonvanit, "Special consideration for big data in IoE or Internet of Everything," in *Proc. 13th Int. Conf. ICT Knowl. Eng.*, 2015, pp. 147–150.
- [4] M. H. Miraz *et al.*, "A review on Internet of Things (IoT), Internet of everything (IoE) and Internet of Nano Things (IoNT)," in *Proc. Internet Technol. Appl.*, 2015, pp. 219–224.
- [5] R. E. Balfour, "Building the Internet of Everything(IoE) for first responders," in *Proc. Long Island Syst., Appl. Technol.*, 2015, pp. 1–6.
- [6] A. Bujari and C. E. Palazzi, "Opportunistic communication for the Internet of Everything," in *Proc. IEEE 11th Consum. Commun. Netw. Conf.*, 2014, pp. 502–507.
- [7] B. Kang, D. Kim, and H. Choo, "Internet of Everything: A large-scale automatic IoT gateway," *IEEE Trans. Multi-Scale Comput. Syst.*, vol. 3, no. 3, pp. 206–214, Jul.–Sep. 2017.
- [8] L. Hu, N. Xie, and Z. Kuang, "Review of cyber-physical system architecture," in *Proc. Int. Symp. Object Compon. Serv. Oriented Real Time Distrib. Comput.*, 2012, pp. 25–30.
- [9] J. Iannacci, "Internet of Things (IoT); Internet of Everything (IoE); tactile Internet; 5G-A (not so evanescent) unifying vision empowered by EH-MEMS (energy harvesting MEMS) and RF-MEMS (radio frequency MEMS)," *Sensors Actuators A, Phys.*, vol. 272, pp. 187–198, 2018.
- [10] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/June 2020.
- [11] Z. Zhao *et al.*, "A novel framework of three-hierarchical offloading optimization for MEC in industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5424–5434, Aug. 2020.
- [12] M. Z. Chowdhury *et al.*, "6G wireless communication systems: Applications, requirements, technologies, challenges, and research directions," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 957–975, 2010.
- [13] Z. Zhang *et al.*, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.
- [14] K. B. Letaief *et al.*, "The roadmap to 6G: AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.

[15] L. Lovén et al., "Edge AI: A vision for distributed, edge-native artificial intelligence in future 6G networks," in *Proc. 1st 6G Wireless Summit*, 2019, pp. 1–2.

[16] B. Zong et al., "6G technologies: Key drivers, core requirements, system architectures, and enabling technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 18–27, Sep. 2019.

[17] M. Pozza et al., "Network-in-a-box: A survey about on-demand flexible networks," *IEEE Commun. Surv. Tut.*, vol. 20, no. 3, pp. 2407–2428, 2018.

[18] Xu L Da, W. He, and S. Li, "Internet of Things in industries: A survey," *IEEE Trans. Ind. Informat.*, vol. 10, no. 4, pp. 2233–2243, Nov. 2014.

[19] P. J. Werbos, "Computational intelligence for the smart grid-history, challenges, and opportunities," *IEEE Comput. Intell. Mag.*, vol. 6, no. 3, pp. 14–21, Aug. 2011.

[20] W. Tong et al., "Artificial intelligence for vehicle-to-everything: A survey," *IEEE Access*, vol. 7, pp. 10823–10843, 2019.

[21] R. V. Kulkarni, A. Forster, and G. K. Venayagamoorthy, "Computational intelligence in wireless sensor networks: A survey," *IEEE Commun. Surv. Tut.*, vol. 13, no. 1, pp. 68–96, 2011.

[22] Y. Wang, W. Kinsner, and D. Zhang, "Contemporary cybernetics and its facets of cognitive informatics and computational intelligence," *IEEE Trans. Syst., Man, Cybern., Part B*, vol. 39, no. 4, pp. 823–833, Aug. 2009.

[23] D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*. Reading, MA, USA: Addison-Wesley, 1989.

[24] D. Ashlock, "The art of artificial evolution: A handbook on evolutionary art and music," *J. Math. Arts*, vol. 2, no. 2, pp. 103–106, 2008.

[25] C. H. Liu and C. K. Ting, "Computational intelligence in music composition: A survey," *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 1, no. 1, pp. 2–15, Feb. 2017.

[26] D. Zhang and W. Zuo, "Computational intelligence-based biometric technologies," *IEEE Comput. Intell. Mag.*, vol. 2, no. 2, pp. 26–36, May 2007.

[27] Q. Xiao, "Technology review-biometrics-technology, application, challenge, and computational intelligence solutions," *IEEE Comput. Intell. Mag.*, vol. 2, no. 2, pp. 5–25, May 2007.

[28] R. D. Labati, A. Genovese, and V. Piuri, "Measurement of the principal singular point in contact and contactless fingerprint images by using computational intelligence techniques," in *Proc. IEEE Int. Conf. Comput. Intell. Meas. Syst. Appl.*, 2010, pp. 18–23.

[29] A. Ghandar et al., "Computational intelligence for evolving trading rules," *IEEE Trans. Evol. Comput.*, vol. 13, no. 1, pp. 71–86, Feb. 2009.

[30] H. Irshad, M. Kamran, and A. B. Siddiqui, "Image fusion using computational intelligence: A survey," in *Proc. 2nd Int. Conf. Environ. Comput. Sci.*, 2009, pp. 128–132.

[31] G. K. Venayagamoorthy, "Potentials and promises of computational intelligence for smart grids," in *Proc. IEEE Power Energy Soc. Gen. Meeting*, 2009, pp. 1–6.

[32] D. Zhao, Y. Dai, and Z. Zhang, "Computational intelligence in urban traffic signal control: A survey," *IEEE Trans. Syst., Man, Cybern., Part C*, vol. 42, no. 4, pp. 485–494, Jul. 2012.

[33] G. K. Venayagamoorthy, "A successful interdisciplinary course on computational intelligence," *IEEE Comput. Intell. Mag.*, vol. 4, no. 1, pp. 14–23, Feb. 2009.

[34] J. Peng, Y. Chen, and R. Eberhart, "Battery pack state of charge estimator design using computational intelligence approaches," in *Proc. 15th Annu. Battery Conf. Appl. Adv.*, 2000, pp. 173–177.

[35] G. N. Yannakakis and J. Togelius, "A panorama of artificial and computational intelligence in games," *IEEE Trans. Comput. Intell. AI Games*, vol. 7, no. 4, pp. 317–335, Dec. 2015.

[36] J. Valls-Vargas, S. Ontanón, and J. Zhu, "Towards story-based content generation: From plot-points to maps," in *Proc. IEEE Conf. Comput. Intell. Games*, 2013, pp. 1–8.

[37] D. Stathakis and A. Vasilakos, "Comparison of computational intelligence based classification techniques for remotely sensed optical image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 8, pp. 2305–2318, Aug. 2006.

[38] T. Back, M. Emmerich, and O. M. Shir, "Evolutionary algorithms for real world applications," *IEEE Comput. Intell. Mag.*, vol. 3, no. 1, pp. 64–67, Feb. 2008.

[39] J. Zhang et al., "Evolutionary computation meets machine learning: A survey," *IEEE Comput. Intell. Mag.*, vol. 6, no. 4, pp. 68–75, Nov. 2011.

[40] F. N. Chowdhury et al., "A survey of neural networks applications in automatic control," in *Proc. 33rd Southeastern Symp. System Theory*, 2011, pp. 349–353.

[41] Y. Jin and B. Hammer, "Computational intelligence in big data," *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 12–13, Aug. 2014.

[42] S. Garg et al., "Edge computing-based security framework for big data analytics in VANETS," *IEEE Netw.*, vol. 33, no. 2, pp. 72–81, Mar./Apr. 2019.

[43] J. Pan and J. McElhannon, "Future edge cloud and edge computing for Internet of Things applications," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 439–449, Feb. 2018.

[44] C. Li et al., "Multiuser overhearing for cooperative two-way multi-antenna relays," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3796–3802, May 2016.

[45] B. Ji et al., "Secrecy performance analysis of UAV assisted relay transmission for cognitive network with energy harvesting," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7404–7415, Jul. 2020.

[46] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78–81, 2016.

[47] T. Taleb et al., "On multi-access edge computing: A survey of the emerging 5G network edge cloud architecture and orchestration," *IEEE Commun. Surv. Tut.*, vol. 19, no. 3, pp. 1657–1681, 2017.

[48] T. Taleb et al., "Mobile edge computing potential in making cities smarter," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 38–43, Mar. 2017.

[49] M. Aazam and E. N. Huh, "Fog computing: The cloud-IoT/IoE middleware paradigm," *IEEE Potentials*, vol. 35, no. 3, pp. 40–44, May/June 2016.

[50] S. Abdelwahab et al., "Enabling smart cloud services through remote sensing: An Internet of Everything enabler," *IEEE Internet Things J.*, vol. 1, no. 3, pp. 276–288, Jun. 2014.

[51] M. Aazam et al., "Cloud of Things: Integrating Internet of Things and cloud computing and the issues involved," in *Proc. Int. Bhurban Conf. Appl. Sci. Technol.*, 2014, pp. 414–419.

[52] J. Pan et al., "HomeCloud: An edge cloud framework and testbed for new application delivery," in *Proc. 23rd Int. Conf. Telecommun.*, 2016, pp. 1–6.

[53] W. Shi et al., "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[54] F. J. De Santos and S. G. Villalonga, "Exploiting local clouds in the internet of everything environment," *Parallel Distrib. Netw. Process.*, vol. 1, pp. 296–300, 2015.

[55] B. Ji et al., "Joint optimization for ambient backscatter communication system with energy harvesting for IoT," *Mech. Syst. Signal Process.*, vol. 135, 2020, Art. no. 106412.

[56] C. Li et al., "Overhearing-based co-operation for two-cell network with asymmetric uplink-downlink traffics," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 3, pp. 350–361, Sep. 2016.

[57] N. Abbas et al., "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[58] K. Dolui and S. K. Datta, "Comparison of edge computing implementations: Fog computing, cloudlet and mobile edge computing," in *Proc. Glob. Internet Things Summit*, 2017, pp. 1–6.

[59] K. E. Skouby and P. Lynggaard, "Smart home and smart city solutions enabled by 5G, IoT, AAI and CoT services," in *Proc. Int. Conf. Contemporary Comput. Inf.*, 2014, pp. 874–878.

[60] B. Ji et al., "Survey on the Internet of Vehicles: Network architectures and applications," *IEEE Commun. Standards Mag.*, vol. 4, no. 1, pp. 34–41, Mar. 2020.

[61] D. Klaus and B. Hendrik, "6G vision and requirements: Is there any need for beyond 5G?," *IEEE Veh. Technol. Mag.*, vol. 13, no. 3, pp. 72–80, Sep. 2018.

[62] K. B. Letaief et al., "The roadmap to 6G – AI empowered wireless networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, Aug. 2019.



Baofeng Ji received the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, China, in 2014.

Since 2014, he has been a Postdoctoral Fellow with the School of Information Science and Engineering, Southeast University, China. He is currently an Association Professor with the Henan University of Science and Technology, Luoyang, China. He has authored more than 40 peer-reviewed papers, authored or coauthored three scholarly books, holds more than five invention patents, and submitted more than five technical contributions to IEEE standards. His current research interests include MIMO wireless communications, cooperative wireless communications, and millimeter wave wireless communications.



Yanan Wang is currently working toward the M.S. degree in information and communication engineering with the School of Information Engineering, Henan University of Science and Technology, Luoyang, China.

Her current research interests include physical layer secure transmission, confidential communication, and 6G reconfigurable intelligent surfaces.



Kang Song received the Ph.D. degree in information and communication engineering from Southeast University, Nanjing, China, in 2016.

In recent years, he has mainly studied the modern signal processing theory and technology of multiantenna broadband wireless communication. His research interests include MIMO wireless communication.



Chunguo Li received the B.S. degree in wireless communication from Shandong University, Jinan, China, in 2005, and the Ph.D. degree in wireless communication from Southeast University, Nanjing, China, in 2010.

In July 2010, he joined the Faculty of Southeast University, where he is currently an Advisor of Ph.D. candidates and Full Professor. From June 2012 to June 2013, he was the Postdoctoral Researcher with Concordia University, Montreal, QC, Canada. From July 2013 to August 2014, he was with DSL Laboratory, Stanford University, Stanford, CA, USA, as a Visiting Associate Professor. From August 2017 to July 2019, he was Adjunct Professor with Xizang Minzu University, Xianyang, China, under the supporting Tibet program organized by China National Human Resources Ministry. His research interests include cell-free distributed MIMO wireless communications and cyberspace security, and machine learning based image or video signal processing.

Dr. Li is an IET Fellow and the IEEE CIS Nanjing Chapter Chair.



Hong Wen received the bachelor's degree in wireless communications from Sichuan University, Chengdu, China, in 1997, and the Ph.D. degree in wireless communications from Southwest Jiaotong University, Chengdu, China, in 2004.

She is currently a Professor with the University of Electronic Science and Technology of China. She has authored more than 70 papers in internationally renowned journals and important international academic conferences, which include IEEE NETWORK, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, IEEE WIRELESS COMMUNICATIONS LETTERS, and other more than ten authoritative international academic journal paper reviewers. Her research interests include communication network security technology, wireless communication physical-layer security technology, and world-integrated network security technology.



Varun G. Menon received the M.Tech. degree in computer and communication from Karunya University, Coimbatore, India, the M.Sc. degree in applied psychology and the M.B.A. degree from Bharatiar University, Coimbatore, India, and the Ph.D. degree in computer science and engineering from Sathyabama University, Chennai, India.

He is currently an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Kochi, India, and the Head with International Partnerships, SCMS Group of Educational Institutions, India. Since January 2018, he has been an Associate Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology. From April 2012 to December 2017, he was an Assistant Professor with the Department of Computer Science and Engineering, SCMS School of Engineering and Technology. From February 2012 to May 2012, he was an Assistant Professor with the Department of Computer Science and Engineering, M.E.T.S. School of Engineering and Technology, Kerala, India, and from May 2008 to July 2009, he was a Software Developer with Global Allies Ltd., Kerala, India. His research interests include Internet of Things, brain-computer interface, mobile adhoc networks, wireless communication, opportunistic routing, wireless sensor networks, fog computing and networking, underwater acoustic sensor networks, information science, scientometrics, and digital library management.

He is a Distinguished Speaker with the Association of Computing Machinery (ACM). He is currently the Guest Editor of the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE SENSORS JOURNAL, the IEEE *Internet of Things Magazine*, and the *Journal of Supercomputing*. He is an Associate Editor for the *IET Quantum Communications*. He is also an Editorial Board Member of the *IEEE Future Directions: Technology Policy and Ethics*.



Shahid Mumtaz received the master's degree from the Blekinge Institute of Technology, Karlskrona, Sweden, in 2006, and Ph.D. degree from the University of Aveiro, Aveiro, Portugal, in 2011, both in electrical and electronic engineering.

Since 2011, he has been with the Instituto de Telecomunicac oes, Aveiro, Portugal, where he is currently an Auxiliary Researcher and adjunct positions with several universities across the Europe-Asian region. He is currently a Visiting Researcher with Nokia Bell Labs, Murray Hill, NJ, USA. He is the author of four technical books, 12 book chapters, and more than 150 technical papers in the area of mobile communications.



