



SCMS SCHOOL OF ENGINEERING & TECHNOLOGY

PUBLICATION DETAILS 2021

SI No:	Name	First Author	Second Author	Third Author	Fourth Author	INDEXING
1	Dr.Sunil Jacob		PEC2101			SCI
2	Vinoj PG	PEC2101				SCI
3	Anandhi V	PEC2104				SCI
4	Dr.Varun Menon	PCS2109(5THAU)		PEC2101		SCI
5				PCS2102	PCS2101,PCS2103, PCS2110,PCS2114(5TH)	SCI
6		PCS2111(6TH AU) PCS2118(6TH)	PCS2115,PCS2117	PCS2116,PEC2104	PCS2112(5TH),PCS2113, PCS2119(5TH),PCS2120 (5TH)	SCI
7	Josna Philomina	PCS2104(5TH AU)				SCI
8	Susmi Jacob	PCS2105				UGC CARE
9	Litty Koshy	PCS2106(5TH AU)				UGC CARE
10	Asha S	PCS2107				SCI
11	Dr.Vidya Chandran	PME2107(SCI)			PME2106	
					PME2105(SCI)	SCI
12	Nikhil Asok N.			PME2110	PME2106:6TH AUTHOR	
13	Jenson Joseph	PME2101				SCOPUS
14	Dr Raghav G R	PME2102(SC0) , PME2106	PME2103(SCO)			SCOPUS
15	Dr.Ratish Menon	PCE2105				SCI
16	Dr. Sam Joshy				PME2109	SCOPUS
17	Sujith R.		PME2102			SCOPUS
18	Sanju A. C.				PME2109:5TH AUTHOR	SCOPUS
19	T M Anup Kumar			PME2109		SCOPUS
20	Suraj R	PME2104	PME2106			SCOPUS
21	Dr. Gibin George	PME2108 (SC0) PME2109(SCO) PME2110(SCI)				
22	Dr. Manikandan H		PME2109			SCOPUS
23	AKhila M	PCE2101(SCI),PCE2103 (SCO)				SCOPUS
24	Anjana Susan John		PCE2102			SCOPUS

25	Sruthy M R			PCE2103		SCOPUS
26	DrJayanand B		PEE 2105			SCOPUS
27	Ms. Sanju Sreedharan	PCE21104				SCI
28	Divya M S	PBSH2101,PBSH2104				UGC CARE
29	Dr.Sreelekha Menon	PBSH2103				UGC CARE
Total Publication for the calender year 2021						39




 DR. PRAVEEN K. C. L.
 PRINCIPAL
 COLLEGE OF ENGINEERING & TECHNOLOGY



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, Vidyannagar 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).

<https://doi.org/10.1016/j.matpr.2021.04.148>

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.



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, eyeful 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.



IMPACT ANALYSIS OF THE COVID19 ON THE ATMOSPHERIC AIR QUALITY AND ELECTRICITY CONSUMPTION PER DAY IN INDIA.

Dr. Sreelekha B.

Associate Professor, SCMS School of Engineering and Technology, Angamaly, Kerala, India.

Nandini Menon

Student, Vellore Institute of Technology, Vellore, Tamil Nadu, India.

ABSTRACT The corona Virus pandemic is an unexpected part of our life. Following the lockdown procedures implemented by the Indian government, the atmospheric air quality and the power sector of India noticed decadent changes. However, the relaxation in lockdown brought situations almost back to normal. The paper focuses on analyzing these changes to review the current scenario according to the published studies. It focuses on electricity consumption, night light intensity, and variations in many pollutants like NO2 and PM. For ease, the main cities chosen for the research are the metropolitan cities of India. The data are from US-EPA, POSOCO, VIIR satellites and so on. The paper helps us understand if this variation is a boon or bairn for the Indian economy.

KEYWORDS : COVID-19, Electricity Consumption, Air Quality, Atmosphere, Energy Consumptions, Night Light Intensity.

INTRODUCTION

On March 25th, Prime Minister of India Narendra Modi announced a nationwide lockdown, accounting for the safety precautions required to fight against the Coronavirus (SARS-CoV-2). Ever since then, there has been an exponential surge in the number of infections reported per day. From April to September 2020, as the number of cases kept increasing, more people restricted themselves from going out to keep themselves safe and well. As of September 7th, 2020, India recorded 41.13 Lakhs of confirmed infections, making her the second most affected country in the world after the USA. On July 15th, 2020, the phase-1 clinical trials for the first indigenous Coronavirus Vaccine, Covaxin, developed by Hyderabad-based pharmaceutical company Bharat Biotech and the National Institute of Virology and Indian Council of Medical Research, starts across the country.

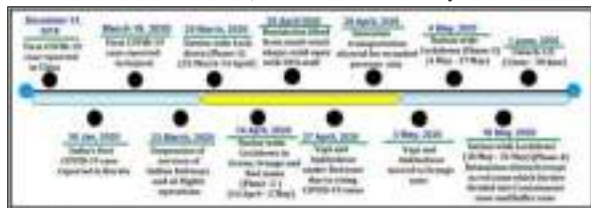


Fig: 1 (6)-Corona-Virus Pandemic India-Timeline.

Simultaneously, many researchers have been working on the impact of lockdown on atmospheric CO2 levels and electricity consumption per day. While e-collaborations positively impact the present climate and environment, it is definite that reduction will be short-lasting, attributing it to the close-down of transport, construction works, and industrial activities.

The Corona-Virus pandemic has an unparallel effect on our everyday life, which will continue until a minimum of the next three years. While research, vaccination, and protocol documentation procedures are currently ongoing, the impact of lockdown on the environment is also a widely inspected topic. A calculated set of restrictions imposed on the economy to reduce the spread of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS- CoV-2) has an overall positive effect on the environment. The beneficial impact includes reduced particulate matter levels in the atmosphere, decreased carbon dioxide (CO2) levels, reduced kerosene or related fuel use, and increased awareness about the importance of the 5Rs. A study conducted by analyzing the data and images collected from the Sentinel satellite-ESA revealed a 45% decline in atmospheric NO2 levels in India(1). Air pollution takes the lives of almost 1.7 million per year. Besides global warming, Air pollution has fueled many recent disasters, namely the Kerala floods of 2018-2019, the Assam floods 2019-2020, the Amazon Forest Fire, the Sydney forest fires, Australia floods 2021, and it keeps going on. Air pollution symptoms include aggravated respiratory diseases like asthma and bronchitis, dry throat, wheezing, nausea, and headache. In India, the Northern parts are the most polluted areas, especially Delhi, mainly due to emissions from Vehicles, brick kilns, coal-based thermal power plants, and crop remnants(2). The total energy consumption (ameasure for the amount of electricity consumed) and

the lights per area (a measure of the intensity of light in the area) are the proxy indicators for consumption level measurements (3). Ever since the nationwide lockdown, there is a decline in daily energy consumption. The official power consumption data captured by POSOCO (Power System Operation Corporation) has recorded a 26% decline since the nationwide lockdown. The table below represents a section of the data released by POSOCO (4).

Date	Energy Consumption (BWh)					All India
	Northern Region	Western Region	Southern Region	Eastern Region	North Eastern Region	
18-Mar-20	817	1187	1108	983	47	3586
22-Mar-20	734 (-12%)	975 (-18%)	975 (-12%)	815 (-18%)	36 (-23%)	3030 (-15%)
23-Mar-20	733 (-12%)	996 (-16%)	1030 (-7%)	825 (-17%)	89 (-9%)	3128 (-12%)
24-Mar-20	805 (-18%)	944 (-20%)	985 (-14%)	834 (-18%)	89 (-7%)	2955 (-17%)
25-Mar-20	965 (-20%)	849 (-20%)	911 (-21%)	820 (-18%)	36 (-23%)	2777 (-23%)
26-Mar-20	828 (-26%)	775 (-34%)	891 (-21%)	827 (-19%)	45 (-16%)	2387 (-28%)

Figures in parentheses indicate percentage change from 18th March 2020

Table 1.1 (4)-POSOCO Data.

The data collected from the different sources suggests a pollution reduction, but the trend is inconsistent. Analyzing these trends and converting them into more consistent data is of the highest priority right now. COVID19 is posing as an opportunity to do the same. The paper has put together a review of the recent researches analyzing this impact.

Atmospheric Air Quality

Researching and identifying air quality index measures of a particular area requires humongous data on the primary sources in the area, optimal pollution levels, meteorology, demographics, geography, and computational capacity (7). PM-2.5 and PM-10 are the most dangerous types of pollutants released into the atmosphere. PM-2.5 is a type of particulate matter whose size is less than 2.5 micro-meter, which is small enough to enter our lungs and bloodstream. Numerous studies link PM-2.5 to various health risks compared to other pollutants.



Fig: 2 (7)-Composition of Particulate Matter.

Table-2 (8)- Percentage Reduction In Emissions During The Lockdown Period 25/03/2020 To 15/04/2020 Compared To 24/03 To 15/04, 2019.

Pollutant	Date	Bengaluru	Chennai	Hyderabad	Mumbai
NO ₂	24, Mar–15 Apr,2019	67.1	47.6	39.7	6.8
	BaU 25, Mar–15 Apr,2020	57.7	28.7	30.1	27.9
PM _{2.5}	24, Mar–15 Apr,2019	45.1	45.7	18.9	42.0
	BaU 25, Mar–15 Apr,2020	45.2	28.7	12.3	39.9
SO ₂	24, Mar–15 Apr,2019	1.7	8.2	9.4	3.4
	BaU 25, Mar–15 Apr,2020	10.7	33.0	-17.2	45.2
CO	24, Mar–15 Apr,2019	23.2	39.6	24.6	-55.1
	BaU 25, Mar–15 Apr,2020	27.6	13.4	9.8	37.1

Change In Pm-2.5 And Pm-10

The inevitable parameter in determining the air quality of a particular area concerning PM is identifying the optimal ambient concentration levels of the pollutant. Table 2 shows the percentage reduction in emissions during the lockdown period 25/03/2020 to 15/04/2020 compared to 24/03 to 15/04, 2019. (Eregowda, n.d.)(8) tabulated the results by collecting data from Bengaluru, Chennai, Hyderabad, and Mumbai. Table-2 shows a 5.1, 45.7, 18.9, and 42% decrease in PM-2.5 pollutant emission. (Ghosh, n.d.) (9) has tabulated the results by collecting data from the Indian Metropolitan cities NCR-Delhi, Mumbai, Chennai, and Kolkata. Figure 3 shows the results obtained by Ghosh, n.d. from the Landsat 8 OLI and TIRS- Derived Data and Mamdani Fuzzy Logic Modelling Approach to understand PM-10 concentration Variation.

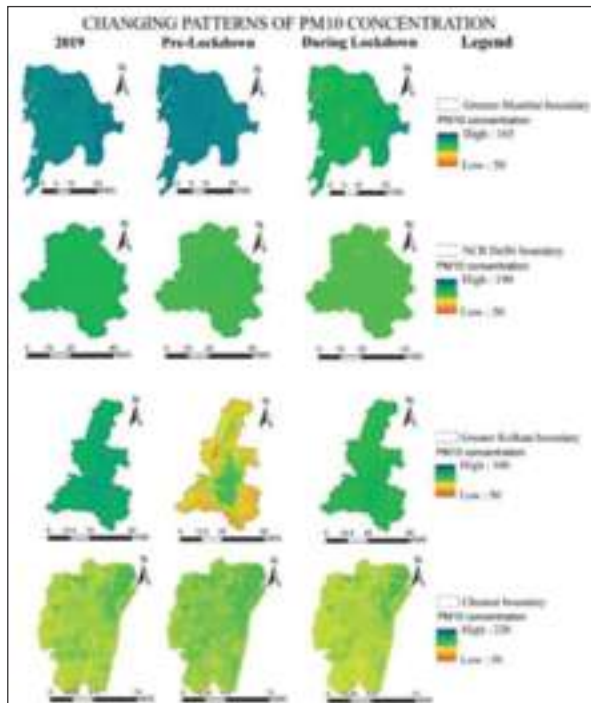


Fig: 3 (9)-Changing PM-10 concentration comparison between 2019, pre-lockdown 2020, and during lockdown 2020.

The PM-10 sources include motor vehicles and construction works. It is the causes numerous health risks, environmental harm, and reduced human comfort levels. While the concentration legends for Mumbai and Chennai are decreasing, Delhi and Kolkata show an increase in PM-10 concentration between pre-lockdown and during the lockdown (9). The variation may be due to the shutdown of industries and

restricted human movement compared to other cities. Mr. P. Singh, in his research work, focused on the Air Quality Index and PM2.5 levels by collecting data from the cities with a branch of the US Embassy in India (10). The Embassies collect the data via US Environmental Protection Agency (EPA) through the Air-Now portal (11).

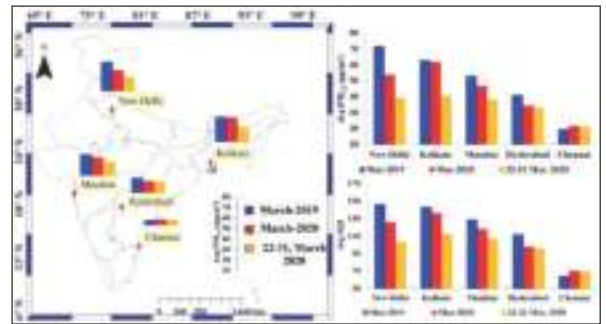


Fig: 4 (10)- The left panel shows the locations for the research work done by Mr. P. Singh. The right panel presents the Average PM2.5 and Air Quality Index (AQI) chart for the area.

Analyzing the data represented in fig-1, it's clear that the PM2.5 and AQI at the areas of study as reduced noticeably during the lockdown period. The pollution levels at Mumbai, Hyderabad, and Chennai have decreased by 19.25%, 3.99%, and 5.40%, respectively. At the same time, Kolkata and Delhi show a considerable reduction of 34.52% and 27.57%, respectively, in the pollutant levels. The northern parts of India, especially the Indo-Gangetic Plains (IGP), have higher levels of PM2.5 throughout the year. The factors included demographic, geographic, seasonal activities, and meteorological parameters. The proximity to the sea for Chennai and Mumbai can cause air mass circulation from the sea surface, which is a possible explanation for the reduced pollution levels. The same is applicable for New Delhi and Kolkata.

Change In No₂

NO₂ is one of the most common pollutants which is highly dependent on the local sources. The dependency is due to its short residence time in the atmosphere. Ms. Eregowda stated in her paper that the NO2 concentration levels at Bengaluru fell from 50 µg/m3 to 10 µg/m3 throughout the lockdown period. The same follows for Chennai, Hyderabad, and Pune. Figure 5 presents the above data graphically.

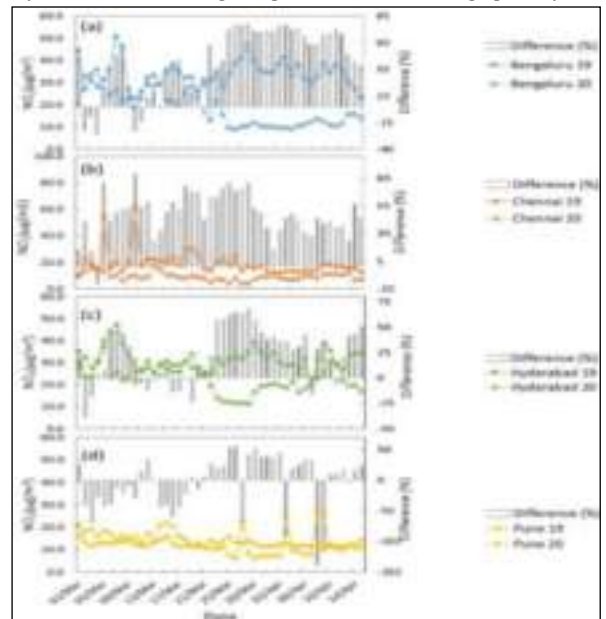


Fig-5 (8)- NO2 Level Graphs With % Difference Representation.

Mr. P. Singh considered the tropospheric NO2 measure by analyzing the data from the Ozone Monitoring Instrument (OMI). The Ozone Monitoring Instrument (OMI) is a section of NASA's A-Train Satellite that measures the levels of various atmospheric gas concentrations. Table-3 represents the box coordinates of the US-embassy locations chosen by Mr. Singh for the research.

Table-3 (10)- The Locations Of The Us-embassies And Their Box Coordinates From OMI.

Location	Latitude	Longitude	Box Coordinates
Delhi	28.59	77.18	W-76.68, S-28.07, E-77.68, N-29.07
Kolkata	22.54	88.35	W-87.86, S-22.08, E-88.86, N-23.08
Mumbai	19.06	72.86	W-72.42, S-18.55, E-73.42, N-19.55
Hyderabad	17.44	78.47	W-77.78, S-17.01, E-78.78, N-18.01
Chennai	13.05	80.25	W-79.76, S-12.56, E-80.76, N-13.56

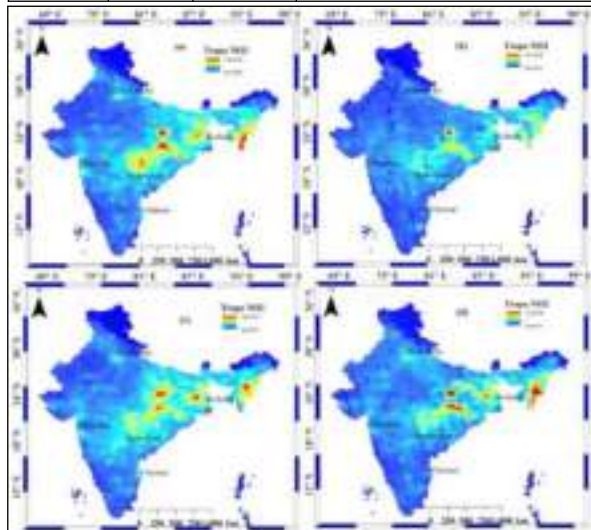


Fig- 6- Tropospheric NO₂ (spatial Variations)- a)10–21March 2019, (b)10–21 March 2020, (c)22–31 March 2019, and (d)22–31 March 2020.

Fig-6 represents the results tabulated by Mr. P. Singh. An HYPSPPLIT Model prepared by Mr. P. Singh shows that long-range air mass transportation affects the air quality at the five selected locations. The westerly air mass transfer is what affects Kolkata, a city located in the eastern Indo-Gangetic Plain. The sources of NO₂ in Mumbai, Delhi, and Kolkata are anthropogenic. While so, the release of NO₂ in Chennai and Hyderabad is due to the burning of biomass. Hence the decline in NO₂ levels during the lockdown is mainly due to the reduction in anthropogenic emissions.

Change In Land Surface Temperature(Lst)

LST is an important factor for environmental health, as it depends on numerous physical and atmospheric parameters (9). Factors like cloud conditions, month, Land Use/Land Cover (LULC) patterns, and so on governs the results. The LST map for 2019, before lockdown 2020 and after lockdown 2020, is given in Figure-7. The transition period in India from winter to summer is from February to March. And the summer season is from March to May.

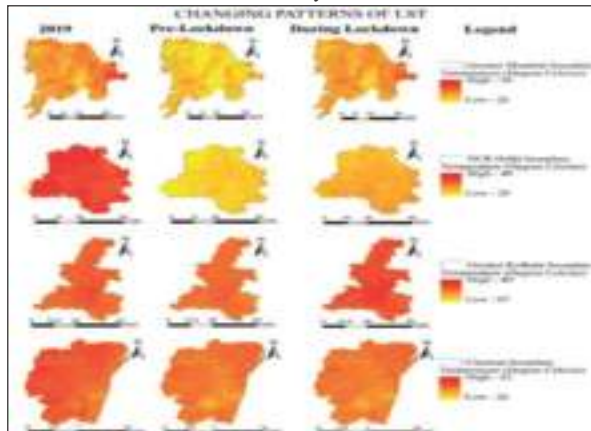


Fig- 7-(9) Changing variations in LST. The LST Map.

These variations in seasonal temperatures can show an increasing trend in the land surface temperature. From the map, it is clear that the temperature map during the pre-lockdown period is less compared to the others. Even though the maps show a similar trend in the four cities, the temperatures in Mumbai, Delhi, and Kolkata are around 48°C to 50°C. Chennai records lower temperatures in the range of 42°C to 44°C.

Electricity Consumption

India is the second most populated (1,391.99 Million) and the seventh-largest country in the world(16). The lockdown has imposed several restrictions on many industrial and everyday goods and other service-related activities. Between January and February of 2019 and 2020, the energy demand increased by 3% and 7%, respectively (17). But, during March 2020, the supply-demand reduced by 3%. Between March 24th and April 19th, the power supply decreased by 25%.

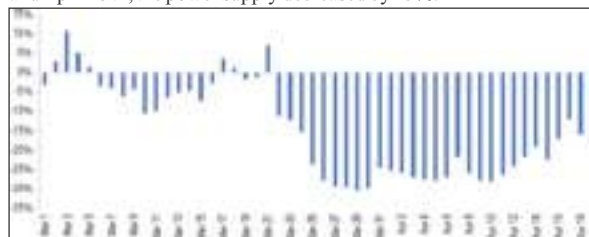


Fig-8 (17)-% change in power supply between March 1st to April 19th.

It is necessary to understand the power consumption in different sectors of India to analyze the power demand variation in India. Fig-9 presents the power consumption in consumer segments of India in 2018-19. Since the government forced the industrial and commercial sectors to shut down during the lockdown, the decrease in demand is self-explained. Table- 4 shows the contribution of different sources to energy production in India.

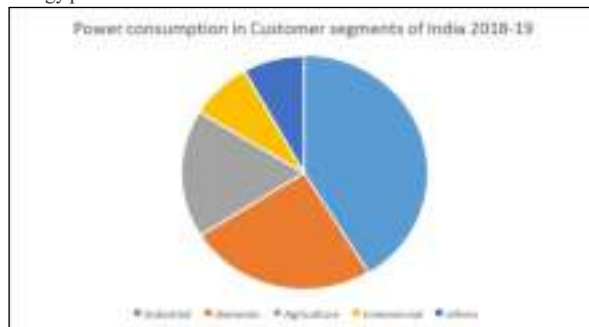


Fig- 9 (17)- Power consumption in Customer segments of India 2018-19.

Table-4 (17)- The contribution of different sources to energy production in India from March 1st to April 19th, 2021.

Energy Sources	Average Generation			Contribution to Total (in %)	
	Mar 1-Mar 24	Mar 25-Apr 19	% change	Mar 1-Mar 24	Mar 25-Apr 19
Coal	2,511	1,873	-25%	72.5%	65.6%
Hydro	302	331	10%	8.7%	11.6%
Renewables (of which)	325	312	-4%	9.4%	10.9%
a. Solar	157	162	3%	4.6%	5.7%
b. Wind	97	96	0%	2.8%	3.4%
Gas, Naptha, Diesel	132	146	11%	3.8%	5.1%
Nuclear	113	114	1%	3.3%	4.0%
Lignite	82	78	-5%	2.4%	2.7%
Total	3,465	2,854	-18%	-	-

Amid the lockdown, the power generation reduced by 25%, compensating for the decrease in demand (17). While considering electricity variations, night light intensity is also a contributing factor. Night light intensity provides information regarding energy consumption in areas with high spatial granularity. Electricity consumption and night light intensity are contributing factors for the analysis of national GDP (18).

For instance, the demonetization in India, 2016 was highly backed up by night light intensity analysis (19). As mentioned, the impact of the nationwide lockdown on India remained even after the release of a few restrictions. The consumption levels were below 14%, and the average monthly fluctuations remain 6% to 10% below the normal (18). Fig-10 shows the trend in electricity consumption from 2013 to 2020.

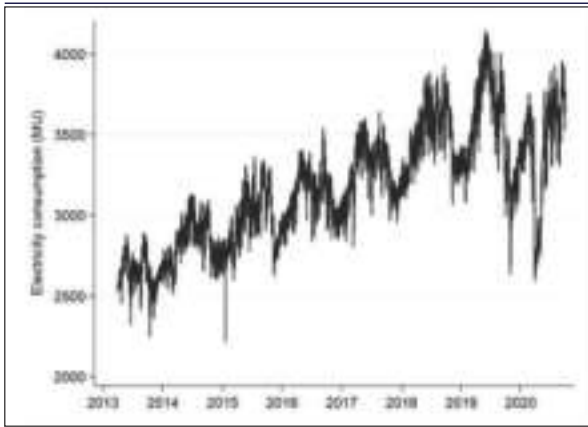


Fig-10 (18) The trend in electricity consumption in India from 2013 to 2020

C. M. Beyer extracted the nighttime light data from the VIIRS-DNB Cloud made available by the Earth Observation Group at the National Oceanic and Atmospheric Administration (NOAA). The data collected was from April 2013 to April 2020 (18). VIIRS satellites have a resolution of 15-arc seconds. Fig-11 shows the changes in the night light time trends in India.

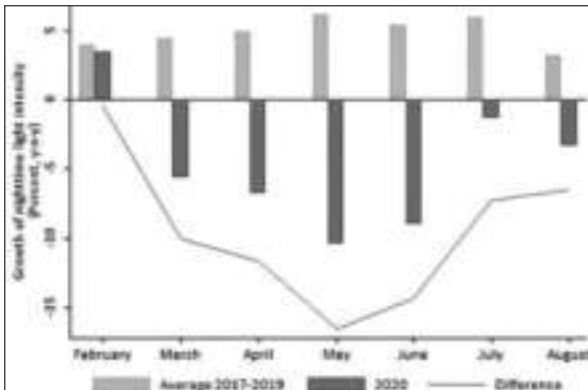


Fig-11 (18) the Changes In The Night Light Time Trends In India.

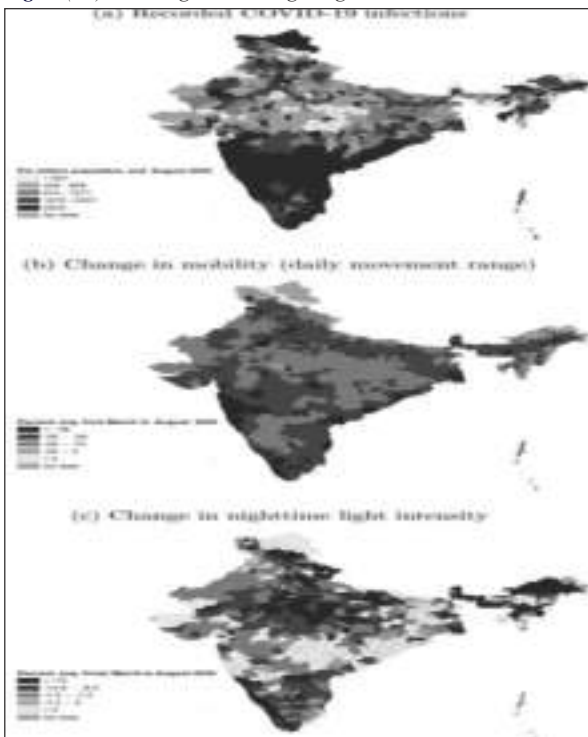


Fig-12 (18)- The spatial variation and impact of COVID-19 across India.

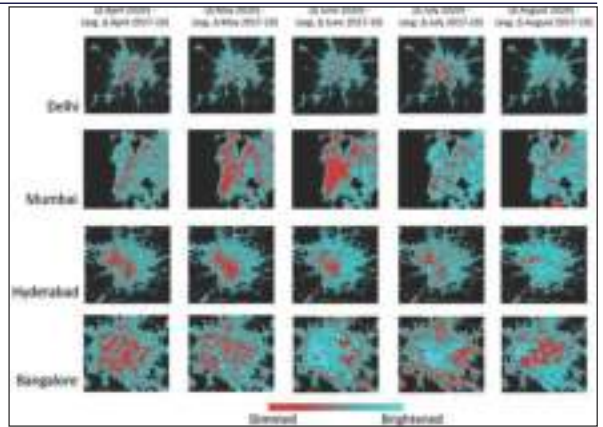


Fig-13-(18)- The changes in night light intensity across India during the lockdown.

CONCLUSION

The impact of lockdown on India is a huge game-changer. The decrease in pollutant levels and electricity consumption rates are posing as an opportunity and a threat at the same time. Through this paper, we have discussed the findings achieved by the researchers for the betterment of our environment. India can rewrite its future to become a sustainable country. The documented results are positive highlighting our potentials.

As the COVID virus spread out, the government keeps extending the lockdown. The positive impact through this episode is huge. Since the energy consumption levels are proportional to household income, the reduced consumption can show the deteriorating levels of the economy. The Corona Virus pandemic will negatively impact all industries, including the power sector of India. But, so far, the impact is positive. Now we get to decide if it will remain the same or not.

REFERENCES

- (1) Shehzad et al., 2020 K. Shehzad, M. Sarfraz, S.G. Meran Shah The impact of COVID-19 as a necessary evil on air pollution in India during the lockdown Environ. Pollut., 115080 (2020), 10.1016/j.envpol.2020.115080 Google Scholar.
- (2) Venkataraman C, Brauer M, Tibrewal K, Sadavarte P, Ma Q, Cohen A, Chaliyakunnel S, Frostad J, Klimont Z, Martin RV, Millet DB (2018) Source influence on emission pathways and ambient PM2.5 pollution over India (2015–2050). Atmos Chem Phys Discuss 8:8017–8039 <https://acp.copernicus.org/articles/18/8017/2018/Article> Google Scholar.
- (3) <https://blogs.worldbank.org/endpovertyinsouthasia/indias-electricity-consumption-data-shows-economic-impact-covid-19>
- (4) <https://energy.economicstimes.indiatimes.com/news/power/coronavirus-impact-within-ten-days-26-per-cent-fall-in-indias-energy-consumption/74854825>
- (5) Positive effects of COVID-19 lockdown on air quality of industrial cities (Ankleshwar and Vapi) of Western India Ritwik Nigam, Kanvi Pandya, Alvarinho J. Luis, Raja Sengupta & Mahender Kothe. <https://www.nature.com/articles/s41598-021-83393-9>
- (6) Air pollution in Indian cities: Understanding the causes and the knowledge gaps Q&A with Dr. Sarath Guttikunda-14th December 2017, <https://www.eprindia.org/news/air-pollution-indian-cities-understanding-causes-and-knowledge-gaps>
- (7) Impact of lockdown associated with COVID19 on air quality and emissions from transportation sector: A case study in selected Indian metropolitan cities. Tejaswini Eregowda, Pritha Chatterjee & Digvijay S. Pawar, *Environment Systems and Decisions* (2021).
- (8) Impact of COVID-19 Induced Lockdown on Environmental Quality in Four Indian Megacities Using Landsat 8 OLI and TIRS-Derived Data and Mamdani Fuzzy Logic Modelling Approach. Sasanka Ghosh, Arijit Das, Tusar Kanti Hembram, Sunil Saha, Biswajeet Pradhan, and Abdullah M. Alamri. [file:///C:/Users/Asus/Downloads/sustainability-12-05464-v2%20\(1\).pdf](file:///C:/Users/Asus/Downloads/sustainability-12-05464-v2%20(1).pdf)
- (9) Impact of Lockdown on Air Quality in India During COVID-19 Pandemic-July 2020 Air Quality Atmosphere & Health 13(13) DOI:10.1007/s11869-020-00863-1, Project: Air Quality measurement during COVID-19, https://www.researchgate.net/publication/342749510_Impact_of_Lockdown_on_Air_Quality_in_India_During_COVID-19_Pandemic
- (10) <https://www.airnow.gov/>
- (11) Air Pollution in India: Real-time Air Quality Index Visual Map. <https://aqicn.org/map/india/>
- (12) AirNow for air quality data. <https://www.airnow.gov/>
- (13) OMI-NASA's A-Train Satellite. <https://giovanni.gsfc.nasa.gov/giovanni/>
- (14) INDIA: AIR QUALITY STANDARDS. <https://www.transportpolicy.net/standard/india-air-quality-standards/>
- (15) India-Wikipedia. <https://en.wikipedia.org/wiki/India>
- (16) Impact of COVID-19 on the Power Sector, Saket Surya. (The PRS Blog-April 23, 2020) <https://www.prsindia.org/theprsblog/impact-covid-19-power-sector>
- (17) Examining the economic impact of COVID-19 in India through daily electricity consumption and nighttime light intensity Robert C.M.Beyera, Sebastian Franco-Bedoyaa, Virgilio Galdoa <https://www.sciencedirect.com/science/article/pii/S0305750X20304149#>
- (18) Beyer et al., 2018 Beyer, R.C., Chhabra, E., Galdo, V., & Rama, M. (2018). Measuring districts' monthly economic activity from outer space (No. 8523). World Bank Policy Research Working Paper. Google Scholar

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.

Study on Treatment of Blood from Abattoir using Microbial Fuel Cell (MFC) Technology with Production of Green Energy

Article type: Research Article

Authors: [Sreedharan, Sanju](https://content.iospress.com:443/search?q=author%3A%28%22Sreedharan, Sanju%22%29) (<https://content.iospress.com:443/search?q=author%3A%28%22Sreedharan, Sanju%22%29>)

Affiliations: Department of Civil Engineering, SCMS School of Engineering and Technology, Karukutty, Ernakulam, Kerala - 683576, India, sanjus@scmsgroup.org (<mailto:sanjus@scmsgroup.org>).

Abstract: Zero energy technologies and sustainable energy production are the two major concerns of present day researches. Microbial fuel cells (MFCs) are bioreactors that extract chemical energy stored in organic compounds, into electric potential, through bio-degradation. The core reason for the high strength of effluent generated from slaughterhouses is animal blood. The current study evaluates the potential of MFC technology to reduce the pollution strength of cattle blood in terms of chemical oxygen demand (COD). The current study was piloted in three stages using lab scale two chambered MFC: The first stage was to determine the best oxidising agent as compared to natural aeration from three accessible options, KMnO₄, diffused aeration and tape grass aquatic plant. KMnO₄ was found to be the superlative with a 30% reduction in COD in 100 hrs batch reactor and a maximum power of 0.97 mW using 125 mL livestock blood. The second stage of the study optimised the concentration of KMnO₄. At 500 mg/L KMnO₄ concentration, 50% COD removal efficiency was acquired in a batch reactor of 60 hrs with an average energy output of 1.3 mW. In the final stage on the addition of coconut shell activated carbon with an Anolyte at a rate of 40 mL/125 mL of substrate COD removal efficiency increased to 74.9%.

Keywords: Adsorption, bio-energy, cattle blood, microbial fuel cell, wastewater treatment

DOI: 10.3233/AJW210053

Journal: [Asian Journal of Water, Environment and Pollution](https://content.iospress.com:443/journals/asian-journal-of-water-environment-and-pollution) (<https://content.iospress.com:443/journals/asian-journal-of-water-environment-and-pollution>), vol. 18, no. 4, pp. 135-140, 2021

Received 20 May 2020 | 23 March 2021 | **Accepted** 23 March 2021 | **Published:** 18 November 2021

Price: EUR 27.50

Effect of Plasticity of Fines on Properties of Uniformly Graded Fine Sand



M. Akhila, K. Rangaswamy, N. Sankar, and M. R. Sruthy

1 Introduction

Even though researchers separate soils based on particle size as sand, silt and clay, in the field, soil always exists as a combination of all these. There are many studies concentrating on the effect of fines on the shear characteristics of sand [1–3] and liquefaction [4–7] but only a few studies have considered the other properties.

Yang and Wei [8] have analysed the change in critical state friction angle for Fujian and Toyoura sands. For clean sand without fines, the critical state friction angle tends to decrease with increasing roundness of sand particles. When those sands were tested with fines (round shape), the critical state friction angle of the mixture tends to decrease with an increase in fines content. But for fines with an angular shape, the critical state friction angle tends to increase with fines content. Phan et al. [9] have conducted one-dimensional consolidation tests on sand–silt mixtures (with low-plastic fines at a constant void ratio and constant relative density) and indicated that the behaviour of the mixtures were similar to those of loose sand. The effect of fines on void ratios was studied by Cubrinovski and Ishihara [10]. The authors reported that the void ratio initially decreases as the fines content increases from 0–20% and above 40% fines, the maximum and minimum void ratios were seen to increase steadily.

It is clear from the literature that the studies on the effect of plasticity of fines on the properties of sand are limited. Hence, the present study is focused on the effect of the amount of fines and the type of fines (or plasticity index of fines) on various properties of sand like specific gravity, limiting void ratios, grain size characteristics, angle of internal friction and compression index.

M. Akhila (✉) · M. R. Sruthy

SCMS School of Engineering and Technology, Ernakulam, Kerala, India

K. Rangaswamy · N. Sankar

NIT Calicut, Kozhikode, Kerala, India

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).

PAPER • OPEN ACCESS

Tribological and Corrosion analysis of Co-20Al-GNSA composites produced through powder metallurgy process

To cite this article: G R Raghav *et al* 2021 *IOP Conf. Ser.: Mater. Sci. Eng.* **1126** 012018

View the [article online](#) for updates and enhancements.

You may also like

- [Ca-Mg Multiple Deoxidation of Ti-50Al-2Cr-2Nb Intermetallic Compound Powder for Additive Manufacturing](#)
Seongjae Cho, Taeheon Kim and Jae-Won Lim
- [Comparative study of hot corrosion behavior of thermal sprayed alumina and titanium oxide reinforced alumina coatings on boiler steel](#)
Gurdeep Singh, Santosh Kumar and Rakesh Kumar
- [Investigations into the effects of volatile biomass tar on the performance of Fe-based CLC oxygen carrier materials](#)
Matthew E Boot-Handford, Nick Florin and Paul S Fennell



ECS **Connect with decision-makers at ECS**

Accelerate sales with ECS exhibits, sponsorships, and advertising!

▶ Learn more and engage at the 244th ECS Meeting!

Tribological and Corrosion analysis of Co-20Al-GNSA composites produced through powder metallurgy process

G R Raghav^{1*}, U Arunachalam², R Sujith³

¹Associate Professor, Department of Mechanical Engineering, SCMS School of Engineering and Technology, Vidyannagar, Karukutty, Ernakulam- 683576, Kerala, India

² Assistant Professor, Department of Mechanical Engineering, University College of Engineering, Nagercoil, 629004, Tamilnadu, India

³ Assistant Professor, Department of Mechanical Engineering, SCMS School of Engineering and Technology, Vidyannagar, Karukutty, Ernakulam- 683576, Kerala, India

* E-mail: raghavmechklnc@gmail.com

Abstract. In this study, Co-20Al-GNSA metal matrix composites were produced using mechanical alloying process. The Co-20Al-GNSA composites were mixed using a high-energy ball mill at a constant speed of 350 rpm for 2 hours. The composite powders were then characterized for their morphological study using Scanning Electron Microscope. The composite powders are then compressed and sintered at 500 MPa and 700°C respectively. The density and compressive strength of the composite materials shows decrement values whereas the wear resistance of the composite materials has increased considerably. The mechanism of wear was identified as abrasive wear. The electrochemical corrosion test also reveals that the Co-20Al-10GNSA composites have better corrosion resistance. The weight-loss corrosion test also shows that the composites with 10GNSA content have better corrosion resistance.