A Comprehensive Survey on Machine Learning-Based Big Data Analytics for IoT-Enabled Smart Healthcare System

Wei Li¹ · Yuanbo Chai¹ · Fazlullah Khan^{2,3} · Syed Rooh Ullah Jan⁴ · Sahil Verma⁵ · Varun G. Menon⁶ · Kavita⁵ · Xingwang Li⁷

Accepted: 22 November 2020 / Published online: 6 January 2021
© Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

The outbreak of chronic diseases such as COVID-19 has made a renewed call for providing urgent healthcare facilities to the citizens across the globe. The recent pandemic exposes the shortcomings of traditional healthcare system, i.e., hospitals and clinics alone are not capable to cope with this situation. One of the major technology that aids contemporary healthcare solutions is the smart and connected wearables. The advancement in Internet of Things (IoT) has enabled these wearables to collect data on an unprecedented scale. These wearables gather context-oriented information related to our physical, behavioural and psychological health. The big data generated by wearables and other healthcare devices of IoT is a challenging task to manage that can negatively affect the inference process at the decision centres. Applying big data analytics for mining information, extracting knowledge and making predictions/inferences has recently attracted significant attention. Machine learning is another area of research that has successfully been applied to solve various networking problems such as routing, traffic engineering, resource allocation, and security. Recently, we have seen a surge in the application of ML-based techniques for the improvement of various IoT applications. Although, big data analytics and machine learning are extensively researched, there is a lack of study that exclusively focus on the evolution of ML-based techniques for big data analysis in the IoT healthcare sector. In this paper, we have presented a comprehensive review on the application of machine learning techniques for big data analysis in the healthcare sector. Furthermore, strength and weaknesses of existing techniques along with various research challenges are highlighted. Our study will provide an insight for healthcare practitioners and government agencies to keep themselves well-equipped with the latest trends in ML-based big data analytics for smart healthcare.

Keywords Sensing · Big data · Data analytics · Internet of things · Healthcare · Machine learning

✉ Fazlullah Khan
fazlullah@tdtu.edu.vn

¹ Faculty of Engineering, Huanghe Science and Technology College, Zhengzhou, China

² Informetrics Research Group, Ton Duc Thang University, Ho Chi Minh City 758307, Vietnam

³ Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 758307, Vietnam

⁴ Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, Pakistan

⁵ Department of Computer Science and Engineering, Chandigarh University, Mohali, Punjab 140413, India

⁶ Department of Computer Science and Engineering, SCMS School of Engineering and Technology, Ernakulam 683576, India

⁷ School of Physics and Electronic Information Engineering, Henan Polytechnic University, Jiaozuo, Henan Province, China

1 Introduction

Over the years, Wireless Sensor Networks (WSNs) have experienced an unprecedented growth in terms of applications, interfacing, scalability, interoperability and data computation. These technological advances along with the innovations in Radio Frequency Identification (RFID), and wireless and cellular networks have laid a solid foundation for the Internet of Things (IoT). The term Internet of Things (IoT) was first coined by Kevin Ashton in 1999 in the context of supply chain management [1]. It refers to a smarter world of objects where every object is connected to the Internet [2]. In IoT, all these objects, also known as entities, have digital identities and are thus organized, managed and controlled remotely and thus having a scope beyond the limits. Due to the growth in the development of smart objects, IoT has enriched almost all aspects of our daily lives and is continuously doing so with diverse range of novel, innovative and intelligent applications

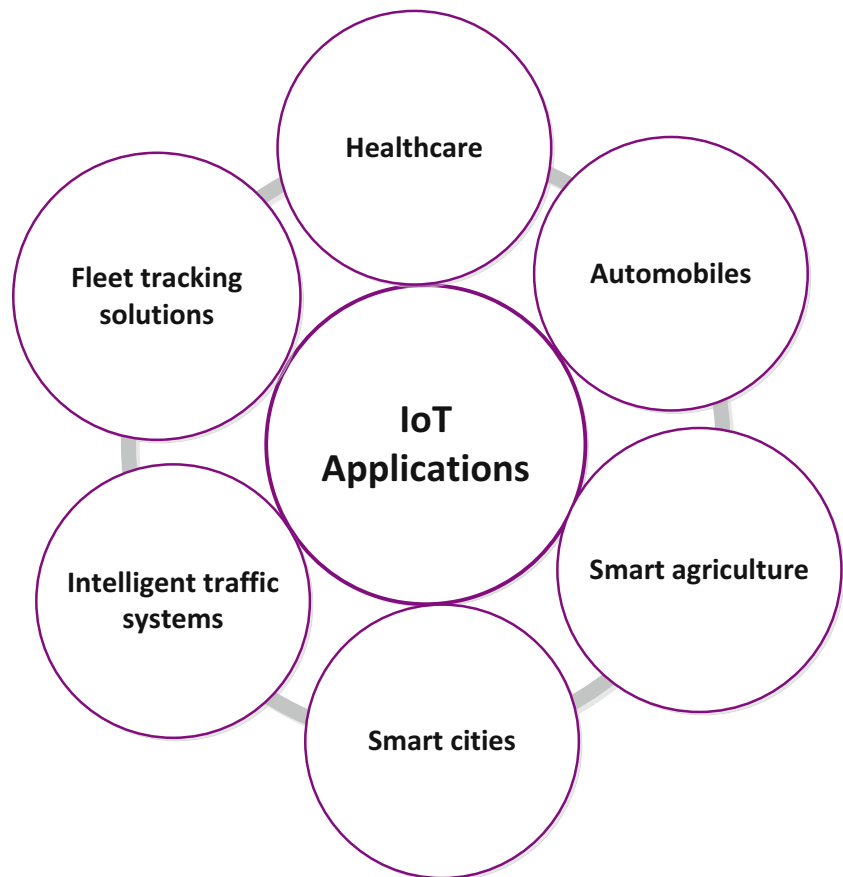
[3, 4]. These applications include smart healthcare [5], smart cities [6], smart agriculture [7], crowd sensing [8, 9], and crowd sourcing [10] etc., as shown in Fig. 1.

These advancements along with innovative applications are highly encouraging and show a bright future of IoT on one side but at the same time, multiple challenges on the other side. Some of these challenges include security, big data analytics, interoperability, Quality of Service (QoS) and energy management [11]. Among them, big data is critical due to the interrelation between IoT objects and plethora of data streams generated by them. A huge amount of information is generated from a vast variety of IoT devices and applications. Various big data analytics are employed to mine such information and improve the decision making. In an IoT context, big data is classified and described by various researchers from different perspectives and various models have been proposed [12–14], however, the most prevalent among them is 5 V model. This model classifies the big data into five categories, based on various attributes associated with them. These attributes are, size of the data (volume), real-time data collection (velocity), heterogeneous data collection from a diverse range of resources (variety), unpredictable data (veracity), and finally the application of such data in various fields, such as industry and academia (value). Recently, we have seen a phenomenal growth in big data research due to its application in various domains. This development is further ignited by the

integration of IoT with big data creating opportunities for the improvement of services for many complicated systems, such as healthcare system. In the IoT literature, there has been a large number of big data technologies that are used for the analysis of large volumes of data from a number of resources in a smart healthcare domain. Among these technologies, machine learning (ML) is a dominant technique that performs complex analysis, intelligent judgments, and creative problem solving on the big data. It is estimated that the economic impact of using ML techniques for big data analytics, i.e., ML-based products and platforms, will range from \$ 5.2 trillion to \$ 6.7 trillion per year by 2025 [15]. This signifies the importance of ML in big data, and particularly in IoT.

There exist numerous comprehensive literature reviews that recognize the research trends in big data, ML, and IoT, respectively. For instance, in [16], the authors discussed the characteristics of big data from various dimensions, i.e. volume, velocity, variety, veracity, variability and value. Moreover, they discussed the current and emerging deep learning architectures and algorithms, specifically designed for big data analytics in various IoT domains. However, the proposed review is generic because it discusses deep learning techniques for big data analysis in multiple domains. Authors in [17] studied the latest machine learning techniques for big data analytics, used for IoT traffic profiling, device identification, security, edge-enabled computing

Fig. 1 Applications of IoT



infrastructure, and network management. However, this survey is restricted to the applicability of ML techniques for big data analysis in a wide range of applications within a specific domain. Similarly, big data technologies across various sectors such as smart health, smart traffic and logistics and smart agriculture were discussed in [18]. This survey enables the readers to choose the most suitable technique from a diverse range of available techniques for data analytics across various domains. Moreover, it also studied the applicability of these techniques in cross domains. However, this survey is limited in scope and pertains only to a single domain. Besides, it partially discussed techniques from each domain. Some surveys, on the other hand, target only a single IoT domain. For instance, the authors in [19] presented a taxonomy of ML-based techniques for smart city domain. However, it does not consider security of the data and the underlying network. All these literature reviews and surveys studied big data and ML from IoT perspective for different applications such as intelligent transportation systems, smart cities, smart agriculture, crowd sensing and smart homes. However, it is evident from the literature that there is a lack of research work that exclusively investigates big data analytics and ML in IoT healthcare domain. Some of the aforementioned surveys dedicated only a single section to this topic, however, there lacks a comprehensive survey on these technologies that identify the most suitable big data technologies and ML techniques for their applicability in IoT healthcare. Moreover, studies that interlink the two cross domains, i.e., big data analytics and healthcare are still in its infancy and thus require further attention from the research community. Similarly, there is no single study that examines the significance of data aggregation and its vital role in this specific domain.

To identify these reach gaps, we have carefully reviewed various papers related to ML techniques for big data analysis. Considering the challenging aspects of big data in the IoT healthcare, in this work, our ultimate objective is to present the state-of-the-art literature on the ML techniques and big data analytics that are exclusively proposed for IoT eHealth. We have also highlighted the strength, weaknesses and future challenges in this context. This will enable the readers to choose the most suitable technique from the available pool of big data analytics tools for healthcare and explore them further in the time ahead. Based on our extensive literature review, this is the first work that targets this particular domain and thus makes it unique from the rest of the papers, available in the literature. The main contributions of this paper are as follows:

- It discusses the relationship between big data and IoT in general, followed by the state of the art big data research in IoT smart health. Finally, a comprehensive discussion is provided on various research challenges that provide further opportunities in this specific domain. This provides the most striking features to all interested parties for further exploration in the years ahead.

- Fundamental concepts of big data and the complex relationship between big data and IoT is explored.
- Big data challenges in IoT healthcare domain are discussed and future research directions are provided in this context.
- A systematic review and study of the existing data aggregation techniques, based on ML and their applicability to IoT smart health are discussed.

The rest of this paper is organized as follows. Section 2 sheds some light on the article classification and our motivation towards researching this specific domain. In Section 3, we provide an introduction of IoT by highlighting its contribution towards various applications. This section exclusively studies the recent developments and transformation of conventional healthcare sector, along with a layered architecture for Wireless Body Sensor Networks (WBSNs). Section 5 discusses the concept of big data challenges, particularly in IoT from smart healthcare perspective. Next, we provide a detailed discussion on the role of ML techniques for the analysis of big data in IoT healthcare in Section 6. A comprehensive and updated literature review on various machine learning techniques for big data analytics in IoT eHealth is provided in Section 7. Research challenges in the field are presented in Section 8. Finally, the paper concludes with Section 9 by stating the limitations and future work for further exploration. The overall structure of this paper is depicted in Fig. 2.

2 Articles classification

In this work, we have examined some of the well-known academic databases and publishers such as Google Scholar, ABI/INFORM Global, Academic Search Premier, Applied Science and Technology Full Text (EBSCO), ACM Digital Library, IEEE Xplore Digital Library, Science direct and general Google search engine. We have used various keywords that include but are not limited to big data, IoT and big data, big data analytics in IoT health, IoT eHealth, and machine learning and big data analytics in IoT healthcare to explore primary challenges and issues in the application of ML to big data analytics in IoT smart health. We were striving for the latest literature including journal papers, conference papers, standards, project reports, patents, white papers and reports from industries. Furthermore, we have restricted our search for the related literature that is published over the past 4 years, i.e., from 2016 to 2020. Among them, particular emphasis was given to papers related to big data research in IoT health care domain. As a result, a total of 361 papers were downloaded, however, only 90 papers among them were selected and thoroughly reviewed, as shown in the Fig. 3. Each paper was carefully analyzed to find the research gaps and clarify our research direction as well as our motivation for carrying out this research. Based on our result, we have selected only 7 out

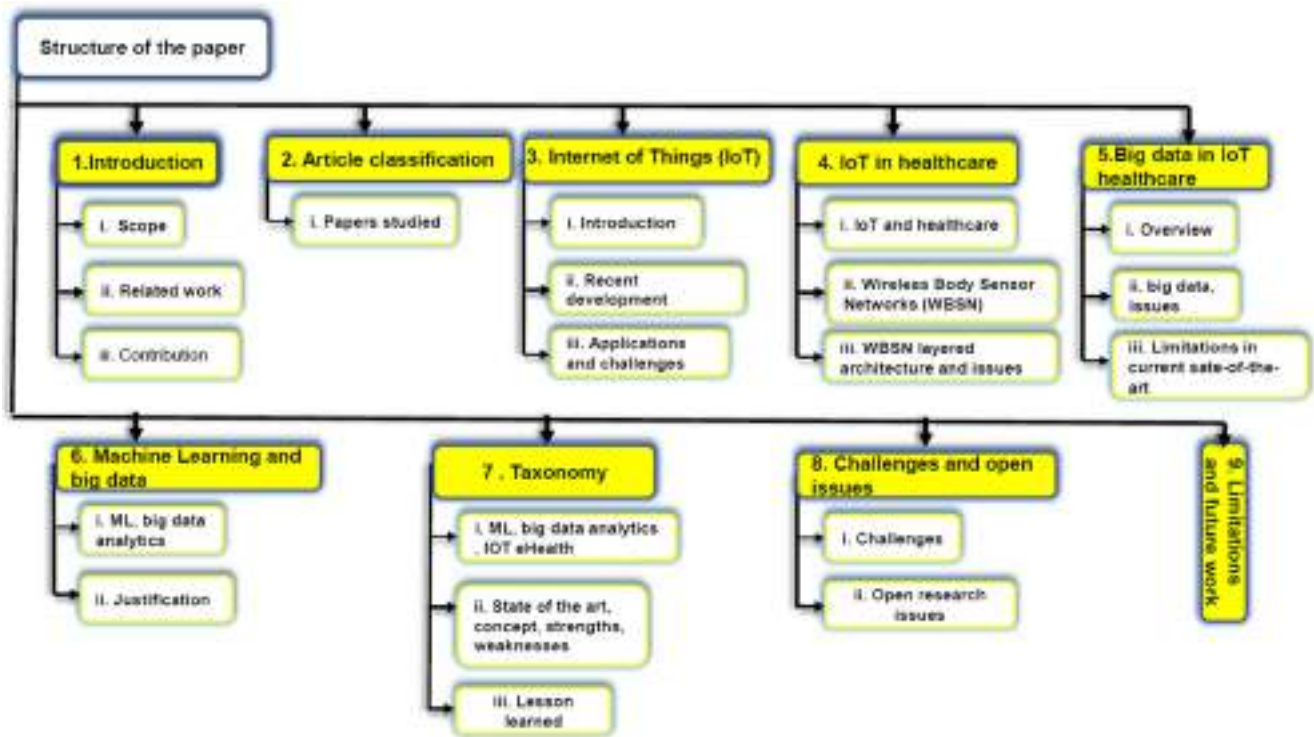


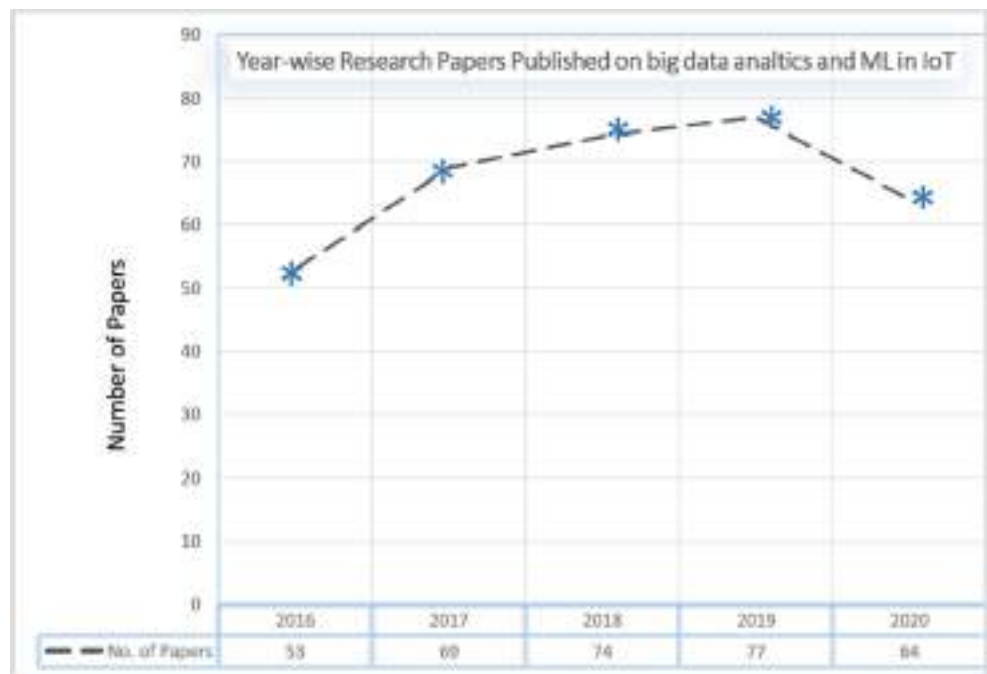
Fig. 2 Structure of the paper

of all research papers, which are [18, 20–25]. A detail discussion on these survey papers was provided in Section 1 that justify as to why we have carried out this research work, and our motivation behind this paper. Moreover, strengths and weaknesses of the aforementioned papers are also provided to justify our work along with the contributions and novelty of this survey.

3 The internet of things

IoT is a web of smart and self-configuring things that can communicate with each other using a global network. It is essentially cyber-physical systems or a network of networks. An informal description for the phrase “IoT” was put forth by IEEE, as “a network of objects each of which is embedded

Fig. 3 Relevant Articles Published over the time



with sensors and these sensors are connected to the Internet” [26]. The seamless communication among participating objects is facilitated using the low-cost sensors installed into a diverse range of objects supporting ubiquitous and pervasive computing applications [27]. Apart from these, other technologies that further stimulated the development of the IoT are wireless technologies, micro-electro-mechanical systems (MEMS) and the Internet. According to the market analysts, around 25 billion sensor-enabled devices will be installed by 2020 [28]. Moreover, the market scope of such devices is expected to be around 2.1 trillion by 2025 [29]. This implies that billions of physical devices or sensor-enabled objects will be connected and will communicate with each other via the Internet. The plethora of objects will generate huge and in most cases, real-time heterogeneous and complex data. It is therefore imperative to extract useful patterns from these raw data in an efficient manner. The raw data gathered from the physical environment need to be analyzed and mined for novel feature extraction and useful information. This become particularly important with the evolution of intelligent IoT applications, where the devices communicate with each other and enable them to share information by making intelligent decisions. As a result, big data analytics using data mining techniques is evolving as a new area of research. In recent years, we have witnessed the development and deployment of a large number of IoT applications [30–32]. These applications include smart cities, smart energy management, smart agriculture, military applications, environmental monitoring and healthcare. IoT has the capabilities to refurbish the current and future scenario of healthcare sector with promising technological, economic, and social prospects. It is estimated that the economic impact of IoT-enabled hardware and software will reach USD 176.82 Billion by 2026 [33]. The healthcare sector alone will constitute about 41%, a major share followed by industrial automation with 33% and energy with 7% of the IoT market [34]. Apart from these, 15% of the IoT market is related to objects and product-related transportation, agriculture, urban infrastructure, security, and retail sectors. These outlooks indicate the remarkable growth of the IoT services to healthcare industry on one side, while, challenges such as big data and other challenges on the other side that the research community will face shortly.

4 IoT in healthcare

With the emergence of eHealth and mHealth, we have witnessed an increasing role of technologies in the healthcare sector. Millions of sensors are attached to the patients that continuously monitor their health using various physiological, environmental and behavioural parameters. In healthcare IoT, i.e., eHealth and mHealth, wireless body sensor networks (WBSN) is a predominant technology for monitoring the

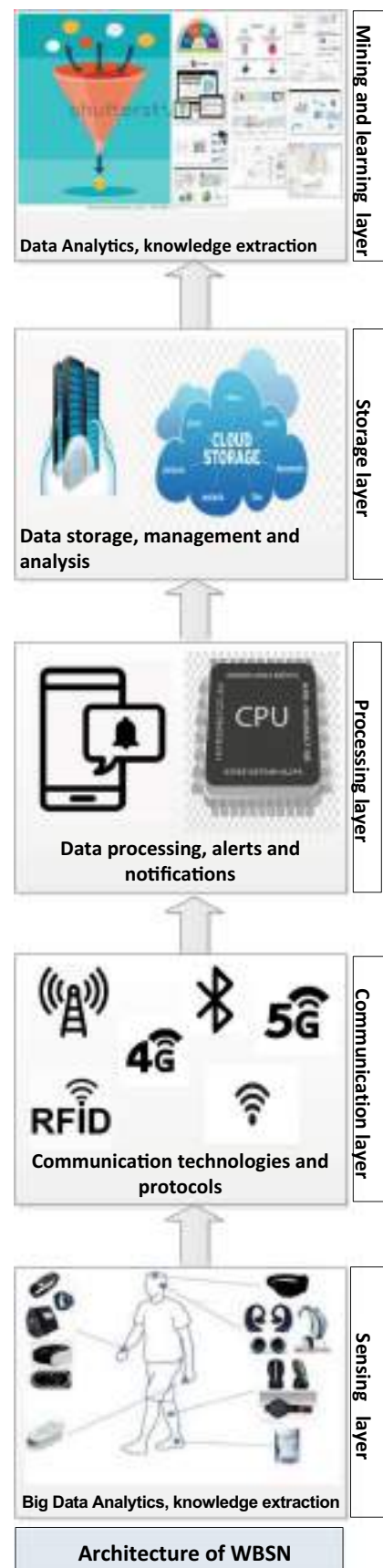


Fig. 4 Layered Architecture of Wireless Body Sensor Network

patients. WBSN consists of sensors that are deployed around the human body [35]. The layered architecture of WBSN comprises of sensing layer, communication layer, processing layer, storage layer, and mining and learning layer as shown in Fig. 4 [36]. Each layer contains various components with their responsibilities. The sensing layer includes various sensing devices, such as wearable sensors and in-body sensors. Recently, medical super sensors (MSS) came into the market that have more memory with improved processing and communication capabilities as compared to the ordinary sensor nodes. These sensors are usually wearables or sometimes implanted inside the patients' skin and can communicate with the network. These sensors gather vital information pertaining to body temperature, blood pressure, heartbeat rate, respiration rate, ECG, and blood glucose for diabetic patients [37]. In recent years, actuators are employed for raising alarms and modifying the environmental parameters, whenever necessary. We have witnessed huge developments in these applications in the form of novel monitoring applications. As a result, a large amount of contextual data is generated from these applications. It is mandatory to consider big data among other challenging issues while designing devices at the sensing layer. Some of these issues are price, size, energy consumption, memory, processing, power, deployment and organization of various devices at this layer. The next layer is the communication layer which is somehow similar to physical layer of the TCP/IP model. This layer is responsible for physical objects to connect and share data in WBSN, using specific communication protocols. It facilitates the inter and intra network communication. Standard and communication protocols defined at this layer provides interoperability in WBSN. These protocols also facilitate the exchange of data with existing infrastructures. There are various standards used by WBSN for intra communication at this layer, such as Bluetooth, ZigBee, RFID, NFC and UWB [38–40]. Each of these standards have their pros and cons and are used based on the specific application's requirements [41]. Various challenges faced at this layer are network management, QoS (congestion, latency and energy efficiency), and security and privacy. Apart from these, data aggregation and big data analytics need to be considered for further exploration. These techniques preserve energy of the resource starving networks by substantially lowering the data transmission across the network. The third layer is the processing layer that analyzes the gathered data, makes decisions, and raises alarms and notifications. The main components of this layer are: (a) the processing unit (b) hardware platforms, and (c) operating system. The challenging issue at this layer is the limited processing capabilities of hardware components. The partially analyzed data at this layer is then passed on to the next layer, i.e., the Storage Layer. In IoT healthcare, a large number of devices can be attached to the human body that generates massive and complex data. It is the responsibility of storage layer to efficiently manage and store

such data for further analysis and usage. IoT-based system are low on memory and are therefore unable to store such data. To overcome this limitation, numerous cloud-based platforms are available for the storage of data such as ThingWorx [41], OpenIoT [26, 42], Google Cloud [43], Amazon [44], Nimbits [45], GENI [46, 47]. These platforms improve the management and storage of data. Data can be reviewed and accessed virtually from anywhere and everywhere. This in turn facilitates the health professionals and researches to explore it further for better understanding and advancement of the field. Finally, the mining and learning layer is responsible for big data analytics and knowledge extraction. Various data mining techniques are available in the literature, however, ML techniques are successfully applied for big data analytics in health care IoT [17, 48]. ML-based techniques can manage huge data set efficiently, learn from the data and improve the learning experience. They are used to mine the vast amount of medical information and extract useful, potentially interesting, and unique and hidden information. The main components of this layer are: clustering, classification, association analysis, time series analysis, and outlier analysis [19, 49]. It is expected in the future that feedback will emerge from this layer, as opposed to present IoT scenario, where it comes from the clinics.

5 Big data challenges in IoT smart healthcare

Despite the hype surrounding the smart applications of eHealth and mHealth in IoT, big data is still a challenging issue. Sensors and various medical devices attached to the patients' bodies generate massive volumes of heterogeneous data, also called Big Data [50]. This huge volume of data contains highly correlated and redundant patterns. It is imperative to mine these data for providing continuous, efficient, and seamless healthcare facilities around the clock. However, the challenging issues are the processing and transmission of such data across the network. These issues not only consume higher energy but also bandwidth of the resource-constrained networks that lead to congestion and reduces the energy and lifetime of the underlying networks [51]. It is therefore imperative to aggregate raw data, using big data analytics, before transmitting it across the network for accurate and timely decision making. Moreover, it becomes a major concern for all stakeholders to process the data within the network intelligently and efficiently. Removing redundant and erroneous data, while identifying and extracting meaningful information and gaining new insights into the large volume of raw captured data is the core utility of big data analytics [52]. These techniques not only improve the performance but also conserve the energy using novel energy management techniques by enabling the long term operation of these networks [20, 51, 53].

6 Machine learning and big data analytics for IoT

In this section, we discuss the application of ML for big data analytics. ML is a subfield of computer science that evolved from pattern recognition and computational learning theory [54]. It is a type of Artificial Intelligence (AI) that provides machines with the ability to learn without explicit programming by making complex decisions [55]. In the past, it has been successfully applied to various domains such as computer vision [56], computer graphics [57], natural language processing (NLP) [58], speech recognition [59], computer networks [60], and intelligent control [61]. In recent years, we have witnessed its vital role in IoT and big data analytics due to its phenomenal growth with a diverse range of innovative applications. As a result, highly correlated data is produced from these heterogeneous and complex data sources, i.e., IoT devices. Thus, data management in these systems becomes extremely difficult that results in numerous challenges for the research community [62–65]. It is important to manage data from these large number of sources with increased velocity and scalability by devising novel big data analysis techniques. Existing techniques are ineffective due to lower accuracy and higher energy consumption that does not cater to these diverse ranges of applications. It is necessary to improve these techniques to cater to various applications. ML techniques play a pivotal role in IoT eHealth [66]. It empowers us to obtain deep analytics from a larger pool of available information. It mines useful information and features hidden in IoT data, and facilitates the decision-making process. Moreover, it helps us in the development of efficient and intelligent IoT applications. An IoT analysis model consists of various components such as data sources, edge/fog computing, and ML techniques for IoT big data analytics. In this model, the potential data sources include wearable devices such as sensors, and body area networks. They capture information related to human health such as temperature, ECG, and environmental data like humidity and camera's images. Various ML techniques are applied to the data captured by these sources for further analysis. It is evident from the literature that ML techniques have successfully been applied for big data analysis in various applications of IoT such as smart traffic [67, 68], smart agriculture [69], smart human activity control [70], smart weather prediction [16, 71], healthcare [72, 73], and smart cities [19]. Big data has been studied in a diverse range of IoT domains. However, it is evident from the literature that there is lack of a comprehensive literature review that exclusively investigates big data analytics in IoT healthcare. Though, some of the aforementioned surveys dedicated only a section to this domain, there is no single study that examines the significance of ML techniques for big data analysis in IoT healthcare. In the next section, we present state-of-the-art literature by reviewing the latest ML

techniques for big data analysis in IoT smart healthcare system. Moreover, strengths and weaknesses along with future challenges are also highlighted. This provides an insight to the readers that enable them to explore it further in the future.

7 A taxonomy of machine learning techniques for big data analysis in IoT smart healthcare system

IoT aims to improve the quality of human lives by automating some of the basic tasks that otherwise humans need to perform manually. In this context, monitoring and decision making is shifted from humans to machines. For instance, in IoT-based assisted living applications, sensors are attached to the health monitoring unit used by the patients. The information gathered by these sensors are transmitted across the network and are made available to all interested parties. This not only helps in timely treatment of the patients but also improves the responsiveness and accuracy of the underlying application [74, 75]. Moreover, the current medicines taken by the patient are monitored and the risk of new medication is evaluated in terms of any allergic reaction [66, 76]. As a result, not only the time is conserved but monetary value remains in place too. In this section, we review only selected ML techniques for big data analytics in IoT eHealth. Moreover, the key concepts along with their similarities and differences, strength and weaknesses are provided, and are summarized in Table 1.

7.1 ML-based recommendation system

In [77], the authors proposed a recommendation system that devised the most feasible IoT wearable devices, based on the needs of an individual. The proposed system initially gathers the available data related to a patient's health, e.g., previous history, demographic information, and retrieval of archived data from the sensors attached to the patient. Various ML-based classification techniques such as decision tree, logistic regression and LibSVM, are used to predict the occurrence of diseases. Finally, a mathematical model is used for recommending a customized IoT solution for each individual. In [78], the authors proposed a disease prediction system by performing the real-time Electrocardiograph (ECG) analysis. Firstly, the proposed approach analyzes and classifies the ECG waveforms that are captured in real-time from the ECG monitoring devices using various ML classifiers such as KNN and bagged tree. Next, any signs of diseases and abnormalities in the ECG are predicted and are then communicated to the cloud in real-time via a purpose-built IoT network, owned by the National Health Services (NHS), UK. Simulation results showed that the precision of the proposed scheme can reach up to 99.4%. However, the precision as well as the

Table 1 Key technological concepts, their similarities and differences

Category	Description
Big data and their characteristics [16]	Discuss IoT from the big data perspective. Also discusses the characteristics of big data from the 6 Vs dimensions, i.e. Volume, Velocity, Variety, Veracity, Variability and Value. Also, analyze and summarize major research attempts that apply deep learning in the IoT domain. Finally, it shed light on some challenges and potential directions for future research in this area.
Machine learning, big data analytics in diverse range of applications [17]	Focuses on the application of machine learning for IoT followed by the relevant techniques, including traffic profiling, IoT device identification, security, edge computing infrastructure, network management and typical IoT applications. Also highlight the most recent advances in machine learning techniques and their diverse applications, challenges and open issues.
Machine learning techniques for big data analysis in smart cities domain [19]	Use case of applying modified Support Vector Machine (SVM) to Aarhus smart city traffic data. Also, present a taxonomy of machine learning algorithms. It further explains the application of these techniques to big data analytics in smart cities domain. The paper is finally concluded with research challenges, and future research directions.
Big data analytics in various IoT application domains [18]	Discuss, analyze and divide latest research related to big data analysis in various IoT application domains. It guides the readers to choose the most suitable technique from a diverse range of available techniques for big data analytics in these domains. A critical view of various big data technologies across these categories are also presented.
Provides a systematic review of the latest data aggregation techniques for IoT [21]	Classify data aggregation techniques based on their underlying topologies, such as, tree, cluster and centralized. It also explores various challenges that these techniques face. A discussion on various performance metrics such as energy efficiency and latency is also provided for the accurate evaluation of these techniques. A comparative study along with their strength and weaknesses of these techniques as well as recommendations for further extension in the future is provided.
Challenges facing IoT [23]	This paper discusses a wide range of technology-based issues and challenges facing IoT. It further explains the vision and various features of this paradigm from different dimensions. The key feature of this work is that it provides a comprehensive and latest survey on a diverse range of IoT enabling emerging technologies. Moreover, It also classifies the existing literature based on different research topics. Finally, an insight into various research challenges and research issues are also provided for further research in the field.
Data redundancy in IoT sensor networks [22]	This paper reviews the challenging aspect of data redundancy and recommends data aggregation as an effectively technique to overcome on this issue. It present cluster based data aggregation techniques. It also classify these techniques based on the location of deployment, their pros and cons and future challenges.
Applications of machine learning for big data analytics in IoT domain [17].	This highlight the application of machine learning technique (supervised and unsupervised) for big data analytics in various application domains. It thoroughly discusses security techniques related to device security and network security. In the end, a comprehensive discussion is provided on various challenges and open research issues.

performance need to be evaluated using other metrics such as time complexity and energy efficiency.