Keywords: Wear, Corrosion, Powder Metallurgy, Microhardness.

1. Introduction

The exploitation of hybrid composites for the potential replacement of conventional metals has been drastically increased in several applications such as automobile industries, commercial industrial applications, and also in aerospace industries where enhanced mechanical, wear and corrosion resistance properties are expected. Hence it was the main objective of the researchers to develop materials with lesser density with better tribological and corrosion performance [1–3].

In the process of developing a composite material, it is very important to select the matrix materials and reinforcements with the good wet ability to improve the bonding of the composite materials. Another important factor is the selection of the fabrication method and its working parameters as per the matrix and reinforcement materials. Now a day the production economy is one of the important factors due to the economic thoughtfulness of the industries [4–7]. They prefer low-cost composites, to reduce the production cost due to the raw materials. The Co is a material with very good mechanical and corrosion resistance property whereas it is very costly. So it is necessary to tailor



the mechanical properties using low-cost reinforcements such as Al and GNSA. The most generally used reinforcements are SiC, Al₂O₃, and TiO₂. The ceramic particles Al₂O₃ and TiO₂ do not have good wetting characteristics [8–10]. Thereby to improve the wettability the Ground Nut Shell Ash (GNSA) is utilized as secondary reinforcement.

The uniform dispersion of reinforcement was another area that has to be addressed while selecting a fabrication process. There are many methods such as stir casting to produce composite materials whereas achieving uniform dispersion is not possible due to cluster formation. But it is possible to produce composite materials with a uniform dispersion of reinforcements using the powder metallurgy process. Another reason for choosing powder metallurgy was their cost-effectiveness and their reliability for the production of high melting point materials [5,8,11,12].

In this study, a range of combinations of Co-20Al-GNSA hybrid composites was formed via the powder metallurgy process. The hybrid composites were then made-up into 10mm cylindrical pellets using a die setup made up of high-speed steel die. The muffle furnace was utilized to harden the compacted green pellets using a sintering operation. The microhardness, density, and compressive strength of the composite materials were studied and reported. The pin on disc apparatus and electrochemical workstation were utilized to study the wear and corrosion resistance properties respectively. Thus the main purpose of this effort is to develop hybrid composite materials with better mechanical, improved wear, and corrosion resistance properties that can be employed in automobile, industrial, and aerospace applications.

2. Materials and Methods

The chemicals used in this work are of research-grade (99.5% purity). The composite powders are synthesized using a high-energy ball mill that comprises tungsten carbide balls. The ball milling process was carried out for a duration of 2 hours at a speed of 350 rpm in the existence of toluene as a process control agent to acquire homogenous hybrid composites. The homogeneously unified composite powders are then packed down using a uniaxial hydraulic press at 500 Mpa and sintered at 700°C to produce a 10 mm cylindrical pellet. The composite materials are characterized using SEM to find out the morphology of the composite materials. The microhardness of the Co-20Al-GNSA hybrid composites was carried out at 0.5 kg load using Vickers hardness equipment and standard deviation values are considered and reported. The density of the Co-20Al-GNSA hybrid composites was calculated based on the Archimedes principle. The 10 mm diameter composite pellets are compressed using the universal testing machine (UTM) at a uniform and gradual speed rate of 3mm min⁻¹ [8,13,14]. The wear analysis was carried out at constant load, constant speed, and sliding distance of 10N, 1.5 m/s, and 1000m respectively. The electrochemical corrosion analysis was carried out using a bio-logic electrochemical workstation. The workstation consists of three electrodes, the platinum counter electrode, Al/AgCl reference electrode and composite pellets as working electrode. The scan was carried out at 5mV/s. The composite pellets are immersed in 3% NaCl solution for 1 hour so as to stabilize the open circuit potential. The weight-loss corrosion analysis was carried out with various corrosive media such as 0.1N HCl, 0.1N H₂SO₄ and 3% NaCl solution. The composite pellets are immersed in the corrosive media for 24 hours. The composite specimens are weighed before and after the test and the weight loss is calculated. [13,15].

3. Results and Discussion

3.1 Density and Microhardness.

The SEM image of The Co-20Al-5GNSA hybrid composites is shown in Figure.1. The SEM image is taken in secondary electron mode. It is observed that there is uniform dispersion of Co, Al, and GNSA particles in hybrid composite material. Figure.2 exhibits the microhardness and density graphs of Co-20Al-GNSA. The density of the Co-20Al-GNSA hybrid composites increases slightly with the addition of GNSA particles. The density of Co-20Al-2.5GNSA composite was 5.2 g/cm³ which increase

slightly to 5 g/cm^3 for Co-20Al-5GNSA hybrid composites. The further addition of GNSA reinforcements resulted in a decrease in the density of the Co-20Al-10GNSA hybrid composites (5 g/cm^3). The microhardness of the Co-20Al-GNSA hybrid composite increases linearly up to 5% GNSA reinforcement further addition of GNSA particles has resulted in a slight reduction in the microhardness of the composite material.

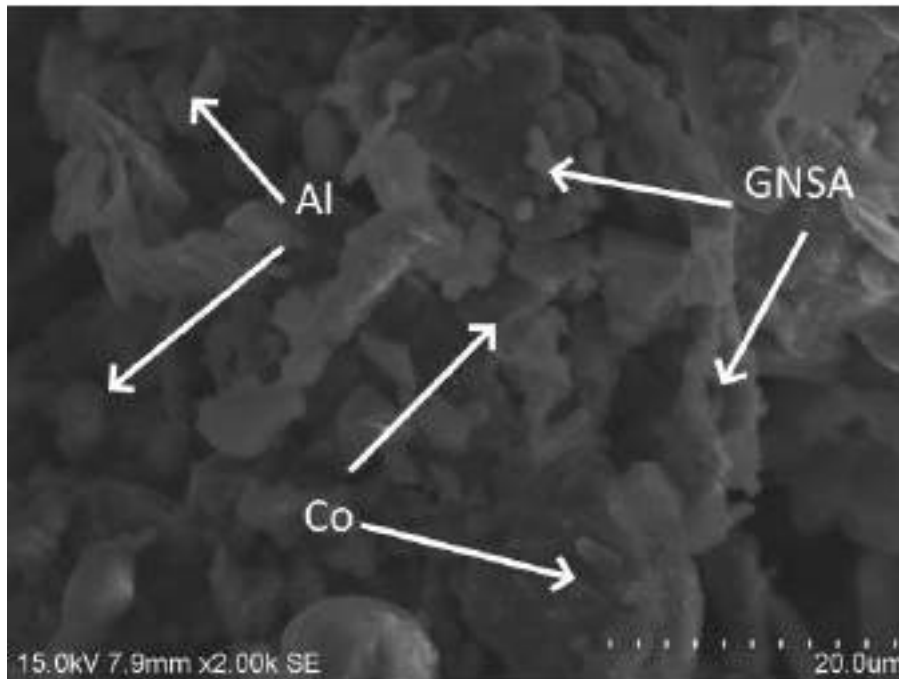


Figure.1 SEM image of Co-20Al-5GNSA hybrid composite

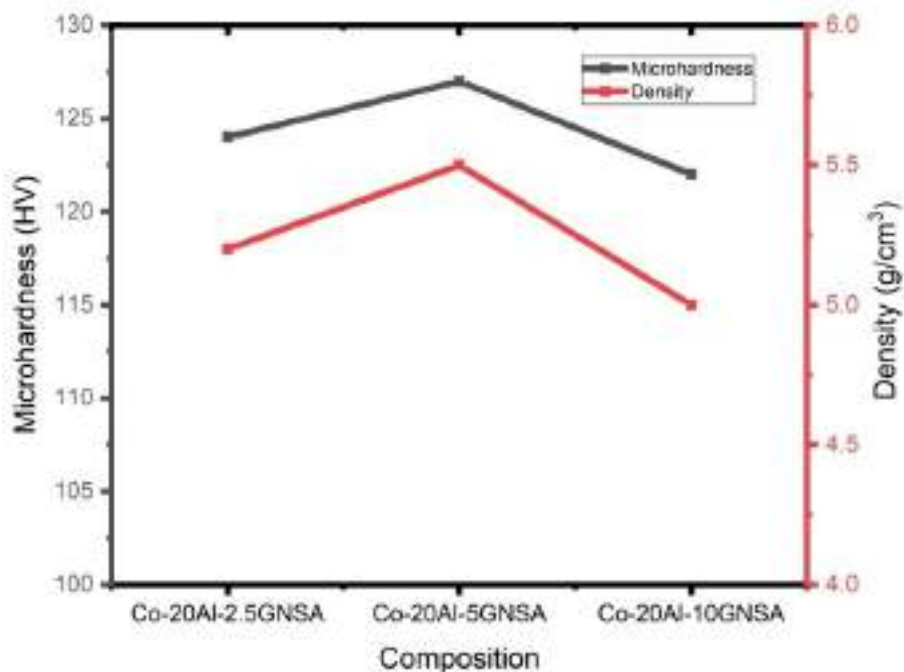


Figure.2 Microhardness and Density of Co-20Al-GNSA hybrid composites

3.2 Compressive Strength

The Compressive strength of Co-20Al-5GNSA hybrid composites is 112 MPa which is better compared to Co-20Al-2.5GNSA (110 MPa) and Co-20Al-10GNSA (108 MPa) hybrid composites. The compressive strength increased gradually with GNSA addition till it reaches 5% GNSA, whereas further addition of GNSA does not influence the compressive strength of the hybrid composites. Figure.3 represents the compressive strength of the Co-20Al-GNSA hybrid composites.

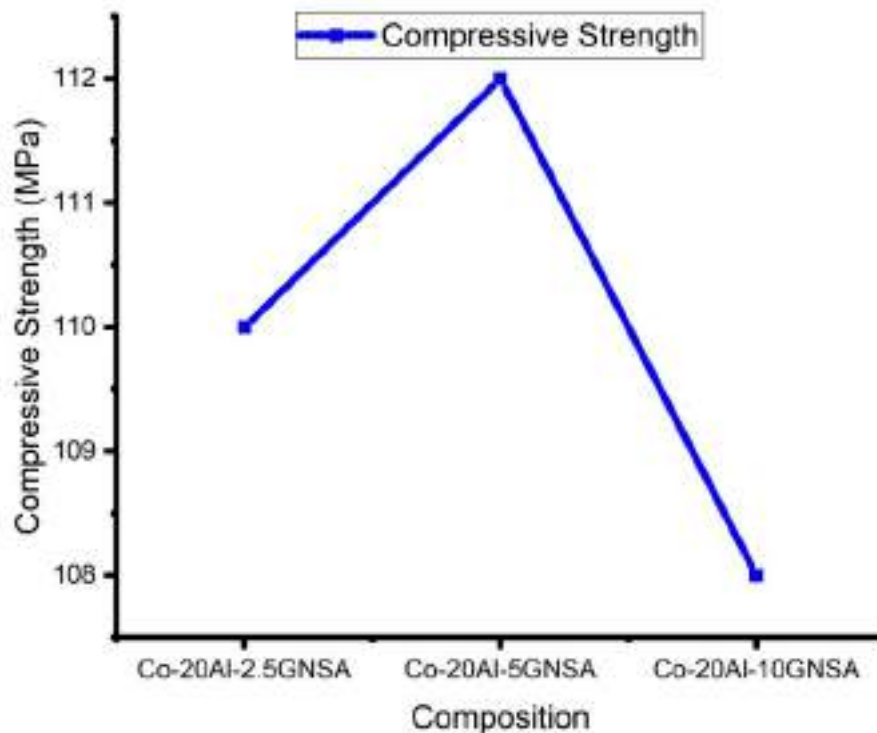


Figure.3 Compressive Strength of Co-20Al-GNSA hybrid composites

3.3 Wear and COF Analysis

The wear analysis results of Co-20Al-GNSA hybrid composites are represented in Figure.4. The wear test was carried out at a constant load of 10N, constant sliding speed of 1.5 m/s, and constant sliding distance of 1000m. The wear loss of the Co-20Al-10GNSA hybrid composites exhibited better wear resistance and COF values compared to that of other specimens. The Co-20Al-2.5GNSA has produced a COF value of 0.9 whereas the COF value of Co-20Al-10GNSA is 0.6. Hence it can be confirmed that the addition of GNSA particles has a good influence in increasing the wear resistance of the hybrid composite materials. Figure.5 represents the SEM image of Co-20Al-10GNSA hybrid composites after the wear test. From the pattern of wear track, it is evident that the major wear mechanism is abrasive wear.

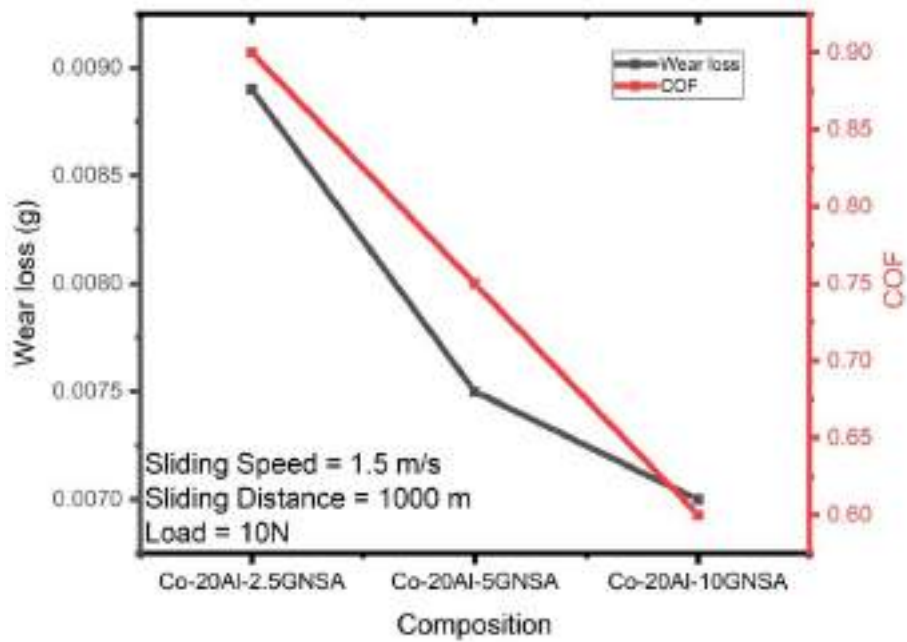


Figure. 4 Wear loss and COF of Co-20Al-GNSA hybrid composites

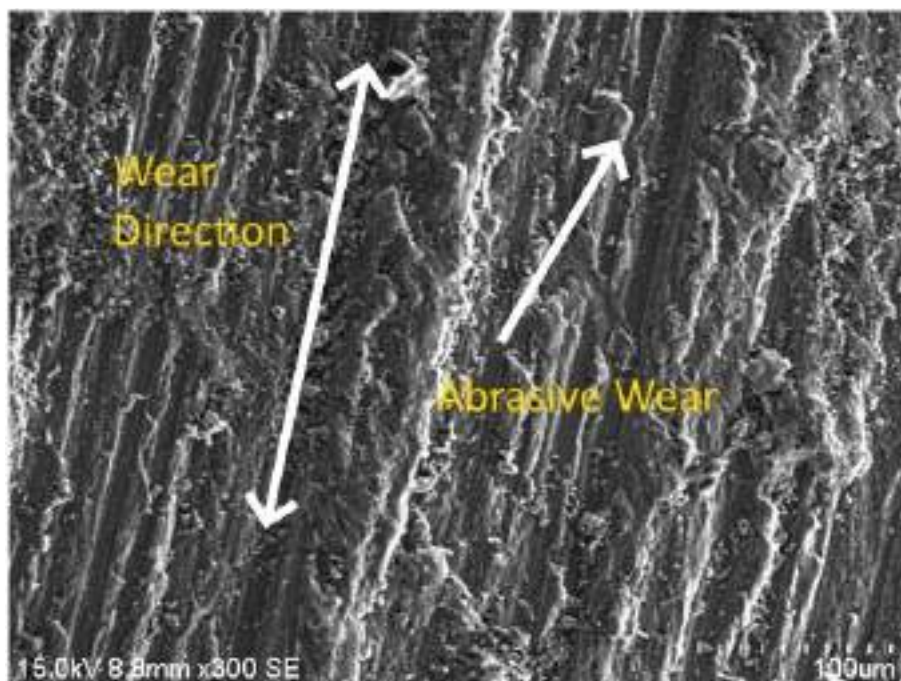


Figure .5 SEM image of Co-20Al-10GNSA hybrid composite after wear test

3.4 Corrosion Analysis

Figure .6 shows the weight-loss corrosion graphs of Co-20Al-GNSA hybrid composites at various Corrosive media such as 3% NaCl, 0.1N HCl, and 0.1N H₂SO₄. From the graph, it can be concluded that the weight loss of the composites decreases with the increase in GNSA content. The weight loss of the Co-20Al-10GNSA hybrid composite is better compared to other combinations in all kinds of corrosive media. The weight loss was maximum for 0.1 N H₂SO₄ for all samples compared to other corrosive media. The weight loss was minimum in 3% NaCl solution. The electrochemical corrosion analysis was carried out using three-electrode systems using the composite pellets as the working electrode. The potentiodynamic polarization results shows that the Co-20Al-10GNSA hybrid composites have exhibited better E_{corr} (-0.442 V) and i_{corr} values (1.5 mA/cm²) compared to that of Co-20Al-5GNSA (-0.448V & 1.7 mA/cm²) and Co-20Al-2.5 GNSA (-0.453V & 1.9 mA/cm²). It is evident that the E_{corr} values are shifted to more positive side and the i_{corr} values decreases with the increase in GNSA content which confirms the increase in corrosion resistance.

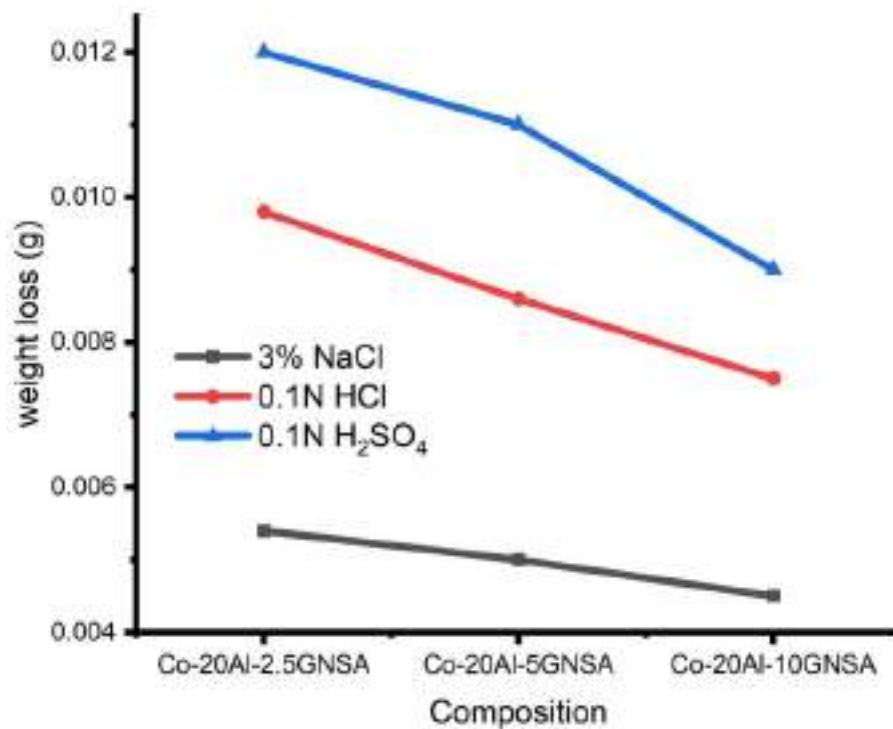


Figure.6 Weight loss corrosion results of Co-20Al-GNSA hybrid composites

4. Conclusions

The Co-20Al-GNSA hybrid composites were amalgamated by employing a high-energy ball mill.

- The Microhardness and density of the Co-20Al-5GNSA hybrid composites have improved compared to other samples.
- The compressive strength of the Co-20Al-5GNSA hybrid composite was 112 MPa and superior compared to other combinations.
- The wear analysis authenticates that Co-20Al-10GNSA hybrid composites exhibited better wear resistance and coefficient of friction.
- The potentiometric polarization analysis shows that the Co-20Al-10GNSA hybrid composites have enhanced corrosion resistance due to the existence of GNSA particles.
- The weight-loss corrosion analysis also proves that the Co-20Al-10GNSA hybrid composites have better corrosion resistance.

From the conclusion of this study, it can be accomplished that the Co-20Al-10GNSA hybrid composites have experienced a slight decrease in density, microhardness, and compressive strength but exhibits better wear and corrosion resistance. Hence it can be concluded that Co-20Al-10GNSA hybrid composites have better tribological and corrosion resistance with decent mechanical properties which may be considered for potential industrial applications.

References:

- [1] Gopinath S, Prince M and Raghav G R 2020 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* **7** 016582
- [2] Raghav G R, Balaji A N, Muthukrishnan D, Sruthi V and Sajith E 2018 An experimental investigation on wear and corrosion characteristics of Mg-Co nanocomposites *Mater. Res. Express* **5** 066523
- [3] Kumar N, Bharti A and Saxena K K 2021 A re-investigation: Effect of powder metallurgy parameters on the physical and mechanical properties of aluminium matrix composites *Mater. Today Proc.*
- [4] Albert T, Sunil J, Christopher A S, Jegan R, Prabhu P A and Selvagesan M 2020 Preparation and characterization of aluminium-titanium carbide (Al-TiC) composite using powder metallurgy *Mater. Today Proc.*
- [5] Karthikeyan N, Krishnan B R, VembathuRajesh A and Vijayan V 2020 Experimental analysis of Al-Cu-Si metal matrix composite by powder-metallurgy process *Mater. Today Proc.* S221478532036525 1
- [6] Li K C, Prior D J, Waddell J N and Swain M V 2015 Comparison of the microstructure and phase stability of as-cast, CAD/CAM and powder metallurgy manufactured Co-Cr dental alloys *Dent. Mater.* **31** e306-15
- [7] Zhou Z, Liu B, Guo W, Fu A, Duan H and Li W 2021 Corrosion behavior and mechanism of FeCrNi medium entropy alloy prepared by powder metallurgy *J. Alloys Compd.* **867** 159094

- [8] Tan L, He G, Li Y, Liu F, Nie Y and Jiang L 2018 Flow behaviors and microstructural evolutions of a novel high-Co powder metallurgy superalloy during hot working *J. Mater. Process. Technol.* **262** 221–31
- [9] Rodrigues W C, Broilo L R, Schaeffer L, Knörnschild G and Espinoza F R M 2011 Powder metallurgical processing of Co–28%Cr–6%Mo for dental implants: Physical, mechanical and electrochemical properties *Powder Technol.* **206** 233–8
- [10] Mihalcea E, Vergara-Hernández H J, Jimenez O, Olmos L, Chávez J and Arteaga D 2021 Design and characterization of Ti6Al4V/20CoCrMo–highly porous Ti6Al4V biomedical bilayer processed by powder metallurgy *Trans. Nonferrous Met. Soc. China* **31** 178–92
- [11] Venkatesh V S S and Deoghare A B 2020 Fabrication and mechanical behaviour of Al-Kaoline metal matrix composite fabricated through powder metallurgy technique *Mater. Today Proc.*
- [12] 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
- [13] Raghav G R, kumar R A, Muthukrishnan D, Nagarajan K J, E S and V S 2020 Synthesis and mechanical characterization of Fe–BN–TiC nanocomposites *Eng. Res. Express* **2** 025036
- [14] Sudha G T, Stalin B, Ravichandran M and Balasubramanian M 2020 Mechanical Properties, Characterization and Wear Behavior of Powder Metallurgy Composites - A Review *Mater. Today Proc.* **22** 2582–96
- [15] Raghav G R, Janardhanan S, Sajith E, Chandran V and Sruthi V 2021 Mechanical and tribological performance of Al-Fe-SiC-Zr hybrid composites produced through powder metallurgy process *Mater. Res. Express* **8** 016533



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-

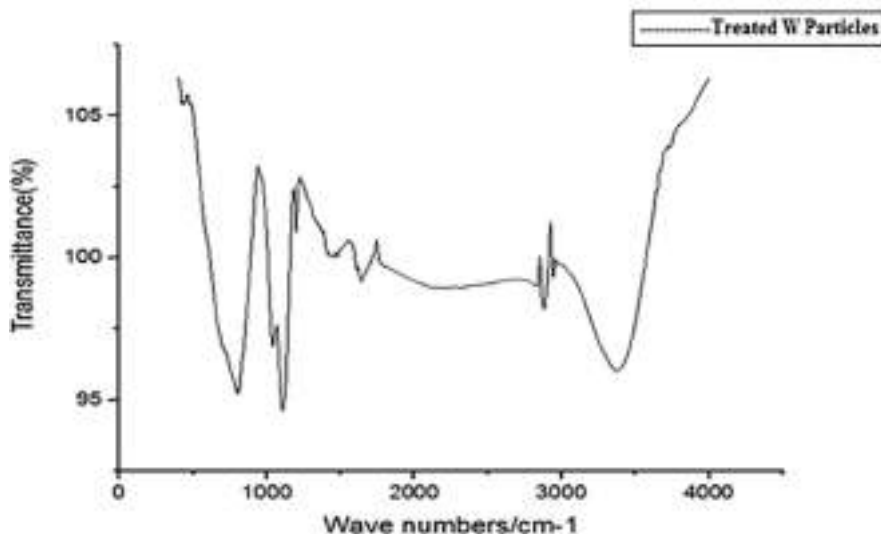


Fig. 1. FTIR Spectra of GPTMS treated W particles.

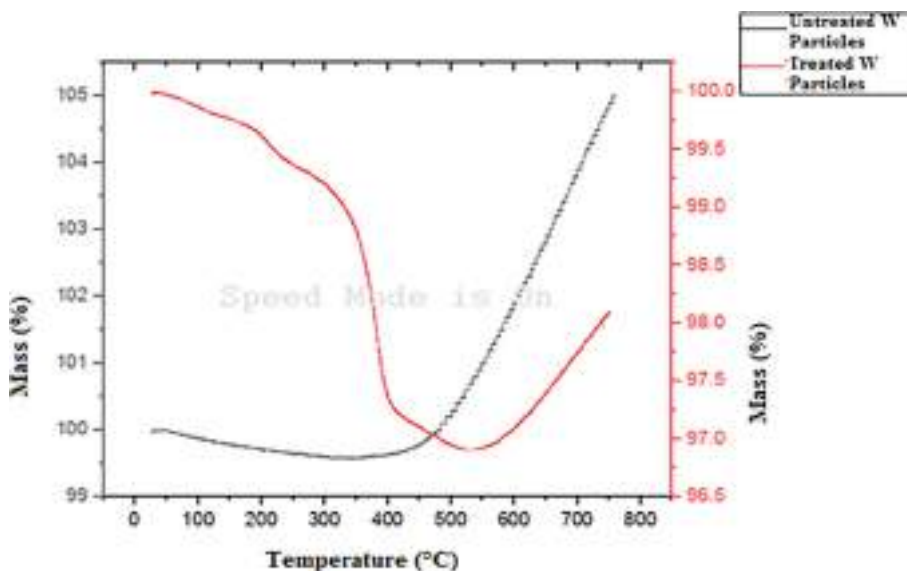


Fig. 2. TGA thermograms of untreated and treated W particles.

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 homogeneously 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.

$$\text{Relative Density} = 1 - \text{Porosity} \quad (1)$$

$$\text{Porosity} = \frac{\text{Theoretical Density} - \text{Actual Density}}{\text{Theoretical Density}} \times 100 \quad (2)$$

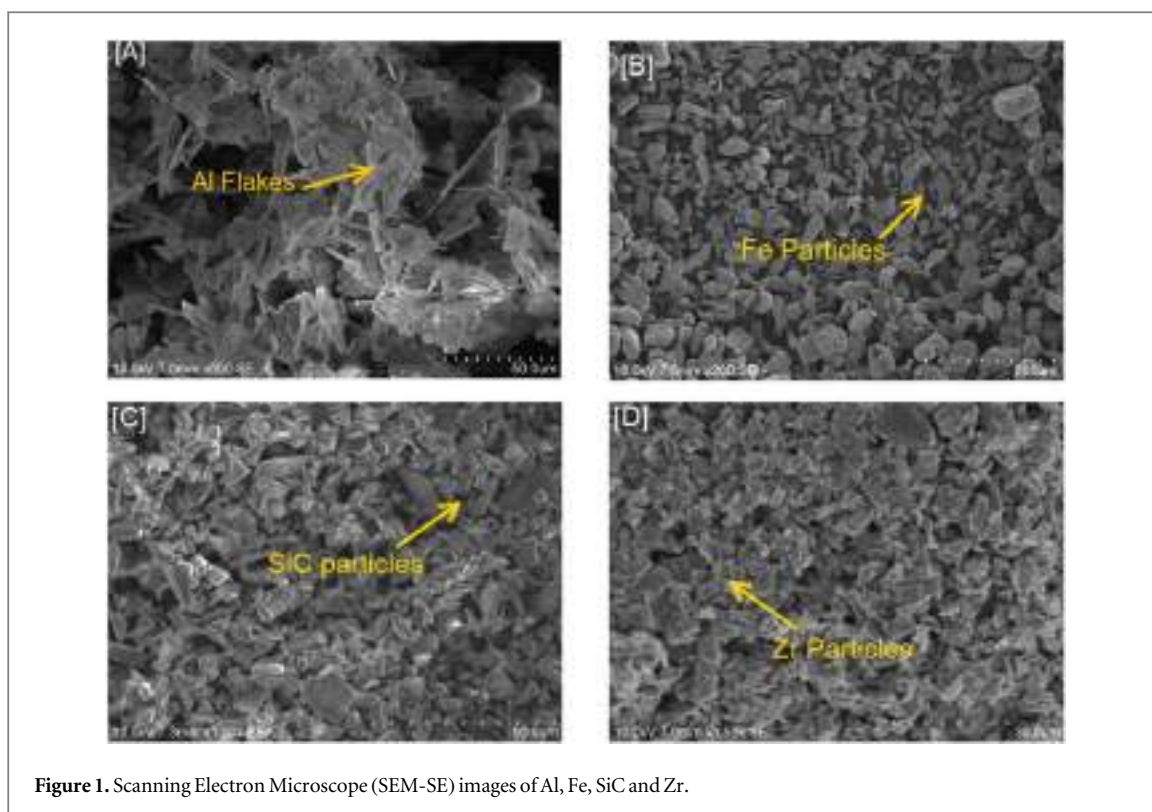


Figure 1. Scanning Electron Microscope (SEM-SE) images of Al, Fe, SiC and Zr.

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

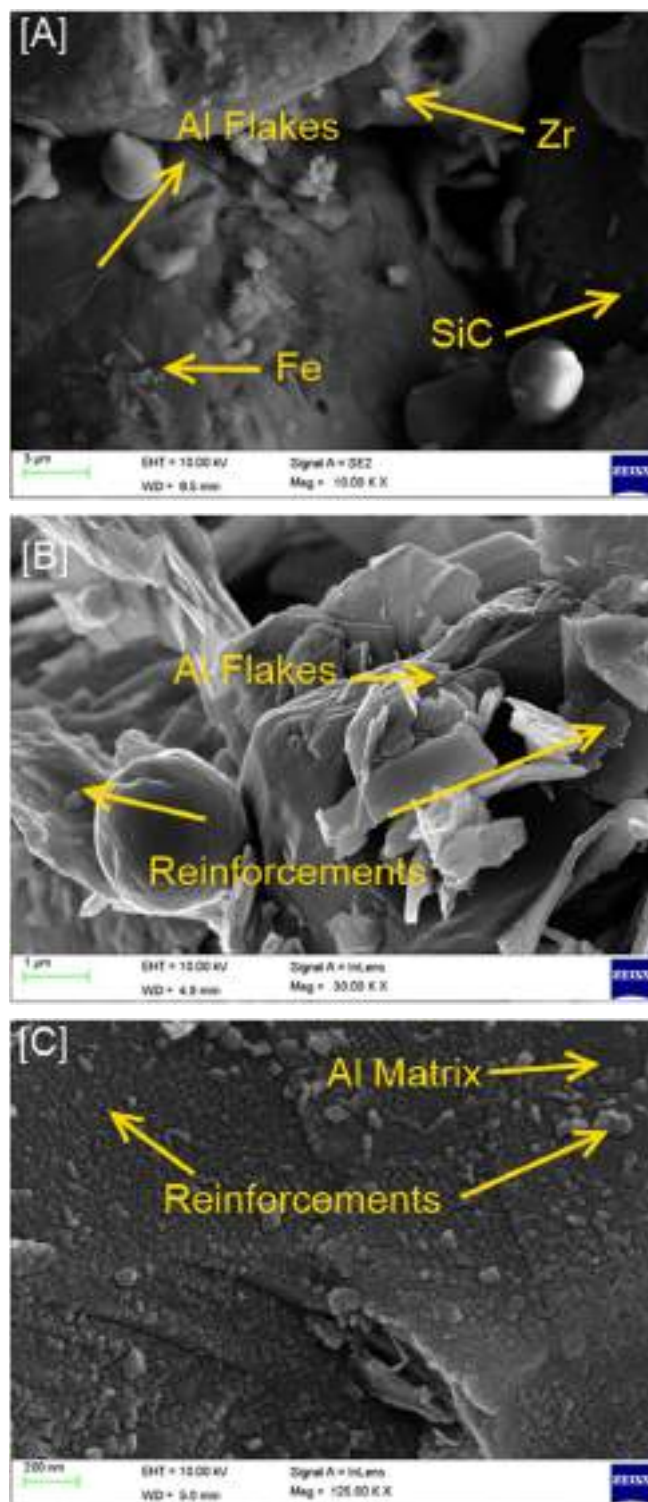


Figure 2. High Resolution Field Emission Scanning Electron Microscope (FE-SEM) image of Al-10Fe- 10SiC- 10Zr hybrid composites.

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 ,

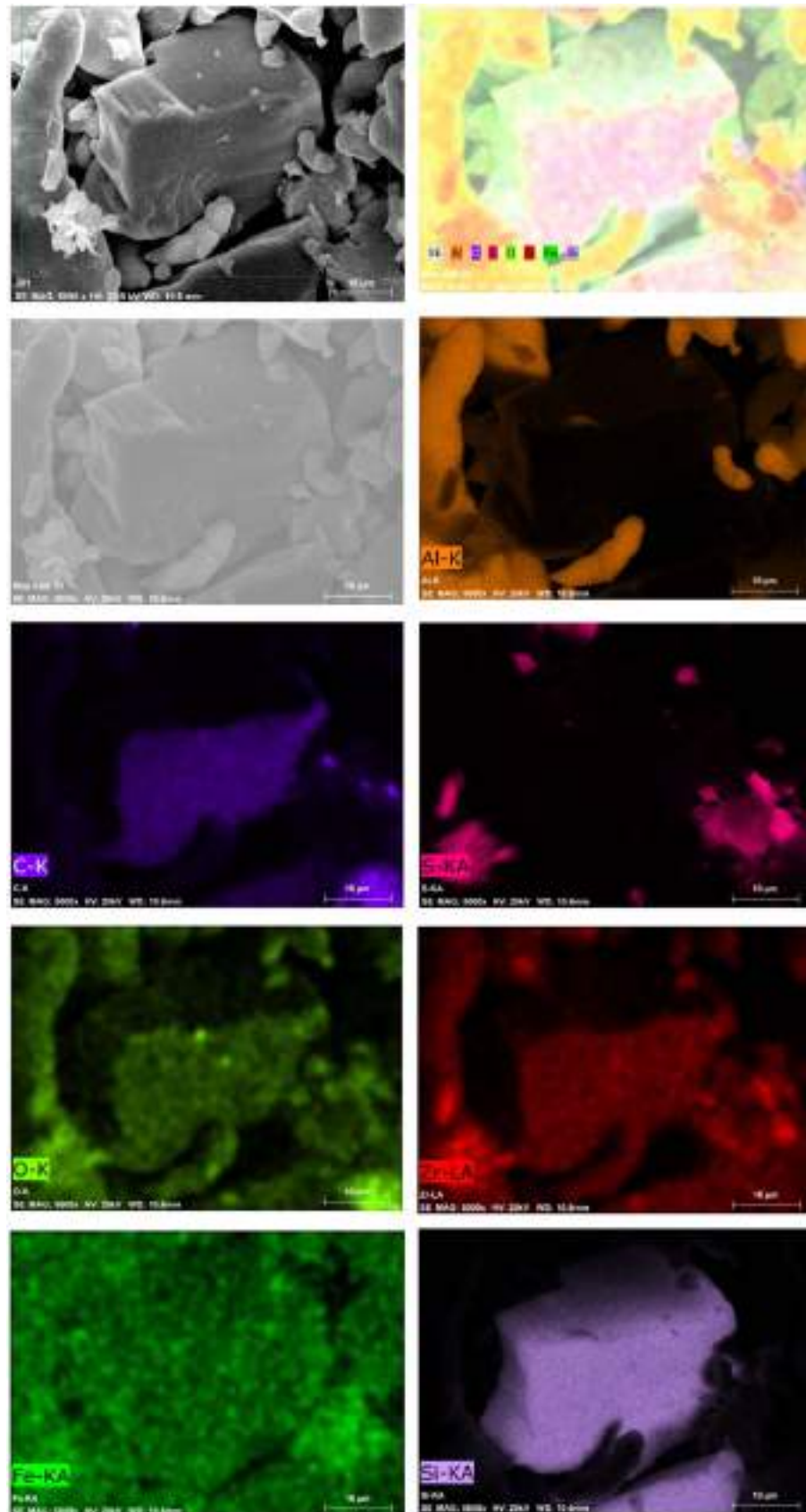


Figure 3. Field Emission Scanning Electron Microscope (FE-SEM) mapping of Al-10Fe-10SiC-10Zr hybrid composites.