In [79], the authors proposed an IoT architecture having five distant but inter-related layers. The first layer is the sensing layer, which includes various sensing devices used for gathering the data. These devices include but are not limited to, sensors, actuators, and a wide range of wearable devices. The second layer is the sending layer, which is somehow similar to the physical layer of the Open Source Interconnection (OSI) model. Its main responsibility is to devise various communication mechanisms for data transmission. This layer discusses communication mechanisms such as Wi-Fi, Bluetooth, ZigBee and Long Term Evolution (LTE) for sending the data to cloud. The third layer is the processing layer, which is concerned with the processing of data, based on some pre-defined criteria. Once the data is processed, notifications and alerts are generated in response. Some of the devices where processing occurs are smart phones, micro-controllers and microprocessors. At the fourth layer, i.e., storage layer, the data is stored at a preferred location such as clouds and hosted servers. Finally, the fifth layer, also known as the mining layer, converts the information into decisions using a diverse range of data mining or ML algorithms for reaching a conclusion. Based on the decision, various suggestions and recommendations are made. In [80], the authors proposed a recommender system Pro-Trip, which allows the users to organize the activities before a trip or on an ongoing trip. Pro-Trip collects all the data from the patients that is used for further recommendations to provide accurate results. The authors also proposed a technique for food RS designed for the healthcare system. The results of Pro-Trip are evaluated based on climate and food datasets that are collected in real-time. In the food recommendation system, they have evaluated the performance with latency, energy efficiency, and security in mind. In [81], the authors proposed a novel recommendation system based on Type-2 fuzzy ontology-aided RS, especially designed for IoT-based healthcare systems. It overcomes the issues faced while monitoring and extracting the optimal value of risk factors in patient's data. Hence, the proposed technique ensures to observe the patient and then recommends the diet with a discrete amount of food and medicines. This approach evaluates the risk faced by the patient, deduces the health state of the patient with the help of wearable devices embedded with sensors, and further suggests the prescription of medicines and food. Authors have amalgamated two techniques: Type-2 fuzzy logic and fuzzy ontology, which remarkably improve the rate of prediction of recommendation. The accuracy, recall, and precision are compared with other ontologies, i.e., Type-1 and classical, which show excellence in the results. The future work could magnify upon the Type-2 fuzzy neural network and sentiment analysis for the RS. In Table 2, we have shown various recommendation systems for smart healthcare.

7.2 ML-based prediction system

In [82], the authors proposed an IoT framework for predicting whether the person under observation is in stress or not by monitoring his/her heart beats. The proposed framework detects the pulse waveforms using a specially designed WiFi equipped board, which forwards the data to a pre-defined server. Next, the data gathered at different time intervals are assembled and stress prediction is evaluated by applying various ML techniques such as SVM and logistic regression. Simulation results showed that precision of the proposed framework can reach up to 68%. However, its precision can be improved further using appropriate classification models. In [83], the authors proposed a smart tele-health monitoring system using speech recognition algorithms. Its design goal is to identify and predict the occurrence of Parkinson's disease using K-mean algorithm. The proposed system is device-independent and can be employed by a variety of wearable devices. The proposed system employs an edge computing framework as the wearable devices are resource-limited. The idea behind using edge computing is to achieve distributed services by reducing the reliance on centralized infrastructure. In [84], a cloud-based IoT framework was proposed for monitoring various diseases. It forecasts the level of these diseases, i.e., from normal to severe among students. It utilizes the concept of computational science on the data collected from the students using sensors and are stored at a repository to predict severity of the disease. Furthermore, various classification algorithms are used to predict the occurrence of such diseases. The proposed approach is evaluated using various performance metrics such as specificity, sensitivity, and F-measure. Simulation results prove that in terms of accuracy, the proposed approach outperforms the traditional approaches. In [85], the authors proposed a smart e-Health Gateway at the edge of the network in Fog-assisted system architecture. The gateway can perform real-time data processing, data mining, and data storage, locally. Moreover, the strength of the proposed architecture is that it can enable us to solve some of the emerging and complex issues faced by the ubiquitous healthcare systems, such as mobility, energy efficiency, scalability, and reliability. Practical demonstration of proposed prototype demonstrated high-level features such as Early Warning Score (EWS) of our health monitoring system. The authors in [86] proposed a three-layer architecture for storing a large amount of sensory data for earlier prediction of heart diseases. In the proposed architecture, the first layer is responsible for data collection. The second layer is concerned with the storage of large volume of sensory data at the cloud. Finally, in the third layer, a prediction model for heart diseases is developed. At this layer, "Receiver Operating Characteristic Curve (ROC) analysis is performed that identifies potential symptoms before the occurrence of heart disease. In [87], the authors discussed the application of IoT in healthcare. They presented a novel

Table 2 ML-based Recommendation Systems for Smart Healthcare

Description	Features	Type	Strengths	Weaknesses
Recommend the most feasible wearable based on the needs of an individual. [77]	Patient’s health, for e.g., previous history, demographic information, and retrieval of archived data	Recommendation System	Efficiency	lacks numerical analysis
Diseases prediction system to perform real time Electrocardiograph [78]	Cardiac abnormalities in real-time	Recommendation System	Accuracy	Portability, real-time Monitoring
IoT architecture having five distant but inter-related layers [79]	Layer1: Sensing layer Layer2: Sending layer Layer3: Processing layer Layer4: Storage layer Layer5: Mining layer	Recommendation System	Accuracy	Quite complex system
Recommender system Pro-Trip [80]	Allows the users to organize the activities before a trip or on an ongoing trip	Recommendation System	Accuracy	Results of Pro-Trip are evaluated based on climate and food datasets
Type-2 fuzzy ontology-aided RS, designed for IoT-based healthcare systems	The proposed technique ensures to observe the patient and then recommends the diet with a discrete amount of food and medicines and evaluates the risk faced by the patient, deduces the health state of the patient.	Recommendation System	Accuracy	Lack of sentiment analysis for the RS.

ML-based model for disease classification in a healthcare monitoring system. Based on the extensive simulations, it was concluded that the proposed framework can extensively enhance the performance and detects diseases with higher accuracy. In [88], the authors proposed a Hierarchical Computing Architecture (HiCH) for the IoT healthcare sector. They proposed and implemented a system, similar to IBM’s MAPE-K model REF for the arrhythmia detection. The proposed system has three distant but interrelated layers of fog computing. They are: sensor devices layer, edge computing devices layer, and cloud computing layer. The responsibility of the first layer, i.e., sensor devices layer, is to sense and monitor the phenomenon of interest. Next, edge computing devices layer is responsible for making a local decision as well as system management. Finally, heavy training procedures are performed at the cloud layer. Simulation results show that the proposed system outperforms the traditional systems in terms of response time, bandwidth utilization, and memory utilization. However, accuracy of the proposed system is lower and may be improved further in the future. In [89], the authors proposed a low-cost, remote monitoring system that detects various fatal diseases such as cardiovascular diseases, diabetic mellitus, hypertension and different chronic degenerative

medical conditions. The proposed system detects these diseases by measuring Heart Rate Variability (HRV), i.e., variation that occurs between consecutive heart beats concerning time. The data from the patients are captured using Zigbee pulse sensor. The captured data is then transmitted to the application server using Message Queuing Telemetry (MQTT), a specially designed IoT protocol. At the application server, the HRV data is further analyzed and visualized that shows any abnormalities for timely actions to be taken. Similarly, in [90], a novel, intelligent system called neuro-fuzzy temporal intelligent medical diagnosis system was proposed. The proposed system uses fuzzy rules that can classify and efficiently predict various fatal diseases. In Table 3, we have shown various ML-based prediction systems for smart healthcare.

7.3 ML-based data aggregation

The authors in [91] proposed a real-time data compression technique, known as Adaptive Learner Vector Quantization (ALVQ). The unique feature of ALVQ is that it works without having prior knowledge of the underlying topology. Initially, data is aggregated at the sensor level by wearables to ensure that only non-correlated data is forwarded towards the cluster

Table 3 ML-based Prediction Systems for Smart Healthcare

Description	Features	Type	Strengths	Weaknesses
Remote monitoring system for cardiovascular activities [89]	Detection of fatal diseases	Prediction System	Low-cost, secured, quick, easy-to-use	Interoperability with the web, low memory
Low-cost heart monitoring system [82]	Cardiovascular stress prediction using SVM	Prediction System	Utility, Accuracy	Efficiency, Privacy
Smart telehealth system using speech recognition. [83]	Parkinson	Prediction System	Lightweight, Energy-efficient	Security, Effectiveness
ROC-based 3-tier prediction model for heart diseases [86]	Cardiovascular	Prediction System	Scalable, availability, high throughput	Energy-efficiency, Accuracy
IBM-based MAPE-K for disease detection. [88]	Arrhythmia	Prediction System	Response-time, Bandwidth, Memory utilization,	Accuracy,
Fuzzy-enabled Intelligent medical diagnosis system [90]	Nervous system	Prediction System	Efficient	Accuracy

head (CH). This not only reduces the computational cost on the CH but also reduces communication cost in the network. However, the proposed technique does not devise any aggregation mechanism at the CH level. Moreover, the applicability of this technique should be evaluated for critical applications with an acceptable level of accuracy. In [92], the authors presented a cluster based self-Organizing data aggregation framework for a healthcare facilitation. A self organizing algorithm is employed that classifies the aggregated healthcare data. The proposed scheme reduces the high-dimensional space into low-dimensional space that lowers the amount of transmitted data in the network and enhances the network lifetime. Moreover, it also enhances the quality of the aggregated data. In [93], the authors eliminated the highly correlated data using big data techniques. Hadoop framework was used to extract the critical information from data captured by sensors detached with the patients. Once redundancy is eliminated, the refined data is forwarded towards the physicians in real-time for timely action. As a result, various services provided by health care professionals are significantly improved. This reduces the amount of data transmitted across the network that in turn improves the responsiveness, accuracy, QoS, energy conservation and network lifetime. In [94], a novel framework known as “health informatics processing pipeline” for big data analytics in IoT was proposed. The proposed framework uses various techniques to extract useful patterns from the raw gathered data. The main features of the proposed framework include data capturing, storage, analysis, and data searching. The proposed framework eliminates the correlated data and transmits only highly refined and useful features. These features enable the framework to decide with the help of a decision support system using various ML techniques. In Table 4,

we have shown various ML-based data aggregation schemes for smart healthcare.

7.4 ML-based living assistance

IoT-based solutions are assisting elderly population in the form of personalized, preventive and collaborative care. In this regard, authors in [95] presented IoT-based living assistance for the aged population. The proposed system monitors and stores the vital information of patients using a cloud-connected wrist band. An alarm is raised during critical situations that assist the patients by informing the healthcare professionals to take the right action and decision. The proposed solution is both energy and cost-efficient. Likewise, in [96], the authors proposed a framework that monitors medicine intake of patients. The key attributes of the proposed system are that: it tracks the medicine intake from the patients history including missed dosage. In case of medication discrepancy, such as missed or over dosage, an alarm is generated alerting both the patients as well as the medical staff. Moreover, in [97], authors designed a patient monitoring system for critically ill patients in the intensive care unit (ICU). The proposed system informs and assists all stakeholders in real time, whenever abrupt changes occurs in the pre-defined conditions for timely action. In [98], the authors has proposed a novel monitoring system based on the patient movement. The proposed system provides emergency services to the patients by evaluating their emergency situation from monitoring their movement. The in-home patient monitoring system relies especially on the proposed monitoring system. In [99], a system that explicitly detects the human presence without using cameras or motion detectors was proposed. Initially, the system

Table 4 ML-based Data Aggregation for Smart Healthcare

Description	Features	Type	Strengths	Weaknesses
Real-time data, Compression using Adaptive-learner Vector quantization [91]	It works without having prior knowledge of the underlying topology.	Data aggregation	Compression, Efficiency, signal reconstruction,	Ignore multidimensional data, Missing data
Self-organizing approach to transform high dimensional space into low dimensional space. [92]	It enhances the quality of the aggregated data	Data aggregation	Reliability, Efficiency, Communication cost	Fault tolerance, Topological support
Hadoop-based framework for spatially and temporally correlated data elimination [93]	Improves the responsiveness, accuracy, QoS, energy conservation and network lifetime.	Data aggregation	Energy-conservation	Flexibility, Efficiency
Health informatics framework for gathering, storing, analysing, and searching data for accurate decisions. [94]	It includes data capturing, storage, analysis, and data searching.	Data aggregation	Accuracy	Optimization

collects interactive data, i.e., reading or writing with a diverse range of devices. Next, the presence of human is detected using various ML classification algorithms such as C4.5 decision tree, linear SVC and random forest. The system was initially trained and tested using a dataset gathered over a period of 3 days from 900 users. Simulation results shows that the precision of the proposed approach may vary from 50 to 99% with varying range of classification algorithms. However, it needs to be tested in real world scenarios within various settings to study its behaviour. In [100], the authors proposed an inexpensive health-care monitoring system for patients. The model is based on lightweight sensor-enabled wearable devices performing sensing, analyzing and sharing of real-time health-care data from the patients. An Arduino-based wearable device with body sensor networks is employed for data collection. Moreover, Labview is integrated with the system to facilitate the remote monitoring of home-bound patients. The proposed system eliminates many deficiencies that exist in manual systems. In Table 5, we have shown various ML-based assisted living approaches for smart healthcare.

7.5 ML-based secured analysis

It is imperative to ensure the security and privacy of health care data due to its sensitive nature. In this regard, authors in [101] presented an on-line healthcare monitoring system. The proposed system collects and analyzes the health-related data from the patients, using sensors

and medical devices, that negate the death circumstances. They fused various techniques such as watermarking and signal enhancements to improve the security and performance, accounting for clinical errors in the proposed scheme. Authors in [102] proposed a uniquely collaborative and intelligent security model for the IoT-based healthcare environment. The main objectives are to reduce security risks posed to a diverse range of IoT-enabled healthcare solutions. The proposed system is designed with a particular emphasis on the recent advances in this field. Various ML techniques are used for secured classification of the patient data. Likewise, authors in [103] presented a WBSN-enabled IoT healthcare solution. The proposed approach monitors the patient using wireless body network that consists of tiny, lightweight sensor nodes. The proposed approach uses various ML techniques to ensure that security is enhanced by protecting WBSN from intruders and various attacks. In [104], the authors proposed a novel mobile cloud computing framework for big data analytics. The main features of the proposed framework are that it offers availability and interoperability of health-care data, which can be shared among all interested parties. Various ML and DL techniques were used for classifying and testing the gathered data from patients. Although privacy and security of the health-care data are thoroughly discussed, they were not evaluated practically. In Table 6, we have shown various ML-based secured analysis approaches for smart healthcare.

Table 5 ML-based Assisted Living Techniques for Smart Healthcare

Description	Feature	Category	Strengths	Weaknesses
Alert-based Cloud-connected monitoring system [95]	Elderly patient assistance	Assisted living	Performance, life time	Limited functionalities, Connectivity,
Medicine-tracking alarm-based patient monitoring system [96]	Medicine intake	Assisted living	Energy-efficiency	Quality of Service
Real-time monitoring system for dynamic changes in the pre-defined health conditions [97]	Healthcare monitoring system for patients in the ICU	Assisted living	Efficient, Accuracy	Availability, Load balancing
Intelligent monitoring system for patients based on their movement. [98]	Motion-awareness	Assisted living	Energy-efficient	Noise, Error
ML-based patient monitoring system [99]	Human presence detection without cameras and motion detectors	Assisted living	Feasible, Inexpensive	Precision, Performance, Interoperability
Arduino-based real-time monitoring system [100]	Lightweight, real-time patient health monitoring	Assisted living	Inexpensive, Simple	Real time, Single subject, Low accuracy

8 Challenges and open research issues

In this section, we provide an insight into various challenges related to ML techniques for big data analytics in the IoT healthcare domain, as shown in Fig. 5. Moreover, research gaps are also provided for researchers to fill them in the future.

8.1 Resource scarcity

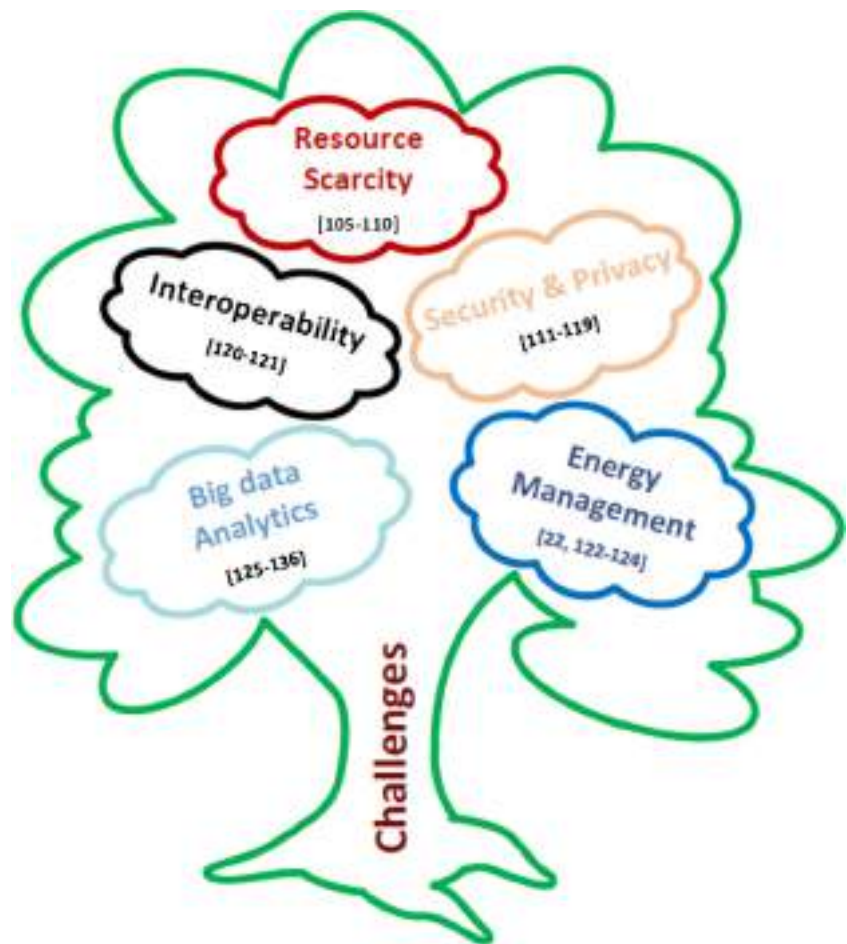
In IoT, most devices such as sensors, smart phones, microcontrollers actuators, RFIDs, and gateways have limited energy with lower computational and processing power [105–107]. Moreover, data generated from these densely

deployed, resource-starved devices contain similar and redundant patterns. Transmitting such correlated data across the network results in high energy consumption, lower QoS and lower throughput [108, 109]. The resource limitation issue is resolved upto some extent by integrating the IoT with the cloud computing paradigm. However, it increases the cost and complexity. Besides, other issues related to resource management such as resource discovery, modeling, provisioning, scheduling, estimation and monitoring are still of higher concern due to the unique nature of IoT networks [110]. Furthermore, optimization within the resource allocation techniques is an area to be explored further in this context. It is compulsory to design novel, lightweight and energy-efficient data aggregation techniques based on ML, as most of the

Table 6 ML-based Secured Analysis for Smart Healthcare

Description	Features	Type	Strengths	Weaknesses
IoT-based real-time monitoring system using watermarking and signal enhancements [101]	Resilient for clinical errors detection	Secured Analysis	Security, Accuracy, QoS	Security Optimization, Implementing and testing on real world patients
Collaborative and intelligent security model for IoT-enabled health care [102]	Ensures the privacy and security of healthcare data	Secured Analysis	Lightweight, Secured	Fault tolerance,
Wireless Body Sensor Network (WBSN)-enabled intelligent monitoring system [103]	Protect the healthcare system from Intruders	Secured Analysis	Security, Efficiency	Performance, Energy-efficiency
Mobile CloudComputing (MCC) framework [104]	Big data analytics for availability & interoperability of health data	Secured Analysis	Availability, Interoperable	Performance, Accuracy

Fig. 5 Challenges faced by ML techniques for big data analytics for IoT healthcare



existing techniques are not energy-efficient. Moreover, novel schemes should be devised that distribute the task among various IoT components that not only matches the resource scarcity of these networks, but also offers an acceptable level of accuracy [105].

8.2 Security and privacy

The application of IoT in healthcare domain is providing personalized facilities, i.e., customized and rapid access to healthcare which was unimaginable earlier. In these applications, both the technology and healthcare devices work with each other to offer a wide range of services. It is forecasted that almost 40% of IoT-related technology will be health-related shortly, more than any other market segment, with a huge market share of USD 136.8 billion by 2021 in [111]. Such developments in this field are revolutionary, however, it should be carefully adopted due to the challenges faced in the context of security, privacy and sensitivity by health-related data [112–114]. Upstream transmission of compromised data not only has a devastating effect on the underlying data aggregation technique but also deteriorates its performance [115]. It exposes the underlying networks to a wide

range of security attacks such as DoS, eavesdropping, Sybil, sinkhole, and sleep deprivation attacks. These threats remain a challenge due to the rapid expansion in the field with an ever-increasing number and complexity of the emerging software and hardware vulnerabilities. Besides, healthcare data containing sensitive and confidential information such as personal details, family history, electronic medical records, and genomic data should be kept confidential. It was predicted that 72% of malicious traffic targeted the healthcare data [116]. It is thus imperative to protect such data from hackers by enforcing privacy and security, both physically and virtually [117]. Other challenges include low security, misconfigured devices, and network settings. Moreover, data from these varying range of devices are mostly heterogeneous in nature and usually managed by third parties and thus governance, security, and privacy of such data become a challenging task [118, 119]. Furthermore, existing security techniques are not a feasible option due to the resource-constrained nature of IoT devices. Designing lightweight and energy-efficient data aggregation techniques that not only secure, but also ensure the confidentiality, security and privacy of the data is an interesting domain for further examination.

8.3 Interoperability

Recently, we have witnessed rapid development both in the hardware and software but the actual challenge is the lack of global standards that are accepted and agreed by public across the globe. Thus, the healthcare IoT devices pose serious interoperability challenges. The designer must not only focus on the development side but at the same time, strive for interoperability among all aspects of IoT eHealth such as smart wearables, body area sensors, and advanced pervasive healthcare to promote healthier life styles [120, 121]. The benefits associated with interoperable devices are increased throughput, minimized unplanned outages, and reduced maintenance costs. Semantic interoperability of the clinical information is an important area for future research.

8.4 Energy management

Energy management is another challenging aspect of IoT healthcare applications. Usually, wearable and sensors attached to the human body are energy-constrained. They are equipped with limited energy supplies [122]. The frequent changes of batteries in these sensors and devices is cumbersome and sometimes impossible. Supplementary healthcare professionals with additional costs will be required to constantly look after these devices and sensors for battery replacement, whenever energy goes beyond certain thresholds. This will result in fatigue and mismanagement due to dynamic environments. Energy efficiency becomes an integral factor that determines the success of the underlying applications [22]. To overcome and improve energy conservation, it is necessary to design low power sensors that do not require frequent changes of batteries while, providing a reliable supply of power at the same time. Moreover, energy optimization algorithms with smarter energy management techniques have seen little attention and therefore need serious consideration from the researchers in IoT healthcare sector [123, 124]. Another area of research is the optimization of routing approaches that exploit the correlation among the captured data before it reaches its final destination, i.e., data aggregation techniques. These techniques eliminate redundancy that lowers the communication cost, conserves the energy and enhances the network lifetime.

8.5 Big data analytics

Another challenging aspect of IoT healthcare is big data analytics that deals with large-scale unstructured data. Recently, we have witnessed significant developments in hardware, software, and a diverse range of innovative IoT applications. Moreover, the growth forecast of IoT in the future is even more exaggerated with a large number of interconnected data sources and platforms with global infrastructure for

information and communication. As a result, huge amount of data is produced. This large volume of mostly redundant data is transmitted across the network for analysis and decision making. Transmitting such large volume of data across the network can adversely affect the network performance. This brings many challenging issues that need to be dealt with utmost care [125]. In this context, it would be interesting to see how to gain insight into this huge volume of data for better decision making and optimized operations using various ML and DL-enabled techniques [126]. It is imperative to design novel big data analytics tools and techniques that perform analysis and extract the required information. Innovative noise removal techniques are needed to enhance the data signal, improve the quality of aggregated data, and conserve the overall energy of the network [127]. More importantly, in healthcare applications, most of the devices perform real-time monitoring and analysis. It would be interesting to see novel ML techniques in the future that apply real-time analytics by monitoring current conditions and respond accordingly. Novel data aggregation techniques with outlier reduction should be devised with improve security, QoS and lowered computation complexity. Furthermore, data aggregation has a stronger relationship with the underlying topology of the network. The performance of these techniques are greatly affected by the underlying topologies [128–131]. In this regard, clustering tends to be more effective in static networks, where network configuration remains the same for longer time. However, they need to be studied in dynamic as well as heterogeneous environments [132–136]. Finding an optimal location for these devices should be further investigated so that IoT can cater for a wide range of emerging healthcare applications in the years ahead.

9 Limitations and future work

In this paper, we have presented a detailed survey of big data analytics in IoT health-care domain. We have thoroughly studied the literature and selected the most relevant and up to date surveys to find research gap. Furthermore, we have also provided a comprehensive and state-of-the-art literature on ML-based techniques for big data analytics in IoT smart health. A detailed discussion of their strengths and weakness was also provided. This provided an insight to the readers in this domain and enable them to start their research by selecting the topic of their choice from available pool of techniques. Various research issues and challenges were discussed that motivate the researchers to exploit them further. Moreover, various issues that raised due to the emerging and cross-domain architectures of IoT, i.e., Internet of Nano-Things (IoNT), and web of Things (WoT) were thoroughly discussed to make a universal IoT vision a reality, a vision that

successfully integrates this technology in almost all domains and that will hopefully flourish our daily lives in the years to come.

Authors' contributions This paper is equally contributed by each author as everyone wrote a section of it. Besides, there was collaborative efforts in brainstorming the idea of this paper, proofread and formatting of this paper.

Data availability Not applicable.

Compliance with ethical standards

Conflicts of interest/competing interests The authors declare that they have no conflict of interest.

Code availability Not applicable.

References

- Evtodjeva TE, Chernova DV, Ivanova NV, Wirth J (2020) The internet of things: possibilities of application in intelligent supply chain management. In: *Digital Transformation of the Economy: Challenges, Trends and New Opportunities*. Springer, Cham, pp 395–403
- Abdollahzadeh S, Navimipour NJ (2016) Deployment strategies in the wireless sensor network: A comprehensive review. *Computer Communications* 91:1–16
- Piccialli F, Jung JE (2017) Understanding customer experience diffusion on social networking services by big data analytics. *Mobile Networks and Applications* 22(4):605–612
- Joe S (2014) Qin. Process data analytics in the era of big data. *AICHE Journal* 60(9):3092–3100
- Baker SB, Xiang W, Atkinson I (2017) Internet of things for smart healthcare: Technologies, challenges, and opportunities. *IEEE Access* 5:26521–26544
- Latif S, Afzaal H, Zafar NA (2018) Intelligent traffic monitoring and guidance system for smart city. In: *International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, pp 1–6
- Babar M, Khan F, Iqbal W, Yahya A, Arif F, Tan Z, Chuma JM (2018) A secured data management scheme for smart societies in industrial internet of things environment. *IEEE Access* 6:43088–43099
- Pouryazdan M, Fiandrino C, Kantarci B, Soyata T, Kliazovich D, Bouvry P (2017) Intelligent gaming for mobile crowd-sensing participants to acquire trustworthy big data in the internet of things. *IEEE Access* 5:22209–22223
- Liu J, Shen H, Narman HS, Chung W, Lin Z (2018) A survey of mobile crowd sensing techniques: A critical component for the internet of things. *ACM Transactions on Cyber- Physical Systems* 2(3):1–26
- Lashkari B, Rezaeadeh J, Farahbakhsh R, Sandrasegaran K (2018) Crowdsourcing and sensing for indoor localization in IoT: A review. *IEEE Sensors Journal* 19(7):2408–2434
- Dehkordi SA, Farajzadeh K, Rezaeadeh J, Farahbakhsh R, Sandrasegaran K, Dehkordi MA (2020) A survey on data aggregation techniques in IoT sensor networks. *Wireless Networks* 26(2):1243–1263
- Rodríguez-Mazahua L, Rodríguez-Enríquez C-A (2016) José Luis Sánchez-Cervantes, Jair Cervantes, Jorge Luis García- Alcaraz, and Giner Alor-Hernández. A general perspective of big data: applications, tools, challenges and trends. *The Journal of Supercomputing* 72(8):3073–3113
- Hashem IAT, Yaqoob I (2015) Nor Badrul Anuar, Salima Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of “bi data” on cloud computing: Review and open research issues. *Information systems* 47:98–115
- Tsai C-W, Lai C-F, Chao H-C, Vasilakos A (2015) Big data analytics: a survey. *Journal of Big data* 2(1):21
- Athey S (2018) The impact of machine learning on economics. In: *The economics of artificial intelligence: An agenda*. University of Chicago Press, pp 507–547
- Mohammadi M, Al-Fuqaha A, Sorour S, Guizani M (2018) Deep learning for iot big data and streaming analytics: A survey. *IEEE Communications Surveys & Tutorials* 20(4):2923–2960
- Cui L, Yang S, Chen F, Ming Z, Lu N, Qin J (2018) A survey on application of machine learning for internet of things. *International Journal of Machine Learning and Cybernetics* 9(8):1399–1417
- Ge M, Bangui H, Buhnova B (2018) Big data for internet of things: a survey. *Future Generation Computer Systems* 87:601–614
- Mahdavinejad MS, Rezvan M, Barekatin M, Adibi P, Barnaghi P, Sheth AP (2018) Machine learning for internet of things data analysis: A survey. *Digital Communications and Networks* 4(3): 161–175
- Firouzi F, Rahmani AM, Mankodiya K, Badaroglu M, Merrett GV, Wong P, Farahani B (2018) Internet-of-Things and big data for smarter healthcare: From device to architecture, applications and analytics. *Future Generation Computer Systems* 78:583–586
- Pourghebleh B, Navimipour NJ (2017) Data aggregation mechanisms in the internet of things: A systematic review of the literature and recommendations for future research. *Journal of Network and Computer Applications* 97:23–34
- Dehkordi SA, Farajzadeh K, Rezaeadeh J, Farahbakhsh R, Sandrasegaran K, Dehkordi MA (2020) A survey on data aggregation techniques in IoT sensor networks. *Springer Wireless Networks* 26(2):1243–1263
- Olaković AČ, Hadžialić M (2018) Internet of things (IoT): A review of enabling technologies, challenges, and open research issues. *Computer Networks* 144:17–39
- Boubiche S, Boubiche DE, Bilami A, Toral-Cruz H (2018) Big data challenges and data aggregation strategies in wireless sensor networks. *IEEE Access* 6:20558–20571
- Ghate VV, Vijayakumar V (2018) Machine learning for data aggregation in wsn: A survey. *International Journal of Pure and Applied Mathematics* 118(24):1–12
- Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M (2015) Internet of things: A survey on enabling technologies, protocols, and applications. *IEEE communications surveys & tutorials* 17(4):2347–2376
- Lee I, Lee K (2015) The internet of things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons* 58(4):431–440
- Shirvanimoghaddam M, Dohler M, Johnson SJ (2017) Massive non-orthogonal multiple access for cellular IoT: Potentials and limitations. *IEEE Communications Magazine* 55(9):55–61
- Aggarwal M, Saxena N, Roy A (2019) Towards connected living: 5g enabled internet of things (IoT). *IETE Technical Review* 36(2): 190–202
- Ghose A, Pal A, Choudhury AD, Chattopadhyay T, Bhowmick PK, Chattopadhyay D (2014) “Internet of things (iot) application development.” U.S. Patent Application 14/286,068, filed November 27, 2014
- Yang C, Shen W, Wang X (2016) Applications of internet of things in manufacturing. *IEEE 20th International Conference on*