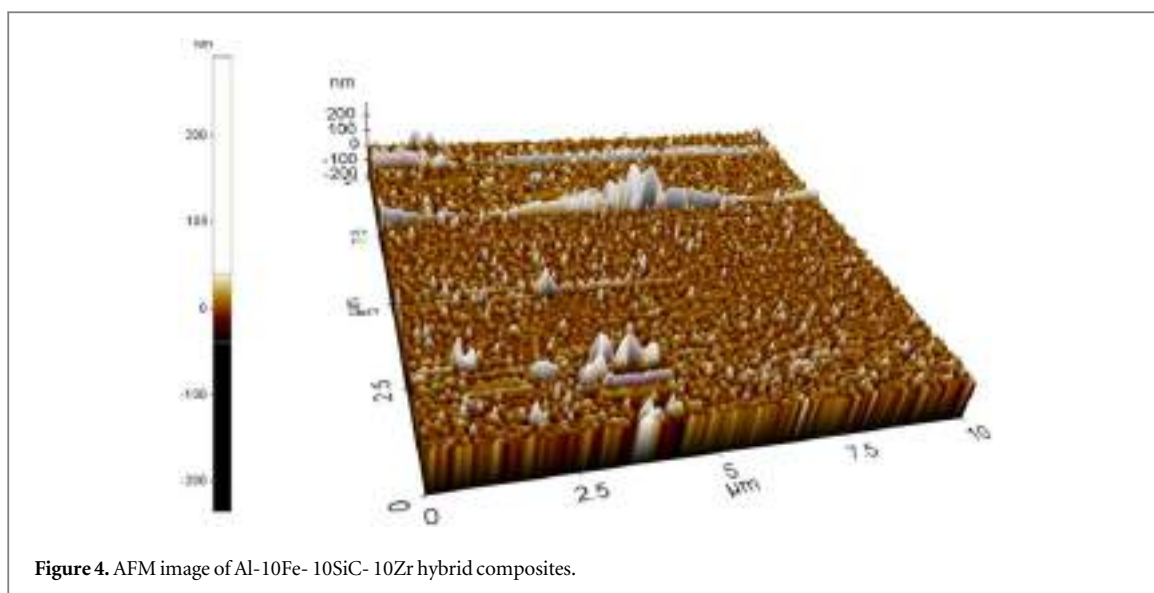


Figure 4. AFM image of Al-10Fe-10SiC-10Zr hybrid composites.

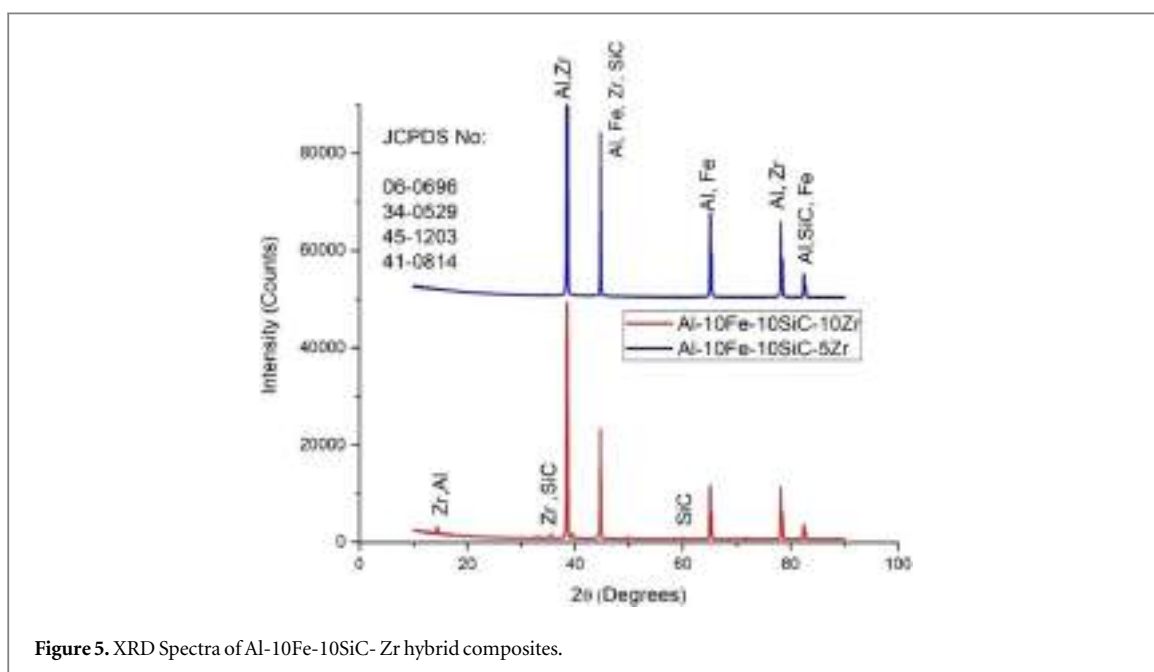


Figure 5. XRD Spectra of Al-10Fe-10SiC-Zr hybrid composites.

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.

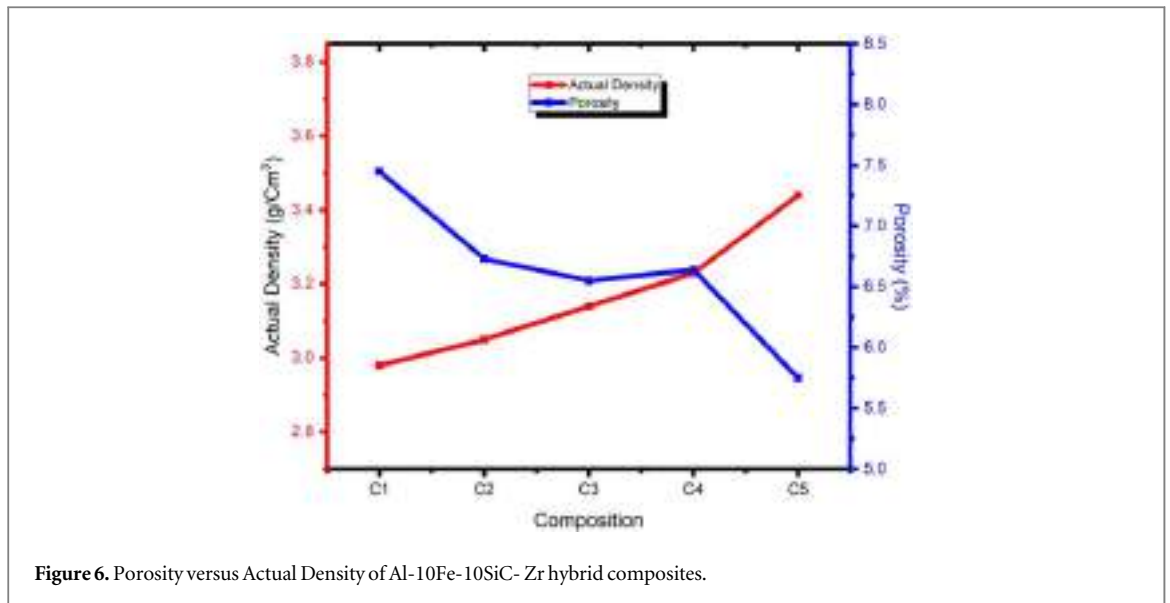


Figure 6. Porosity versus Actual Density of Al-10Fe-10SiC- Zr hybrid composites.

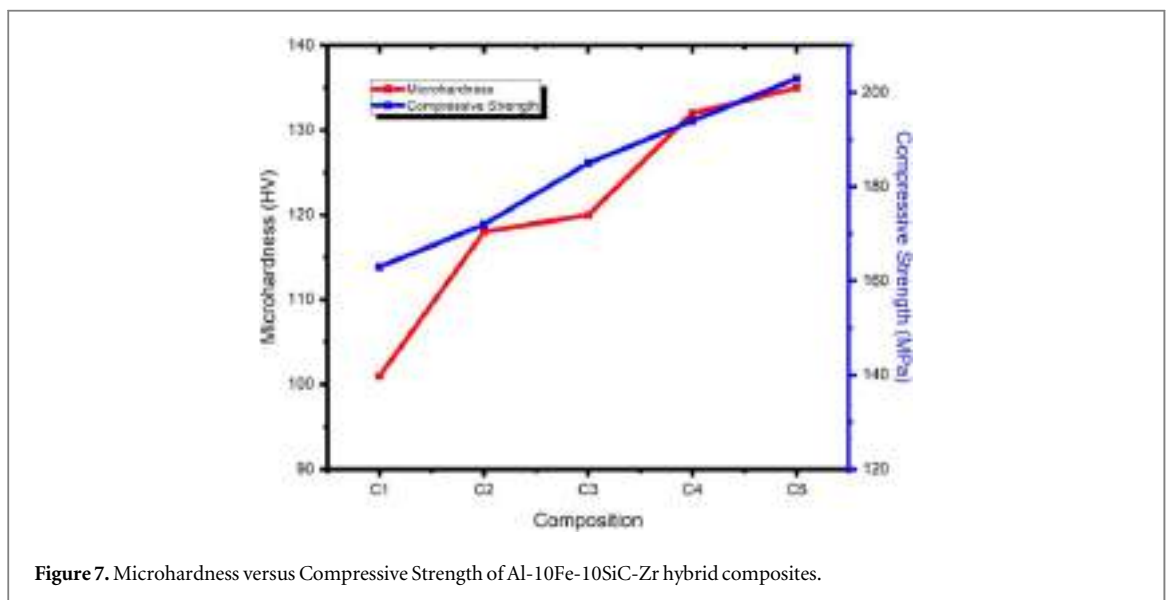


Figure 7. Microhardness versus Compressive Strength of Al-10Fe-10SiC-Zr hybrid composites.

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

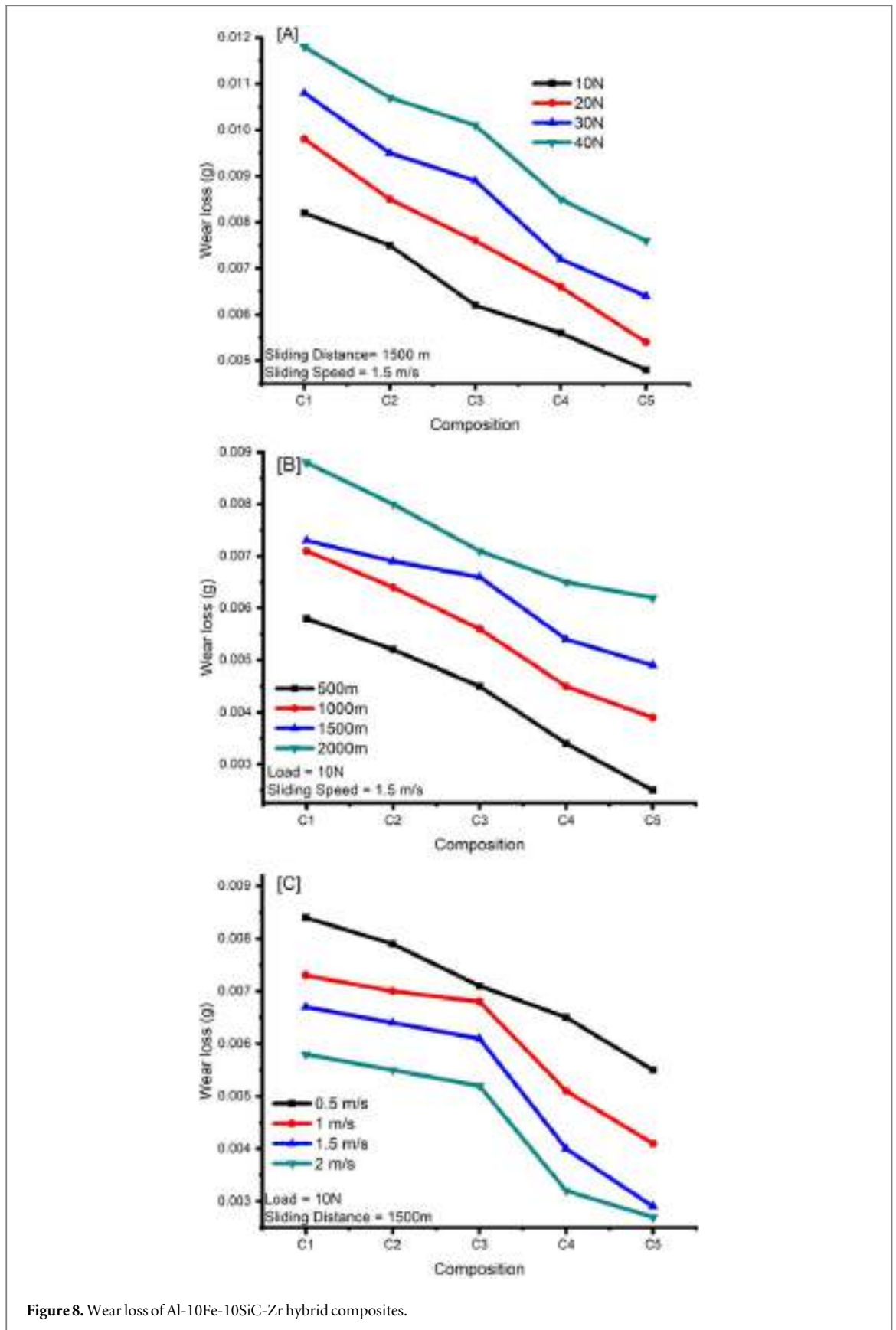
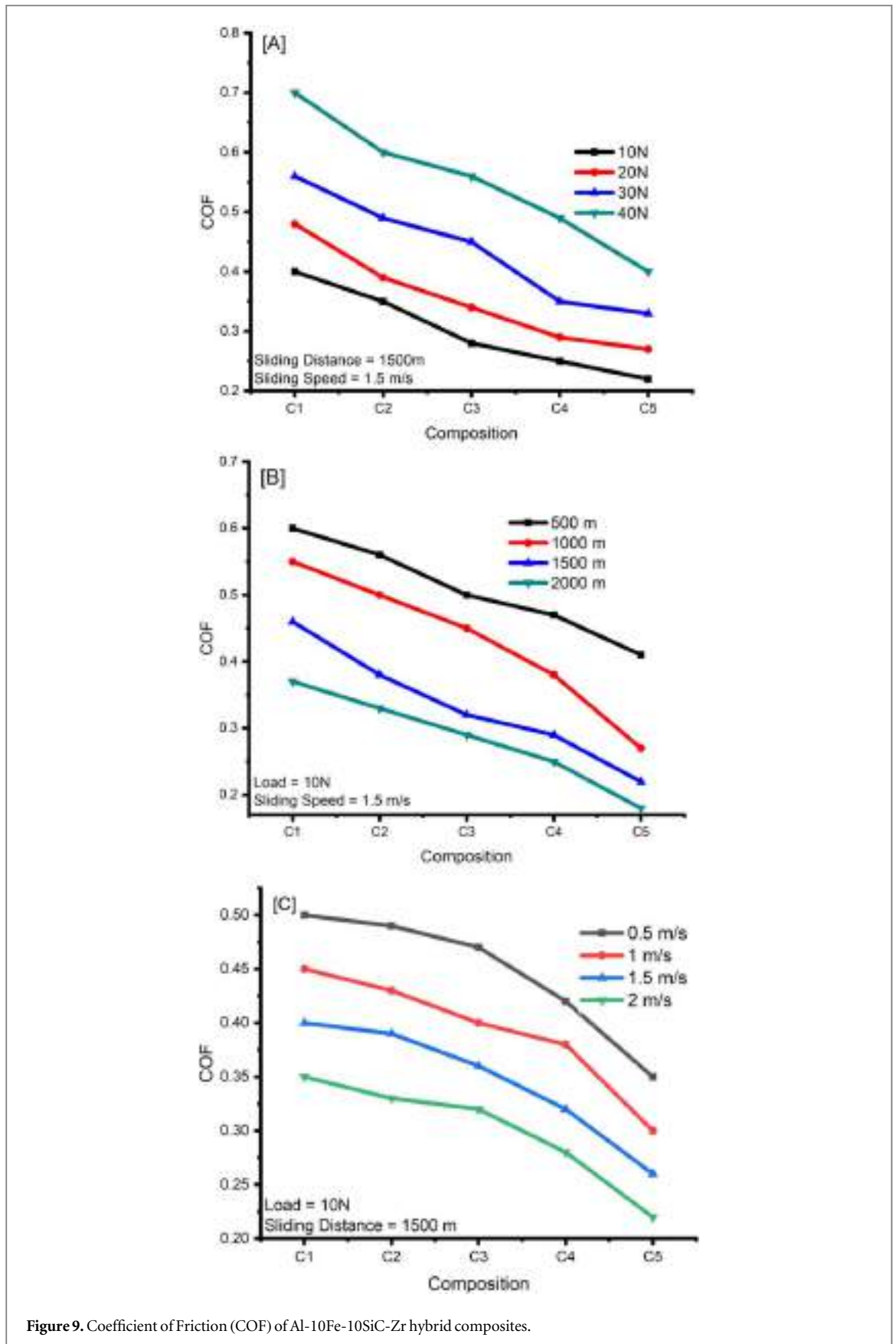


Figure 8. Wear loss of Al-10Fe-10SiC-Zr hybrid composites.

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



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

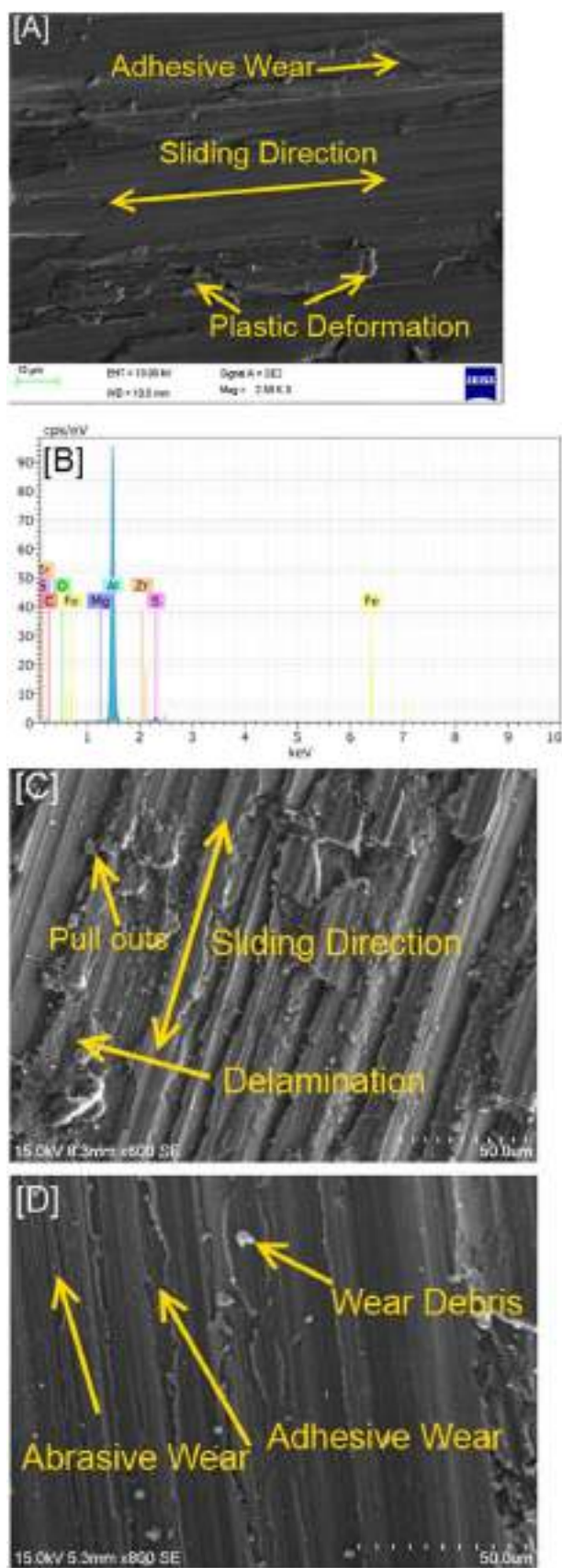


Figure 10. [A] FESEM image of Al-10Fe-10SiC-10Zr worn out Sample, [B] EDAX Spectra of Al-10Fe-10SiC-10Zr worn out Sample [C] SEM image of Al-10Fe worn out sample [D] SEM image of Al-10Fe-10SiC worn out Sample.

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

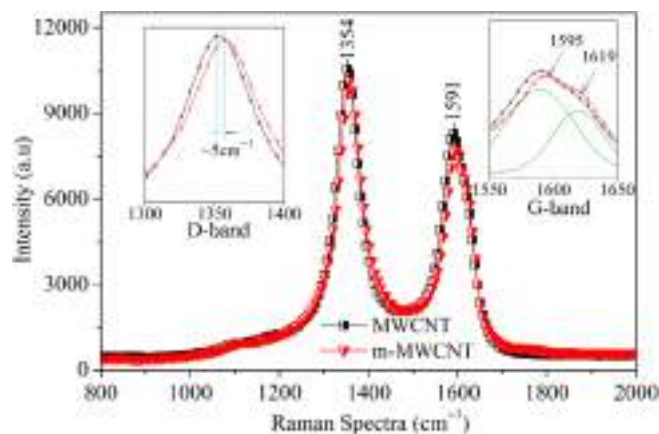


FIGURE 2 Raman spectra of pristine multiwalled carbon nanotube (MWCNT) and modified MWCNT [Color figure can be viewed at wileyonlinelibrary.com]

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

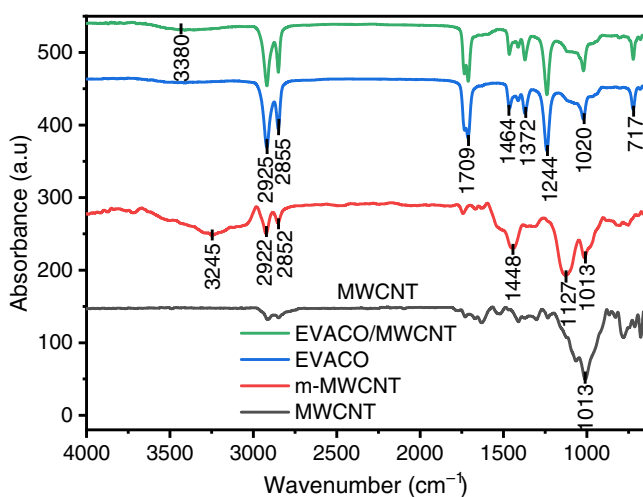
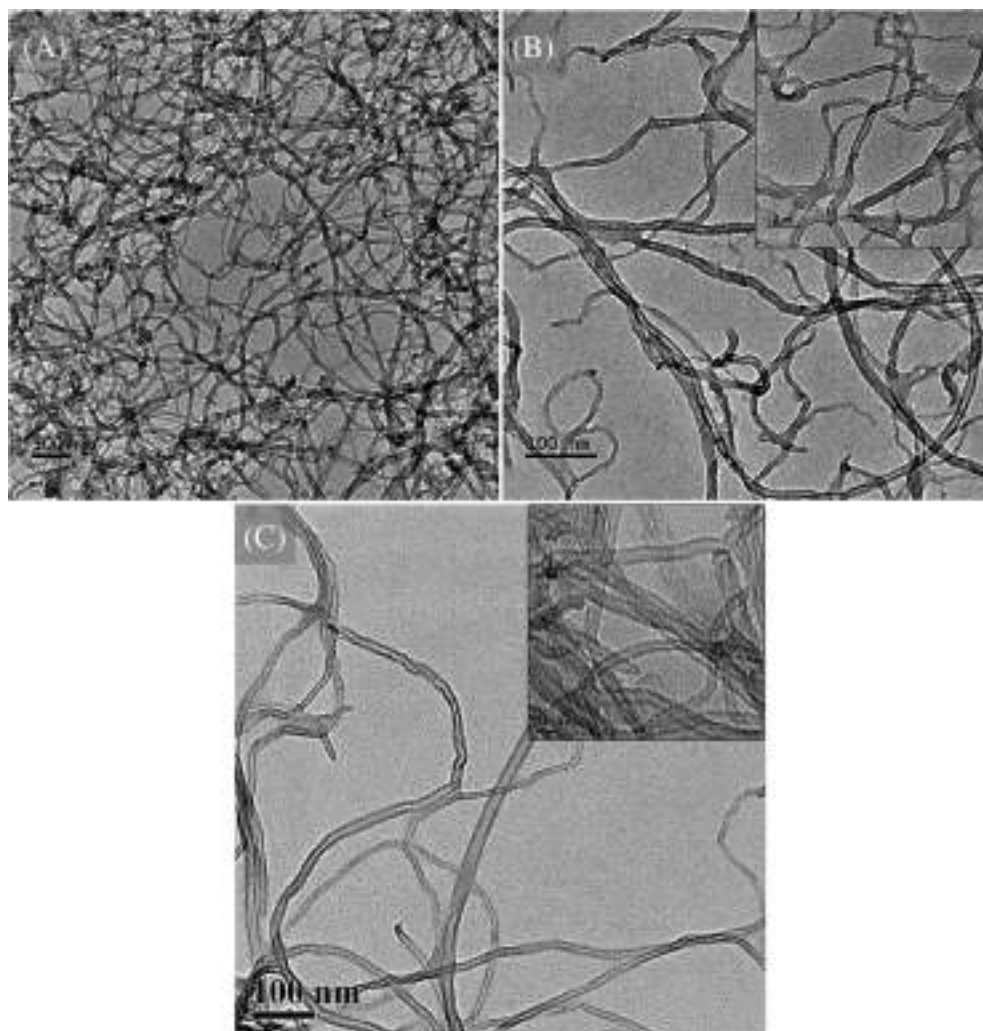


FIGURE 4 Fourier transform infrared spectra of pristine MWCNT, modified MWCNT, neat EVACO, and 0.25% m-MWCNT-loaded EVACO. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

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

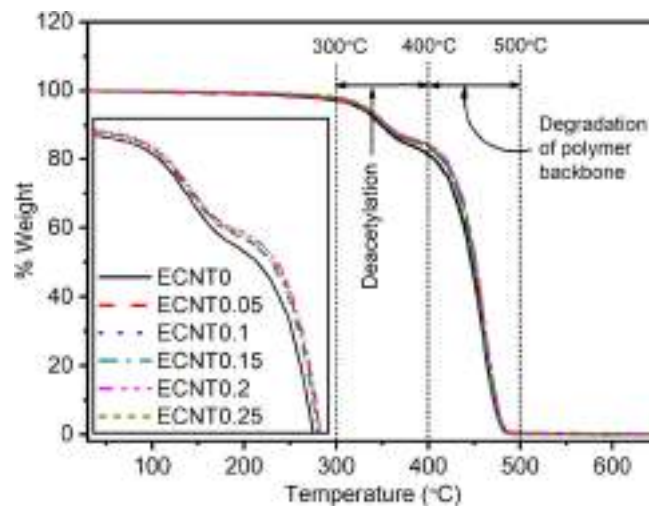


FIGURE 5 Thermogravimetric analysis results of pristine EVACO and EVACO/m-MWCNT (ECNT) composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

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

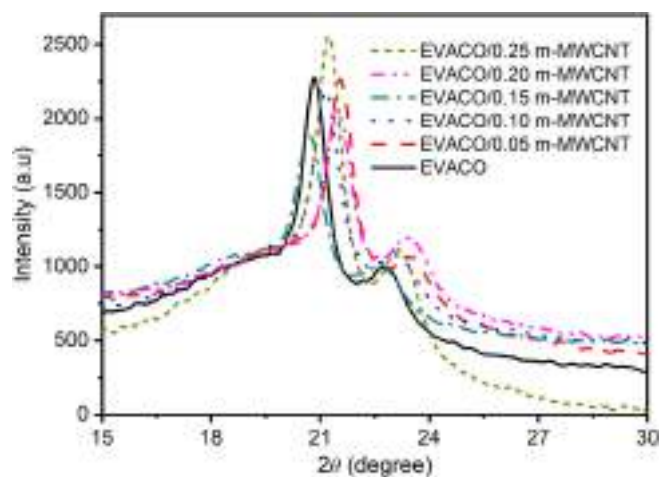


FIGURE 6 X-ray diffractograms of neat EVACO and EVACO/m-MWCNT composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

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.

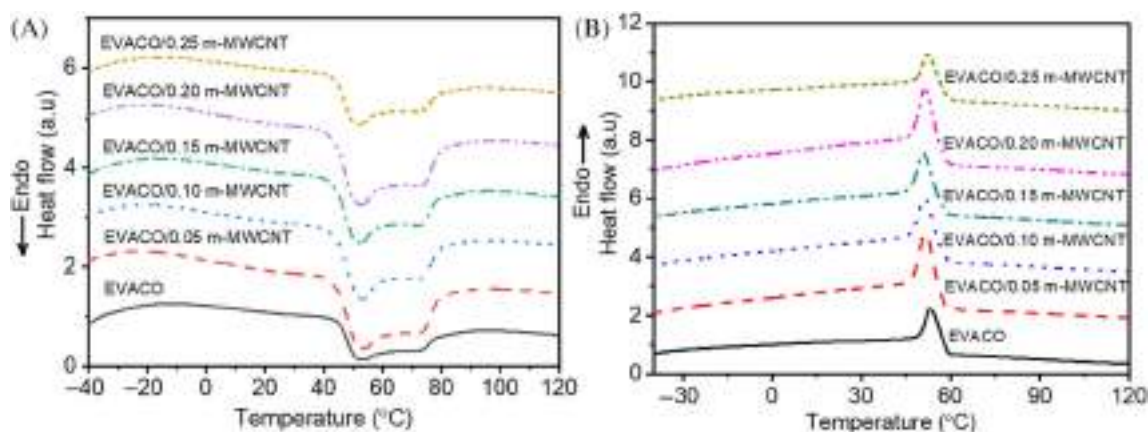


FIGURE 7 DSC curves of neat EVACO and EVACO/m-MWCNT composite (A) heating and (B) cooling. DSC, Differential scanning calorimetry; EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

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

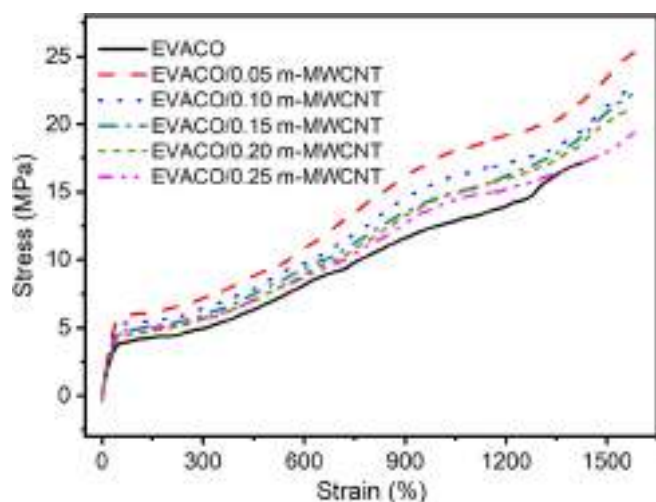


FIGURE 8 Stress versus strain curves of neat EVACO and EVACO/m-MWCNT composites. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube [Color figure can be viewed at wileyonlinelibrary.com]

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

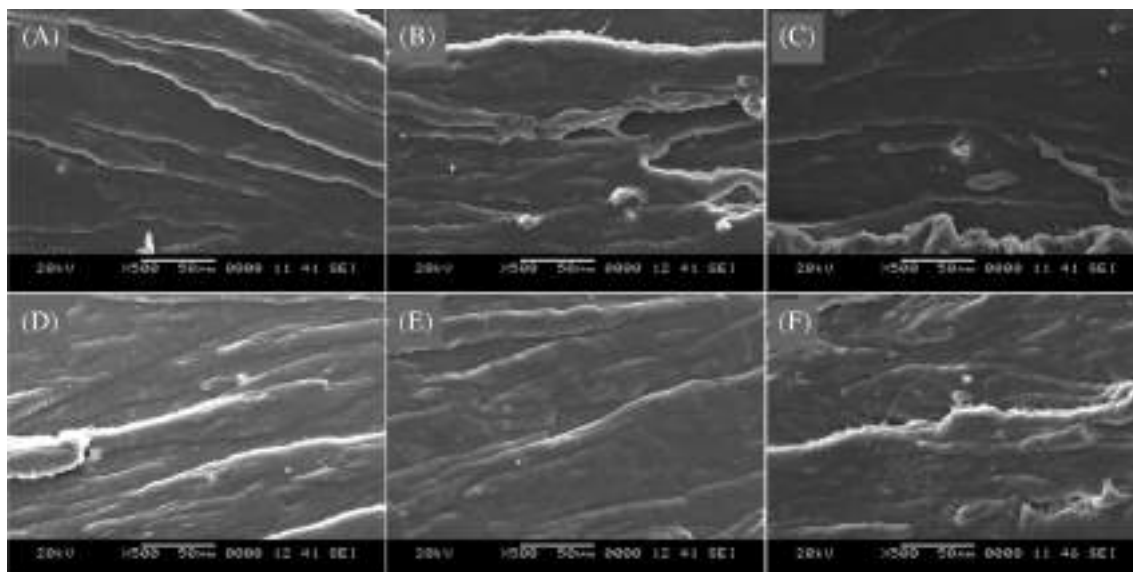


FIGURE 9 The scanning electron microscopy micrographs of EVACO with different m-MWCNT loadings: (A) 0%, (B) 0.05%, (C) 0.1%, (D) 0.15%, (E) 0.2%, and (F) 0.25%. EVACO, poly(ethylene vinyl acetate-co-carbon monoxide); m-MWCNT, modified-multiwalled carbon nanotube

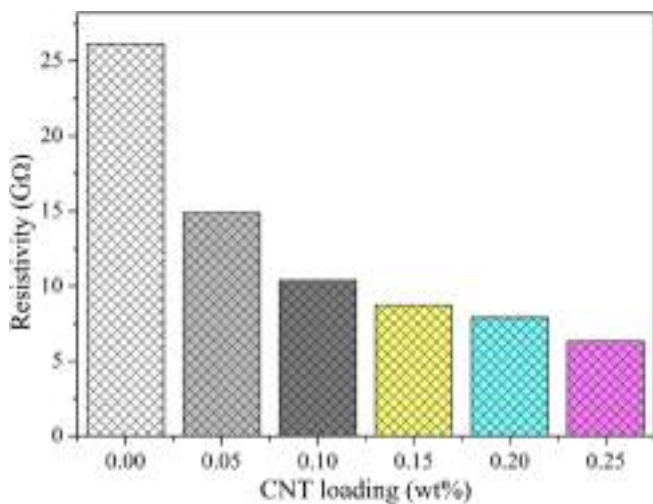


FIGURE 10 Electrical volume resistivity of the composites [Color figure can be viewed at wileyonlinelibrary.com]

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. Szczyzyk, 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>

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/352993955>

In-Line and Cross-Flow Response Interactions during Vortex Induced Vibration of Marine Risers

Article in MER - Marine Engineers Review · July 2021

CITATIONS

0

READS

96

2 authors:



Vidya Chandran

SCMS School of Engineering and Technology, Karukutty

24 PUBLICATIONS 31 CITATIONS

SEE PROFILE



Sheeja Janardhanan

Indian Maritime University

66 PUBLICATIONS 77 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Seaplanes [View project](#)



Bio-Inspired Propulsion Systems [View project](#)



In-Line and Cross-Flow Response Interactions during Vortex Induced Vibration of Marine Risers

Vidya Chandran¹, Sheeja Janardhanan²

¹Department of Mechanical Engineering, SCMS School of Engineering and Technology, Karukutty, Kerala, India

²School of Naval Architecture and Ocean Engineering, Indian Maritime University, Visakhapatnam, Andhra Pradesh, India

ABSTRACT

The paper presents a simplified method for understanding the interaction between in-line and cross-flow responses using computational fluid dynamics simulations. Interaction between the responses in the in-line and cross-flow directions in vortex induced vibrations of cylindrical risers in the marine environment is still not fully understood. The trends of variation of hydrodynamic and structural parameters as well as pattern of shedding have been determined numerically to understand the effect of the in-line degree of freedom as well on the riser response and hydrodynamic force coefficients and the results show that a single degree of freedom riser is more susceptible to lock in vibration.

KEYWORDS: Vortex Induced Vibration, In-line, Cross-Flow, Force Coefficients, Response

1. INTRODUCTION

Drilling riser is a pipe laid vertically from the oil well at the ocean bed to the offshore drilling platform. It conveys the drilling fluid and mud to and from the drill site. Marine drilling risers are used especially with floating rigs which are less stable and in particular cases where disconnection of the platform from the seafloor may be required quite often. **Figure 1** shows various layouts of marine risers depending on the constructional specification of platforms. **Figure 2** shows different

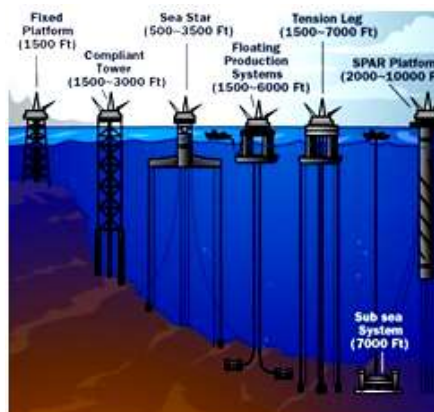


Figure 1 Constructional variation of offshore platforms with drilling depth [1]

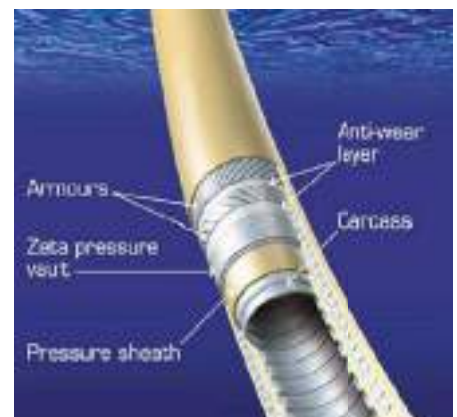


Figure 2 Cross section of a typical flexible riser [1]

Marine drilling risers are used especially with floating rigs which are less stable and in particular cases where disconnection of the platform from the sea floor may be required quite often

layers used in the construction of a flexible riser.

The marine risers, which are classified as **bluff bodies**, and when encountering fluid flow, alternate vortices are shed in the wake of the structure due to boundary layer separation. This alternate vortex shedding happens at a discrete frequency depending upon the flow Reynolds Number (Re). When the vortex shedding frequency matches with the natural frequency of the riser structure, it resonates with high amplitude of oscillation. These large amplitude vibrations, that occurs during "lock-in" of risers are catastrophic and needs to be arrested for the safety of crew working on the floating platforms and also for extending operational life of the risers. Vortex induced vibration (VIV) of marine risers poses all the challenges in the deployment and operation of marine risers.

There have been lot of research in the recent past to understand their behaviour under various sub-sea flow conditions. But most of the studies have concentrated on understanding the wake characteristics and estimating hydrodynamic loading and response of either stationary cylinder or cylinder with a single degree of freedom (1DOF)[2].

Few results have been reported for study of hydrodynamic response of cylinder with two degrees of freedom (2DOF) in both in-line (IL) and cross-flow (CF) directions. IL vibrations have significant impact on the shedding pattern and also on the amplitude of CF vibrations [3].

The first of its kind discussions were reported in the case of flow around cylinder with 2DOF [4]. They established the effect of reduced velocity (U_r) on the effect of forced and free 2DOF response [4]. The effect of IL response on CF response depends on the ratio of natural frequencies in both the directions

$$\left(\eta_b = \frac{f_{n,IL}}{f_{n,CF}} \right)$$

During lock in, if the natural frequency in the IL direction is twice that in the CF direction, resonance occurs in both directions leading to premature failure of the riser [5]. Also it has been observed that IL response amplitude is a function of U_r and stability parameter, whereas the CF response amplitude is a function of U_r and flow velocity [6]. Wake characteristics, hydrodynamic force coefficients and response vary significantly when both IL and CF vibrations occur simultaneously. Hence there is a need for prediction of response that hold good for the combined IL and CF vibration.

2. PROBLEM DESCRIPTION

In the present paper a riser model with outer diameter 0.076 m has been numerically analysed using two dimensional (2D) computational fluid dynamics (CFD). Specifications of the riser and the flow condition in listed in **Table 1**. The incoming flow velocity is fixed as 0.5 m/s to maintain the flow regime uniform at $Re = 3.8 \times 10^4$ which corresponds to the ocean condition encountered by a real marine riser used for petroleum extraction in offshore industries [7]. In this paper an effort has been made to study the effect of IL vibration on the amplitude of CF vibration and also on the wake characteristics.

2.1. Mathematical Model

The riser has been modelled as a 2D cylinder with 2DOF in the CF and IL directions. The equations of motion for the riser can be represented as Eq. (1) and (2)

$$m\ddot{Y} + c\dot{Y} + kY = F_L(t) \tag{1}$$

$$m\ddot{X} + c\dot{X} + kX = F_D(t) \tag{2}$$

Where Y is the displacement in CF direction and X is the displacement in

Properties	Values	Units
Diameter (D)	0.076	m
Aspect ratio (L/D)	13.12	-
Flow velocity (V)	0.5	m/s
Reynolds Number of flow (Re)	3.8×10^4	-
Mass ratio (m^*)	0.66	-

Table 1 Riser model specifications and flow characteristics

the IL direction. The excitation forces are lift force, $F_L(t)$ and drag force, $F_D(t)$. The excitation forces are periodic in nature due the alternate shedding of vortices, which causes the riser to oscillate in CF as well as IL directions. The riser is observed to oscillate with frequency equal to frequency of vortex shedding (f_v) in the CF direction and at double the frequency in the IL direction during lock in. Lock in can be defined as the resonance condition during which the vortex shedding frequency lock on to the natural frequency of the riser in the cross flow direction. A simple representation of the mathematical model of riser with 2DOF is represented in **Figure 3**.

The riser is modelled with zero structural damping in the CF and IL directions. k_x and k_y are stiffness coefficients in the IL and CF directions respectively. In the present study $k_x = k_y$. For such a specific case the natural frequencies in both directions will be same and hence $\eta_b = 1$

2.2. Fluid Domain Extends

Figure 4 (a) shows the computational domain for the CFD simulation of VIV of an elastically mounted cylinder with 2DOF. The origin of the Cartesian coordinate system is located at the centre of the cylinder. The length of the

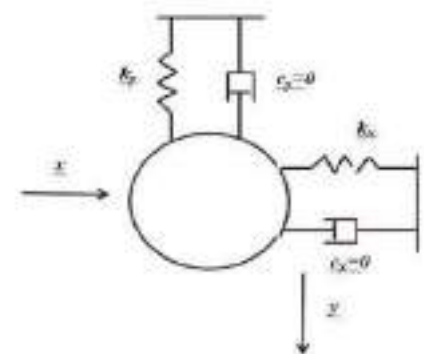


Figure 3 Representation of mathematical model of riser with 2DOF

domain is $40D$ with the cylinder located at $10D$ away from the inlet boundary. The cross flow width of the domain is $20D$ with the centre of the cylinder at the middle.

Detailed views of the mesh around the cylinder along with the computational domain after meshing have been shown in Figure 4 (b) and (c) respectively. There are 307 nodes around the circumference of the cylinder and the minimum element size near the rigid wall boundary has been computed from boundary layer theory to be $0.0001D$ [8].

The non-dimensional element size represented as y^+ , next to the cylinder surface is found to be less than unity. For cylinder wall a no slip boundary condition has been applied assuming the surface to be smooth. Inlet boundary has been treated as velocity-inlet with inflow velocity, $V = 0.5$ m/s. Outlet boundary has been treated as pressure outlet, the gradients of fluid velocity are set to zero and the pressure with zero reference pressure. On the two transverse boundaries, symmetry boundary condition has been applied. Grid independency study has been carried out for the present grid in the previous work done by the authors [9].

2.3. Flow Model

Numerically this problem has been treated as a case of two-way fluid structure interactions (2way FSI). Modeling and meshing has been performed in ANSYS ICEM CFD and solving using ANSYS FLUENT. Flow around the cylinder is modeled using the transient, incompressible Reynolds Averaged Navier-Stokes equation (RANSE) based solver with $k-\omega$ SST as the turbulence model. RANSE solver does the virtual averaging of velocities over an interval of time and hence for a specific interval, the velocity vector appears to be constant in a RANSE solver. In the present work an optimised fine grid is used to compensate for this drawback of the solver enabling it to capture the physics of Von-Karman Street eddies.

The governing equations are discretised using finite difference method. Non iterative time advancement (NITA) scheme with fractional time stepping method (FSM) has been chosen for pressure-velocity coupling of the grid. A least-squares- cell (LSC) based scheme has been used for gradient in spatial discretisation and a second order upwind scheme as convective scheme.

2.4. Structural Model

An elastically mounted cylinder can be mathematically represented by Eq. (1) and (2). These equations of motion are solved using a six degrees of freedom solver (6DOF), an integral part of the main solver by defining the cylinder as an object with 2DOF in transverse direction. A user defined function (UDF) compiled in C programming language has been hooked to the cylinder dynamic boundary conditions. The governing equations for the motion of the centre of gravity of the cylinder in the CF and IL directions are solved in the inertial coordinate system. Velocity in the CF and IL directions are obtained by performing integration on Eq. (3) and (4).

$$\ddot{x} - \frac{1}{m} \sum F_x \quad (3)$$

$$\ddot{y} - \frac{1}{m} \sum F_y \quad (4)$$

Where \ddot{x} and \ddot{y} , are accelerations in the IL and CF direction respectively, m is the mass of the cylinder and F , resultant fluid force acting on the cylinder in the respective direction. Position of the centre of gravity of the cylinder (CG) is updated after solving the equations of motion of a spring mass system represented by Eq. (1) and (2). Mass of the cylinder is given in the UDF as in Eq. (5) and (6).

$$m = m_b + m_a \quad (5)$$

$$m_a = (1 + C_a)m_b \quad (6)$$

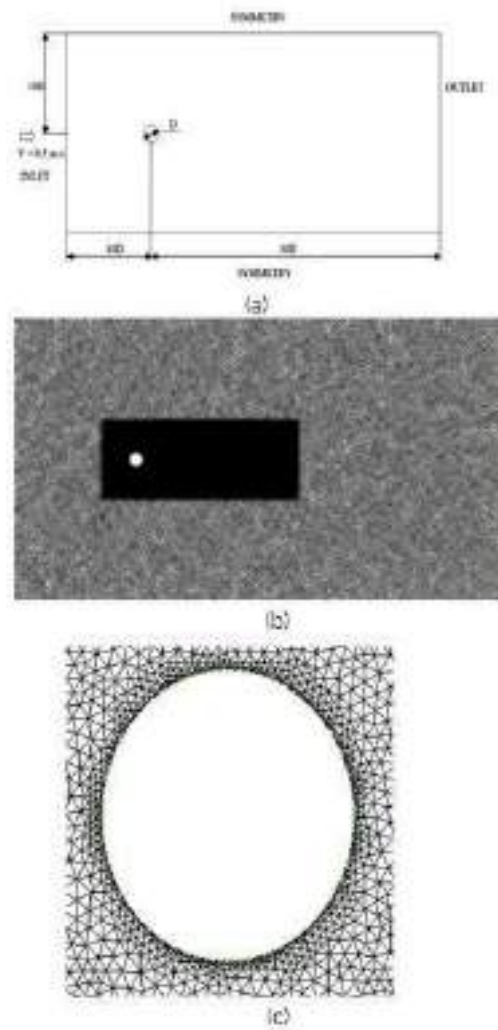


Figure 4 (a) Computational domain
(b) computational mesh
(c) mesh around the cylinder

Where m_a is the added mass and m_b is the mass of the cylinder. Added mass coefficient C_a for the aspect ratio of the present model is found to be equal to 1.0 [10].

Analysis has been performed assigning the cylinder 2DOF with $k_x = k_y$ so that the natural frequencies of the cylinder in both directions remain equal. The results are compared with the case when the cylinder has only 1DOF in the CF direction. Amplitudes of CF response are compared with existing results [9] and also the shedding patterns in both cases have been analysed.

3. RESULTS AND DISCUSSIONS

From the numerical analysis of cylinder with 2DOF it has been observed that the hydrodynamic force coefficient in the CF direction, C_L shows an increase of 17.4% than that for 1DOF case. This result is comparable with the findings of previous research in the field which shows an increase in the lift coefficient value by permitting an extra degree of freedom [11].

RMS value of C_D is constant for both cases with a very small decrease of 4% with 2DOF case. C_L oscillates about zero with almost equal frequencies for both the cases. But the frequency of oscillation of C_D is lesser by 7.2% for 2DOF case. The values of important hydrodynamic and structural parameters of both cases are shown in **Table 2**.

The non-dimensional amplitude in the CF direction obtained with 2DOF is 11.3% more than that with 1DOF. X/D is approximately 0.2. Time histories of major parameters obtained from the 1DOF analysis are shown in **Figures 5(a) – (d)** and that for 2DOF in **Figures 6(a) – (d)**. Frequency of oscillation of the cylinder in the CF direction obtained from 1DOF case is found to be more closer to the theoretical value of vortex shedding frequency obtained from the normal value of $St = 0.2$ ($f_v=1.3$). For the 2DOF case, the frequency of oscillation deviates from the vortex shedding frequency.

For 2DOF case, the frequency of oscillation of C_L and the oscillation frequency of cylinder in the CF direction remains same. In 1DOF case, C_L oscillation frequency remains same as that in the 2DOF case, but the cylinder vibration frequency in the CF shifts towards the natural frequency of cylinder in CF direction.

In the present analysis, the natural frequency in both directions are specifically fixed to be equal to the theoretical value of vortex shedding frequency. Hence the phenomenon can be looked upon as the lock-in of vortex shedding frequency on to the natural frequency of the cylinder. It can be concluded that a cylinder with 1DOF is more prone to lock in vibration compared to that with 2DOF.

This observation can be related to the shifting of the vortex shedding pattern from two singles (2S) to two pairs (2P) mode when motion in IL direction is arrested. The shedding patterns for 1DOF and 2DOF cases are shown in **Figure 5(a)** and **Figure 6(a)** respectively. St obtained also is with the range of normal value for cylinders during lock in. Even though the values of C_D for both cases are almost same, the oscillating frequency varies significantly.

The trajectory of oscillation of cylinder in 2DOF case is represented in **Figure 7**. A clear eight figure trajectory is observed which is typical for VIV of cylinders [11]. Also it has been observed that the motion the IL direction lags behind that in CF direction by a phase angle 30° . The represented trajectory in **Figure 7** corresponds to 30° phase lag [12].

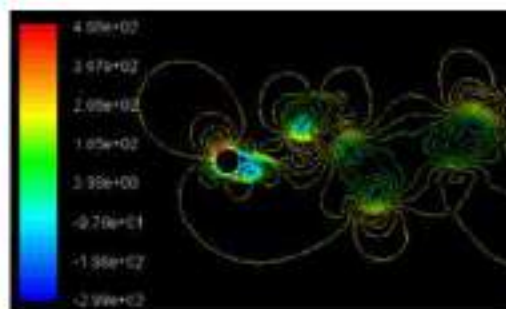
4. CONCLUSIONS

Accounting for an additional degree of freedom seems to have significant effect on the magnitude of lift coefficient but the frequency of oscillation of C_L remains constant for both the cases. C_D is independent of the degree of freedom of the cylinder but the frequency of oscillation varies significantly. Oscillation amplitude of the cylinder in the CF direction is more in 2DOF case which can be related to the increase in C_L .

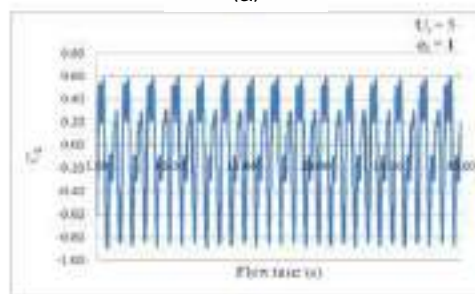
It has been clearly observed that with 1DOF, the cylinder is more susceptible to lock in vibration since the vortex shedding frequency locks on to the natural frequency of the cylinder in the CF direction. But with 2DOF, no such shifting of frequency is observed. Shedding pattern shifts from 2S during 2DOF motion to 2P when motion in IL direction is arrested. An eight figure trajectory typical for VIV is obtained from the

Parameters	1DOF	2DOF ($\eta_b = 1$)
C_L	0.57	0.69
C_D	1.49	1.43
$f_{osc\ C_L}(f_v)$	1.16	1.14
$f_{osc\ C_D}$	2.5	2.32
f_{CF}	1.26	1.15
f_{IL}	-	3.28
St	0.18	0.17
Y/D	1.06	1.2
X/D	-	0.17

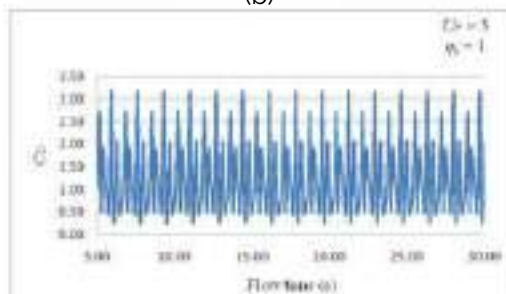
Table 2 Hydrodynamic and structural parameter off cylinder with 1DOF and 2DOF



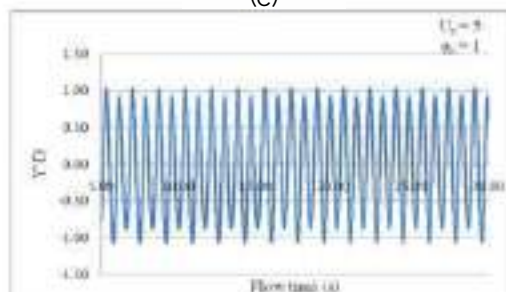
(a)



(b)



(c)



(d)

Figure 5 Pressure contours and Time histories of various hydrodynamic and structural parameters (a) Vortex shedding pattern behind cylinder with 1DOF showing 2P mode (b) C_L of cylinder with 1DOF (c) C_D of cylinder with 1DOF (d) Motion history of cylinder with 1DOF

It can be concluded that a cylinder with 1DOF is more prone to lock in vibration compared to that with 2DOF

2D simulation. Hence the efficacy of 2D CFD as a tool to predict response of cylinder with 2DOF under VIV is accomplished. The observations made above are definitely strong inputs in the design deployment and operation of marine risers.

5. REFERENCES

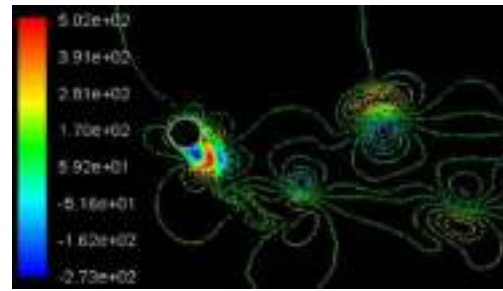
1. Bai, Y., and Qiang B., *Subsea engineering handbook*. Gulf Professional Publishing, Vol. 1, pp. 28, 2018.
2. A. Khalak, C. H. K. Williamson, 'Investigation of the relative effects of mass and damping in vortex induced vibration of a circular cylinder', *Journal of Wind Engineering. Ind. Aerodyn.* vol. 69-71, pp. 341 – 350, 1997.
3. Jauvtis, N. and Williamson, C. H. K., 'The effect of two degrees of freedom on vortex-induced vibration at low mass and damping', *Journal of Fluid Mechanics* 509, 23 – 62, 2004.
4. Moe, G. and Wu, Z. J., 'The lift force on a cylinder vibrating in a current', *Journal of Offshore Mechanics and Arctic Engineering* 112, 297- 303, 1990.
5. T. Sarpkaya., 'Hydrodynamic damping, flow-induced oscillations, and biharmonic response', *ASME Journal of Offshore Mechanics and Arctic Engineering*, 117:232-238, 1995.
6. Yin, D., Experimental and Numerical Analysis of Combined In-line and Cross-flow Vortex Induced Vibrations. *Ph D Thesis, Norwegian University of Science and Technology* pp 7, 2013.
7. Narendran, K., Murali, K., Sundar, V., 'Vortex-induced vibrations of elastically mounted circular cylinder at Re of the O (10^5)', *Journal of Fluids and Structures*, 54, 503 – 521, 2015.
8. Schlichting, H., 'Boundary layer theory', *McGraw-Hill Book Company, New York*, 1979.
9. Chandran, V., Sekar, M., Janardhanan, S., Menon, V., 'A numerical study on the influence of mass and stiffness ratios on the vortex induced motion of an elastically mounted cylinder for harnessing power', *Energies*, 11, 2580, 2018.
10. Naudascher, E.; Rockwell, D., 'Flow induced vibration – An engineering guide', *Dover Publications Inc, Mineola, New York, USA*, 2005.
11. Williamson, C.H.K., Govardhan, R., 'Vortex induced vibrations' *Annual review of fluid mechanics*, vol. 36 pp 413 – 455, 2004.
12. W, Jie., L, Halvor., M, L, Larsen., L, Stergios., B, Rolf. 'Vortex-induced vibration of a flexible cylinder: Interaction of the in-line and cross-flow responses', *Journal of Fluids and Structures* 63 238–258, 2016.

ABOUT THE AUTHORS

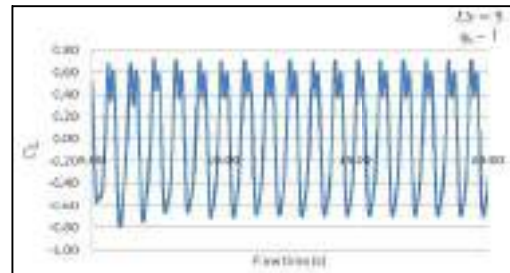
Dr. Vidya Chandran is Associate Professor in the Department of Mechanical Engineering, SCMS School of Engineering and Technology, Cochin. She has a PhD in Vortex Induced Vibrations from Karunya University, India. Her research interests include Vortex Induced Vibrations, Marine Clean Energy and Under Water Robotics.

Dr. Sheeja Janardhanan is Associate Professor in the School of Naval Architecture and Ocean Engineering, Indian Maritime University, Vishakhapatnam, India. She has a PhD in Numerical Ship Hydrodynamics from Indian Institute of Technology Madras, India. Formerly she worked as Professor and Head, Department of Mechanical Engineering, SCMS School of Engineering and Technology, Ernakulam and also as Surveyor in the Research and Rule Development Division of Indian Register of Shipping, Mumbai, India. Her research interests include controllability of surface ships, underwater robotics, vibrations of risers and computational fluid dynamics.

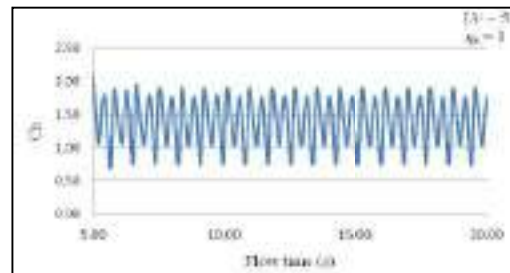
Email: sheejaj@imu.ac.in



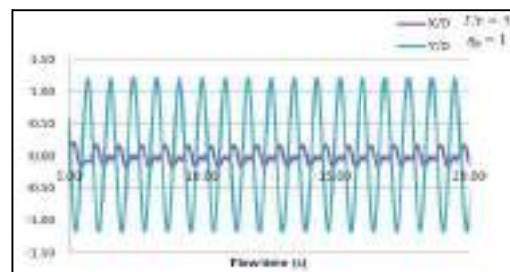
(a)



(b)



(c)



(d)

Figure 6 Pressure contours and time histories of various hydrodynamic and structural parameters (a) Vortex shedding pattern behind cylinder with 2DOF showing 2S mode (b) C_L of cylinder with 2DOF (c) C_D of cylinder with 2DOF (d) Motion history of cylinder with 2DOF

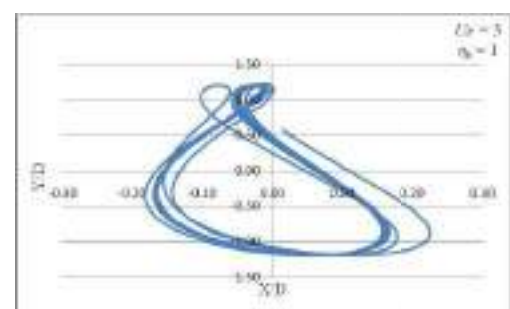


Figure 7 Trajectory of the cylinder with 2DOF motions under VIV

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 prominently displayed in the upper 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.

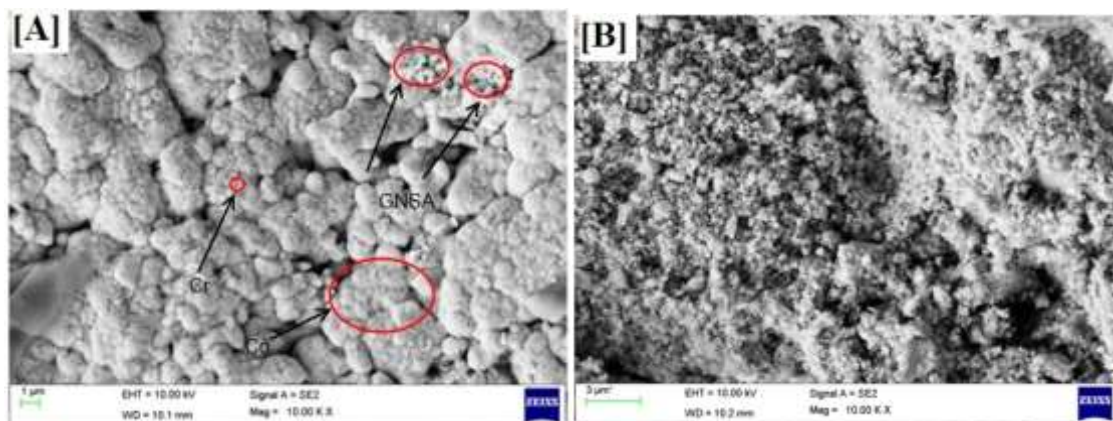


Figure 1.FESEM images of Co-10Cr- 3.5 GNSA & Co-10Cr-5GNSA hybrid Composite.

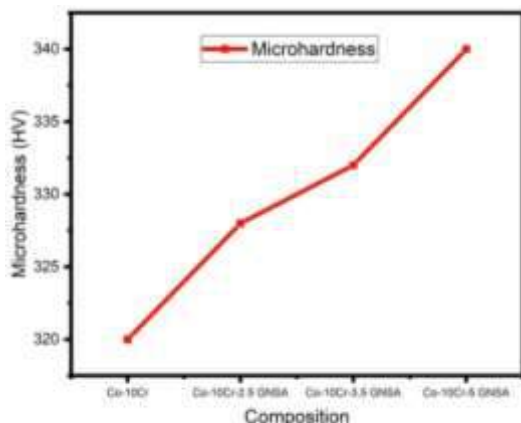


Figure 2.Graphical Representation of Co-10Cr-GNSA hybrid composites.

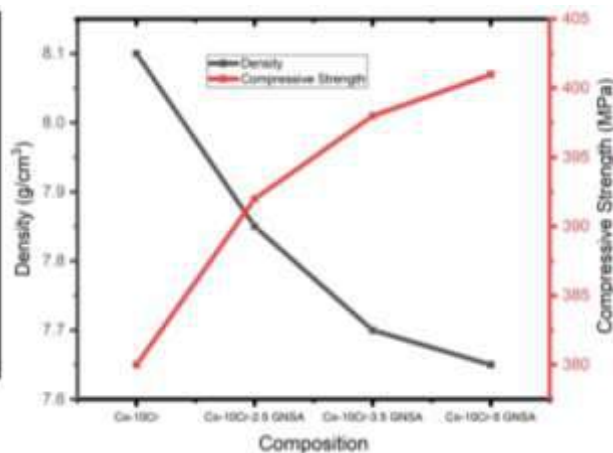


Figure 3.Comparison of Density and Compressive Strength of the Co-10Cr-GNSA hybrid composites.

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.

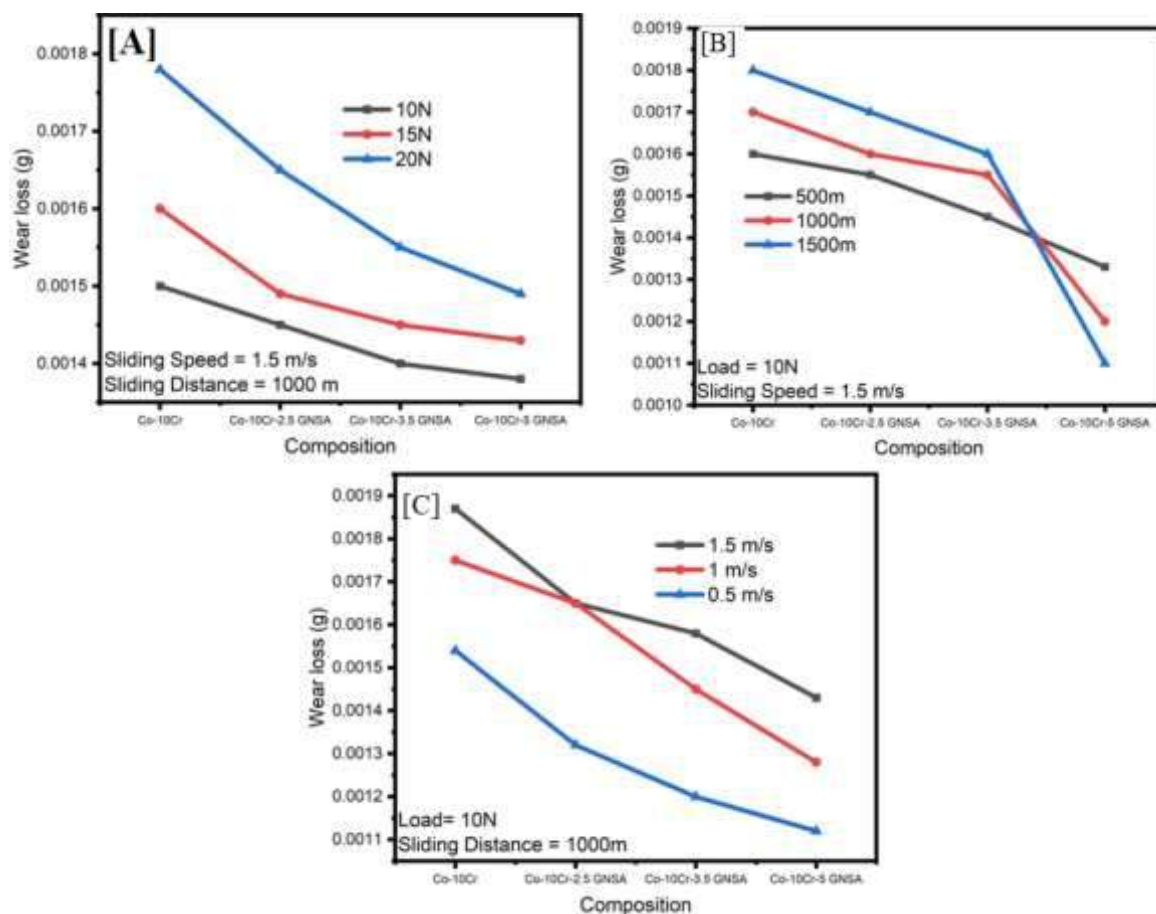


Figure 4. Wear Loss plot of Co-10Cr-GNSA hybrid composites.

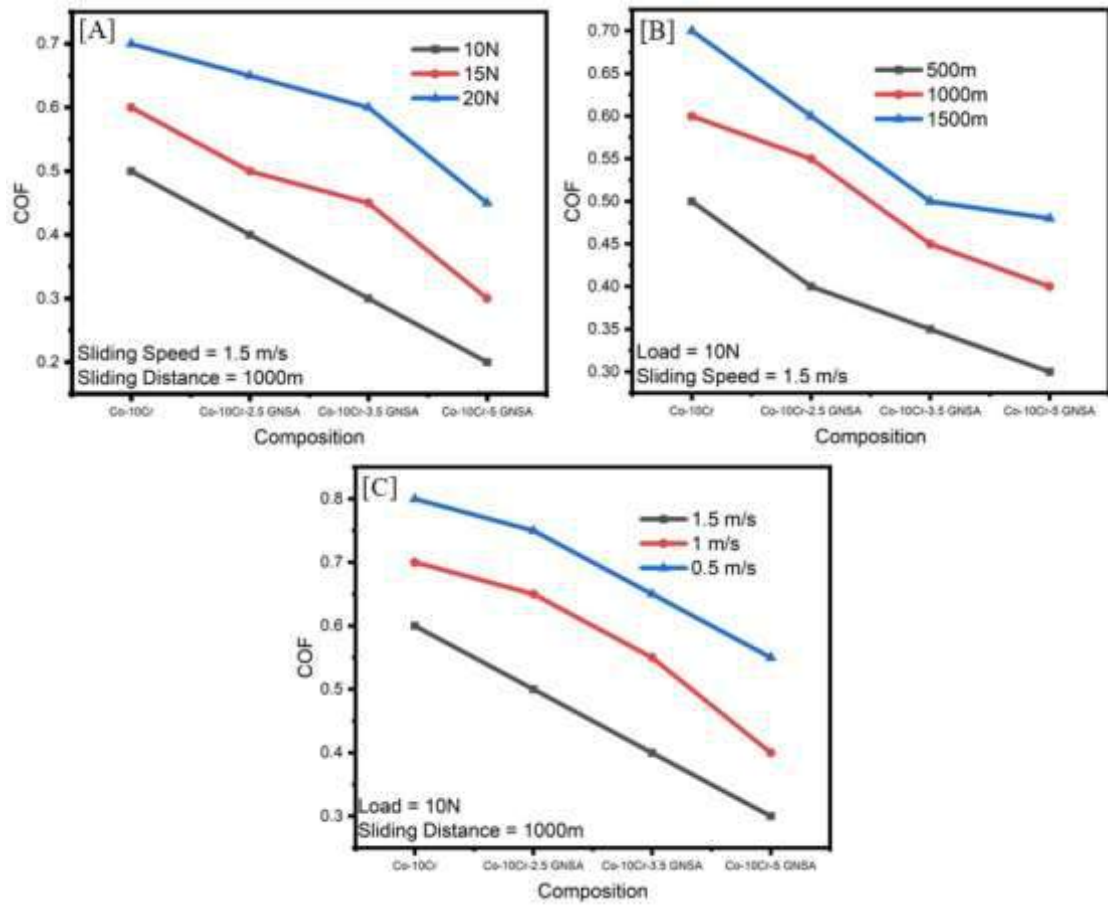
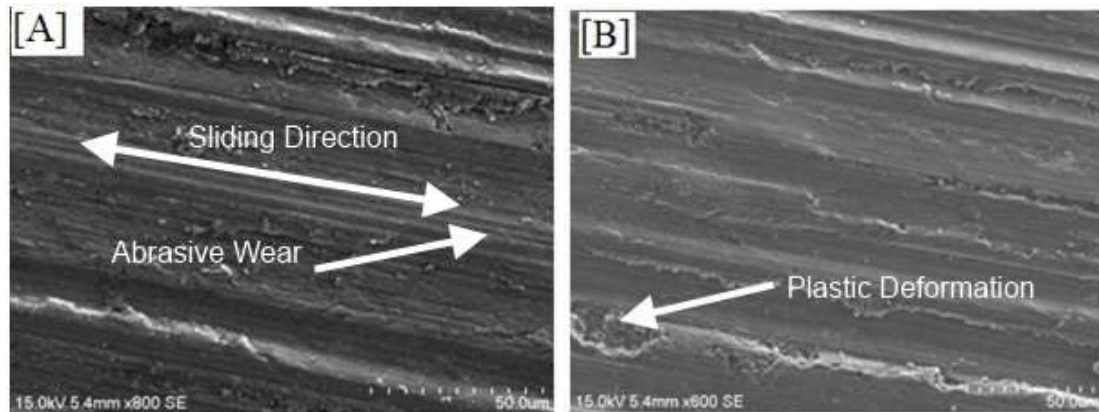


Figure 5. COF plot of Co-10Cr-GNSA hybrid composites.



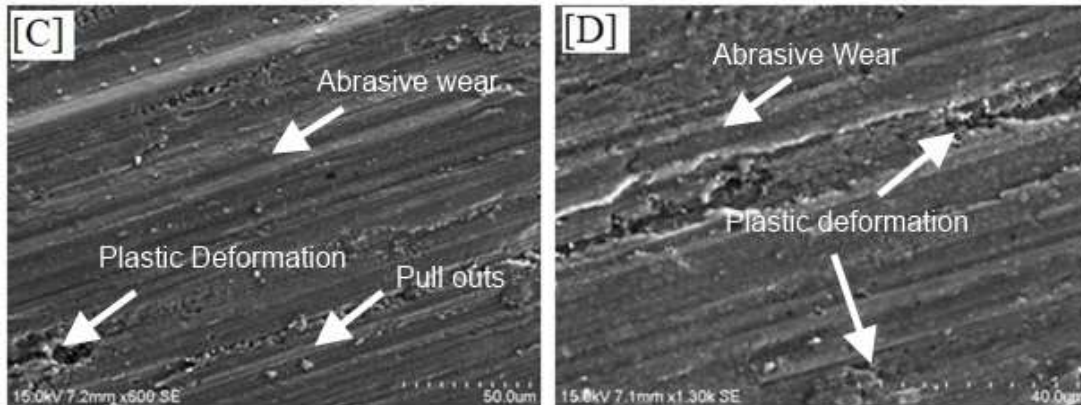


Figure 6.Worn out Surface analysis of Co-10Cr-GNSA hybrid composites.

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

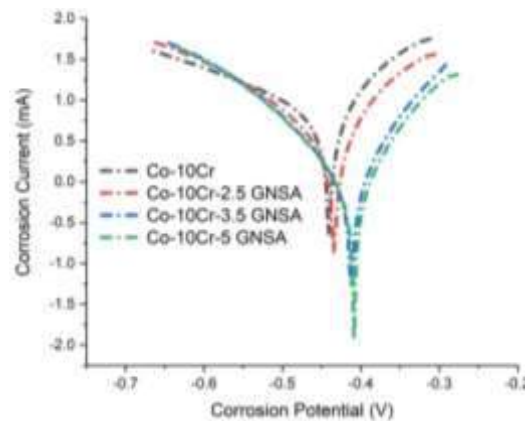


Figure 7.Potentiodynamic polarization plot of Co-10Cr-GNSA hybrid composites

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, Knörnschild 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) (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

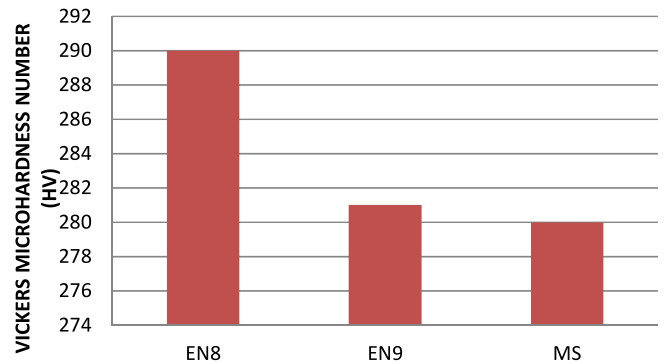


Fig. 5. Variation of hardness with respect to the material.

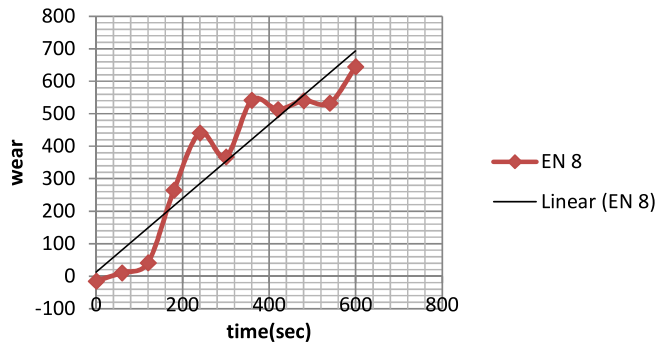


Fig. 2. Variation of wear with respect to time for EN8.

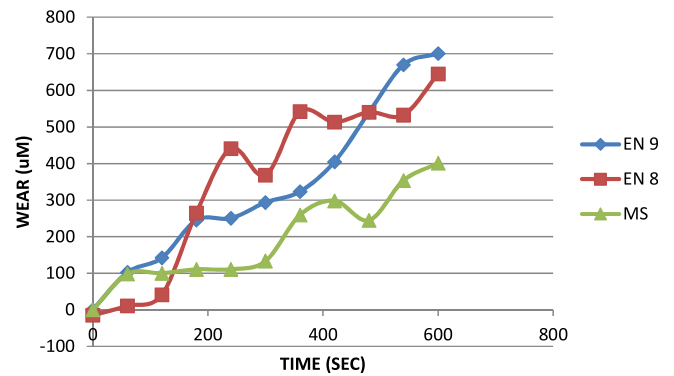


Fig. 6. Comparison of wear for EN 8, EN 9, MS.