- Computer Supported Cooperative Work in Design (CSCWD). IEEE 670–675
32. Ansari S, Aslam T, Poncela J, Otero P, Ansari A (2020) Internet of Things-Based Healthcare Applications. In: IoT Architectures, Models, and Platforms for Smart City Applications. IGI Global, pp 1–28
 33. Shah S, Ververi A (2018) Evaluation of Internet of Things (IoT) and its Impacts on Global Supply Chains. In: 2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD). IEEE, pp 160–165
 34. Enginkaya E, Akgül AK (2018) The consumers'life simplifiers: Innovative developments and transformations. Business Studies 83
 35. Alkhayyat A, Thabit AA, Al-Mayali FA, Abbasi QH (2019) WBSN in IoT health-based application: toward delay and energy consumption minimization. Journal of Sensors, Hindawi
 36. Nguyen HH, Mirza F, Naeem MA, Nguyen M (2017) A review on IoT healthcare monitoring applications and a vision for transforming sensor data into real-time clinical feedback. In: 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE, pp 257–262
 37. Abdullah A, Ismael A, Rashid A, Abou-ElNour A, Tarique M (2015) Real time wireless health monitoring application using mobile devices. International Journal of Computer Networks & Communications (IJCNC) 7(3):13–30
 38. Yuehong YIN, Zeng Y, Chen X, Fan Y (2016) The internet of things in healthcare: An overview. Journal of Industrial Information Integration 1:3–13
 39. Ramathulasi T, Rajasekhara Babu M (2020) Comprehensive Survey of IoT Communication Technologies. In: *Emerging Research in Data Engineering Systems and Computer Communications*. Springer, Singapore, pp 303–311
 40. Al-Garadi MA, Mohamed A, Al-Ali A, Du X, Ali I, Guizani M (2020) A survey of machine and deep learning methods for internet of things (IoT) security. IEEE Communications Surveys & Tutorials
 41. Shah JL, Bhat HF (2020) CloudIoT for Smart Healthcare: Architecture, Issues, and Challenges. In: *Internet of Things Use Cases for the Healthcare Industry*. Springer, Cham, pp 87–126
 42. Aman W, Khan F (2019) Ontology-based Dynamic and Context-aware Security Assessment Automation for Critical Applications. In: the IEEE 8th Global Conference on Consumer Electronics (GCCE). IEEE, pp 644–647, Japan
 43. Jayaraman PP, Perera C, Georgakopoulos D, Dustdar S, Thakker D, Ranjan R (2017) Analytics-as-a-service in a multi-cloud environment through semantically-enabled hierarchical data processing. Software: Practice and Experience 47(8):1139–1156
 44. Pflanzner T, Kertész A (2016) A survey of iot cloud providers. Croatian Society for Information and Communication Technology Electronics 730–735
 45. Ray PP (2016) A survey of iot cloud platforms. Future Computing and Informatics Journal 1(1-2):35–46
 46. Khan F, Yahya A, Jan MA, Chuma J, Tan Z, Hussain K (2019) A Quality of Service-Aware Secured Communication Scheme for Internet of Things-Based Networks. *MDPI Sensors* 19(19):4321
 47. Bowya M, Karthikeyan V (2020) A Novel Secure IoT Based Optimizing Sensor Network for Automatic Medicine Composition Prescribe System. In: *Inventive Communication and Computational Technologies*. Springer, Singapore, pp 1109–1118
 48. Qian B, Jie S, Wen Z, Jha DN, Li Y, Guan Y, Puthal D et al (2020) Orchestrating the development lifecycle of machine learning-based iot applications: A taxonomy and survey. *ACM Computing Surveys (CSUR)* 53(4):1–47
 49. Babu GC, Shantharajah SP (2018) Survey on data analytics techniques in healthcare using IoT platform. International Journal of Reasoning-based Intelligent Systems 10(3-4):183–196
 50. Jagadeeswari V, Subramaniaswamy V, Logesh R, Vijayakumar VJHIS (2018) A study on medical internet of things and big data in personalized healthcare system. Health information science and systems 6(1):14
 51. Elhayatmy G, Dey N, Ashour AS (2018) Internet of Things based wireless body area network in healthcare. In: *Internet of things and big data analytics toward next-generation intelligence*. Springer, Cham, pp 3–20
 52. Wang Y, Kung LA, Byrd TA (2018) Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change 126: 3–13
 53. Vassakis K, Petrakis E, Kopanakis I (2018) Big data analytics: applications, prospects and challenges. In: *Mobile big data*. Springer, Cham, pp 3–20
 54. Huang X-L, Ma X, Hu F (2018) Machine learning and intelligent communications. Mobile Networks and Applications 23(1):68–70
 55. Khattak MI, Edwards RM, Shafi M, Ahmed S, Shaikh R, Khan F (2018) “Wet environmental conditions affecting narrow band on-body communication channel for WBANs.” 40, 297–312
 56. Kremer J, Stensbo-Smidt K, Gieseke F, Pedersen KS, Igel C (2017) Big universe, big data: machine learning and image analysis for astronomy. IEEE Intelligent Systems 32(2):16–22
 57. Choo J, Liu S (2018) Visual analytics for explainable deep learning. IEEE computer graphics and applications 38(4):84–92
 58. Young T, Hazarika D, Poria S, Cambria E (2018) Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine 13(3):55–75
 59. Nassif AB, Shahin I, Attili I, Azzeh M, Shaalan K (2019) Speech recognition using deep neural networks: A systematic review. IEEE Access 7:19143–19165
 60. Ayoubi S, Limam N, Salahuddin MA, Shahriar N, Boutaba R, Estrada-Solano F, Caicedo OM (2018) Machine learning for cognitive network management. IEEE Communications Magazine 56(1):158–165
 61. Sheikhnejad Y, Gonçalves D, Oliveira M, Martins N (2020) Can buildings be more intelligent than users?-the role of intelligent supervision concept integrated into building predictive control. Energy Reports 6:409–416
 62. Chen H, Chiang RHL, Storey VC (2012) Business intelligence and analytics: From big data to big impact. MIS quarterly, 1165–1188
 63. Karim A, Siddiq A, Safdar Z, Razzaq M, Gillani SA, Tahir H, Kiran S, Ahmed E, Imran M (2020) Big data management in participatory sensing: Issues, trends and future directions. Future Generation Computer Systems 107:942–955
 64. Diène B, Rodrigues JJPC, Diallo O, Ndoye ELHM, Korotaev VV (2020) Data management techniques for internet of things. Mechanical Systems and Signal Processing 138:106564
 65. Firouzi F, Farahani B, Weinberger M, DePace G, Aliee FS (2020) IoT Fundamentals: Definitions, Architectures, Challenges, and Promises. In: *Intelligent Internet of Things*. Springer, Cham, pp 3–50
 66. Farahani, Bahar, Farshad Firouzi, and Krishnendu Chakrabarty. “Healthcare iot.” In *Intelligent Internet of Things*, pp. 515-545. Springer, Cham, 2020.
 67. Malakis S, Psaros P, Kontogiannis T, Malaki C (2020) Classification of air tra_c control scenarios using decision trees: insights from a field study in terminal approach radar environment. Cognition, Technology & Work 22(1):159–179
 68. Lee S, Kim Y, Kahng H, Lee S-K, Chung S, Cheong T, Shin K, Park J, Kim SB (2020) Intelligent tra_c control for autonomous

- vehicle systems based on machine learning. *Expert Systems with Applications* 144:113074
69. Crane-Droesch A (2018) Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters* 13(11):114003
 70. Stupar S, Čar MB, Kurtović E, Vico G (2020) Theoretical and Practical Aspects of Internet of Things (IoT) Technology. In: *International Conference “New Technologies, Development and Applications”*. Springer, Cham, pp 422–431
 71. Alsharif MH, Kelechi AH, Yahya K, Chaudhry SA (2020) Machine learning algorithms for smart data analysis in internet of things environment: taxonomies and research trends. *Symmetry* 12(1):88
 72. Obermeyer Z, Emanuel EJ (2016) Predicting the Future—Big Data, Machine Learning, and Clinical Medicine. *New England Journal of Medicine* 375(13):1216–1219
 73. Ker J, Wang L, Rao J, Lim T (2017) Deep learning applications in medical image analysis. *Ieee Access* 6:9375–9389
 74. Dantu R, Dissanayake I, Nerur S (2020) Exploratory Analysis of Internet of Things (IoT) in Healthcare: A Topic Modelling & Co-citation Approaches. In: *Information Systems Management*. Taylor & Francis, pp 1–17
 75. Mehta N, Pandit A, Kulkarni M (2020) Elements of healthcare big data analytics. In: *Big Data Analytics in Healthcare*. Springer, p 23
 76. Balakrishna S, Thirumaran M, Solanki VK (2020) IoT sensor data integration in healthcare using semantics and machine learning approaches. In: *A Handbook of Internet of Things in Biomedical and Cyber Physical System*. Springer, Cham, pp 275–300
 77. Asthana S, Megahed A, Strong R (2017) A recommendation system for proactive health monitoring using IoT and wearable technologies. In: *2017 IEEE International Conference on AI & Mobile Services (AIMS)*. IEEE, pp 14–21
 78. Yao W, Yahya A, Khan F, Tan Z, Rehman AU, Chuma JM, Jan MA, Babar M (2019) A secured and efficient communication scheme for decentralized cognitive radio-based Internet of vehicles. *the IEEE Access* 7:160889–160900
 79. Moosavi SR, Gia TN, Rahmani A-M, Nigusie E, Virtanen S, Isoaho J, Tenhunen H (2015) SEA: A Secure and Efficient Authentication and Authorization Architecture for IoT-Based Healthcare Using Smart Gateways. *Procedia Computer Science* 52:452–459
 80. Subramaniyaswamy V, Manogaran G, Logesh R, Vijayakumar V, Chilamkurti N, Malathi D, Senthilselvan N (2019) An ontology-driven personalized food recommendation in IoT-based healthcare system. *The Journal of Supercomputing* 75(6):3184–3216
 81. Ali F, Islam SMR, Kwak D, Khan P, Ullah N, Yoo S-j, Kwak KS (2018) Type-2 fuzzy ontology-aided recommendation systems for iot-based healthcare. *Computer Communications* 119:138–155
 82. Khan F, Rehman AU, Zheng J, Jan MA, Alam M (2019) Mobile crowdsensing: A survey on privacy-preservation, task management, assignment models, and incentives mechanisms. *Future Generation Computer Systems* 100:456–472
 83. Borthakur D, Dubey H, Constant N, Mahler L, Mankodiya K (2017) Smart fog: Fog computing framework for unsupervised clustering analytics in wearable internet of things. In: *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, pp 472–476
 84. Verma P, Sood SK (2019) A comprehensive framework for student stress monitoring in fog-cloud IoT environment: m-health perspective. *Medical & biological engineering & computing* 57(1):231–244
 85. Rahmani AM, Gia TN, Negash B, Anzanpour A, Azimi I, Jiang M, Liljeberg P (2018) Exploiting smart e-health gateways at the edge of healthcare internet-of-things: A fog computing approach. *Future Generation Computer Systems* 78:641–658
 86. Kumar PM, Gandhi UD (2018) A novel three tier internet of things architecture with machine learning algorithm for early detection of heart diseases. *Computers & Electrical Engineering* 65: 222–235
 87. Gelogo YE, Oh J-W, Park JW, Kim H-K (2015) Internet of things (IoT) driven u-healthcare system architecture. *8th International Conference on Bio-Science and Bio-Technology (BSBT)*, 24–26
 88. Azimi I, Anzanpour A, Rahmani AM, Pahikkala T, Levorato M, Liljeberg P, Dutt N (2017) Hich: Hierarchical fog assisted computing architecture for healthcare IoT. *ACM Transactions on Embedded Computing Systems (TECS)* 16(5):1–20
 89. Kirtana RN, Lokeswari YV (2017) An IoT based remote HRV monitoring system for hypertensive patients. In: *2017 International Conference on Computer, Communication and Signal Processing (ICCCSP)*. IEEE, pp 1–6
 90. Ganapathy K, Sethukkarasi R, Yogesh P, Vijayakumar P, Kannan A (2014) An intelligent temporal pattern classification system using fuzzy temporal rules and particle swarm optimization. *Sadhana* 39(2):283–302
 91. Alsheikh MA, Lin S, Niyato D, Tan H-P (2016) Rate-distortion balanced data compression for wireless sensor networks. *IEEE Sensors Journal* 16(12):5072–5083
 92. Qiu T, Liu X, Lin F, Yu Z, Zheng K (2016) An efficient tree-based self-organizing protocol for internet of things. *IEEE Access* 4: 3535–3546
 93. Khan F, Rehman AU, Jan MA, Rahman IU (2019) Efficient resource allocation for real time traffic in cognitive radio internet of things. In: *In the International Conference on Internet of Things (iThings)*. IEEE, pp 1143–1147
 94. Fang R, Pouyanfar S, Yang Y, Chen S-C, Iyengar SS (2016) Computational health informatics in the big data age: a survey. *ACM Computing Surveys (CSUR)* 49(1):1–36
 95. Pinto S, Cabral J, Gomes T (2017) We-care: An IoT-based health care system for elderly people. In: *2017 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, pp 1378–1383
 96. Li J, Cai J, Khan F, Rehman AU, Balasubramaniam V, Sun J, Venu P (2020) A Secured Framework for SDN-Based Edge Computing in IoT-Enabled Healthcare System. *IEEE Access* 8: 135479–135490
 97. Prajapati B, Parikh S, Patel J (2017) An Intelligent Real Time IoT Based System (IRTBS) for Monitoring ICU Patient. In: *International Conference on Information and Communication Technology for Intelligent Systems*. Springer, Cham, pp 390–396
 98. Kim S-H, Chung K (2015) Emergency situation monitoring service using context motion tracking of chronic disease patients. *Cluster Computing*, Springer 18(2):747–759
 99. Jan MA, Zhang W, Usman M, Tan Z, Khan F, Luo E (2019) SmartEdge: An end-to-end encryption framework for an edge-enabled smart city application. *Journal of Network and Computer Applications* 137:1–10
 100. Vippalapalli V, Ananthula S (2016) Internet of things (IoT) based smart health care system. In: *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*. IEEE, pp 1229–1233
 101. Khan F, Jan MA, Rehman A u, Mastorakis S, Alazab M, Watters P (2020) “A Secured and Intelligent Communication Scheme for IIoT-enabled Pervasive Edge Computing”, in *IEEE Transaction on Industrial Informatics*. Early Access
 102. Khan F, Rehman A u, Usman M, Tan Z, Puthal D (2018) Performance of cognitive radio sensor networks using hybrid automatic repeat ReQuest: Stop-and-wait. *Mobile Networks and Applications* 23(3):479–488
 103. Gope P, Hwang T (2015) Bsn-care: A secure IoT-based modern healthcare system using body sensor network. *IEEE sensors journal* 16(5):1368–1376