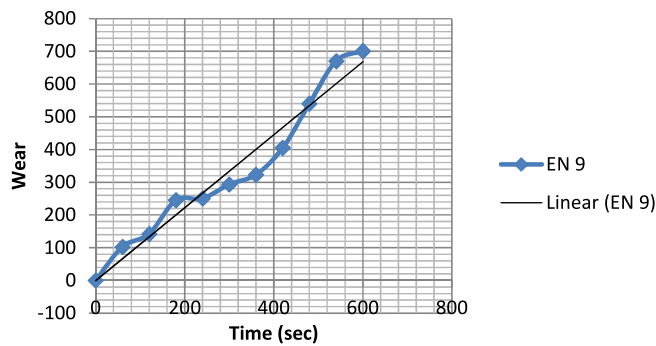


Fig. 3. Variation of wear with respect to time for EN9.

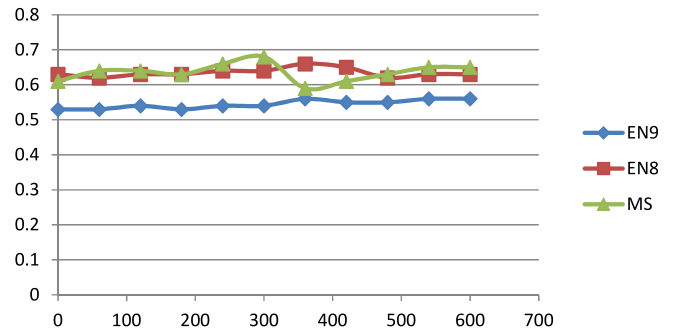


Fig. 7. Comparison of frictional coefficients for the materials..

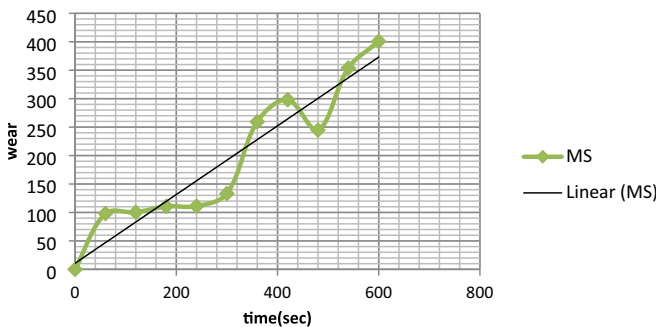


Fig. 4. Variation of wear with respect to time for EN24.

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. Microstructure 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 Quantimet 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

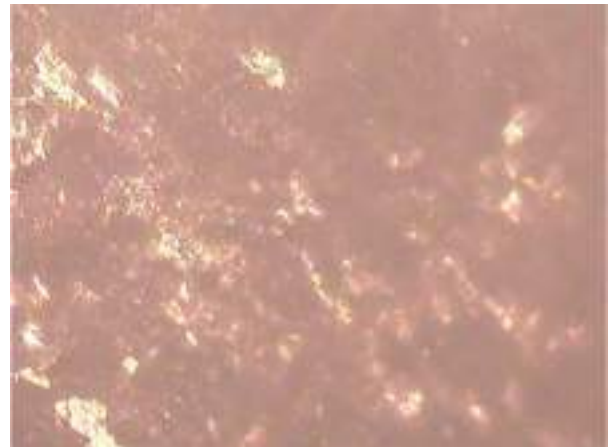


Fig. 12. Microstructure of worn surface of EN 9.



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

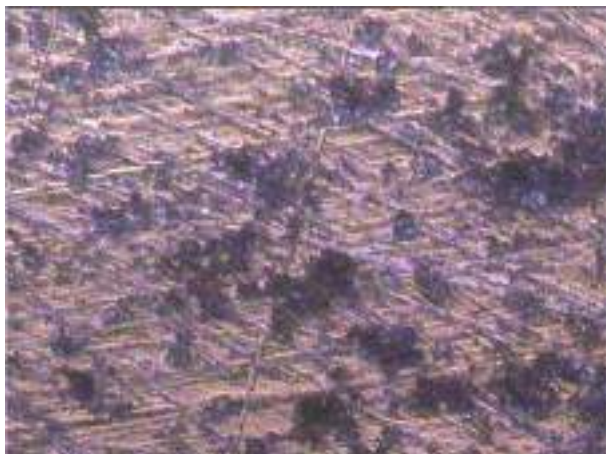


Fig. 14. Microstructure of welded area of EN 8.



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

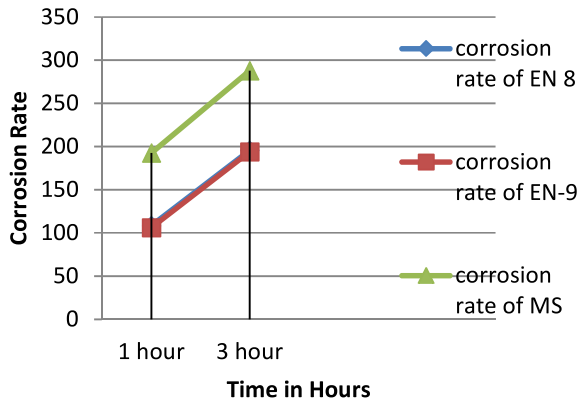


Fig. 17. Corrosion rate.

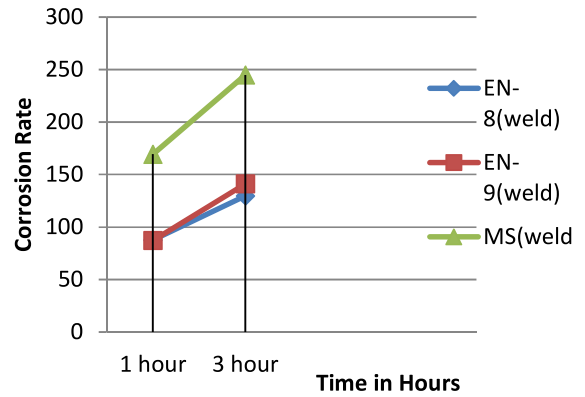


Fig. 21. Comparison of Welded Materials.

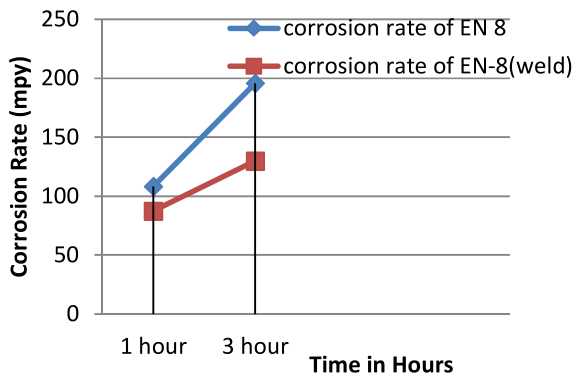


Fig. 18. Comparison of Corrosion Rate of Parent Material and Weld of EN-8.

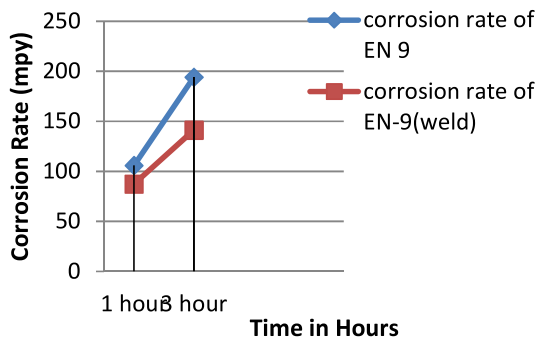


Fig. 19. Comparison of Corrosion Rate of Parent Material and Weld of EN-9.

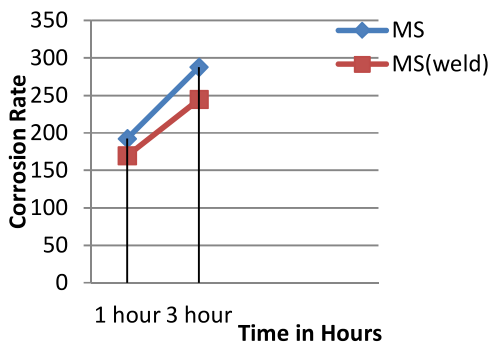


Fig. 20. Comparison of Corrosion Rate of Parent Material and Weld of MS.

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), anandtm@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 SiO₄ and AlO₆ 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 °C min⁻¹. Differential scanning calorimetric measurements (DSC) (Mettler Toledo DSC, USA) were carried out in a nitrogen atmosphere between 0 and 150 °C at a heating rate of 10 °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]:

$$\sigma = \sqrt{\frac{1}{n} \sum_i (N_{Ai} - \bar{N}_A)^2} \quad (4)$$

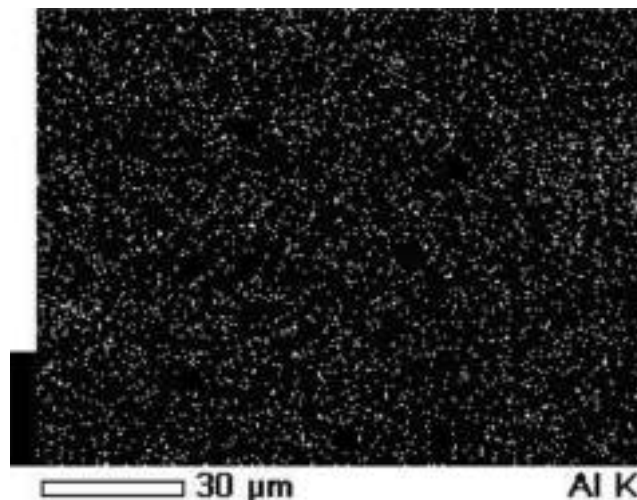


Fig. 1. Representative Aluminium elemental map of the composite with 1 wt.% HNT loading.

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

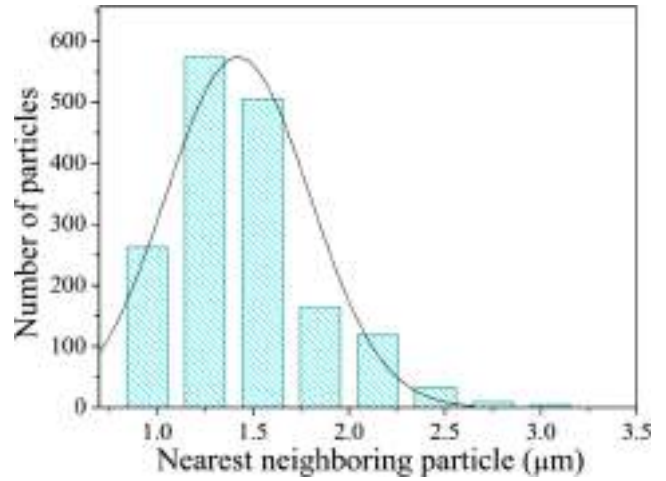


Fig. 3. Nearest neighbor distribution of 1 wt.% HNT filled nanocomposite.

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

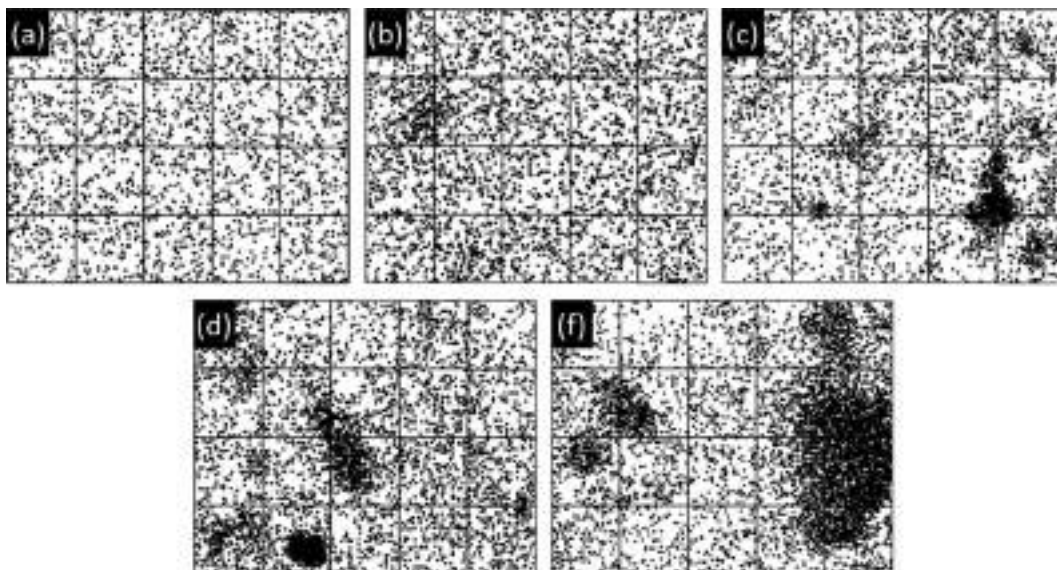


Fig. 2. Sparse sampled aluminum maps of HNT loaded composites, (a) 1 wt.%, (b), 3 wt.% (C), 5 wt.% (d) 7 wt.% and (e) 10 wt.%.

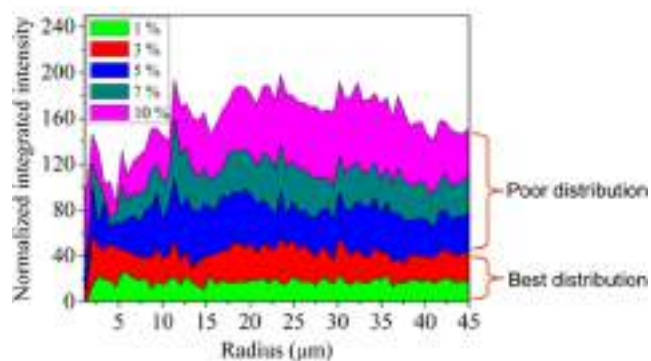


Fig. 4. Radial distribution of HNTs in EVACO/HNT composites.

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

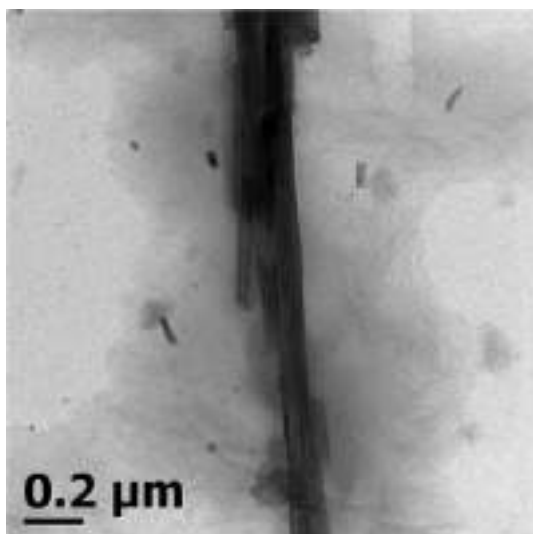


Fig. 5. TEM micrograph of EVACO/CNT composite with 1 wt.% HNT loading.

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.

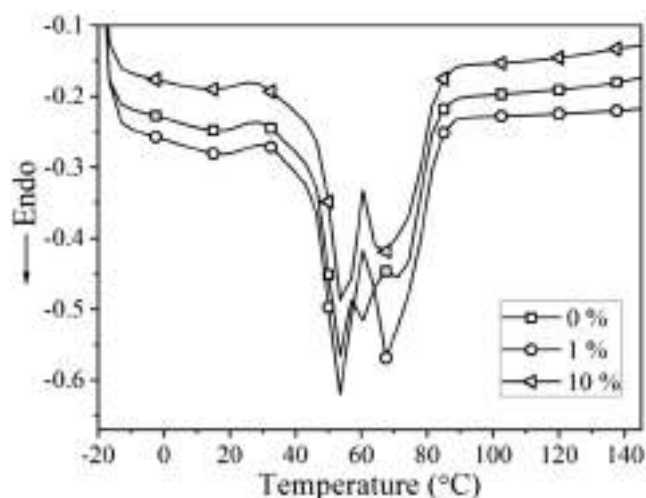


Fig. 6. DSC first heating curves of pristine EVACO and representative composites.

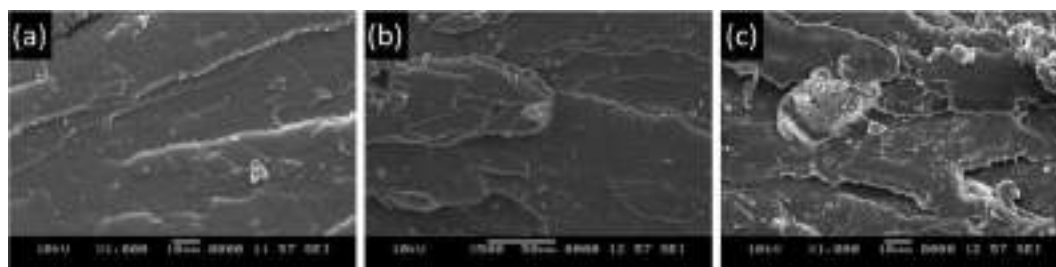


Fig. 7. The fracture surface of (a) EVACO, (b) EVACO/HNT composite with 1 wt.% HNT loading, and (c) 7 wt.% HNT loading.

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. Talbot, 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, Pallissery, 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 \approx 100) (Fig. 1a), followed by HNTs (AR \approx 20) (Fig. 1b), and the lowest in the case of nano-ATH (AR \approx 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

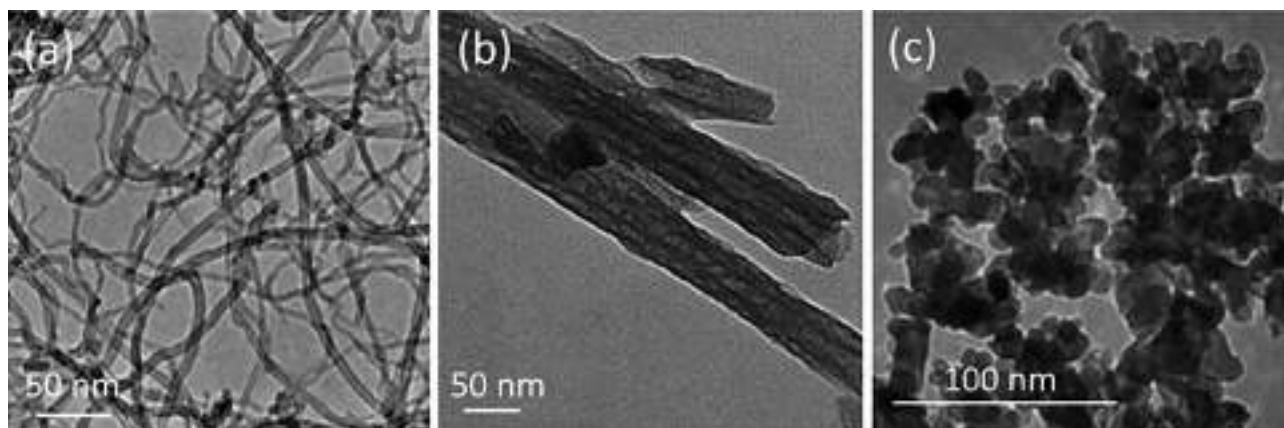


Fig. 1. TEM images of (a) MWCNTs, (b) HNTs and (c) nano-ATH.

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

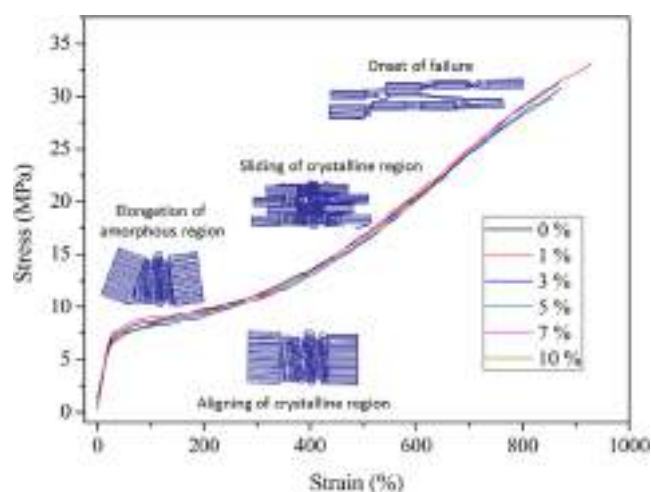


Fig. 2. Representative stress-strain curve of EVACO/nanofiller composite.

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 (kN 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
MWCNTs	0	31.5	14.3	944
	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

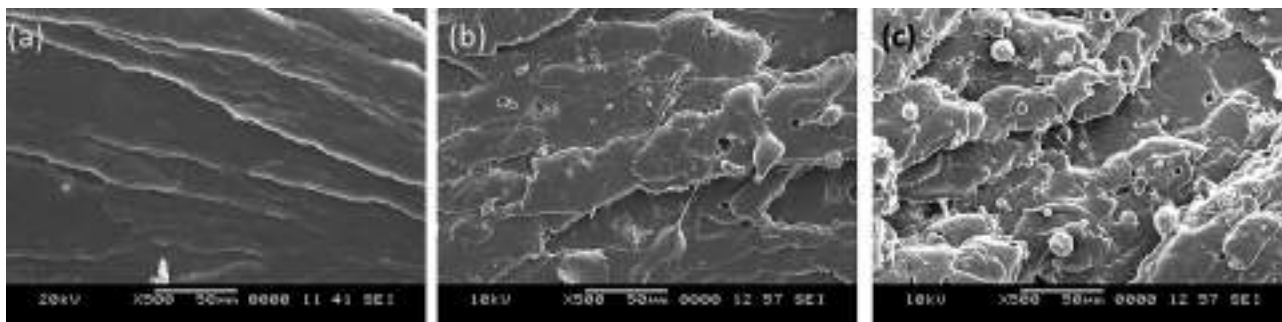


Fig. 3. The fracture surface of (a) EVACO, (b) EVACO/1 wt% ATH composite, and (c) EVACO/10 wt% ATH composite.

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 crystallite 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.



Access through SCMS School of Engineering...

Purchase PDF

Access the

Article preview

Abstract

Introduction

Section snippets

References (13)

Cited by (1)

materialstoday:
PROCEEDINGS



Volume 47, Part 15, 2021, Pages 4945–4949

Synthesis and characterization of Co-5Cr-RHA hybrid composite using Powder metallurgy

U. Anusachalam¹, G.R. Raghav¹, S. Dharmesh¹

Show more

+ Add to Mendeley | Share | Cite

<https://doi.org/10.1016/j.matpr.2021.04.148>

Get rights and content

G.R. Raghav

View in Scopus

Department of Mechanical Engineering, SCMS School of Engineering and Technology, Vidyasagar Karukutty, Ernakulam 683576, India

Corresponding author.

raghavmech1nce@gmail.com

More documents by G.R. Raghav

Provided by Scopus

Surface texture characterization of selective laser melted Ti-6Al...

Proceedings of the Institution of Mech...

Akshay V., ... Devedula, S. Activate Windows

Reuse of used paper and carton boxes as a source to FEEDBACK

Received June 19, 2021, accepted July 6, 2021, date of publication July 9, 2021, date of current version July 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3096125

Scalar and Vector Controlled Infinite Level Inverter (ILI) Topology Fed Open-Ended Three-Phase Induction Motor

A. HAREESH¹, (Member, IEEE), AND B. JAYANAND², (Member, IEEE)

¹Electrical and Electronics Engineering Department, Government Engineering College, Thrissur, Affiliated to APJ Abdul Kalam Technological University, Thrissur, Kerala 680009, India

²Electrical and Electronics Engineering Department, SCMS School of Engineering and Technology, Affiliated to APJ Abdul Kalam Technological University, Karukutty, Kerala 683576, India

Corresponding author: A. Hareesh (harielayur@gmail.com)

ABSTRACT The design and performance analysis of an open-ended three-phase induction motor, driven by an Infinite Level Inverter (ILI) with its speed control using scalar and direct vector control techniques are presented in this paper. The ILI belongs to an Active-Front-End (AFE) Reduced-Device-Count (RDC) Multi-level Inverter (MLI) topology. The fundamental structure of this inverter topology is a dc-to-dc buck converter followed by an H-bridge. This topology performs a high-quality power conversion without any shoot-through issues and reverse recovery problems. The performance of the proposed topology is validated using a resistive load. The THD of output voltage waveform obtained is 1.2%. Moreover, this topology has exhibited a high degree of dc-source voltage utilization. ILI considerably reduces the switching and conduction losses, since only one switch per phase is operated at high frequency, and other switches are operated at power frequency. The overall efficiency of the inverter is 98%. The speed control performance of the ILI topology using three-phase open-ended induction motor has been further validated through scalar and direct vector control techniques. Results obtained from simulation studies are verified experimentally.

INDEX TERMS Active-front-end, multi-level inverters, reduced-device-count, scalar and direct vector control, three-phase infinite level inverter.

AFE	Active Front End	qZSI	Quasi Z- Source Inverters
CHB	Cascaded H- Bridge	RDC	Reduced Device Count
ESR	Equivalent Series Resistance	RSA	Reduced Switch Asymmetrical
FOC	Field Oriented Control	RSM	Reduced Switch Modified
FC	Flying Capacitor	RSS	Reduced Switch Symmetric
GaN	Gallium Nitride	Si	Silicon
IGBT	Insulated Gate Bipolar Transistor	SiC	Silicon Carbide
ILI	Infinite Level Inverter	SM	Sub Modules
M	Modulation Index	SPWM	Sinusoidal Pulse Width Modulation
MIC	Module Integrated Converter	SVPWM	Space Vector Pulse Width Modulation
MLI	Multi-Level Inverter	THD	Total Harmonic Distortion
MMC	Modular Multilevel Converter	WBG	Wide Band Gap
NPC	Neutral Point Clamped	ZVS	Zero Voltage Switching
PWM	Pulse Width Modulation	C	Filter Capacitance
qCHB-FLBI	Quasi Cascaded H- Bridge Five Level Boost Inverter	D	Diode
		f_c	Carrier Frequency
		f_m	Modulating Frequency
		f_s	Fundamental Frequency
		f_{sw}	Switching Frequency
		$I_{as}^*, I_{bs}^*, I_{cs}^*$	Stator Phase Current References

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaofeng Yang¹.

i_c	Current through the Capacitor
$I_{Cripple}$	Capacitor Ripple Current
I_{Cpeak}	Peak Current through the Capacitor
i_D	Current through the Diode
I_F^*, I_T^*	Flux And Torque Producing Components of Current
$I_{Drating}$	Current Carrying Capacity of Switch
i_L	Current through the Inductor
I_{Lavg}	Average Inductor Current
I_{Lpeak}	Peak Current through the Inductor
$I_{Lripple}$	Inductor Ripple Current
I_m	Maximum Current
i_{ph}	Phase Current
I_{rr}	Reverse Recovery Current
I_s	Stator Current
I_s^*	Phasor Reference Current
i_{SW}	Current through the Switch
$I_{SWrating}$	Current Carrying Capacity of Switch
L	Inductance
P_D	Switching Power Dissipation
$Qc1, Qc2, Qc3$	High Frequency Operating Switches
$Q1, Q2, Q3, Q4$	Power Frequency Operating Switches
R_L	Load Resistance
T_e	Actual Torque
T_e^*	Reference Torque
T_{OFF}	Turn OFF Time
T_{ON}	Turn ON Time
t_{rr}	Reverse Recovery Time
T_s	Sampling Time
V_c	Voltage across the Capacitor
V_{co}	Average Output Voltage across Buck Capacitors.
V_{ce_sat}	Collector Emitter Voltage at Saturation
V_{Cmax}	Maximum Voltage across the Capacitor
$V_{Cripple}$	Capacitor Voltage Ripple
V^D	Voltage across the Diode
$V_{Dstress}$	Voltage Stress across the Diode
V_f	Forward Voltage of Diode
V_{GS}	Gate Source Voltage
V_i	Input DC- Source Voltage.
V_L	Voltage across the Inductor
V_m	Maximum Voltage
V_o	Sinusoidal Output Voltage
V_{oa}, V_{ob}, V_{oc}	Output Phase Voltages
V_{ph}	Phase Voltage
V_s	Standing Voltage across the Switch
V_{SW}	Voltage across the Switch
$V_{SWstress}$	Switching Voltage Stress
δ	Sinusoidal Varying Modulation Index
ω	Angular Frequency
ΔI_L	Linear Current Change through Inductor
ω_r	Actual Speed
ω_r^*	Reference Speed

λ^*	Reference Rotor Flux Linkage
λ_r	Flux Linkage
θ_T	Torque Angle
\mathcal{L}	Laplace transform

I. INTRODUCTION

Nowadays dc-to-ac inverters are playing a crucial role in power electronic areas such as electric drives, electric vehicles/hybrid electric vehicles, uninterruptable power supplies, HVDC power transmission, renewable energy integration, Flexible AC Transmission Systems (FACTS) and static VAR compensators. Based on the development and the nature of output voltage waveform, the inverters are broadly classified as two-level or square wave inverters, quasi-square wave inverters, two-level PWM inverters and multilevel inverters (MLI). The major problems associated with conventional or two-level inverters include the requirement of semiconductor devices of higher power ratings. To obtain the required voltage/current capacity, many devices need to be connected in series/parallel strings. As a result, these inverters generate low power quality output waveforms along with more conduction loss. Hence to overcome the aforementioned drawbacks, MLI can be chosen as a better alternative [1].

During the past three decades, MLIs [2]–[7] attracted wide attention both in the scientific as well as in industrial fields, because of its improved power conversion capabilities such as better power quality, control and efficiency as compared to all other conventional inverters. The basic concept of MLI is to generate a higher number of voltage levels with less distortion. The power conversion is performed by various dc low voltage levels along with several low rated power semiconductor switches. Each level is defined as the portion of voltage waveform where the magnitude of voltage remains constant for a defined duration which leads to the generation of a staircase voltage waveform. If the number of voltage levels increases, then the power quality of the output voltage waveform also increases. In modern industrial applications, demand for MLIs have increased due to its viable technology to implement controlled speed drive and to maintain power quality in high-power applications. The basic fundamental topologies of MLIs are classified into three categories; they are i. Cascaded H-Bridge (CHB) [2], [8], [9], ii. Neutral Point Clamped (NPC) or Diode Clamped [10] and iii. Flying capacitor (FC) [11]. They are popularly known as classical topologies. Out of these fundamental topologies, first two are required to have a single dc source, while the third one requires multiple dc-sources. But, to realize those topologies, more numbers of semiconductor devices and passive components are required. Hence the system becomes bulky and complex. One of the most desirable properties of the multilevel structure is modular structure or modularity. Among the aforementioned topologies, the CHB due to its modularity exhibits higher output voltage, power level and reliability. Practically, NPC and FC topologies are feasible up to 5-levels only, beyond this limit, their structure and

control implementation become more complex and also the device count for different ratings increases a lot. On the other hand, for higher levels, CHB is the best solution, since its devices are of the same power rating. However, the control implementation is more complex due to the requirement of coordination among too many power semiconductor devices and the increased number of dc-sources required for generating a higher number of levels [9]. Apart from the classical topologies, several novel application-oriented MLI topological developments have been reported in various literatures. They are basically derived from classical topologies or their hybrids. However, none of the new generation topological developments can be claimed to be absolutely advantageous. Most of them are designed for specific applications.

In general, the application-oriented topologies have been focussed on increasing number of output voltage levels and power quality, with reduced number of switching devices, passive component counts and cost. Some of the application-oriented approaches are discussed in [12]–[19]. Another significant development in MLIs is fault-tolerant topology. Chen *et al.* [20] has proposed a fault-tolerant topology to obtain uncompromised multilevel voltage waveform in the event of partial failure(s) in the power circuit. Such topologies maintain the output voltage waveform utilizing control signal modification along with redundancy offered by multi switching states. Another important contribution in MLI topologies was introduced by Glinka [21] presented a new single DC source-based multilevel topology, known as ‘Modular Multilevel Converter (MMC)’. The salient features of this MMC are its modularity and scalability. This topology is simpler than CHB inverter. It could conceptually meet any voltage level requirements with reduced harmonic issues, lower converter components ratings, and also with improved efficiency. The proposed topologies have caught the attention for medium/high-power energy conversion systems, industrial applications including FACTS, HVDC transmission systems, medium-voltage variable-speed drives and medium/high voltage dc-to-dc converter applications. Suman Debnath *et al.* [22] presented a review article regarding MMC. It is highlighted with a general overview on the basics of operation, control challenges, state-of-the-art control strategies and application challenges. Majority of MLIs are originated by arranging different configurations of Sub Modules (SM), which are able to connect to each other, that can create conventional and modern MLIs. Vijeh *et al.* [23] presented a general review of MLIs based on main submodules.

In the last couple of decades, the researchers are taking efforts and are being directed to reduce the device count (RDC) in MLI topologies. The reduction of semiconductor switches and passive components in MLIs can improve the efficiency and reliability of the system as well as the overall loss, cost, size and complexity. A large number of different novel RDC-MLI topologies have been reported in different literatures [24]–[34]. The aforementioned topologies have their own merits and demerits from application requirements.

Bana *et al.* [35] presented a comprehensive review on some recently developed RDC-MLI topologies, which are more suitable in different applications such as machine drives, renewable energy systems and FACTS. These topologies can be used in grid-tied as well as in standalone applications. The RDC-MLI topologies are broadly classified into three types. They are reduced switch symmetric H-bridge type MLI (RSS-MLI) [36], [37] and reduced switch asymmetrical H-bridge type MLI (RSA-MLI) [38]. The general structure of the configurations mentioned above is an RDC-MLI coupled with an H-bridge, whereas reduced switch modified MLI (RSM-MLI) [39]–[41] topologies are without H-bridge.

In recent years, a number of novel topologies have been noticed in various literatures. The general structure of such existing topologies are Active Front End (AFE)-RDC-MLI topology. These existing MLI topologies varies depending on the dc-to-dc converter used. Those topologies consist of an AFE converter (dc-to-dc converter) followed by a synchronized H-bridge. The main role of the AFE converter is generation of voltage levels, and that of H-bridge topology is polarity generation (inverter). Different AFE- RDC-MLI topologies are discussed in various literatures [42]–[53].

This paper presents the design and performance analysis of a three-phase AFE-RDC-MLI topology for open-ended three-phase induction motor drive application. In this proposed topology dc-to-dc buck converter is used for generating different voltage levels, which is operating at high switching frequency. The generation of output voltage levels depends upon the switching frequency of the AFE converter. If the switching frequency is increased, then the number of output voltage levels also increases. Due to the higher number of voltage levels, the output voltage waveform becomes smooth sinusoidal. If the switching frequency of the AFE converter is increased to almost infinity, the levels generated in the output voltage waveform becomes infinite. Hence, the power quality of the output voltage waveform is enhanced. Therefore, the proposed topology is called as an infinite level inverter (ILI) topology. The performance of the proposed topology has been validated with a resistive load as well as with an open-ended three-phase induction machine. Results obtained from the simulation studies are verified experimentally. The taxonomy of MLIs are briefly discussed in TABLE-1. The proposed MLI topology has several advantages.

- 1) Only one switch per phase operates at high frequency.
- 2) H-bridge inverter circuit operates at power frequency.
- 3) This topology is free from shoot through menace because of the presence of inductors between the voltage source and inverters.
- 4) The output filtering capacitors in AFE converters can be an electrolytic capacitor which is smaller and less expensive as compared to ac capacitors.
- 5) This topology is suitable for implementing sophisticated algorithms.
- 6) The advance control strategies developed by the researchers related to dc-to-dc converter can be directly

TABLE 1. Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
1	[1], Feb, 2020.	Review paper; Reduced Switch Count (RDC) MLI	A comprehensive review of some of the recently proposed RDC -MLI topologies. Pd: Total power semiconductor switch count, number of DC sources, passive component requirement, highest switch voltage rating, total standing voltage	A qualitative and quantitative summary of various new MLI topologies have also been discussed and a comparison is made.
2	[2], Aug, 2002	Diode-clamped, capacitor-clamped and cascaded multicell with separate dc-sources	A survey of topologies, controls, and applications. Pd: High-voltage high-power devices, optical sensors and other opportunities for future development, Control and modulation methods.	Different industrial applications and technological aspects are discussed
3	[3], Jan/Feb 1999	Six-level back-to-back 10-kW diode-clamped converter prototype	The multilevel diode-clamped converter; obtains well-balanced voltages across the dc-link capacitor with less THD. Pd: THD, efficiency and power factor.	Simulation and experiment performed using 32-bit digital signal processor. Application: high-power and/or high-voltage electric motor drives.
4	[4], May/June, 1996	Diode-clamp, flying capacitors and cascaded-inverters with separate dc-sources.	The operating principle, features, constraints, and potential applications of these converters will be discussed. Pd: Harmonics, EM1 problems, voltage stress, components and topology comparison.	Simulation and experiment. Application: Reactive power compensation, back-to-back inertia, utility compatible adjustable speed drives.
5	[5], Aug, 2010	Review paper; Multilevel converters.	Recent contributions on topologies, modulation, and recent advances and industrial applications of multilevel converters. Pd: Current state in industrial applications, Modulation and control of multilevel converters, Nontraditional applications of MLI, Future trends and challenges	High-performance multilevel converter applications based on topologies are tabulated.
6	[6], July, 2010	A survey - Cascaded MLIs	A survey of different topologies, control strategies and modulation techniques used by cascaded MLI. Pd: High degree of modularity, the possibility of connecting directly to medium voltage, high power quality, both input and output, high availability, and control of power flow in the regenerative version.	Application: Pumps, fans, STATCOM, traction, LNG plant.
7	[7], June, 2008	A Review: Potential in current and future power applications	Review and analyse the most relevant characteristics of MLI and provides an introduction of the modelling techniques, operational and technological issues. Pd: Modeling techniques, modulation strategies, operational and technological issues.	Multilevel converter-driven applications overview.
8	[8], Dec, 2007	Review; Voltage-source-converter topologies. 1) 2L VSCs, 2) NPC, 3) CHB, and 4) FC multilevel VSCs.	Review of voltage-source-converter topologies. Pd: Operating principle of each topology and Modulation methods. The latest advances and future trends of the technology are discussed.	Application: Industrial medium-voltage drives.
9	[9], Oct, 2002	Cascade multilevel converter	Charge balance control schemes used for cascade multilevel converter in order to maintain equal charge/discharge rates from the dc sources (batteries, capacitors, or fuel cells) in an HEV. Pd: Output voltage, THD, efficiency and power factor.	Simulation and experimental. Application: Hybrid electric vehicles
10	[10], July, 2000	Diode Clamping MLI	Diode clamping inverter, which works without the series association of the clamping diodes. Pd: Blocking Voltage Distribution, switch and diode clamping mechanism, auxiliary clamping, operation principle.	Experimental. Application: Diode clamping MLI in large power conversion area.

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
11	[11], Aug, 2002	Flying Capacitor MLIs	The requirements imposed by a direct torque control (DTC) strategy on multilevel inverters are analyzed. Pd: Flux and torque.	Simulation and experimental. Application: Motor drive
12	[12], Jan, 2021	A Novel Asymmetrical 21-Level Inverter for Solar PV Energy System With Reduced Switch Count	The proposed topology achieves 21-level output voltage without H-bridge using asymmetric DC sources. Pd: RDC, THD, maximum power point tracking, total standing voltage, switches count, and sources count, gate driver boards, the number of diodes and capacitor count and component count level factor.	Simulation and experimental. Application: Solar PV Energy System
13	[13], 2020.	Hybrid multilevel dc-to-ac converter using single-double source unit	The hybrid topologies compared with the conventional CHB converter, and the best topologies recommended for medium voltage applications. Pd: Number of power switches, THD, symmetric and asymmetric inverter, high and low voltage comparison included.	Experimental. Application: Medium voltage applications. Grid-connected photovoltaic system.
14	[14], 2020.	Transistor clamped T-type multilevel H-bridge inverter	Investigation of a transistor clamped T type H-bridge MLI with Inverted double reference single carrier PWM technique (IDRSCPWM) are discussed. Pd: RDC topology , THD	Experimental. Application: Renewable energy applications
15	[15], 2020.	Cascaded MLI	Design and implementation of MLIs for fuel cell energy conversion system. Pd: RDC topology, THD, quality of output and stress, efficiency and power loss.	Simulation and experimental. Application: Fuel cell energy conversion system
16	[16], 2020.	Binary hybrid MLI	The DSOGI controller is implemented to control a binary hybrid MLI by which real power control, reactive power control, harmonic elimination and source current balancing. Pd: RDC, THD, efficiency, Power quality.	Simulation and experimental. Application: Grid integrated solar energy conversion system.
17	[17],2017.	Cascaded H-bridge MLI	Design and testing of 5-level symmetric cascaded MLI-CHB inverter with 6 switches for photovoltaic system. Pd: THD, power loss, implementation complexity, cost, advantages.	Simulation and experimental. Application: Photovoltaic system.
18	[18], 2020.	Three Phase MLI	A new three phase MLI with reduced number of components count is proposed. Pd: RDC, THD, compared number of levels, active switches, diodes, capacitors and total standing voltage.	Simulation and experimental. Application: Low and medium power photovoltaic systems.
19	[19], 2020.	Dual source MLI (DS-MLI)	DS-MLI with fewer power switches are proposed for solar PV power conversion systems. Pd: RDC. THD, efficiency, loss, Small signal model is discussed. The proposed one is compared with its conventional counterpart.	Simulation and experimental. Application: Domestic solar PV applications
20	[20], Mar, 2005.	Multilevel converter topology	A multilevel converter topology with fault-tolerant ability is presented. Pd: Voltage levels balance, fault-tolerant ability	Simulation and experimental.

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
21	[21], 2004.	Modular-multilevel converter	A 2MW-17-level prototype of the new MMLC family is presented. Pd: Behaviour at steady state, transient conditions and fault conditions.	Simulation and experimental (Stratix FPGA), dSpace (ds1103). Application: Wide range of high voltage applications, traction converters, operating directly on the power line, and network inerties.
22	[22], Jan, 2015	A review; Modular Multilevel Converter	A general overview of the basics of operation of the MMC along with its control challenges are discussed. Pd: Modulation, design, control, and modelling of the MMC	Application: HVDC Systems, Variable-Speed Drives, Dynamic Braking Chopper, flexible ac transmission system
23	[23],2018	General review; MLI based on Main Submodules: Structural point of view	This paper presents five Main Submodules (SMs) to be used as the basic structures of MLIs and has widely reviewed almost all presented multilevel inverters in these manuscripts based on the proposed SMs with variety connections. Pd: Number of components, the ability to create inherent negative voltage, working in regeneration mode and using single DC source. Comparison details of asymmetrical MLI included.	Application: Wide range of power electronics applications.
24	[24], 2020.	Topological Review: voltage source MLI with reduced device count	Summarize the recently developed multilevel topologies with a reduced component count, based on their advantages, disadvantages, construction, and specific applications. Pd: A detailed comparison in terms of switch, diode, capacitor, inductor, transformer count was performed and systematically summarized in tables.	Giving guidelines to further improve the current multilevel topologies more efficiently and compactly.
25	[25], 2015.	A Review: MLI topologies:	Recently proposed multilevel inverter topologies with reduced power switch count are reviewed. Pd: A detailed comparison of recently proposed MLI topologies is presented	Application: High and low power applications.
26	[26], 2017.	High-level inverter topology	A new topology with a lower voltage rating component to improve the performance by remedying the mentioned disadvantages are discussed. Pd: DC, efficiency, Power loss, Voltage stress, load power factor. The proposed topology has been compared with other similar topologies.	Simulation and experimental (dSPACE). / The proposed topology can be operated with a different real-time environment with good performance.
27	[27], 2016.	A 9-level inverter is developed by stacking two-hybrid 5-level inverter.	Generation of higher number of voltage levels by stacking inverters of lower multilevel structures with low voltage devices for drives is presented. Pd: Speed control of induction motor, A 9-level inverter is compared with other existing 9-level topologies and tabulated.	Simulation and experimental. (FPGA,TMS320F28335). Application: Induction motor drives.
28	[28], 2020.	Switched Capacitor Converter (SCC)	RDC based Basic Cell (BC) of a novel Switched-Capacitor Converter (SCC) and its generalized structure has been proposed. The detailed analysis of capacitor selection procedure for 13 levels SCMLI is presented. Pd: Switching loss, reduced devices, voltage balance,	Simulation and experimental. (dSPACE 1104). / The proposed SCMLI is verified for asymmetric dc-source configuration with resistive load condition.

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
29	[29], 2019.	Asymmetrical MLI topology	<p>A new single-phase MLI topology has been proposed in this paper to reduce the number of switches in the circuit and obtain higher voltage level at the output.</p> <p>Pd: Reduced Switch Count, power loss.</p>	<p>Simulation(PLECS software)and experimental.(dSPACE 1104). /</p> <p>Application: MLI topology with multiple extension capabilities.</p>
30	[30], 2017.	Packed U-Cell (PUC), a new reduced-structure multilevel converter	<p>The issues associated with PUC are addressed and two approaches as remedy are presented.</p> <p>Pd: Reduce component count, total blocking voltage, switch ratings and extending its performance to high voltage applications.</p>	<p>Simulation and experimental (TMS320-F28335).</p> <p>Application: High and low voltage applications.</p>
31	[31], 2017.	Three phase five-level inverter topology	<p>An optimized topology of three-phase, MLI topology configuration for five or higher-level operation with reduced switch count is discussed.</p> <p>Pd: Reduced switching device count, conduction and switching losses, efficiency, THD.</p>	<p>Simulation and experimental (MS320LF28069).</p> <p>Application: Solar PV, drives applications.</p>
32	[32], 2019.	Five-level inverter topology	<p>In this paper, five-level inverter topology based on the stacked cell approach to MLIs with reducing the number of switches, flying capacitors to cut size, weight and costs while facilitating higher reliability, simpler wiring and lesser electromagnetic interference are discussed.</p> <p>Pd: Fault-tolerant operation, loss steering and common-mode elimination. Comparison of the proposed topology with existing methods and highlight its merits and features.</p>	<p>Simulation and experimental.</p> <p>Application: High reliability electric drives (induction motor drive)</p>
33	[33], 2019.	Single-phase symmetrical and Asymmetrical MLI topology.	<p>The presented topologies can generate 9-level output voltage in a symmetrical configuration, 13-level and 17-level in asymmetrical configuration with a single cell. A detailed comparison has been done of the presented topology with recently proposed topologies in terms of DC sources.</p> <p>Pd: RDC, THD, semiconductor switches, capacitor and total blocking voltage.</p>	<p>Simulation and experimental (dSPACE-1103).</p> <p>Application: High power/ high voltage applications with equal and unequal DC voltage sources.</p>
34	[34], 2017.	Single DC source nine-level switched-capacitor boost inverter topology.	<p>In this paper proposed boost inverter topology with nine level output voltage waveform using a single dc-source and two switched capacitors.</p> <p>The number of devices and cost are highlighted by comparing the recent and conventional inverter topologies.</p> <p>The total voltage stress of the proposed topology is lower and has a maximum efficiency of 98.25%.</p> <p>Pd: Step-up inverter, switched capacitor, reduce switch count, efficiency. Comparison of different PWM techniques is presented.</p>	<p>Simulation(PLECS) and experimental.</p> <p>Application: Low and medium voltage applications.</p>

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
35	[35], 2019.	Review paper; Reduced Switch(RS) MLI Topologies, Symmetrical H-bridge based RS MLI, Asymmetrical H-bridge based RS MLI, and Modified RS MLI topologies.	This review paper focuses on a number of recently developed MLIs used in various applications. Each topology has been reviewed carefully based on the number of switch count, number of dc sources used, PIV, TSV, and applications. Pd: Recently developed RS MLI topologies, comprehensive analysis and comparative evaluation. RDC, comparison based on RSS MLI, RSA MLI and RSM MLI topologies included. / Switching loss, THD	Simulation and experimental. Application: Renewable energy integration and Drives Application.
36	[36], 2014.	Cascaded MLI with series connection of H-bridge; basic units	In order to generate all voltage levels (even and odd) single-phase cascaded multilevel inverter based on H-bridge unit is proposed. Nine different algorithms are proposed to determine the magnitudes of dc-voltage sources and it compared to the conventional topologies. Pd: RDC, blocked voltage.	Simulation and experimental. (89C52 microcontroller by ATMEL). Application: Drive and control of electrical machines, connection of renewable sources, FACTS devices.
37	[37], 2015.	Symmetric multilevel converter	Design of symmetric multilevel converter to enhance the circuit's performance is presented . Pd: proposed inverter is compared to the conventional topologies, conventional cascaded inverter, semi cascaded inverter and cross-switched inverter.	Simulation(MATLAB/Simulink) and experimental (ATMega 64 microcontroller by ATMEL). Application: Suitable for medium-voltage applications and where a higher number of output levels are required.
38	[38], Oct, 2014	Symmetric, Asymmetric, and cascade switched-Diode multilevel converter	Symmetric, asymmetric, and cascade switched-diode multilevel converter are proposed, which can produce many levels with minimum number of power electronic switches, gate driver circuits, power diodes, and dc voltage sources. Pd: RDC,THD, efficiency, standing voltage on switches.	Simulation and experimental. Application: High-voltage applications.
39	[39], Mar, 2018.	Seven-level Pack U-Cell inverter	In this paper present single-phase Pack U-Cell (PUC) MLI. The output voltage has a higher amplitude than the maximum dc-link value used in the topology as a boost operation. Pd: THD, efficiency, Power losses.	Simulation and experimental (dSpace 1103). Application: Power Quality, Renewable Energy Conversion, Photovoltaic Applications.
40	[40], June 2013.	Cascaded MLI topology	This suggested topology requires less switches and relevant gate driver circuits realising the same level in output voltage compared with traditional cascaded inverter. Pd: RDC,symmetric, asymmetric structures, PIV and power losses.	Simulation (MATLAB/Simulink) and experimental (DSPIC30F4011). Validate the practicability of the proposed symmetric and asymmetric multilevel inverters, which can generate all voltage steps for a test case nine-level symmetric and 17-level asymmetric inverter.
41	[41], 2016.	An Envelope Type (E-Type) Module:	The proposed E-Type module which can generate 13 levels with reduced components, It can be used in high voltage high power applications with unequal DC sources. Pd: RDC, THD, Power losses.	Simulation(MATLAB/Simulink) and experimental (Microcontroller ATMEGA16). Application: High voltage high power applications with unequal DC sources.

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
42	[42], June, 2011	Single-phase seven-level Grid-connected inverter	A single-phase seven-level inverter for grid-connected photovoltaic systems, with a PWM control scheme is proposed in this paper. Pd: RDC, THD.	Simulation(MATLAB Simulink) and experimental (TMS320F2812). Application: Photovoltaic system; grid-connected PV inverters.
43	[43], April 2015.	Boost dc-link cascaded MLI	A structure of single-phase seven-level boost dc-link cascaded MLI (BDCLMLI) is proposed. Pd: Reduction of voltage stress across the inverter switches, RDC, Power loss, THD.	Simulation (MATLAB)and experimental. Application: Uninterrupted Power Supply.
44	[44], 2014.	SHE-PWM Cascaded MLI with adjustable dc-voltage levels control	New Multilevel Selective Harmonic Elimination Pulse Width Modulation (MSHE-PWM) technique based transformerless Static Synchronous Compensator (STATCOM) system employing cascaded H-bridge inverter configuration is discussed. Pd: efficiency, THD, Reactive power (VAR) compensation.	Simulation(Matlab/Simulink)and experimental(dSPACE DS1104). Application: Reactive power(VAR) compensation; STATCOM.
45	[45], 2014.	Buck-Boost MLI	The proposed inverter is derived from Buck-Boost converter, which is with the ability to buck/boost the input voltage. Pd: The current sharing with modulation strategy is discussed.	Simulation and experimental. Application: Wide range varied dc-link voltage, such as renewable energy generation.
46	[46], 2019.	Switched-Capacitor (SC) based (2n+1)-level single-phase inverter.	This paper presents a novel Switched-Capacitor (SC) based (2n+1)-level single-phase inverter with a reduced number of components and input dc-voltage supply. Pd: quasi-resonant switching, efficiency, comparative analysis, thermal modelling and design guidelines, THD, loss calculation.	Simulation(MATLAB-Simulink-PLECS) and experimental (sb-RIO GPIC). Application: 500 VA prototype (Level Step-up Single-Phase Inverter) are presented.
47	[47] , 2016	Single phase step-up five-level inverter	The proposed topology can realize the multilevel inversion with high step-up output voltage, simple structure and reduced number of power switches. Pd: RDC, THD, summarize comparisons of the popular five-level inverter topologies and the proposed inverter.	Simulation and experimental. Application: Photovoltaic(PV) systems
48	[48],Sept. 2019.	Single dc-source-based seven-level Boost inverter	This paper discussed about an effective circuit arrangement of an MLI that can increase the number of output voltage levels with a lesser number of device count. Pd: RDC, THD. Device count-based comparative analysis is presented.	Simulation (MATLAB/Simulink) and experimental (FPGA-NEXYS 4). Application: Suitable for electric vehicles as less number of series-connected battery cells are required.
49	[49], Jan. 2009.	Single-phase five-level Photovoltaic (PV)inverter topology	A single-phase MLI utilizes two reference signals and a carrier signal to generate PWM switching signals for PV application. Pd: THD, efficiency.	Simulation and experimental (DSP TMS320F2812). Application: Grid-connected PV systems

TABLE 1. (Continued.) Taxonomy of multilevel inverters.

Sl. No.	[Ref.], Year	Topology	Objective and Parameter discussed (Pd)	Simulation / Experiment / Suggestions / Application
50	[50], 2017.	Modified quasi-Z-source cascaded hybrid five-level inverter	<p>This paper proposes a single-phase modified quasi-Z-source (MqZS) hybrid three-level inverter. In addition, a single-phase modified quasi-Z-source cascaded hybrid five-level inverter (MqZS-CHI) is designed by connecting two three-level PWM switching cells in series for producing a nine-level output voltage.</p> <p>Pd: THD, efficiency, boost factor.</p>	<p>Simulation and experimental (TMS320F28335).</p> <p>Application: Recommended for high output voltage with lower THD applications.</p>
51	[51], Oct. 2011.	Z-source-based MLI with reduction of switches	<p>This study presents a new inverter topology based on a mixture of cascaded basic units and one H-bridge unit. The cascaded basic units produce positive and zero-voltage levels and at the same time suggested inverter obtains positive, zero- and negative voltage levels.</p> <p>Pd: THD, RDC.</p>	<p>Simulation (MATLAB/Simulink) and experimental.</p> <p>Application: DVR: Power quality.</p>
52	[52], 2019.	Integrated semi-double stage based MLI with voltage boosting scheme	<p>This paper proposes a single-phase, seven-level, transformerless inverter. The proposed configuration achieves voltage boosting using a non-isolated interleaved buck-boost converter, which is fused with the inverter configuration through two switched capacitors (SC).</p> <p>Pd: Stray capacitor voltage, the common-mode voltage (CMV), Leakage current minimization, Semi-double stage system, Voltage boosting, power transmission ratio, efficiency.</p>	<p>Simulation (MATLAB) and experimental.</p> <p>Application: Photovoltaic Systems.</p>
53	[53], 1998	Switch-Mode dc-to-ac inverter using buck converter	<p>A dc-to-ac inverter of bidirectional buck topology using non-linear robust controller is proposed to achieve dynamic robust performances for both resistive and reactive loads.</p> <p>Pd: The proposed inverters are immune to the large disturbances in input voltage and load current and the output voltage remains steady state and dynamic stable.</p>	<p>Simulation and experimental./</p> <p>Dynamic performance of the proposed inverter is validated for both resistive and reactive loads.</p>
54	[54], Aug. 2013.	40-kVA SiC JFET Inverter	<p>This paper describes the concept, design, construction, and experimental investigation of a 40-kVA inverter with silicon carbide junction field-effect transistors (JFETs). The inverter was designed to reach an efficiency exceeding 99.5%.</p> <p>Pd: Silicon carbide (SiC), Power losses, efficiency measurements.</p>	<p>Simulation and experimental.</p> <p>Application: High-efficiency inverter; exceeding 99.5%</p>

applied to the system. Since the AFE converter operates at high switching frequency ranges, it is more flexible to control.

- 7) The system exhibits high dynamic performance. Hence, the output voltage remains dynamically unchanged when subjected to large disturbances in supply voltage or load currents.

- 8) The dc-to-dc converter based topology is highly compatible for implementing closed-loop control system.

The main outcomes of this article are as follows.

- Design, analysis and performance of a three-phase infinite level inverter-driven induction motor is performed.

- This topology has been tested with a resistive load and found to possess very good quality voltage and current waveforms in terms of THD.
- The dc-voltage requirement for generating a fixed ac-voltage output is much less than that required by other similar topologies, which is validated by SPWM control.
- The third harmonic injection modulation scheme has also been performed using this inverter and found that the dc-source utilization can be improved further.
- The efficiency of the inverter has also been found to be better since only one switch per phase is operated at a high frequency. All the switches in conventional inverters are operated at high frequency.
- Scalar and vector control of induction motor has also been performed using this topology. It has been found that both the controls exhibits better dynamic performance with this topology. Moreover, the ILI has been found to impart better performance to an induction motor drive.
- In conventional inverter circuits, the voltage applied across the terminals of the motor would be of absolute discrete values like V_{dc} , $-V_{dc}$, $V_{dc}/2$, $-V_{dc}/2$, etc. Hence, the instantaneous error voltage between the applied voltage and the desired sinusoidal voltage would be large. This error voltage manifests as torque ripples in the motor, deteriorating the performance of the motor.
- In the buck converter-based topology presented in this paper, the applied voltage is near to sinusoid and hence the error in the applied voltage is almost negligible. This results in negligible torque pulsations in the motor, thereby obtaining better performance from the motor. Open ended windings are used so that all the three buck converters can be connected to the same voltage source.

Now-a-days power, electronic industries and their application fields are changing from traditional silicon (Si) power semiconductor devices to potentially superior and high frequency operable counterpart, i.e., Wide Band Gap (WBG) devices, like Silicon Carbide (SiC) and Gallium Nitride (GaN) [54]. With the advent of WBG materials, standard Si technology is being replaced by high-frequency switches. If the Si semiconductor devices used in ILI topology are replaced with WBG devices, then there will be more improvement in efficiency and performance. Thereby, the system becomes more compact and operable at high frequency ranges.

This paper is organized as follows. In section II, a brief description of the basic structure of the proposed three-phase infinite level inverter topology is discussed. Principle of operation of ILI is discussed in section III. In section IV, modes of operations. In section V, the mathematical model of the proposed topology is discussed. In section VI, design of ILI. Criteria for selection of components are discussed in section VII. The efficiency calculation is discussed in section VIII. In section IX, the third harmonic injection method is discussed. In section X, scalar and vector control of ILI fed induction motor are presented. Simulation results

are presented in section XI. In Section XII, the comparison of ILI with conventional H-bridge inverters and AFE-RDC-MLI topologies are presented. In section XIII, discusses about the experimental setup and results. Finally, section XIV, concludes the paper and XV, provides with future scope.

II. THREE-PHASE INFINITE LEVEL INVERTER TOPOLOGY

The basic structure of infinite level inverter is a buck converter followed by an H-bridge. The proposed topology has the sole objective of reducing the count of passive elements and power semiconductor components without any loss of power quality in power conversion. Meanwhile, it reduces the switching and conduction losses, size and control complexity of the circuit. Three individual ILI circuits are combined to obtain a three-phase ILI topology which is shown in Fig. 1. The proposed topology consists of one high-frequency operated switch for every buck circuit and four low-frequency operated switches for every H-bridge; hence, one inductor and one capacitor per phase. This topology has a high degree of dc-source voltage utilization and low voltage stress across the switches as compared to traditional two-level and other similar inverter topologies. Other attributes of this circuit are the absence of shoot-through issue, less reverse recovery loss and body diode conduction loss in semiconductor switches.

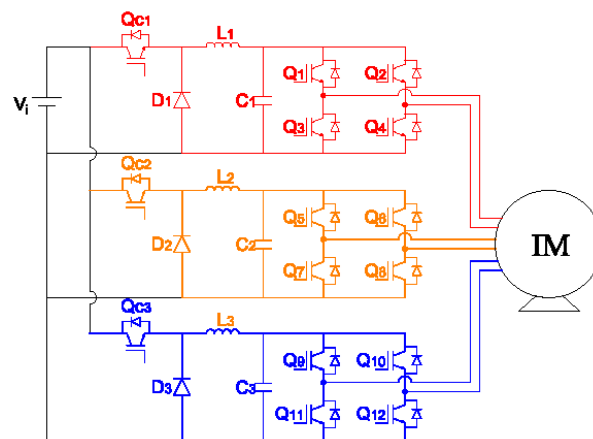


FIGURE 1. Three phase Infinite Level Inverter topology. Basic structure of the proposed topology is a buck converter (AFE converter) followed by an H-bridge. This topology consists of one high-frequency operated switch for every buck circuit and four low-frequency operated switches for every H-bridge; hence, one inductor and one capacitor per phase.

III. PRINCIPLE OF OPERATION

In an MLI each level is defined as the portion of voltage waveform where the magnitude of voltage remains constant for a defined duration. In the ILI, the voltage varies at every instant, and there is no time span where the voltage remains constant. If the time span in which the voltage remains constant tends to zero, the voltage waveform becomes a pure sine wave. The AFE converter generates a fully rectified infinite level voltage waveform by varying the duty cycle of the buck circuit in a fully rectified sinusoidal manner. The

quality of output voltage depends on the switching frequency of the AFE converter. Fig. 2 shows the SPWM control logic of ILI. The number of voltage levels developed across the buck capacitor is given by

$$V_{Level} = \frac{f_c}{f_m} \tag{1}$$

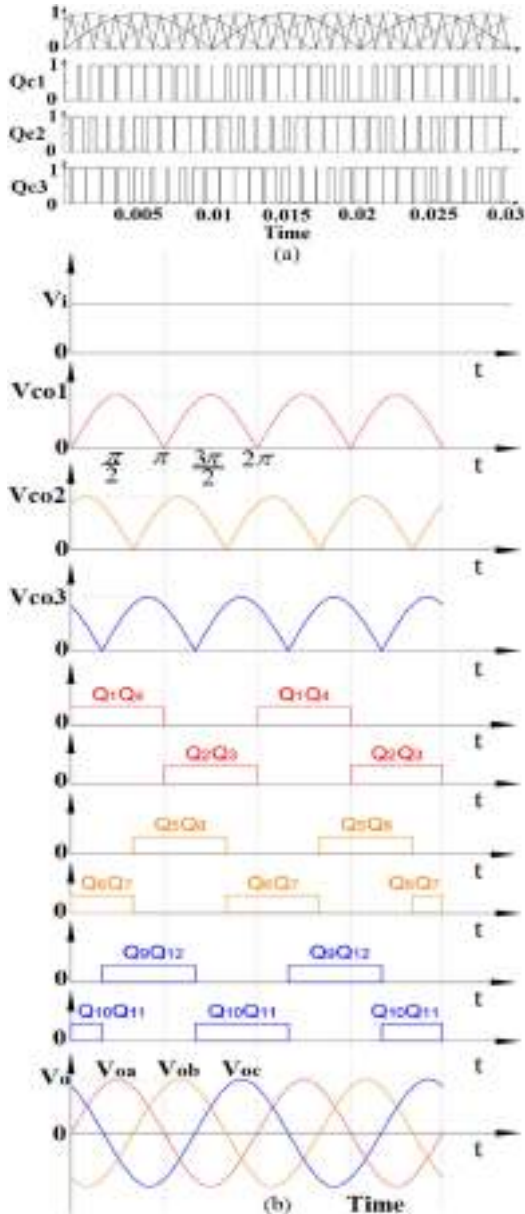


FIGURE 2. SPWM control logic of ILI. (a) High-frequency switching pulses of AFE converter (b) Power frequency switching patterns of polarity generation-part (H-bridge inverter).

where, ' f'_c ' is the carrier frequency and ' f'_m ' is the modulating signal. The duty cycle of the AFE converter is denoted by ' δ ' which varies in a rectified sinusoidal fashion.

$$\delta = M |\sin(\omega t)| \tag{2}$$

where, ' M ' is the modulation index. The output voltage of the buck converter is given by,

$$V_{co} = \delta V_i \tag{3}$$

The fully rectified sinusoidal output voltage across ' C'_1 ' is given by,

$$V_{co1} = V_i M |\sin(\omega t)| \tag{4}$$

where, $\omega = 2\pi f$ and ' V'_i ' is the input dc-source voltage. The rectified sine wave is unfolded by the H-bridge circuit to get a sinusoidal voltage across the load (7). The switching pulses for the H-bridge are synchronised with the modulating signal of the AFE converter. Therefore, the output voltage becomes

$$+V_i M |\sin(\omega t)| \quad 0 \leq \omega t \leq \pi \tag{5}$$

$$-V_i M |\sin(\omega t)| \quad \pi \leq \omega t \leq 2\pi \tag{6}$$

$$V_o = V_i M \sin(\omega t) \quad 0 \leq \omega t \leq 2\pi \tag{7}$$

The inversion process occurs at the natural zero-crossing point of the output voltage, resulting in zero voltage switching (ZVS) and hence reduces the switching power loss. In a three-phase ILI, the duty ratios of dc-to-dc converter switches are displaced by 120° . The three-phase voltage generation across the buck converters are fully rectified sine waveforms (8), (9), (10).

$$V_{co1} = V_i M |\sin(\omega t)| \quad 0 \leq \omega t \leq 2\pi \tag{8}$$

$$V_{co2} = V_i M |\sin(\omega t - 120)| \quad 0 \leq \omega t \leq 2\pi \tag{9}$$

$$V_{co3} = V_i M |\sin(\omega t + 120)| \quad 0 \leq \omega t \leq 2\pi \tag{10}$$

These rectified voltage waveforms are unfolded into the sinusoidal waveforms by appropriate switching of H-bridge.

$$V_{oa} = V_i M \sin(\omega t) \quad 0 \leq \omega t \leq 2\pi \tag{11}$$

$$V_{ob} = V_i M \sin(\omega t - 120) \quad 0 \leq \omega t \leq 2\pi \tag{12}$$

$$V_{oc} = V_i M \sin(\omega t + 120) \quad 0 \leq \omega t \leq 2\pi \tag{13}$$

IV. MODES OF OPERATION

There are four modes of operation for this inverter. They are discussed with the help of Fig. 3 Mode-1: During the period δT , the switch Q_{c1} is turned ON, and the inductor current starts rising. The diode is reverse biased. The output capacitor is charged exponentially. Mode-2: During the period $(1 - \delta) T$, switch Q_{c1} is turned OFF, and the parallel diode starts conducting. The current through the inductor falls, and it freewheels through the diode. In modes 1&2, Q_1 and Q_4 are conducting, so that a positive voltage is obtained across the load. Hence it generates the output voltage as $+V_i M |\sin(\omega t)|$ across the load side. Mode-3: is same as mode-1, except for the fact that Q_2 and Q_3 are switched ON. Mode-4: is same as mode-2, except for the fact that Q_2 and Q_3 are switched ON. Hence it generates the output voltage as $-V_i M |\sin(\omega t)|$ across the load side. The output voltage developed across the inverter load terminal is $V_o = V_i M \sin(\omega t); 0 \leq \omega t \leq 2\pi$.

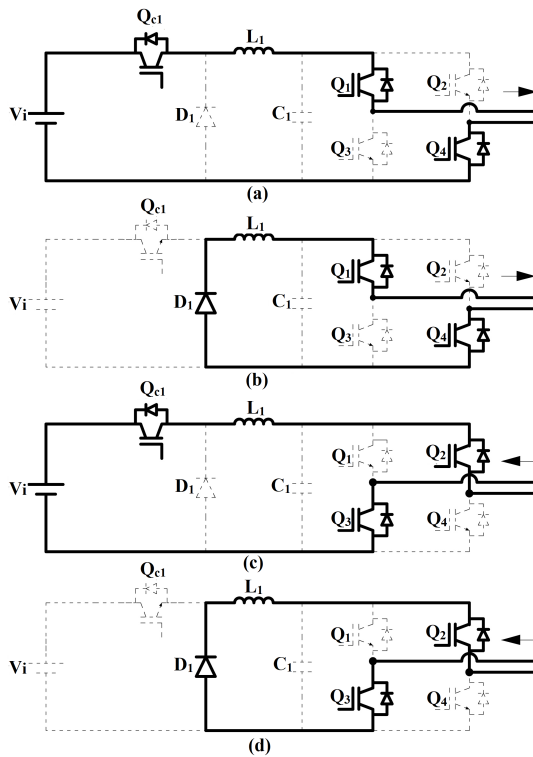


FIGURE 3. Mode of operations. (a) Mode-1, (b) Mode-2, (c) Mode-3, (d) Mode-4.

V. MATHEMATICAL MODEL OF ILI

The ILI is a buck–H bridge cascade. The buck converter is represented using the conventional transfer function of a dc-to-dc converter. The input to the transfer function is a rectified sine wave, acting as its variable duty cycle. The output for this transfer function is a rectified voltage, which would be available from the buck converter provided with such a duty cycle. This rectified voltage is unfolded with an H-bridge. This process is represented using a switch, which passes the input as such to the output when the input is positive and passes the negative of the input when the input is negative. This is equivalent to inversion process. The main part of the proposed converter is a dc-to-dc buck converter is shown in Fig.5. Here the buck converter plant is modelled to attain a simplified and linearized system around the equilibrium point using feedback control technique. The dynamic averaged equations of buck converter are Inductor current,

$$\frac{di_L}{dt} = \frac{-V_{CO}}{L} + \frac{V_d}{L} \tag{14}$$

Capacitor voltage,

$$\frac{dV_{CO}}{dt} = \frac{i_L}{C} - \frac{V_{CO}}{RC} \tag{15}$$

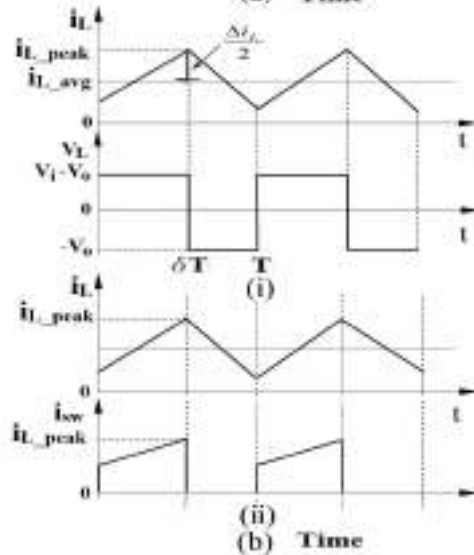
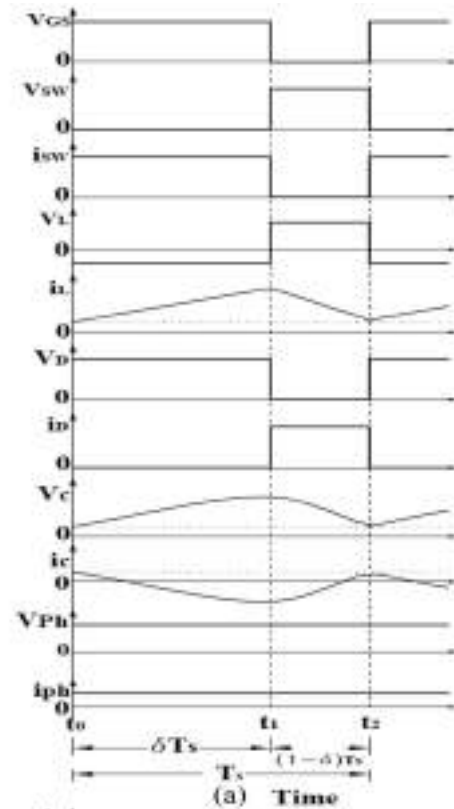


FIGURE 4. (a) AFE converter switching characteristics wave forms and (b) Waveforms of voltage and current of (i) Inductor and (ii) High-frequency switch.

On applying these equations into single input single output system model, the standard variables are used.

$$\left. \begin{aligned} \frac{dx_1}{dt} &= \frac{-1}{L}x_2 + \frac{V_i}{L}U \\ \frac{dx_2}{dt} &= \frac{1}{C}x_1 - \frac{1}{RC}x_2 \end{aligned} \right\} \tag{16}$$

where, system input 'U', system output 'y', and system state 'X'. On rewriting the system equations in terms of the

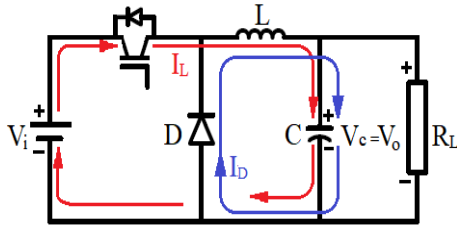


FIGURE 5. DC-to-dc buck converter topology.

standard variables, the inductor current, $i_L = x_1$, capacitor voltage, $V_{co} = x_2$ and duty ratio is the input 'u' and Output, $y = V_{co} = x_2$. Thereby the dynamic averaged equations can be rewritten with the new variables. Hence the state equations are (16). This equations satisfies linearity. By using sinusoidal PWM switching technique, the input dc-source voltage (V_i) is transformed into a fully rectified sinusoidal voltage waveform. Its frequency is same as the reference signal. In case of the proposed topology, the duty cycle is in a rectified sinusoidal fashion. Hence, the transfer function of input is

$$U(s) = \frac{1}{s^2 + 1} \tag{17}$$

To obtain the transfer function, from the state equation apply Laplace transform

$$\left. \begin{aligned} CY \left(S + \frac{1}{RC} \right) &= X_1 \\ SCY \left(S + \frac{1}{RC} \right) &= \frac{-1}{LC} Y + \frac{V_i}{CL} U \end{aligned} \right\} \tag{18}$$

On rearranging the equation,

$$Y \left(S^2 + \frac{1}{RC} S + \frac{1}{LC} \right) = \frac{V_i}{CL} U \tag{19}$$

The transfer function of the buck converter is given by the equation.

$$\frac{Y(s)}{U(s)} = \frac{V_i/LC}{S^2 + \frac{1}{RC} S + \frac{1}{LC}} \tag{20}$$

The rectified sinusoidal output voltage across buck capacitor is inverted using an H-bridge inverter. The converter output transfer function is mathematically expressed as

$$y(s) = \frac{V_i/LC}{S^2 + \frac{1}{RC} S + \frac{1}{LC}} \mathcal{L}(|\sin \omega t|) \tag{21}$$

For mathematical modelling, the following parameters are used. The design values are taken as $L = 9.7mH$, $C = 0.23\mu F$, $R_L = 50\Omega$. The numerator and denominator coefficients of the transfer function are $V_i/LC = 1519 \times 10^8$ and $s^2 + 8.7 \times 10^4 s + 4.5 \times 10^8$ respectively. Mathematical model of ILI and its output response are shown in Fig. 6.

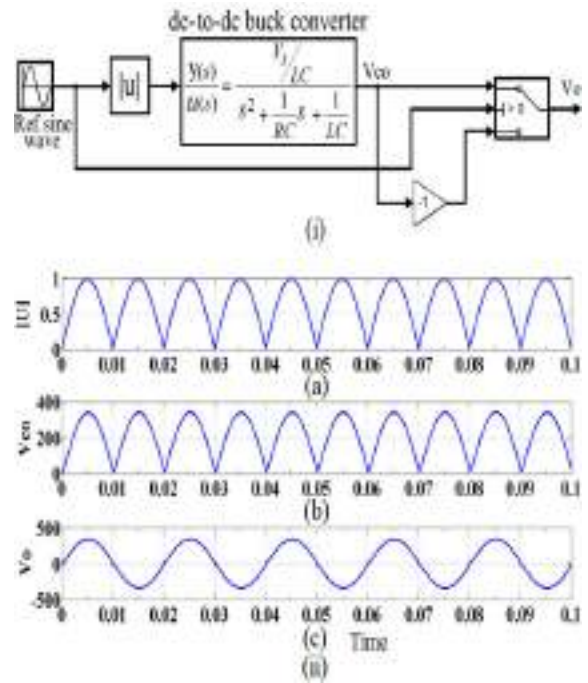


FIGURE 6. i) Mathematical model of proposed converter, (ii) Simulated output waveforms (a) reference signal, (b) rectified sinusoidal output voltage waveform generated by means of dc-to-dc buck converter transfer function, (c) sinusoidal output voltage waveform obtained from the inverter.

VI. DESIGN OF ILI

A. INDUCTOR (L)

In the ILI, the design of inductors plays a crucial role. The inductors are placed in the high-frequency operating part of the circuit, and it is fabricated using E65 ferrite core. When the switch 'Qc1' is turned ON for time period 'δT', the voltage across the inductor (V_L) is given by the difference between supply voltage and the output voltage.

$$V_L = L \frac{di_L}{dt} \tag{22}$$

The linear change in current (ΔI_L) through the inductor is,

$$\Delta I_L = \frac{V_i - V_o}{L} \delta T \tag{23}$$

Fig. 4(a) shows the voltage and current waveforms of the inductor and Fig. 4(b) shows those wave forms across the high frequency switch. The peak current through the inductor considered is $I_{Lavg} + \frac{\Delta I_L}{2}$. When the high frequency switch Q_{c1} is turned OFF, the freewheeling diode conducts during $(1 - \delta)T$ period and voltage of $-V_o$ appears across the inductor. Hence, voltage stress across the switch during this period is equal to V_i . The inductor current is controlled by two voltages which are appearing at both ends of the inductor (L). One is a high-frequency PWM pulsating voltage appearing across the diode (D) and the other is the voltage across the capacitor (C), which is fully rectified sinusoidal voltage waveform of 10 kHz. Output voltage is a pure sinusoid super imposed with ripple.