104. Essa YM, Attiya G, El-Sayed A, ElMahalawy A (2018) Data processing platforms for electronic health records. *Health and Technology* 8(4):271–280
105. Khan IH, Khan MI, Khan S (2020) Challenges of IoT Implementation in Smart City Development. In: *Smart Cities—Opportunities and Challenges*. Springer, Singapore, pp 475–486
106. Ishtiaq M, Rehman AU, Khan F, Salam A (2019) Performance Investigation of SR-HARQ transmission scheme in realistic Cognitive Radio System. In: *the IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, pp 0258–0263
107. Hussain F, Hassan SA, Hussain R, Hossain E (2020) Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges. *IEEE Communications Surveys & Tutorials* 22(2):1251–1275
108. Naha RK, Garg S, Chan A, Battula SK (2020) Deadline-based dynamic resource allocation and provisioning algorithms in fog-cloud environment. *Future Generation Computer Systems* 104: 131–141
109. Zhou J, Cao Z, Dong X, Vasilakos AV (2017) Security and privacy for cloud-based IoT: Challenges. *IEEE Communications Magazine* 55(1):26–33
110. Ali SA, Ansari M, Alam M (2020) Resource Management Techniques for Cloud-Based IoT Environment. In: *Internet of Things (IoT)*. Springer, Cham, pp 63–87
111. Marr B (2018) Why the internet of medical things (IoMT) will start to transform healthcare
112. Kaur H, Atif M, Chauhan R (2020) An Internet of Healthcare Things (IoHT)-Based Healthcare Monitoring System. In: *Advances in Intelligent Computing and Communication*. Springer, Singapore, pp 475–482
113. Almolhis N, Alashjaee AM, Duraibi S, Alqahtani F, Moussa AN (2020) The Security Issues in IoT-Cloud: A Review. In: *2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, pp 191–196
114. Bansal S, Kumar D (2020) IoT Ecosystem: A Survey on Devices, Gateways, Operating Systems, Middleware and Communication. *International Journal of Wireless Information Networks*:1–25
115. Sharma D, Tripathi RC (2020) Performance of internet of things based healthcare secure services and its importance: Issue and challenges. Technical report, EasyChair
116. Jan MA, Khan F, Alam M, Usman M (2019) A payload-based mutual authentication scheme for Internet of Things. *Future Generation Computer Systems* 92:1028–1039
117. Bhattacharjya A, Zhong X, Wang J, Li X (2020) Present Scenarios of IoT Projects with Security Aspects Focused. In: *Digital Twin Technologies and Smart Cities*. Springer, Cham, pp 95–122
118. Flynn T, Grispos G, Glisson W, Mahoney W (2020) “Knock! Knock! Who is there? Investigating data leakage from a medical internet of things hijacking attack.” In Proceedings of the 53rd Hawaii International Conference on System Sciences
119. Williams PAH, McCauley V (2016) Always connected: The security challenges of the healthcare Internet of Things. In: *2016 IEEE 3rd World Forum on Internet of Things (WF-IoT)*. IEEE, pp 30–35
120. Khan F (2014) Fairness and throughput improvement in multihop wireless ad hoc networks. In: *the IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE)*. IEEE, pp 1–6
121. Qadri YA, Nauman A, Zikria YB, Vasilakos AV, Kim SW (2020) The Future of Healthcare Internet of Things: A Survey of Emerging Technologies. *IEEE Communications Surveys & Tutorials* 22(2):1121–1167
122. Park J, Bhat G, Geyik CS, Ogras UY, Lee HG (2020) Energy per operation optimization for energy-harvesting wearable IoT devices, Multidisciplinary Digital Publishing Institute. *Sensors* 20(3):764
123. Selvaraj S, Sundaravaradhan S (2020) Challenges and opportunities in IoT healthcare systems: a systematic review. *SN Applied Sciences* 2(1):139
124. Mittal M, Tanwar S, Agarwal B, Goyal LM (eds) (2019) *Energy Conservation for IoT Devices: Concepts, Paradigms and Solutions*, vol 206. Springer
125. Yang K, Shi Y, Zhou Y, Yang Z, Fu L, Chen W (2020) Federated machine learning for intelligent IoT via reconfigurable intelligent surface. arXiv preprint arXiv:2004.05843
126. Gill SS, Buyya R (2019) Bio-inspired algorithms for big data analytics: a survey, taxonomy, and open challenges. In: *Big Data Analytics for Intelligent Healthcare Management*. Academic Press, pp 1–17
127. Wan R, Xiong N, Hu Q, Wang H, Shang J (2019) Similarity-aware data aggregation using fuzzy c-means approach for wireless sensor networks. *EURASIP Journal on Wireless Communications and Networking* 2019(1):59
128. Qi G, Wang H, Haner M, Weng C, Chen S, Zhu Z (2019) Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation. *CAAI Transactions on Intelligence Technology* 4(2):80–91
129. Li X, Zhao M, Liu Y, Li L, Ding Z, Nallanathan A (2020) “Secrecy Analysis of Ambient Backscatter NOMA Systems under I/Q Imbalance,” *IEEE Transactions on Vehicular Technology*, accepted for publication, Jun. 2020
130. Wiens T (2019) Engine speed reduction for hydraulic machinery using predictive algorithms. *International Journal of Hydromechanics* 2(1):16–31
131. Li X, Wang Q, Liu Y, Tsiftsis TA, Ding Z, Nallanathan A (2020) UAV-Aided Multi-Way NOMA Networks with Residual Hardware Impairments. In: *IEEE Wireless Communications Letters*
132. Shokri M, Tavakoli K (2019) A review on the artificial neural network approach to analysis and prediction of seismic damage in infrastructure. *International Journal of Hydromechanics* 2(4): 178–196
133. Xue X, Lu J, Chen J (2019) Using NSGA-III for optimising biomedical ontology alignment. *CAAI Transactions on Intelligence Technology* 4(3):135–141
134. Ma J (2019) Numerical modelling of underwater structural impact damage problems based on the material point method. *International Journal of Hydromechanics* 2(4):99–110
135. Khan F, Rehman A u, Jan MA (2020) A secured and reliable communication scheme in cognitive hybrid ARQ-aided smart city. *Computers & Electrical Engineering* 81:106502
136. Yu T, Wang J, Wu L, Xu Y (2019) Three-stage network for age estimation. *CAAI Transactions on Intelligence Technology* 4(2): 122–126

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Automated Power Depiction System Using Iot Platform

Ananya Rajesh¹, Geetika Gopinath¹, Tejas C J¹, Amal Prasannan¹, Dr. Parvathy M²

¹ B. Tech students,
SCMS School of Engineering and Technology

² Associate Professor, Department of ECE
SCMS School of Engineering and Technology
meetparu.parvathy@gmail.com

Abstract—

Precise and well-timed expertise of electricity consumption is an imperative requirement for implementing any power saving strategy. The primary function of energy management systems is to measure the electricity utilization in the homes. Monitoring energy usage in homes indirectly offers knowledge about consumer preferences, types of activities carried out by residents, and forecasting potential energy consumption, energy saving methods and lots of other proposes. Electricity consumers are oblivious to their daily unit consumption of electricity. They know about their power consumption only when they receive the electricity bill on a cycle of 2 months to 3 Months which give consumers no possible option to control the usage of electricity as there are only limited number of methods for the consumers to know about how much units of electricity they have consumed. So, we are going to come up with a solution for the same so as to develop a method which would easily replicate the Electricity Meter reading and would also provide us with the power consumption of each individual appliance on to your phone. This helps us to compare the energy consumption and also helpful for the users to understand if their electricity usage is more or less on a need basis and can compensate or reduce electricity usage in turn reducing electricity bill.)

Keywords—*Internet of things (IOT), Android smart phone, NodeMCU.*

I. INTRODUCTION

The advancement of the IOT (Internet of Things) have become an additional standard as a result of its contribution to economical solutions for several real time applications. It conjointly revolves the association between M2M that are embedded with electronics, software, sensors etc., which assist users in observing and controlling devices with efficiency. In an IoT system, objects and living beings are provided with unique

identifiers with the ability to transfer the data. Nowadays, IoT is being applied in several areas like gas, water, electricity etc. to make our lives automated. Electricity is a crucial invention and its demand is also increasing at a relentless rate and is being utilized for numerous purposes such as agriculture, industries, hospitals etc. So, it is becoming additionally sophisticated to handle the electricity requirements and maintenance. Clearly there ought to be a necessity for measuring the consumed electricity. Therefore, it is indispensable to execute a technique of taking energy meter readings automatically, which can realize the power consumption management to the customers to be adjustable and manageable to save the electrical energy. Also, to take the reading of the meter, a human operator has to go turn by turn to every resident & commercial building, hence this will increase the work and potency. Therefore, the operating hours also increases to realize the complete area data reading. So, to achieve an efficient energy meter reading, reduce billing error, and operation cost, we need a system that can read the meter reading automatically at each time interval.

The right way to cope up with the problem is to understand the situation thoroughly. At first, we need to realize the quantity of energy that we tend to consume so that we begin to limit it. So, as a solution to it, we intended to construct an application using IOT, which would display the overall power consumed by the household together with the power consumed by individual appliances and conjointly enables the user to set a threshold which upon surpassing will notify the user with an alert message thus helping the user to monitor the power consumption in their house. Thus, this device uses IoT to automate the purpose of measuring consumption of power in homes, allowing for web access and digital technology.

II. LITERATURE SURVEY

In recent years, numerous papers have proposed the design of energy meter monitoring system. In [2], the author proposed a Prepaid Energy Meter based on IOT. The circuit mainly composed of an ADE7758 meter circuit, microcontroller i.e.; Atmega328p, and finally Wi-Fi module. The meter keeps the track of the number of units consumed and sends the information, as well as the cost, through the internet. If the usage of the user is approaching the set point, it will alert the user a warning. If consumption exceeds the set point, the device will switch off the battery.

In [3], the author presented a paper based on the consumption of current by using an Infrared sensor device. The Infrared transmitter is mounted in the EB meter's revolving assembly. The receiver photodiode is positioned in a specific location to determine the number of rotations thereby getting the current consumption. rotations. By determining the rotations, the current consumption can be obtained. After obtaining the current consumption, ARM processor will limit the unit given for a particular user. If the device is reduced to its bare minimum, it will intimate the consumer via alarm and LCD unit. If a person decides to update more units for him, he must contact the EB section and send a request. The required cost will be dispatched to the ARM controller through a GSM modem. Then the unit is incremented by the processor in the memory.

The author presented a remote device monitoring system on a smart phone GUI which is built on an Android Smartphone in [8]. A client logs into the app and gently presses the buttons in order to send the message instructions from the GUI to the home information center with the aid of the GSM network. The ATmega processor acknowledges the detailed command and uses wireless radio frequency to control the home appliance switches, allowing for eventual remote control of the appliances. This lecture emphasizes on the configuration of an Android terminal, the contact between the PIC and the GSM module, the implementation of the wireless module device's driver, and the difficulty of providing the required low- DC voltage for the MCU and wireless module using just one live cable. Consumers can control appliances at any time, allowing our homes to become increasingly smart and automated.

In [12], the author proposed a mobile internet monitoring of a domestic electrical power meter using wireless communication. This is accomplished with the WSN platform's embedded active RFID tag module. The electrical power meters are monitored and identified using active RFID using the ZigBee protocol. The modules which are embedded in the power meters function as wireless sensors, thereby monitoring the power meter's electricity consumption value and transmitting the data value to the portable reader through an RF signal. The relayed signal is used to speed up certain everyday tasks, save time, and minimize the cost and error in information systems that humans can create.

A Smart Home System (SHS) is a home with a communication network that links services and electrical appliances so they can be monitored managed and accessed remotely. SHS uses a variety of methods to accomplish a variety of goals, ranging from improving everyday comfort to allowing elderly and handicapped people to live more independently. In [10], the author presented a report based on the key 4 fields for SHS: domestic automation and faraway control, temperature and humidity monitoring, fault monitoring and management, and eventually monitoring of health. The machine is constructed round the Microcontroller by using the MIKRO C program, several active and passive sensors, and wi-fi web services, all of which are used in a variety of monitoring and manage processes.

III. PROPOSED METHODOLOGY

In the proposed technique, we propose a system that tracks the power consumption of domestic devices, allowing consumers to better control their energy consumption by tracking their use over time.

Figure: 3.1 Block Diagram of Iot based power depiction system.

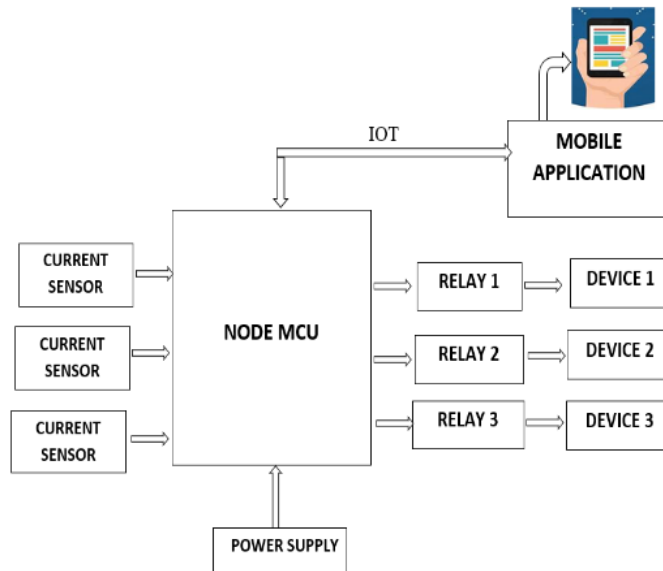


Figure : 3.1 Block Diagram of Iot based power depiction system.

The main component is the Node MCU to which the devices are interfaced. The Node MCU is a web-based software and hardware environment based on the ESP8266, a low-cost System on a Chip (SoC). The current sensors are attached to this module to determine the amount of power used by individual appliances. We're using an ACS 712 current sensor, which can work in both DC and AC and provides separation between the load (AC/DC load) and also the measuring unit, which is the microcontroller. It is a sensor which runs on 5 volts and outputs an analogue voltage and is proportional to the current measured. This provides us with the individual power consumed by the appliances. Both the values (ie; total power and power consumed by individual appliances) are then sent to the user through this module to a mobile application using IOT from where the user can easily know about the power consumption pattern in his/her household and also about the appliance that is consuming more power just by comparing the total power and individual power of appliances displayed in the app. Hence this would help them in adopting certain energy conservation methods of their choice thus reducing the usage and electricity bill. Also, relays are connected to the devices as output from the Node MCU as a provision for the user to switch on or off these appliances, when necessary, through this app.