The average duty ratio is taken to be 50%. The rms value of the fundamental component of inductor current ripple ($I_{Lripple}$) can be calculated as

$$\frac{4}{\sqrt{2\pi}} \left[\frac{V_i}{2} \right] = (2\pi f_s L) I_{Lripple} \quad (24)$$

The duty cycle is getting varied continuously from 0 to δ_{max} in every half cycle of sinusoidal output wave. Depending upon the modulation index, δ_{max} will also vary. Hence, 50% duty cycle is taken as an average operating point. The maximum permitted amount of ripple is 5%. Therefore, $I_{Lripple} \leq 5\%I$ where I is rms value of the load current. The inductor value can be calculated as

$$L \geq \frac{\left(\frac{4}{\sqrt{2\pi}}\right)\left(\frac{V_i}{2}\right)}{(2\pi f_s) I_{Lripple}} \quad (25)$$

B. CAPACITOR (C)

The ripple in the load current is considered as negligible. Hence, the ripple current through the capacitor ($I_{Cripple}$) is equal to the ripple current of the inductor current. So, the ripple voltage of the capacitor is

$$V_{Cripple} = \frac{I_{Cripple}}{C\omega} \quad (26)$$

where $V_{Cripple} \leq 5\%V_o$ and in turn,

$$C \geq \frac{I_{Cripple}}{2\pi f_s (V_{Cripple})} \quad (27)$$

In order to ensure that output waveform is unaffected by the resonance caused by L and C , their values must be chosen such that

$$LC \leq \left(\frac{1}{2\pi 10f_s}\right)^2 \quad (28)$$

VII. SELECTION OF ILI COMPONENTS

A. SWITCHES

ILI has two different topological stages. (i) AFE Converter operating in power frequency and (ii) H-Bridge operating at high switching frequency.

Stage-I: In AFE converters, the high-frequency switches and diodes (D) are subjected to the same voltage. The voltage stress experienced across these devices is the same as the source voltage (V_i). During the turn OFF period $(1 - \delta)T$, the voltage stress across the high-frequency switch becomes equal to V_i , at the same time D is turned ON. During δT period, high frequency switches are turned ON and D is turned OFF. The diodes act as an open switch. Hence, the voltage stress across diode is V_i . In order to ensure safe operation of the switches, their ratings are taken to be higher than ' V_i '.

Stage-II: After every half cycle a couple of switches are switched ON, while the other switches are switched OFF. The effective voltage stress across the H-bridge switches is equal to the load voltage (V_o). Hence their voltage ratings must be higher than V_o . Current passes through the switches only when it is turned ON. The maximum current flowing

through the switches is equal to the peak current through the inductor (I_{Lpeak}).

$$I_{SWrating} = I_{Lpeak} \quad (29)$$

Voltage stress ($V_{SWstress}$) experienced across the high frequency switches, during the turn OFF period is

$$V_{SWstress} = V_i \quad (30)$$

B. DIODE

Current passes through the diode only when the high frequency operated switch is turned OFF. Its rating is chosen as

$$I_{Drating} = I_{Lpeak} \quad (31)$$

When the diode becomes reverse biased the voltage stress exerted across it is equal to ' V_i '.

$$V_{Dstress} = V_i \quad (32)$$

C. OUTPUT CAPACITOR

The maximum voltage (V_{Cmax}), applied across the capacitor is equal to V_i . The capacitor carries only the ripple component of current. Hence,

$$I_{Cpeak} = \Delta I_L \quad (33)$$

VIII. EFFICIENCY

Efficiency of any power electronic circuit is determined by the losses taking place in switches as well as in passive components. Losses in switching devices include switching losses and conduction losses. Switching loss depends on switching frequency and conduction loss depends on the load current. Since only one switch per phase is operated in high frequency, switching loss in this topology will be less. Moreover, total number of switches in this topology is much less than that in any other multilevel circuit, making conduction losses much lesser in this topology.

The simplified mathematical expression for the conduction and switching losses of IGBT switch in AFE converter is based on equation (34)

$$P_D = V_{ce_sat} I_m + \frac{1}{2} [(T_{ON} \frac{V_s I_m}{2}) + (T_{OFF} \frac{V_s I_m}{2})] f_{sw} \quad (34)$$

where P_D is the switching power dissipation, I_m is the collector current passing through the switch, V_{ce_sat} is the on-state drop, f_{sw} is the switching frequency, T_{ON} is the turn-on period, T_{OFF} is the turn-off period and V_s is the standing voltage of the switch. In case of diodes, the forward conduction and switching losses are considered. Meanwhile, the inductor and capacitor loss of the AFE converter is calculated by equation (35). ESR is also considered for the calculation of capacitor loss.

$$\left\{ (V_f I_m) + [V_D I_{rr} t_{rr} f_{sw}] \right\} + \left\{ \left(I_{Lpeak} + \frac{I_{ripple}}{2} \right) R_L \right\} + \{ 0.1 I_L ESR \} \quad (35)$$

where V'_D is the voltage across the diode. In polarity generation part, the H-bridge inverter switches are operating in soft-switching manner at fundamental frequency. The conduction and switching losses of H-bridge inverter comes to

$$(V_{ce_sat} I_m) N_{sw} + \frac{1}{2} \left[\left(T_{ON} \frac{V_s I_m}{2} \right) + \left(T_{OFF} \frac{V_s I_m}{2} \right) \right] f_L N_{sw} \quad (36)$$

The calculated value of switching losses per phase for both high frequency operating switch and H-bridge switches are 1.26W and 0.025W respectively. In theoretical loss calculation, the efficiency of the inverter is found to be 98%.

IX. THIRD HARMONIC INJECTION METHOD

The SPWM control technique used in conventional VSI, can obtain sinusoidal output voltage, with less harmonic distortion, however the maximum obtained output fundamental amplitude, is only 78.54 %. while, $M = 1$ (Line voltage = $0.6123V_iM$). To expand the output fundamental amplitude beyond from this limit, the ‘ $M > 1$ ’ have to be taken more than one. To improve the dc-bus voltage utilization of the inverter, without entering the over modulation, the third harmonics injection logic can be used. The third harmonics injection in the leg voltage references leads to an increase in the linear modulation range of three-phase VSI, by reducing the peak of the leg reference voltages and hence the modulation index can be pushed beyond the value of one without entering over modulation. By injecting the third harmonics with amplitude of one sixth of the fundamental harmonics, the maximum of the fundamental can be increased by 15.47% as compared to the SPWM control scheme (Line voltage = $1.4142V_iM$) The same concept is applied to the proposed topology, consequently the linear modulation range is extended. Hence, this increases the maximum fundamental output voltage without moving into the over modulation region. Moreover the third harmonic component does not affect the output phase voltages. Fig. 7 shows the logic for implementing the third harmonic injection PWM. It helps to improve the dc-bus voltage utilization of ILI as well as the fundamental amplitude of output line-to-line voltage by 15.47%. Thereby, effectively improve the dc-bus voltage utilization. The duty ratio of an ILI is

$$\delta_{ref} = |d \sin(\omega t)| \quad (37)$$

A third harmonic component of $k \sin(3\omega t)$ is superimposed with the sinusoidally varying duty cycle. The optimum point obtained is at $d = 2/\sqrt{3}$ and $k=1/6$ without any over modulation, Hence the reference signal is modified as equation (38)

$$\delta_{ref} = |d \sin \omega t + k \sin(3\omega t)| \quad (38)$$

The instantaneous value of the output voltage across the converter capacitance is

$$V_{co_fund} = |1.1547V_m \sin(\omega t)| + |\frac{1}{6}V_m \sin(3\omega t)| \quad (39)$$

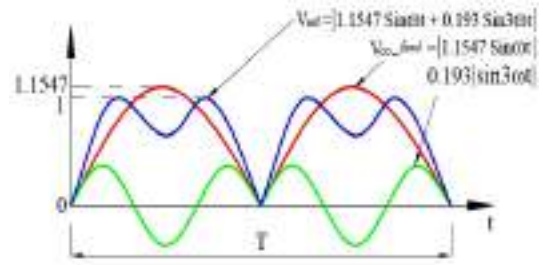


FIGURE 7. Third harmonic injection PWM control implementation logic.

The output line voltage across the load is

$$V_{o_fund} = 1.1547V_m \sin(\omega t) \quad (40)$$

The magnitude of three-phase output voltage becomes (41), (42), and (43).

$$V_{o1_fund} = 1.1547V_m \sin(\omega t); \quad 0 \leq \omega t \leq 2\pi \quad (41)$$

$$V_{o2_fund} = 1.1547V_m \sin(\omega t - 120); \quad 0 \leq \omega t \leq 2\pi \quad (42)$$

$$V_{o3_fund} = 1.1547V_m \sin(\omega t + 120); \quad 0 \leq \omega t \leq 2\pi \quad (43)$$

X. SCALAR AND VECTOR CONTROL OF ILI FED INDUCTION MOTOR

This section discusses about the speed control of open-ended three-phase induction motor using different control schemes, such as scalar (V/f) and direct vector control. Using the proposed ILI topology, the induction motor speed can be controlled with high dynamic performance.

A. SCALAR (V/F) CONTROL METHOD

The V/f control technique is an attractive method to control the induction machine speed because of its simplicity and user-friendly nature. Block diagram of scalar control implementation logic is shown in Fig. 8. Here, the air-gap flux of the induction machine can be controlled at desired value by proportionally varying stator voltages and frequency. Consequently, the machine retains its torque/ampere capacity at any speed. Meanwhile, the machine speed can be accurately maintained at any desired value under steady-state condition. However, at low-speed range, the torque capability is limited because of the voltage drop across the stator winding resistance, which is dominant.

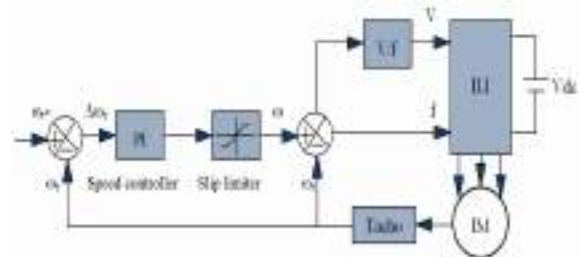


FIGURE 8. Scalar control of ILI fed induction motor implementation logic.

B. DIRECT VECTOR CONTROL METHOD

The block diagram of the direct vector control system is shown in Fig. 9. Here the field angle is calculated using

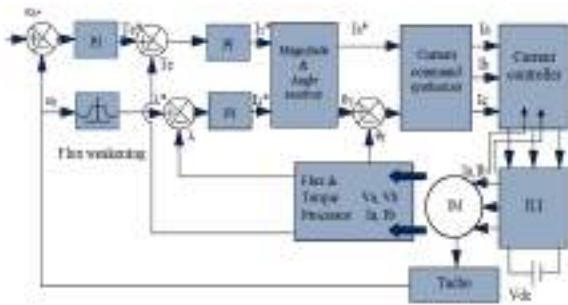


FIGURE 9. Direct vector control of ILI fed induction motor.

the terminal voltages and currents. The actual speed ω_r is compared with reference speed ω_r^* , the error is amplified and limited to generate the required reference torque T_e^* . The rotor flux linkage λ^* is kept at 1pu. Beyond 1pu, it is varied as a function of speed. This is to ensure that the rotor speed extends beyond the base speed by weakening the rotor flux linkages. The torque and flux references are compared to the actual torque T_e and flux linkage λ_r , calculated from the measured voltages and currents. The errors are amplified and limited to generate the required flux and torque producing components of current I_F^* and I_T^* . Phasor addition of these components yields the phasor reference current I_s^* and torque angle (θ_r) can be calculated from (52). The sum of torque angle and field angle gives the position of the stator current phasor I_s . Together with I_s^* , this generates the stator phase current references $I_{as}^*, I_{bs}^*, I_{cs}^*$.

The PWM control scheme is implemented using phase current control loops. The following equations represent the stator phase based calculation of various quantities.

$$V_{ds} = (R_s + L_s * p) I_{ds} + (L_m * p) I_{dr} \quad (44)$$

$$V_{qs} = (R_s + L_s * p) I_{qs} + (L_m * p) I_{qr} \quad (45)$$

$$I_{dr} = \frac{1}{L_m} \left\{ \int (V_{ds} - R_s I_{ds}) dt - L_s I_{ds} \right\} \quad (46)$$

$$I_{qr} = \frac{1}{L_m} \left\{ \int (V_{qs} - R_s I_{qs}) dt - L_s I_{qs} \right\} \quad (47)$$

$$\lambda_{dr} = L_r I_{qr} + L_m I_{qs} \quad (48)$$

$$\lambda_{qr} = L_r I_{dr} + L_m I_{ds} \quad (49)$$

Flux,

$$\lambda_r = \sqrt{\lambda_{dr}^2 + \lambda_{qr}^2} \quad (50)$$

Flux position,

$$\theta_f = \tan^{-1} \frac{\lambda_{qr}}{\lambda_{dr}} \quad (51)$$

$$\theta_T = \tan^{-1} \frac{I_T^*}{I_F^*} \quad (52)$$

Stator current phase angle $\theta = \theta_f + \theta_T$ Torque equation,

$$T_e = \frac{3P}{2} L_m (I_{qs}^* I_{dr} - I_{ds}^* I_{qr}) \quad (53)$$

Reference current is calculated as

$$I_s^* = \sqrt{(I_{ds}^*)^2 + (I_{qs}^*)^2} \quad (54)$$

Three phase reference currents are

$$\left. \begin{aligned} I_{as}^* &= I_s^* \sin(\theta) \\ I_{bs}^* &= I_s^* \sin(\theta + 120) \\ I_{cs}^* &= I_s^* \sin(\theta - 120) \end{aligned} \right\} \quad (55)$$

XI. SIMULATION RESULTS OF ILI

A. SPWM CONTROL

The proposed inverter topology and its control are realized using MATLAB/Simulink environment. A three-phase balanced resistive load of 50Ω is taken for the simulation analysis. A switching frequency of 10 kHz is used for sine PWM pulse generation. Fig.10-13 shows the simulated waveforms of ILI, using resistive load. The power quality of the output voltage can be improved by increasing the AFE converter operating frequency. It significantly reduces the THD, which is less than the permissible limit as per IEEE-519 standards. The FFT analysis is shown in Fig. 14. This scheme requires only a 338V dc-source for obtaining 415V ac 50 Hz line-to-line output voltage.

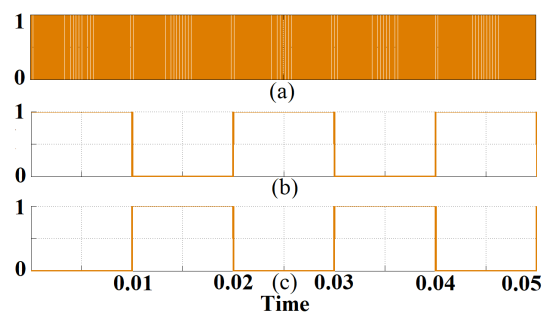


FIGURE 10. Simulated waveforms of ILI. (a) High frequency, (b,c) Low frequency switching pulses.

B. THIRD HARMONIC INJECTION PWM CONTROL

The dc-source voltage utilization of the proposed inverter has been again validated using third harmonics injection PWM method. It requires only 293V dc-source voltage, in order to obtain 415V ac, 50Hz line-to-line output voltage. The simulated waveform of the third harmonic injection scheme is shown in Fig. 15. The proposed inverter is compared to a traditional two-level inverter topology. For a similar output voltage, a SPWM control scheme requires 654V dc-source voltage and an ordinary third harmonic injection scheme requires 564V dc-source. Fig.16 shows the comparison of dc-source voltage utilization between ILI and traditional inverter. The proposed topology has very high dc-source voltage utilization, compared to the traditional two-level inverter.

C. PERFORMANCE EVALUATION OF ILI TOPOLOGY USING THREE PHASE INDUCTION MOTOR

1) SPEED CONTROL USING SCALAR (V/F) CONTROL METHOD

The variable speed control performance of the proposed topology is validated using a three-phase induction motor. The speed control logic is implemented using constant V/f

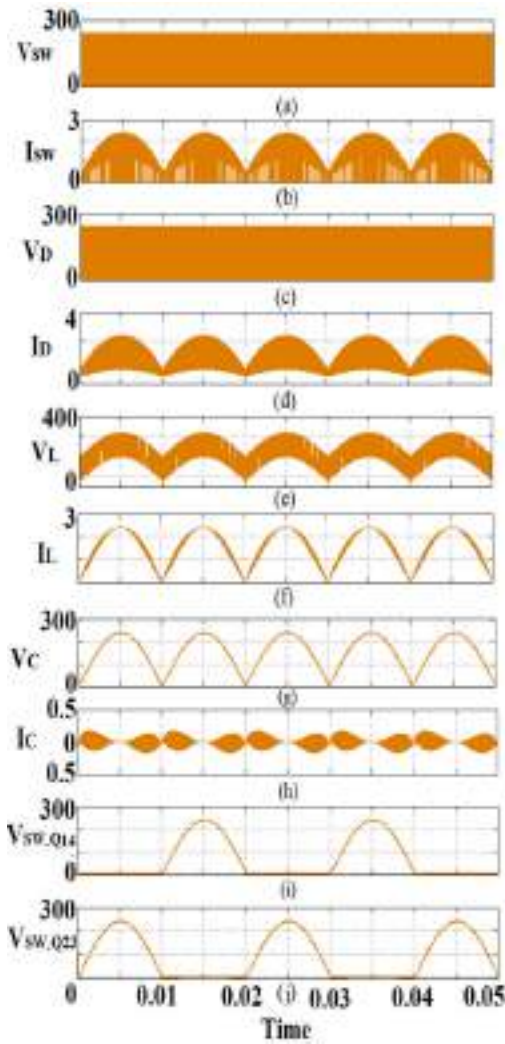


FIGURE 11. Simulated waveforms of ILI using resistive load. Voltage and current wave forms across the AFE converter components. (a,b) High frequency switch, (c,d) diode,(e,f) inductor, (g,h) capacitor,(i,j) voltage across low frequency operating switches.

scheme. The dynamic responses of the simulated output voltage waveforms using the V/f control are shown in Fig. 17. To verify the rotor start-up performance, the motor runs from standstill condition to a pre-defined speed of 300rad/s by applying 415V, 50Hz source voltage. Thereby, the rotor speed is attained and critically stabilized at that pre-defined speed with in a fraction of second.

2) SPEED CONTROL USING DIRECT VECTOR CONTROL METHOD

Direct vector control operation was simulated as per the block diagram mentioned in Fig.9. Reference speed was increased as well as decreased. The dynamic response of output voltage waveforms as well as the voltage across the buck capacitor is given in Fig. 18 for acceleration as well as deceleration.

XII. COMPARISON OF ILI WITH CONVENTIONAL H-BRIDGE INVERTERS AND AFE-RDC-MLI TOPOLOGIES

TABLE-2 compares the active components, passive components and THD of the 2-level H-bridge, 3-level H- bridge,

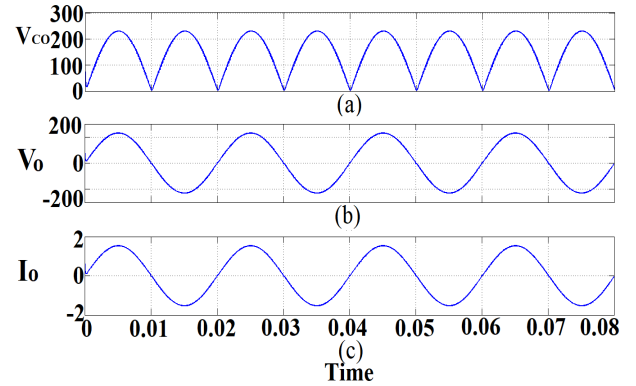


FIGURE 12. Simulated waveforms of ILI using resistive load. (a) Voltage waveform across the buck capacitor. (b) Voltage, (c) current waveforms across the load resistance.

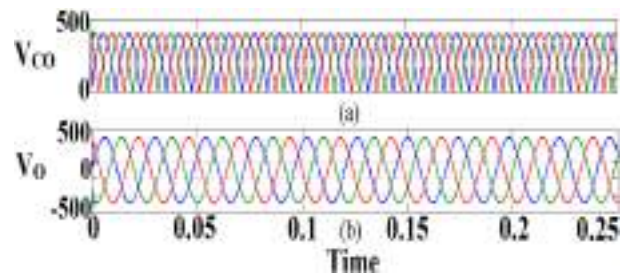


FIGURE 13. Simulated waveforms of ILI using resistive load. (a) Three-phase output voltage waveforms across the buck capacitor, (b) Three-phase output voltage wave form across the load resistance.

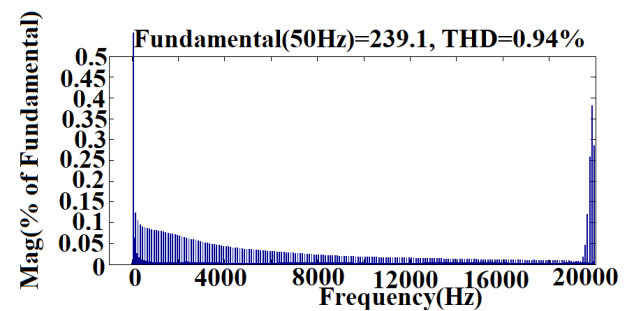


FIGURE 14. FFT analysis of output voltage waveform of the ILI.

5-level cascaded H-bridge MLI with the proposed ILI. Fig. 19 shows the simulation results of the output voltage wave forms of conventional and proposed inverter topologies. TABLE-4 shows the comparative analysis of different AFE-RDC-MLI (single phase circuit) topologies with proposed inverter. In comparison with conventional topologies, the proposed topology uses one high frequency operated switch and four power frequency operating switches per phase, thereby reduces the switching loss of the proposed inverter. Moreover, AFE converter circuit is a part of the proposed inverter that can generate non-finite number of voltage levels, which can improve the THD of the output voltage.

XIII. EXPERIMENTAL SETUP AND RESULTS

The hardware setup for both scalar and vector control implementation using ILI topology is shown in Fig. 20.

TABLE 2. Comparison between the proposed ILI and the conventional H-bridge inverters, (* Output voltage levels depends on switching frequency).

Sl. No.	Topology Type	High frequency operating switches (Sw)	Power frequency operating switches	Diode (D)	Inductor (L)	Capacitor (C)	dc-Source	Output voltage Levels	Switching frequency (fsw)	Control	THD of output voltage
1	H-Bridge (2-Level)	4	0	0	0	0	1	2	10 KHz	SPWM	58.34%
2	H-Bridge (3-Level)	4	0	0	0	0	1	3	10 kHz	SPWM	37.04%
3	Cascaded H-Bridge MLI (5-Level)	8	0	0	0	0	2	5	10 kHz	SPWM	35.38%
4	ILI (Proposed)	1	4	1	1	1	1	*	10 kHz	SPWM	1.20%

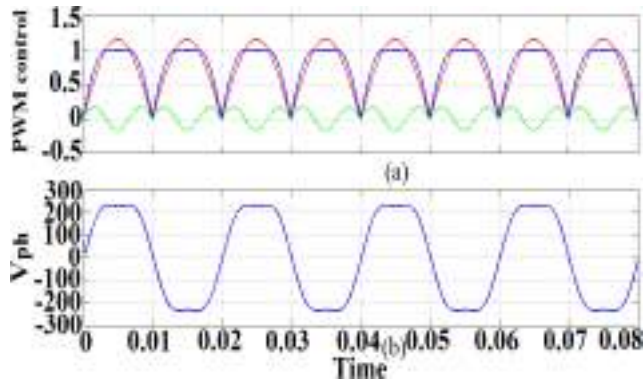


FIGURE 15. Simulated waveforms of (a) third harmonic injection PWM control implementation logic, (b) phase voltage waveform of the ILI using resistive load.

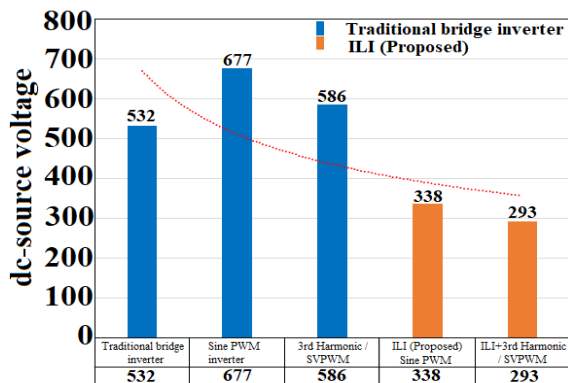


FIGURE 16. Comparison of dc-source voltage utilization of ILI with traditional bridge inverter.

TABLE-3 shows the parameters of ILI used for the experiment. For hardware tuning nearest available component values have been selected. Experiment was conducted on ILI topology under loads such as resistive and open-ended induction motor. Speed control of induction motor was carried out under constant V/f condition as well as under direct vector control method.

A. PERFORMANCE EVALUATION OF ILI TOPOLOGY USING SPWM CONTROL

Initially the performance of ILI topology was experimentally verified for a resistive load. NI PCIe-6351 card is used for data acquisition and control signal generation under

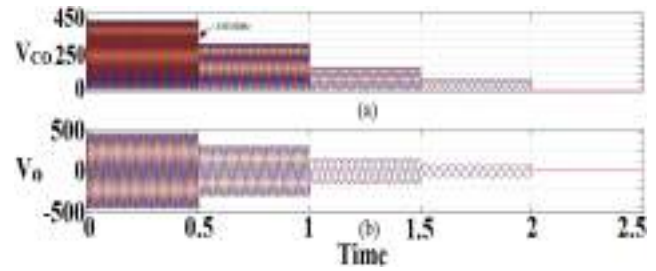


FIGURE 17. The dynamic responses of the simulated output voltage waveforms using V/f control. (a) Voltage waveform across the buck capacitor, (b) Line-to-line voltage across the load.

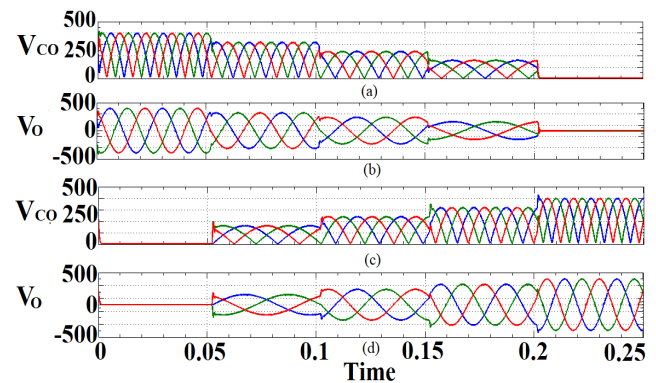


FIGURE 18. The dynamic responses of the simulated output voltage waveforms using direct vector control. (a,c) Voltage waveform across the buck capacitor, (b,d) Line-to-line voltage across the load.

TABLE 3. Parameters used for ILI.

Unit	Parameters	Values
kVA	Power rating	1
V	Inverter output voltage	240/415
Hz	frequency	50
mH	Inductor	9.7
μ F	Capacitor	0.23
Ω	Load resistance	50
kHz	Switching frequency	10

MATLAB Real-Time Windows environment. SPWM control is implemented using this setup. The voltage level of the pulses generated by the PCI card is only 3.3V. In order to make this voltage sufficient enough to drive an IGBT, a voltage level shifter card is used as a buffer circuit. SEMIKRON IGBT modules- SKM50GB12T4 and MUR 860G ultrafast power diodes had been used for the hardware realization.

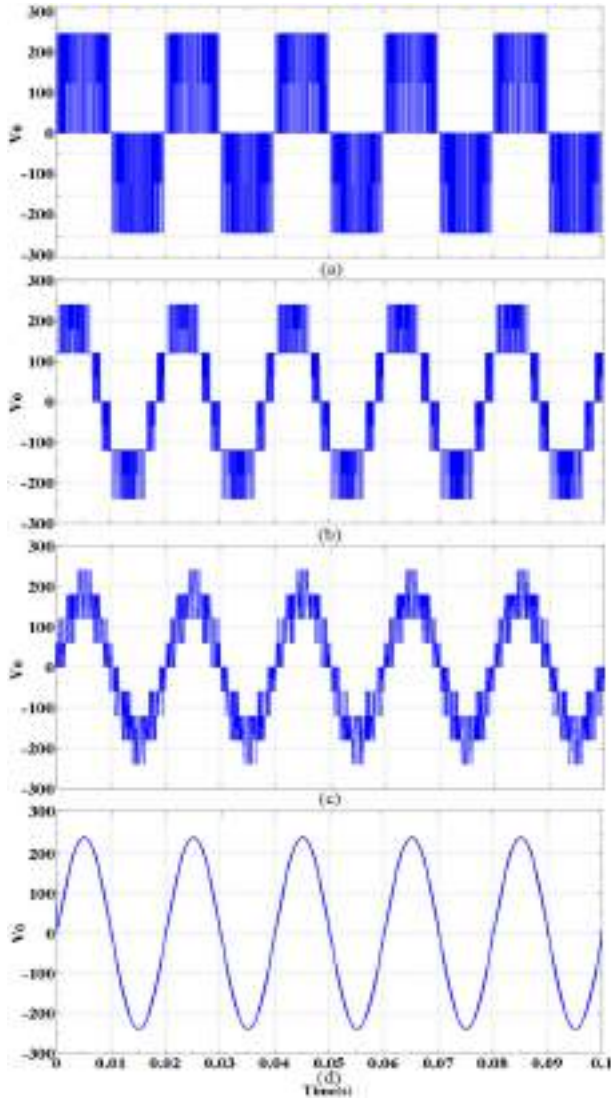


FIGURE 19. The simulated output voltage waveforms using resistive load (a) conventional 2-level H-bridge inverter, (b) 3-level H-bridge inverter, (c) 5-level cascaded H-bridge MLI, (d) Proposed topology.

Fully rectified sinusoidal output voltage is generated across the buck capacitor. This is then converted into sinusoidal voltage waveform using the H-bridge. The load voltage and current were measured using sensors LA 55-P and LV-25-1000, respectively. Fig. 21 shows the voltage and current waveforms of the AFE converter components. Fig. 22 shows the experimental results of ILI using resistive load and Fig. 23 shows the THD of ILI output voltage waveform using FLUK 434 energy analyser, and it is found to be 1.2%. The peak value of the pole voltage in a conventional inverter is $V_{dc}/2$. In case of ILI, the presence of the buck converter makes the peak value of per phase voltage equal to V_{dc} itself. In this case, this results in the reduction of dc-source requirement. The inherent filtering present with the buck converter makes the voltage a smoothly varying one, which results in the THD reduction.

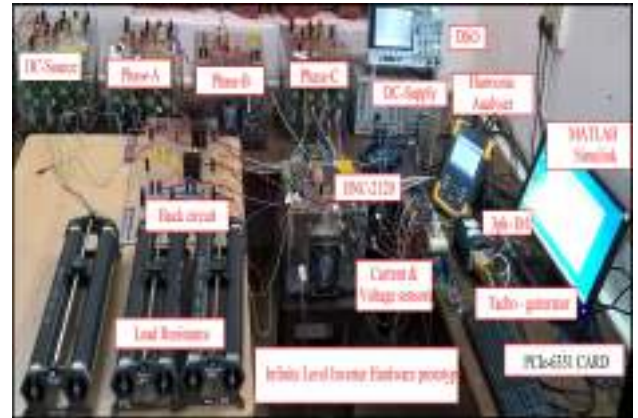


FIGURE 20. Experimental setup of an ILI to control an open-ended three-phase induction motor implemented using scalar and vector control logic.

B. PERFORMANCE EVALUATION OF ILI TOPOLOGY USING THIRD HARMONIC INJECTED PWM CONTROL

In order to increase the output voltage further without resorting to over modulation, harmonic injection control is used. Since triplen harmonic waves are in phase with all the three phase voltages, they cancel out at the line to line voltages. This property makes it possible to have a modulation index higher than one, without any over modulation of the modulating signal with respect to the carrier wave. DC-source voltage utilization of the proposed inverter is experimentally validated using third harmonic injection PWM method, which is almost similar to SVPWM technique. Fig. 24 shows the experimental voltage waveforms under this condition. The main advantage of third harmonic injection PWM is that, the fundamental amplitude of output line-to-line voltage increases by 15.4%, thereby effectively improving the dc-source voltage utilisation of the proposed MLI topology. The test results prove that, while it requires 338V dc-source voltage supply for developing 415V output voltage under SPWM control, third harmonic injected PWM control requires only 293V for maintaining the same voltage level.

C. PERFORMANCE EVALUATION OF ILI TOPOLOGY USING OPEN-ENDED THREE PHASE INDUCTION MOTOR

1) SPEED CONTROL USING V/f CONTROL METHOD

In case of conventional inverters, switching signals for V/f control are generated by comparing the modulation signal with a carrier wave. The two switches in the same leg are operated in a complimentary manner. In case of ILI, the modulation signal is rectified and compared with the carrier signal for generating switching signals for a buck converter. The practical testing of proposed ILI is carried out using a three-phase open-ended induction machine with NI-data acquisition card prototyping system. Fig. 25 shows the block diagram of real-time implementation. The actual rotor speed of the induction motor is measured using electronic tacho-generator (analog output of 10V-dc at 1500 rpm),

TABLE 4. Comparative analysis of different AFE-RDC-MLI (single phase circuit) topologies with proposed inverter. (Output voltage levels depends on switching frequency).**

Sl. No.	Ref. No.	Topology Type	Sw	D	L	C	dc-source	AFE-converter	Output voltage levels	fsw	Polarity generation	Control	Efficiency	THD of output voltage	RDC	Applications
1	[49]	Conventional three-level PWM inverter	5	1	1	1	1	Boost	3	20kHz	H-Bridge	PWM	90%	8.00%	Yes	PV-Grid connected
2	[44]	Cascaded MLI	10	2	2	2	2	Buck	5	1.6kHz	H-Bridge	Multilevel Selective Harmonic Elimination -PWM	97%	23.45%	Yes	Power system , STATCOM
3	[45]	Buck-Boost MLI	8	0	2	1	1	Buck-Boost	5	24kHz	H-Bridge	Phase Opposition Disposition - PWM	-	22.65%	Yes	Renewable energy generation
4	[46]	Five-level inverter	8	1	1	3	1	Boost	5	20kHz	H-Bridge	Level Shifted - SPWM	96.50%	2.10%	Yes	For low voltage de-source
5	[47]	Step-Up Five-Level inverter.	6	3	1	2	1	Boost	5	15kHz	H-Bridge	Level Shift Multicarrier-PWM	-	4.80%	Yes	-
6	[50]	Modified quasi-Z-source cascaded hybrid inverter.	12	6	4	8	2	quasi-Z-source	5	10 kHz,	H-Bridge	Alternative phase opposition disposition	91%	3.60%	Yes	For high output voltage with lower THD
7	[49]	Five-level inverter	6	5	1	2	1	Boost	5	20kHz	H-Bridge	PWM	86%	5.40%	Yes	PV-Grid connected
8	[48]	Seven-level boost inverter.	8	4	1	2	1	Boost	7	5kHz	H-Bridge	Phase shifted -SPWM	-	14.92%	Yes	Electric vehicle
9	[42]	Seven-Level Grid-Connected Inverter	7	10	1	4	1	Boost	7	-	H-Bridge	PWM	Yes	3.90%	Yes	PV-Grid connected
10	[43]	Boost DC-link cascaded MLI	10	2	2	2	2	Boost	7	-	H-Bridge	Multicarrier-SPWM	-	4.16%	Yes	UPS and AC-Drive
11	[51]	Z-source-based MLI	10	3	6	6	3	Z source	7	-	H-Bridge	PWM	-	24.05%	Yes	Power quality; DVR
12	[52]	Integrated Semi-Double Stage based MLI	12	2	2	3	1	Interleaved -Boost	7	20 kHz,	H-Bridge	PWM	91.04%	1.26%	yes	PV-Grid connected
13	[38]	Twenty-Five-Level Cascade Switched-Diode Converter	12	4	0	0	4	Switched -Diode	25	-	H-Bridge	Fundamental frequency-switching	93.99%	2.19%	Yes	Renewable energy sources and medium -voltage applications
14	*	ILI	5	1	1	1	1	Buck	**	10kHz	H-Bridge	SPWM	98%	1.20%	Yes	Induction motor drive

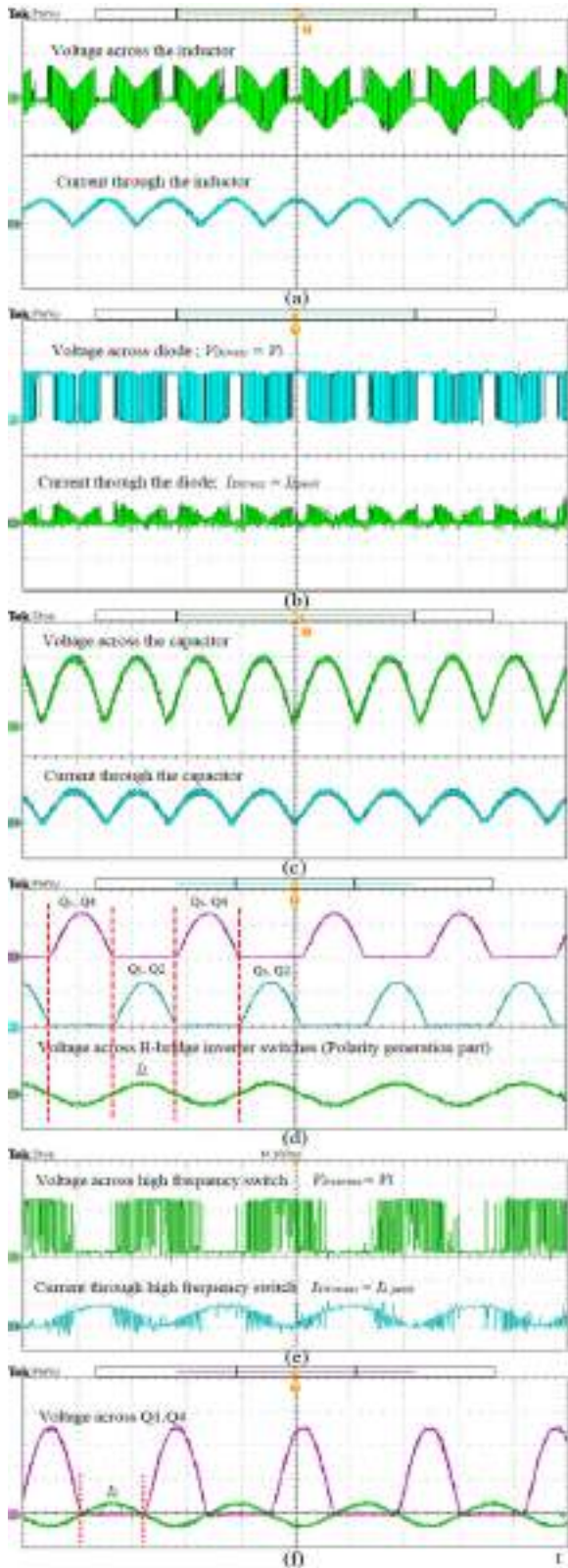


FIGURE 21. Hardware experimental results of ILI using resistive load. Voltage and current performance wave forms across the AFE Converter components (a) inductor, (b) diode, (c) capacitor, (d) voltage across low frequency operating switch, (e) voltage across high frequency operating switch, (f) Voltage and current waveform through Q1 and Q4.