IV. CIRCUIT DIAGRAM

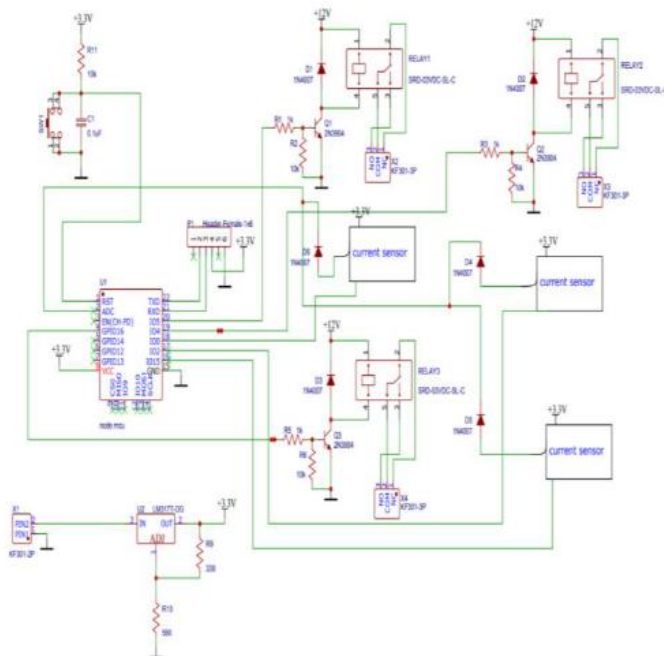


Figure 4.1 Circuit diagram of IOT based power depiction system.

4.1. WORKING

The central portion of the circuit is the Node Mcu, an open-source microcontroller to which the devices are interfaced. Note that the MCU isn't solely a microcontroller however additionally a microcontroller with a Wi-Fi indication where the Wi-Fi is already constitutional. Node MCU gives access to the General-Purpose Input/Output pins. It has 17 GPIO pins (0- 16), out of that solely 11 of them can be used because the other pins (GPIO 6 -11) are used for connecting the non-volatile memory chip. ESP8266 has an incorporated 10-bit ADC which has only one ADC channel to where the sensing element is connected. Although the Node MCU has just one ADC pin it does not limit the quantity of analog detector to 1 per module. Multiple sensors can be connected to the module through the method of multiplexing.

During this arrangement, only a sensor at a time will have an entire circuit, or only one sensor will be able to operate at a time (that is, switch on a sensor), take the values needed from the sensor, shut down that sensor then move to consecutive sensor. For this we tend to connect diodes to the sensors as they direct the current towards one direction as a result, the sensor circuits are isolated from being read. When the reading is taken, the General-purpose pin is made HIGH by sending 3V to the sensor, thus finishing the circuit. Other pins are set to LOW, so they don't send any voltage, resulting in a ground to-ground link with no current. The ADC pin is read after the voltage is sent, and the value is written, which can also be stored in a variable.

After the reading is taken the sensor is switched off and we move to the next sensor for taking the reading. Except when taking readings from analogue sensors, all GPIOs being used with them should be made to LOW. The signals given to the node MCU uses three types of output because of the presence of three relays for the purpose of switching On or Off the devices. These relays are connected to the GPIO pins of the MCU. The resistor going to the input of the relay is the current limiting resistor, a pull-down resistor is connected to the ground, and finally a transistor.

The transistor is set to switch from 3 volts to 12 volts. Next, we can sense the current by passing the phase line of ac with it. So that we can know how much energy that device has used. Then there is a 3.3 voltage regulator, which is connected to the power supply. And finally, there is a reset switch at the top. It is used to reset the microcontroller.

4.2 FLOW CHART

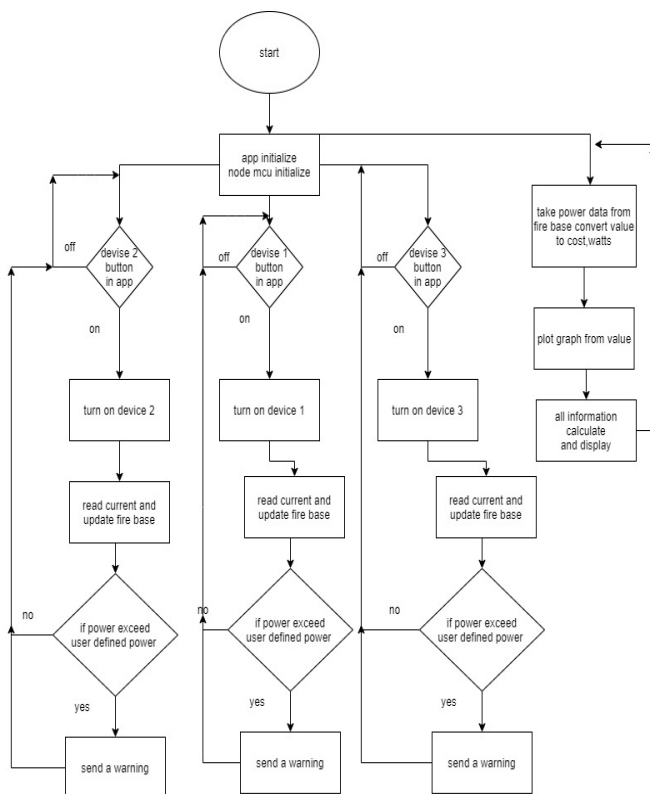
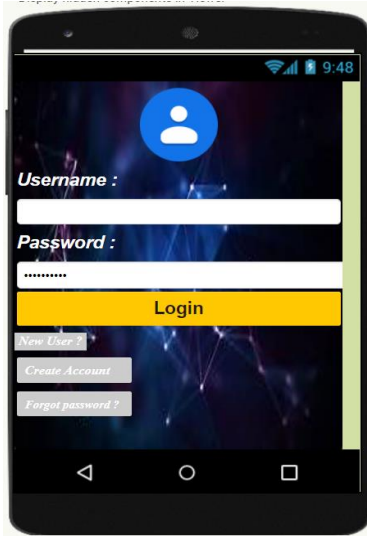


Figure 4.2 Flow chart of IOT based power depiction system .

V. MOBILE APP DESIGN .



App Inventor Blocks Editor Screen 1

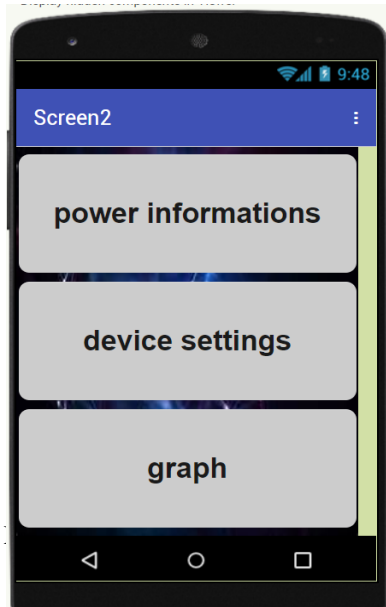


Fig. 5.2: Screen 2.

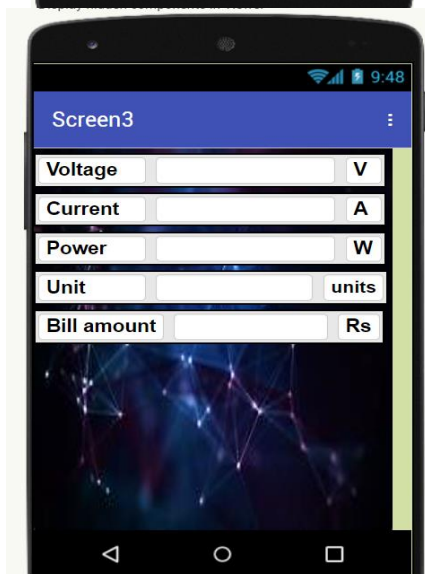


Figure. 5.3: MIT App Inventor Blocks Editor Screen 3 – User can monitor the voltage, current, power consumption unit , and also the bill amount in the text box provided.

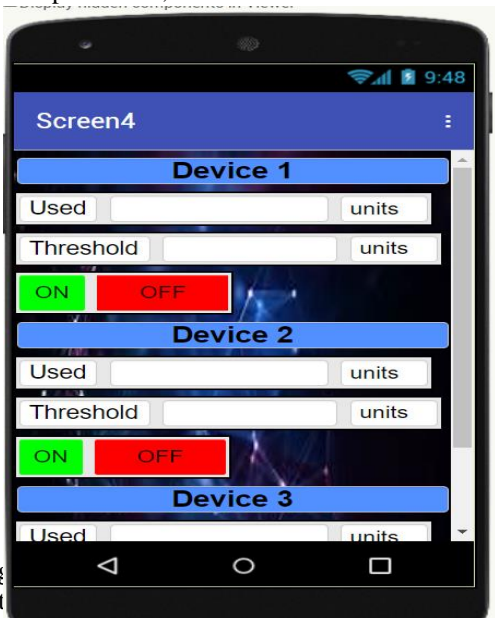


Fig 5.3 .
In the app, when the power consumption exceeds the threshold value, the user will receive an alert message.

VI. RESULTS AND DISCUSSIONS

Our Project IOT based power depiction system on Smartphone is being designed such that whenever the usage of energy exceeds the threshold value which is set by user, it will give an alert message. The monthly billing status through SMS is also being send onto the user's mobile. This implementation illustrates the definition and implementation of a modern system with low running costs, increased data protection, and reduced manpower requirements. As a result, it not only fixes the issue of traditional meter reading, but it also adds new functionality to cell phones.

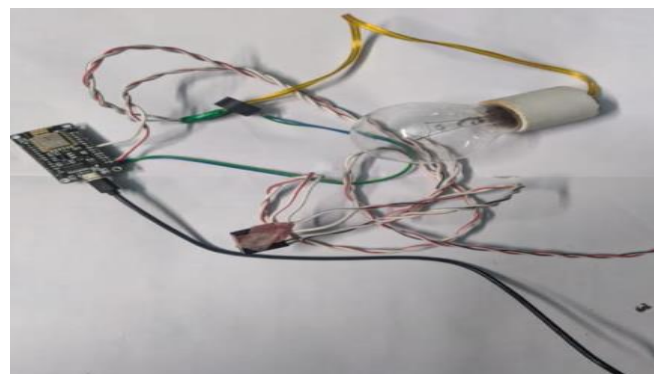
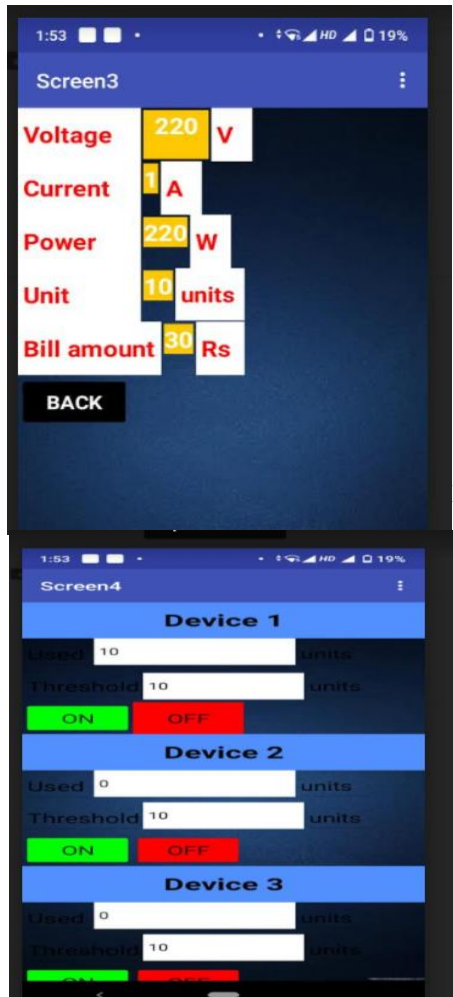


Figure 6.1 Hardware Implementation.



power information's of the devices used.

Figure 6.3 Illustration for setting threshold of devices .



Figure 6.4 Image of alert message popup when threshold exceeds.

VII. CONCLUSION & FUTURE SCOPE

In this paper, the design of IOT Based Power Consumption Depiction System is proposed. The application which we are developing helps us to give the information regarding the individual energy consumption and also the total energy consumed by all the devices which would be easier for the user to identify which device consumes more power and there is a provision to set a threshold for each appliance and hence it will notify the user when threshold exceeds and can switch off the device when needed. Hence, it will reduce the wastage of energy and bring awareness among all.

ACKNOWLEDGMENT

We feel really grateful and would like to take this as an opportunity to thank our project guide Dr. Parvathy M, Associate Professor, Department of Electronics and Communication for giving us immense support and guidance we required. Her indispensable encouragement and valuable suggestions were very helpful during the process.

REFERENCES

- [1] Umakanta Nanda, D Nayak, S K Swain, S K Das (2020) ,“Arduino GSM Based Power Theft Detection and Energy Metering System” 2020, IEEE International Conference, pp: 14- 30.
- [2] Vinayalk Rangrao Patil, Manoj D Patil, Anupam Tanaji Khude (2020) , “IOT Based Prepaid Energy meter”, 5 th International Conference on Devices, Circuits and Systems (ICDCS), March 2020.
- [3] Suresh Kavuri, Kadiri Vijaya Lakshmi (2019) “GSM Based Automatic Energy Meter Reading System with Instant Billing”, January 2019, IEEE Conference, pp: 65-72.
- [4] Dodwinsla. megaha, "System Design and Simulation results of Single Phase Intelligent Prepaid", Innovative Systems Design and Engineering, vol. 4, no. 1, pp. 1-10, November 2018.
- [5] Ad. Ahmed Naik, A. R. U. Haque, "Implementation of Design and implementation of using usb gprs/edge modem", Electronics Computer Technology (SOC) 2012 3rd International Conference on, volume. 7, pp. 220-225, 2018.
- [6] W.O. Veerababu, W. rani, W. samuel Abednego,"Electrical Power Measurement Using Arduino Mega Microcontroller and LabVIEW” 2013", 3rd International Conference on Instrumentation Communications(ICdJICI-BMEdk) 226 Bandung, November 7-8, 2016
- [7] Jain kyan and T.S.Remmy “Algorithm of bill generation system automatically”, IEEE transaction volume. 7. Pp.no 70-78, 2015..
- [8] Sachin Kishor Khadke (2015) ,” Home Appliances Control System Based on Android Smartphone”, IOSR Journal of Electronics and Communication Engineering,9(3):5-6, July 2015.
- [9] Alsibai, M.H.; Siang, H.M. (2015) “A smart driver monitoring system using android application and embedded system” In Proceedings of the 5th IEEEInternational Conference on Control Systems, Computing and Engineering (ICCSCE 2015), Penang, Malaysia, 27– 29 November 2015; pp. 242–247.
- [10] Mohamed Abd El-Latif Mowad (2014), “Smart Home Automated Control System Using Android Application and Microcontroller”, International Journal of Scientific & Engineering Research,5(5):5, May 2014
- [11] Ashna.k, Sudhish N George (2014), “GSM Based Automatic Energy Meter Reading System with Instant Billing”, 2014, IEEE Conference, pp: 65- 72
- [12] Wasana Boonsong , Widad Ismail (2014), “Wireless Monitoring of Household Electrical Power Meter Using Embedded RFID with Wireless Sensor Network Platform”, Hindawi Publishing Corporation International Journal of Distributed Sensor Networks Volume 2014, Article ID 876914
- [13] Rajeev Piyare (2013) , “Internet of Things: Ubiquitous Home Control and Monitoring System using Android based Smart Phone”, International Journal of Internet of Things, 2(1): 5-11, 2013..
- [14] Wibhada Naruephiphat , Sadit Satiempaisarn, Ridnarong PromYa (2012), “Applying Wireless Sensor Network for Power Consumption Monitoring”, 2012,IEEE, pp: 1- 4.

- [15] Md. Kamal Hossain, Md. Mortuza Ali, Md. Rafiqul Islam Sheikh, (2011) “Microcontroller Based Single Phase Digital Prepaid Energy Meter for Improved Metering and Billing System” ,International Journal of Power Electronics and Drive Systems, October 2011.