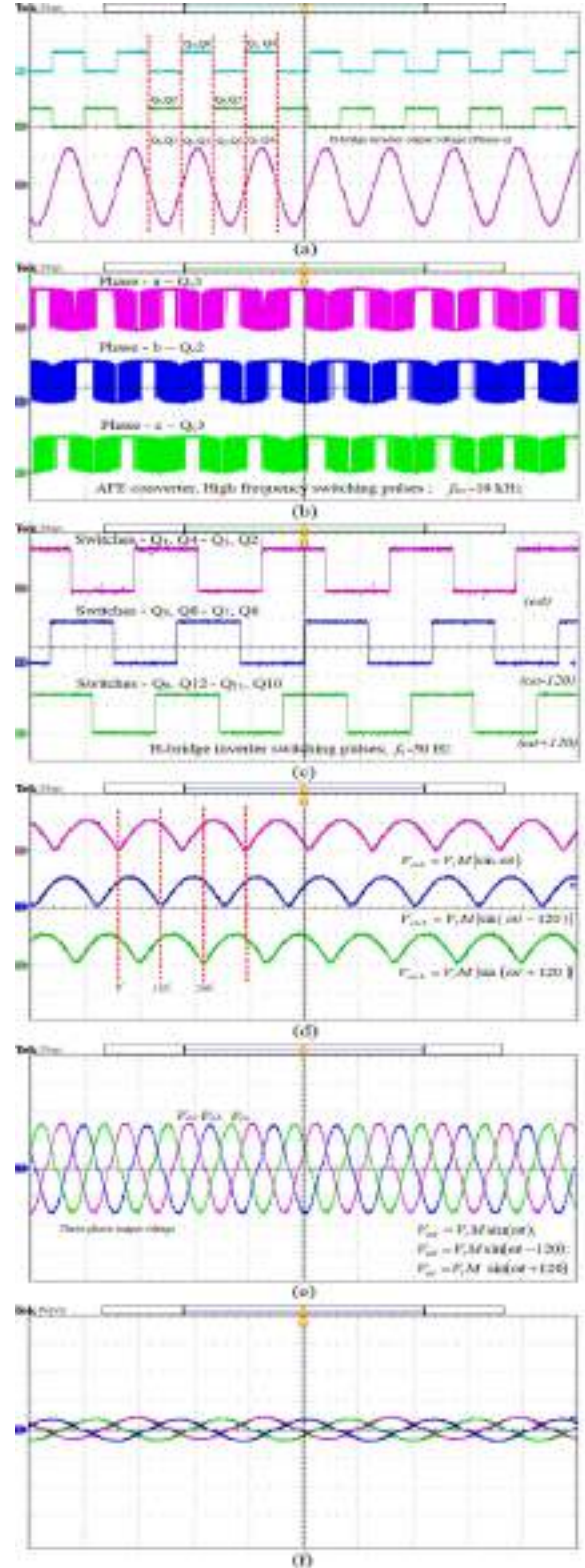


FIGURE 22. Hardware experimental results of three-phase ILI using resistive load. (a) H-bridge switching pulses and single-phase ILI output voltage waveform, (b, c) High and Low frequency switching pulses, (d) voltage waveforms across buck capacitors, (e) output voltage waveforms across the load terminals, (f) current waveforms.



FIGURE 23. Hardware experimental result; THD of output voltage waveform of the ILI.

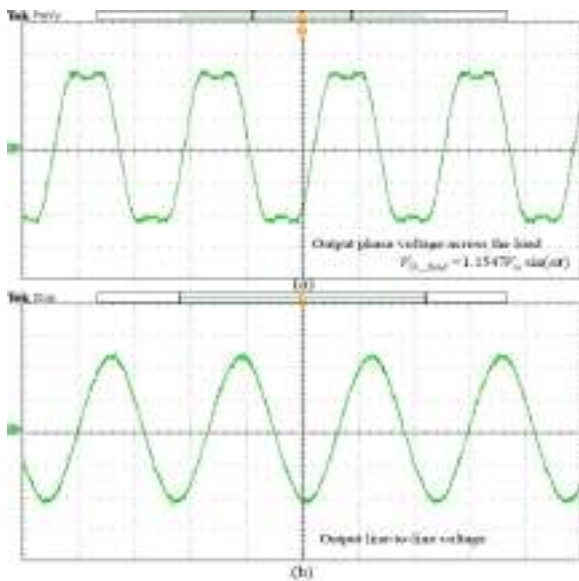


FIGURE 24. Hardware experimental results of ILI under resistive load, obtained by third harmonic injection PWM method. (a) Phase voltage waveform. (b) Line-to-line voltage waveform.

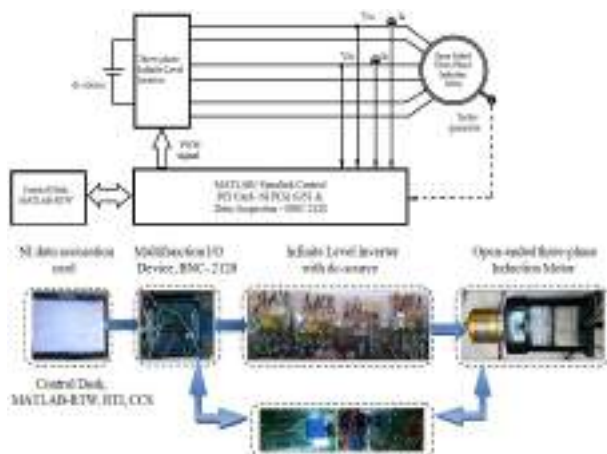


FIGURE 25. Real-time implementation block diagram.

and the corresponding dc-voltage value is fed into the computer through NI-PCIE-6351 Card. The capability of ILI for

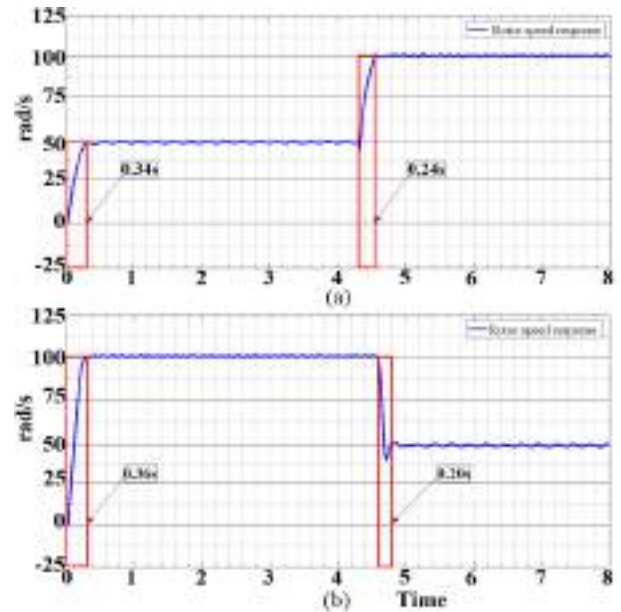


FIGURE 26. Experimental results of dynamic speed responses obtained from V/f control. (a) Increase of rotor speed from initial condition to pre-defined speed range 0 rad/s - 50 rad/s in 0.34s and then shifted to 100rad/s in 0.24s. (b) Rotor speed shifted from 0 rad/s to 100 rad/s in 0.36s and stepped-down speed transitions of varying rotor speed from 100 rad/s to 50 rad/s in 0.20s.

TABLE 5. Experimental data.

Motor speed responses	Speed control method V/f	
	Fig.26(a)	Fig.26(b)
Speed (Initial) in rad/sec.	0	0
Speed (Pre-fixed) in rad/sec.	50	100
Start-up time in sec.	0.34	0.36
Speed (Pre-fixed) in rad/s.	100	50
Speed (transition time) in Sec.	0.24	0.2
	Speed control method FOC	
	Fig.27	Fig.28
Speed (Initial) In rad/sec.	0	0
Speed (Pre-fixed) in rad/sec.	50	75
Start-up time in sec.	0.24	0.25
Speed (Pre-fixed) in rad/s.	100	135
Speed (transition time) in Sec.	0.12	0.12
Speed (Pre-fixed) in rad/sec.	50	75
Speed (transition time) in Sec.	0.11	0.12

running the induction machine under variable speed condition using V/f control scheme was experimentally verified. 10 kHz carrier signal was used for SPWM signal generation. In order to verify the rotor start-up performance, the motor was made to run from standstill condition to different pre-defined speed ranges. Fig. 26. (a) and (b) shows the MATLAB plot of speed responses of the induction motor. Rotor speed increased from 0 to 50 rad/s in 0.34s. It was further increased to 100 rad/s in 0.24s and got reduced to 50 rad/s in 0.2s.

2) SPEED CONTROL USING DIRECT VECTOR CONTROL METHOD

The induction motor speed can be more accurately controlled using the direct vector control logic. This scheme was implemented using the MATLAB/Simulink environment

TABLE 6. Comparison of simulation and experimental results.

Particulars	Simulation results	Figure Nos.	Hardware results	Figure Nos.	Parameters
Mathematical model of proposed converter. Fig. 6.	Simulated waveforms of ILI using resistive load. (a) Voltage waveform across the buck capacitor. (b) Voltage waveforms across the load resistance.	12(a,b)	Hardware experimental results using resistive load. (a) single-phase ILI output voltage waveform	22 (a)	Design values $L = 9.7mH$, $C = 0.23\mu F$, $R_L=50 \Omega$. $V_o = 240V$ $f = 50Hz$, $V_i = 338V$
Different switching pulses of ILI.	(a) High frequency, (b,c) Low frequency switching pulses	10 (a,b,c)	High and Low frequency switching pulses,	22 (b,c)	Switching frequency at High: $10kHz$, Low: $50Hz$.
Voltage and current wave forms across the AFE converter components. (Under resistive load)	(a, b) High frequency switch, (c, d) diode, (e, f) inductor, (g, h) capacitor, (i, j) voltage across low frequency operating switches.	11(a-j)	(a) inductor, (b) diode, (c) capacitor, (d) voltage across low frequency operating switch, (e) voltage across high frequency operating switch, (f) Voltage and current waveform through low frequency operating switch	21 (a-f)	The voltage stress across: High-frequency switches ($V_{SW\ stress} = V_i$) H-bridge switches = V_o . Diodes: ($V_{D\ stress} = V_i$) Inductors: = V_i , $V_o = 240/415V$
ILI output voltage waveforms across the load resistance using SPWM control.	(a) Voltage waveform across the buck capacitor. (b) Voltage and (c) current waveforms across the load resistance.	12(a,b,c)	(a) ILI output voltage (per phase) waveform.	22(a)	Output voltage(V_{ph}) = $240V$, $f = 50Hz$, $V_o = 1.4142V_i M$ $V_i = 338V$ for $V_o = 415V$ In traditional two-level inverter: • SPWM control scheme requires $654V$ dc-source voltage.
ILI three-phase output voltage waveforms (Under resistive load)	(a) Three-phase output voltage waveforms across the buck capacitor, (b) Three-phase output voltage waveform across the load resistance.	13 (a,b)	(d) Voltage waveforms across buck capacitors (e) Output voltage waveforms across the load terminals, (f) Current waveforms.	22(d-f)	ILI output voltage: $240/415V$
FFT analysis	FFT analysis of output voltage waveform of the ILI.	14	THD of output voltage waveform of the ILI	23	Simulation - 0.94% Hardware - 1.2%
i). Third harmonic injection PWM control (Under resistive load). ii). comparison of dc-source voltage utilization. (traditional two-level inverter v/s ILI)	i). (a) Third harmonic injection PWM control implementation logic. (b) Phase voltage waveform of the ILI. ii). dc-source voltage utilization between ILI and traditional inverter.	i).15(a,b) ii). 16	Third harmonic injection PWM method. (a) Phase voltage waveform. (b) Line-to-line voltage waveform	24(a,b)	i) $V_i = 293V$ for $V_o = 415V$ $V_o = 0.6123V_i M$ ii) In traditional two-level inverter: • An ordinary third harmonic injection scheme requires $564V$ dc-source voltage.
Dynamic speed responses obtained from V/f control. (Open-Ended Three-Phase Induction Motor load)	(a) Voltage waveform across the buck capacitor, (b) Line-to-line voltage across the load.	17(a,b)	(a) Increase of rotor speed from initial condition to pre-defined speed range. (b) Rotor speed shifted (retardation)	26(a,b)	(a) $0\ rad/s$ - $50\ rad/s$ in $0.34s$ and then shifted to $100rad/s$ in $0.24s$. (b) $0\ rad/s$ to $100\ rad/s$ in $0.36s$ and stepped-down speed transitions of varying rotor speed from $100\ rad/s$ to $50\ rad/s$ in $0.20s$
Dynamic speed responses obtained from direct vector control. (Open-Ended Three-Phase Induction Motor load)	(a, c) Voltage waveform across the buck capacitor, (b, d) Line-to-line voltage across the load.	18(a-d)	i). Increase of rotor speed from initial condition to a pre-defined speed ii). Increase of rotor speed from initial condition to a pre-defined speed. iii). (a, b). a.(i). Phase current and b.(ii). phase voltage across the motor winding.	i). 27 ii). 28 iii). 29	i). $0\ rad/s$ to $50\ rad/s$ in $0.24s$, then shifted to $100rad/s$ in $0.12s$ finally stepped-down to $50\ rad/s$ in $0.11s$. ii). $0\ rad/s$ to $75\ rad/s$ in $0.25s$, then shifted to $135\ rad/s$ in $0.12s$ finally stepped-down to $75\ rad/s$ in $0.12s$. iii). The rotor speed shifted from $130\ rad/s$ to $95\ rad/s$ by changing the phase voltage and frequency from $210V$, $45Hz$ to $150V$, $30Hz$ at $3.3s$. and vice versa.

with the same ILI hardware setup. The reference current components I_d and I_q obtained from flux and torque errors respectively are converted to reference phase voltages after subjecting them through axes transformations. These reference voltages form the modulation signals in the conventional

method. Here, these modulation signals are rectified and given as switching pulses to the buck converter. This rectified buck voltage is inverted using the H-bridge. Fig. 27 shows the dynamic speed responses obtained from the tachogenerator output of ILI fed induction motor. In order to validate the rotor

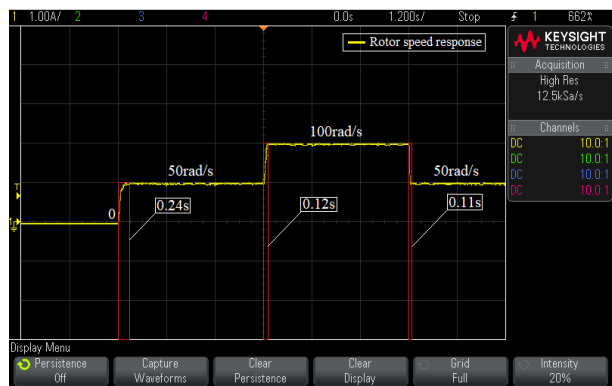


FIGURE 27. Experimental results of dynamic speed responses obtained from direct vector control. Increase of rotor speed from initial condition to a pre-defined speed of 0 rad/s to 50 rad/s in 0.24s, then shifted to 100rad/s in 0.12s finally stepped-down to 50 rad/s in 0.11s.

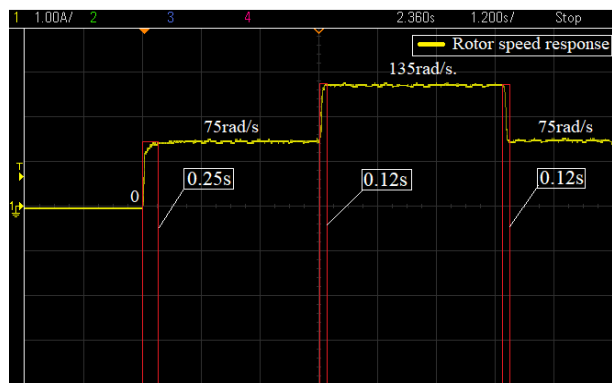


FIGURE 28. Experimental results of dynamic speed responses obtained from direct vector control. Increase of rotor speed from initial condition to a pre-defined speed of 0 rad/s to 75 rad/s in 0.25s, then shifted to 135 rad/s in 0.12s finally stepped-down to 75 rad/s in 0.12s.

start-up performance and acceleration, the motor was made to start from standstill condition to a pre-defined reference speed value of 50 rad/s. The rotor quickly attained the reference speed and critically stabilized in 0.24s. Again, the rotor was accelerated by changing the reference speed value to 100 rad/s. The rotor attained this speed in 0.12s. To test the retardation capability, reference speed was changed to 50 rad/s and the motor attained this speed in 0.11s. The step-up and step-down speed transition experiments were also conducted at various speed ranges starting from standstill condition to 75 rad/s in 0.25s, 135 rad/s in 0.12s and retarded back to 75 rad/s in 0.12s, as shown in Fig. 28. The transient and steady-state closed-loop performance of the inverter system over a wide range of speed is observed. Fig. 29.(a) shows the dynamic response of the phase current and voltage waveforms during retardation and Fig. 29.(b) shows the dynamic response of the phase current and voltage waveforms during acceleration. Here, the sensor conversion scales are 1:1 and 1:222 for current and voltage measurements, respectively. To validate the dynamic response and quality of power conversion of the ILI, the rotor speed is critically changed from 130 rad/s to 95 rad/s. It was repeated in reverse direction too.

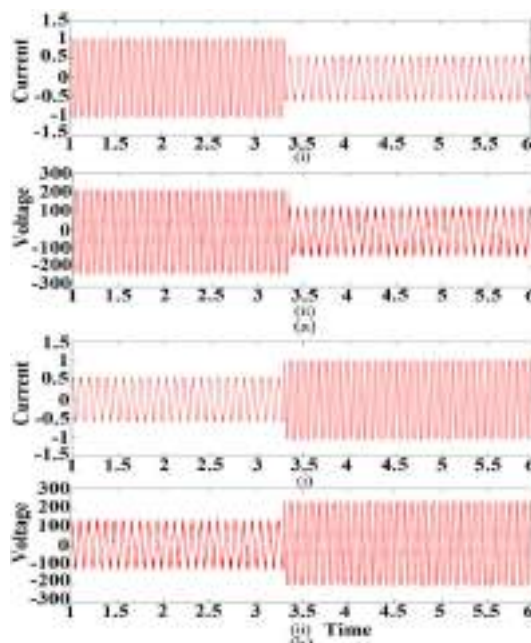


FIGURE 29. (a, b) The dynamic response of ILI fed induction motor obtained from direct vector control. (i) Phase current and (ii) phase voltage across motor winding. The rotor speed shifted from 130 rad/s to 95 rad/s by changing the phase voltage and frequency from 210V, 45Hz to 150V, 30Hz at 3.3s. and the vice versa. (Current and voltage sensor conversion scale is 1:1 and 1:222 respectively).

In TABLE-5 illustrate the experimental data of dynamic speed response and settling time of three-phase open-ended induction motor using ILI topology.

In comparison to scalar control technique the direct vector control exhibits an improved performance in settling time for speed responses. The experimental results prove that the quality of voltage and current waveforms remain constant before and after the speed transitions. Comparison of simulation and experimental results are summarised and presented in the TABLE- 6.

XIV. CONCLUSION

Design and analysis of the performance of an infinite level inverter driven induction motor have been discussed in this paper. ILI has been found to impart better performance to an induction motor drive. The ILI which belongs to an AFE-RDC-MLI topology has been tested with a resistive load and found to possess very good quality voltage and current waveforms in terms of THD. While conventional inverter topologies contain tens of percentage of THD, the topology mentioned in this paper contains a THD as low as 1.2%. Moreover, the dc- voltage requirement for generating a fixed ac-voltage output has been found to be much less than that required by other similar topologies, making the dc-source utilization better with this topology. While it is required to have a dc-voltage requirement of 677V in a conventional inverter working in sine PWM mode, the requirement of dc-voltage in the new inverter is only 338V.

Use of third harmonic injection modulation scheme has also been performed using this inverter and found that the

dc-source utilization can be improved further. Efficiency of inverter has also been found to be better, since only one switch per phase is operated at high frequency. All the switches in conventional inverters are operated at high frequency. Scalar and vector control of induction motor have also been performed using this topology. It has been found that the dynamic performance is better with this topology. This has been validated by accelerating and decelerating the machine with different reference speeds. Since the harmonic content in current has been very less, torque pulsations experienced by the motor would be negligible. Requirement of de-rating associated with induction motors fed by conventional inverters is not present in this case. Since there is no shoot-through menace, the chances of the inverter getting damaged is less, which results in better life and reliability of the drive system.

XV. FUTURE SCOPE

Silicon based power electronic devices are getting replaced with Wide Band Gap (WBG) devices. WBG devices are capable of switching at frequencies in the range of several megahertz. ILI can be a perfect fit for WBG devices, since the output voltage waveform can approach a pure sinusoidal status at these switching frequency levels. The performance of the drive system with advanced control schemes can also be studied.

APPENDIX

A. DESIGN OF INDUCTOR (L)

The rms value of the fundamental component of current ripple ($I_{Lripple}$) is obtained from,

$$\frac{4}{\sqrt{2\pi}} \left[\frac{V_i}{2} \right] = (2\pi f_s L) I_{Lripple} \quad (56)$$

where, V_i is the input voltage, $I_{Lripple}$ is the inductor current ripple, f_s is the switching frequency. The maximum permitted amount of ripple is considered as 5%. Therefore, ($I_{Lripple} \leq 5\%I$), Where, 'I' is the rms value of load current. Hence, the value of desirable inductor (L) is

$$L \geq \frac{\left(\frac{4}{\sqrt{2\pi}}\right)\left(\frac{V_i}{2}\right)}{2\pi f_s (0.05I)} \quad (57)$$

B. DESIGN OF CAPACITOR (C)

For designing the value of capacitor, assume the value of ripple in the load current as very small and hence, can be negligible. Here, the ripple current of the capacitor ($I_{Cripple}$) is equal to the ripple current of the inductor current ($I_{Lripple}$). The maximum permitted amount of the ripple voltage of the capacitor is considered as 5%. Therefore, $V_{Cripple} \leq 5\%V_o$.

$$V_{Cripple} = \frac{I_{Cripple}}{2\pi f_s C} \leq 0.05V_o \quad (58)$$

On rearranging the above equation, the value of the capacitor (C) is

$$C \geq \frac{I_{Cripple}}{2\pi f_s (0.05V_o)} \quad (59)$$

REFERENCES

- [1] P. Omer, J. Kumar, and B. S. Surjan, "A review on reduced switch count multilevel inverter topologies," *IEEE Access*, vol. 8, pp. 22281–22302, Jan. 2020.
- [2] J. Rodríguez, J.-S. Lai, and F. Z. Peng, "Multilevel inverters: A survey of topologies, controls, and applications," *IEEE Trans. Ind. Electron.*, vol. 49, no. 4, pp. 724–738, Aug. 2002.
- [3] L. M. Tolbert, F. Z. Peng, and T. G. Habetler, "Multilevel converters for large electric drives," *IEEE Trans. Ind. Appl.*, vol. 35, no. 1, pp. 36–44, Jan./Feb. 1999.
- [4] J.-S. Lai and F. Z. Peng, "Multilevel converters—A new breed of power converters," *IEEE Trans. Ind. Appl.*, vol. 32, no. 3, pp. 509–517, May 1996.
- [5] S. Kouro, M. Malinowski, K. Gopakumar, J. Pou, L. G. Franquelo, B. Wu, J. Rodríguez, M. A. Pérez, and J. I. Leon, "Recent advances and industrial applications of multilevel converters," *IEEE Trans. Ind. Electron.*, vol. 57, no. 8, pp. 2553–2580, Aug. 2010.
- [6] M. Malinowski, K. Gopakumar, J. Rodríguez, and M. A. Pérez, "A survey on cascaded multilevel inverters," *IEEE Trans. Ind. Electron.*, vol. 57, no. 7, pp. 2197–2206, Jul. 2010.
- [7] L. G. Franquelo, J. Rodríguez, J. I. Leon, S. Kouro, R. Portillo, and M. A. M. Prats, "The age of multilevel converters arrives," *IEEE Ind. Electron. Mag.*, vol. 2, no. 2, pp. 28–39, Jun. 2008.
- [8] J. Rodríguez, S. Bernet, B. Wu, J. O. Pontt, and S. Kouro, "Multilevel voltage-source-converter topologies for industrial medium-voltage drives," *IEEE Trans. Ind. Electron.*, vol. 54, no. 6, pp. 2930–2945, Dec. 2007.
- [9] L. A. Tolbert, F. Z. Peng, T. Cunnyngham, and J. N. Chiasson, "Charge balance control schemes for cascade multilevel converter in hybrid electric vehicles," *IEEE Trans. Ind. Electron.*, vol. 49, no. 5, pp. 1058–1064, Oct. 2002.
- [10] X. Yuan and I. Barbi, "Fundamentals of a new diode clamping multilevel inverter," *IEEE Trans. Power Electron.*, vol. 15, no. 4, pp. 711–718, Jul. 2000.
- [11] M. F. Escalante, J. C. Vannier, and A. Arzandé, "Flying capacitor multilevel inverters and DTC motor drive applications," *IEEE Trans. Ind. Electron.*, vol. 49, no. 4, pp. 809–815, Aug. 2002.
- [12] S. R. Khasim, C. Dhanamjayulu, S. Padmanaban, J. B. Holm-Nielsen, and M. Mitolo, "A novel asymmetrical 21-level inverter for solar PV energy system with reduced switch count," *IEEE Access*, vol. 9, pp. 11761–11775, 2021.
- [13] D. J. Almkhles, J. S. M. Ali, S. Padmanaban, M. S. Bhaskar, U. Subramaniam, and R. Sakthivel, "An original hybrid multilevel DC-AC converter using single-double source unit for medium voltage applications: Hardware implementation and investigation," *IEEE Access*, vol. 8, pp. 71291–71301, 2020.
- [14] M. S. Bhaskar, D. Almkhles, S. Padmanaban, D. M. Ionel, F. Blaabjerg, J. He, and A. R. Kumar, "Investigation of a transistor clamped T-type multilevel H-bridge inverter with inverted double reference single carrier PWM technique for renewable energy applications," *IEEE Access*, vol. 8, pp. 161787–161804, 2020.
- [15] C. Dhanamjayulu, S. R. Khasim, S. Padmanaban, G. Arunkumar, J. B. Holm-Nielsen, and F. Blaabjerg, "Design and implementation of multilevel inverters for fuel cell energy conversion system," *IEEE Access*, vol. 8, pp. 183690–183707, 2020.
- [16] C. M. N. Mukundan, P. Jayaprakash, U. Subramaniam, and D. J. Almkhles, "Binary hybrid multilevel inverter-based grid integrated solar energy conversion system with damped SOGI control," *IEEE Access*, vol. 8, pp. 37214–37228, 2020.
- [17] S. Shuvo, E. Hossain, T. Islam, A. Akib, S. Padmanaban, and M. Z. R. Khan, "Design and hardware implementation considerations of modified multilevel cascaded H-bridge inverter for photovoltaic system," *IEEE Access*, vol. 7, pp. 16504–16524, 2019.
- [18] M. H. Mondol, M. R. Tür, S. P. Biswas, M. K. Hosain, S. Shuvo, and E. Hossain, "Compact three phase multilevel inverter for low and medium power photovoltaic systems," *IEEE Access*, vol. 8, pp. 60824–60837, 2020.
- [19] P. Ponnusamy, P. Sivaraman, D. J. Almkhles, S. Padmanaban, Z. Leonowicz, M. Alagu, and J. S. M. Ali, "A new multilevel inverter topology with reduced power components for domestic solar PV applications," *IEEE Access*, vol. 8, pp. 187483–187497, 2020.
- [20] A. Chen, L. Hu, L. Chen, Y. Deng, and X. He, "A multilevel converter topology with fault-tolerant ability," *IEEE Trans. Power Electron.*, vol. 20, no. 2, pp. 405–415, Mar. 2005.

- [21] M. Glinka, "Prototype of multiphase modular-multilevel-converter with 2 MW power rating and 17-level-output-voltage," in *Proc. IEEE 35th Annu. Power Electron. Spec. Conf.*, vol. 4, Dec. 2004, pp. 2572–2576.
- [22] S. Debnath, J. Qin, B. Bahrani, M. Saedifard, and P. Barbosa, "Operation, control, and applications of the modular multilevel converter: A review," *IEEE Trans. Power Electron.*, vol. 30, no. 1, pp. 37–53, Jan. 2015.
- [23] M. Vijeh, M. Rezanejad, E. Samadaei, and K. Bertilsson, "A general review of multilevel inverters based on main submodules: Structural point of view," *IEEE Trans. Power Electron.*, vol. 34, no. 10, pp. 9479–9502, Oct. 2019.
- [24] A. Salem, H. Van Khang, K. G. Robbersmyr, M. Norambuena, and J. Rodríguez, "Voltage source multilevel inverters with reduced device count: Topological review and novel comparative factors," *IEEE Trans. Power Electron.*, vol. 36, no. 3, pp. 2720–2747, Mar. 2021.
- [25] K. K. Gupta, A. Ranjan, P. Bhatnagar, L. K. Sahu, and S. Jain, "Multilevel inverter topologies with reduced device count: A review," *IEEE Trans. Power Electron.*, vol. 31, no. 1, pp. 135–151, Jan. 2016.
- [26] M. D. Siddique, S. Mekhilef, M. Rawa, A. Wahyudie, B. Chokaev, and I. Salamov, "Extended multilevel inverter topology with reduced switch count and voltage stress," *IEEE Access*, vol. 8, pp. 201835–201846, 2020.
- [27] R. V. Nair, S. A. Rahul, R. S. Kaarthik, A. Kshirsagar, and K. Gopakumar, "Generation of higher number of voltage levels by stacking inverters of lower multilevel structures with low voltage devices for drives," *IEEE Trans. Power Electron.*, vol. 32, no. 1, pp. 52–59, Jan. 2017.
- [28] T. Roy and P. K. Sadhu, "A step-up multilevel inverter topology using novel switched capacitor converters with reduced components," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 236–247, Jan. 2021.
- [29] M. D. Siddique, S. Mekhilef, N. M. Shah, A. Sarwar, A. Iqbal, M. Tayyab, and M. K. Ansari, "Low switching frequency based asymmetrical multilevel inverter topology with reduced switch count," *IEEE Access*, vol. 7, pp. 86374–86383, 2019.
- [30] A. N. Babadi, O. Salari, M. J. Mojibian, and M. T. Bina, "Modified multilevel inverters with reduced structures based on PackedU-cell," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 6, no. 2, pp. 874–887, Jun. 2017.
- [31] A. Hota, S. Jain, and V. Agarwal, "An improved three-phase five-level inverter topology with reduced number of switching power devices," *IEEE Trans. Ind. Electron.*, vol. 65, no. 4, pp. 3296–3305, Apr. 2018.
- [32] A. Karthik and U. Loganathan, "A reduced component count five-level inverter topology for high reliability electric drives," *IEEE Trans. Power Electron.*, vol. 35, no. 1, pp. 725–732, Jan. 2020.
- [33] A. Chappa, S. Gupta, L. K. Sahu, S. P. Gautam, and K. K. Gupta, "Symmetrical and asymmetrical reduced device multilevel inverter topology," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 9, no. 1, pp. 885–896, Feb. 2021.
- [34] M. D. Siddique, S. Mekhilef, N. M. Shah, N. Sandeep, J. S. M. Ali, A. Iqbal, M. Ahmed, S. S. M. Ghoneim, M. M. Al-Harthi, B. Alamri, F. A. Salem, and M. Orabi, "A single DC source nine-level switched-capacitor boost inverter topology with reduced switch count," *IEEE Access*, vol. 8, pp. 5840–5851, 2020.
- [35] P. R. Bana, K. P. Panda, R. T. Naayagi, P. Siano, and G. Panda, "Recently developed reduced switch multilevel inverter for renewable energy integration and drives application: Topologies, comprehensive analysis and comparative evaluation," *IEEE Access*, vol. 7, pp. 54888–54909, 2019.
- [36] E. Babaei, S. Laali, and S. Alilu, "Cascaded multilevel inverter with series connection of novel H-bridge basic units," *IEEE Trans. Ind. Electron.*, vol. 61, no. 12, pp. 6664–6671, Dec. 2014.
- [37] M. R. J. Osukee, E. Salary, and S. Najafi-Ravadanegh, "Creative design of symmetric multilevel converter to enhance the circuit's performance," *IET Power Electron.*, vol. 8, no. 1, pp. 96–102, 2015.
- [38] R. S. Alishah, D. Nazarpour, S. H. Hosseini, and M. Sabahi, "Novel topologies for symmetric, asymmetric, and cascade switched-diode multilevel converter with minimum number of power electronic components," *IEEE Trans. Ind. Electron.*, vol. 61, no. 10, pp. 5300–5310, Oct. 2014.
- [39] H. Vahedi, M. Sharifzadeh, and K. Al-Haddad, "Modified seven-level pack U-cell inverter for photovoltaic applications," *IEEE Trans. Emerg. Sel. Topics Power Electron.*, vol. 6, no. 3, pp. 1508–1516, Sep. 2018.
- [40] A. Ajami, M. R. J. Osukee, A. Mokhberdoran, and A. van den Bossche, "Developed cascaded multilevel inverter topology to minimise the number of circuit devices and voltage stresses of switches," *IET Power Electron.*, vol. 7, no. 2, pp. 459–466, Feb. 2014.
- [41] E. Samadaei, S. A. Gholamian, A. Sheikholeslami, and J. Adabi, "An envelope type (E-type) module: Asymmetric multilevel inverters with reduced components," *IEEE Trans. Ind. Electron.*, vol. 63, no. 11, pp. 7148–7156, Nov. 2016.
- [42] N. A. Rahim, K. Chaniago, and J. Selvaraj, "Single-phase seven-level grid-connected inverter for photovoltaic system," *IEEE Trans. Ind. Electron.*, vol. 58, no. 6, pp. 2435–2443, Jun. 2011.
- [43] R. Uthirasamy, U. S. Ragupathy, and V. K. Chinnaiyan, "Structure of boost DC-link cascaded multilevel inverter for uninterrupted power supply applications," *IET Power Electron.*, vol. 8, no. 11, pp. 2085–2096, Nov. 2015.
- [44] L. K. Haw, M. S. A. Dahidah, and H. A. F. Almurib, "SHE-PWM cascaded multilevel inverter with adjustable DC voltage levels control for STATCOM applications," *IEEE Trans. Power Electron.*, vol. 29, no. 12, pp. 6433–6444, Dec. 2014.
- [45] L. Wang, G. Ma, S. Zhao, and Y. Diao, "Study of a new buck-boost multilevel inverter," in *Proc. Int. Power Electron. Appl. Conf. Expo.*, Nov. 2014, pp. 78–81.
- [46] M. N. H. Khan, M. Forouzes, Y. P. Siwakoti, L. Li, and F. Blaabjerg, "Switched capacitor integrated $(2n + 1)$ -level step-up single-phase inverter," *IEEE Trans. Power Electron.*, vol. 35, no. 8, pp. 8248–8260, Aug. 2020.
- [47] F. Gao, "An enhanced single-phase step-up five-level inverter," *IEEE Trans. Power Electron.*, vol. 31, no. 12, pp. 8024–8030, Dec. 2016.
- [48] N. K. Pilli, M. Raghuram, A. Kumar, and S. K. Singh, "Single DC-source-based seven-level boost inverter for electric vehicle application," *IET Power Electron.*, vol. 12, no. 13, pp. 3331–3339, Nov. 2019.
- [49] J. Selvaraj and N. A. Rahim, "Multilevel inverter for grid-connected PV system employing digital PI controller," *IEEE Trans. Ind. Electron.*, vol. 56, no. 1, pp. 149–158, Jan. 2009.
- [50] A.-V. Ho and T.-W. Chun, "Single-phase modified quasi-Z-source cascaded hybrid five-level inverter," *IEEE Trans. Ind. Electron.*, vol. 65, no. 6, pp. 5125–5134, Jun. 2018.
- [51] M. R. Banaei, R. Alizadeh, H. Khounjahan, E. Salary, and A. R. Dehghanzadeh, "Z-source-based multilevel inverter with reduction of switches," *IET Power Electron.*, vol. 5, no. 3, pp. 385–392, Mar. 2012.
- [52] S. Dhara and V. T. Somasekar, "An integrated semi-double stage-based multilevel inverter with voltage boosting scheme for photovoltaic systems," *IEEE J. Emerg. Sel. Topics Power Electron.*, vol. 8, no. 3, pp. 2326–2339, Sep. 2020.
- [53] P. C. Sen and Z. Yang, "A new DC-to-AC inverter with dynamic robust performance," in *Proc. IEEE Region Int. Conf. Global Connectivity Energy, Comput., Commun. Control (TENCON)*, vol. 2, Dec. 1998, pp. 387–390.



His research interests include power electronics, drives, multilevel inverters, power quality, and power systems.



He had also functioned as the Chairman of the Board of Electrical Engineering streams of Calicut University and Kerala Technological University, Kerala. He had also associated with consultancy projects for industries. His research interests include power electronics, drives, neural network applications to power electronics, FACTS, and real time DSP implementations. He had received the Career Award for Young Teachers of AICTE, in 2000.

...



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

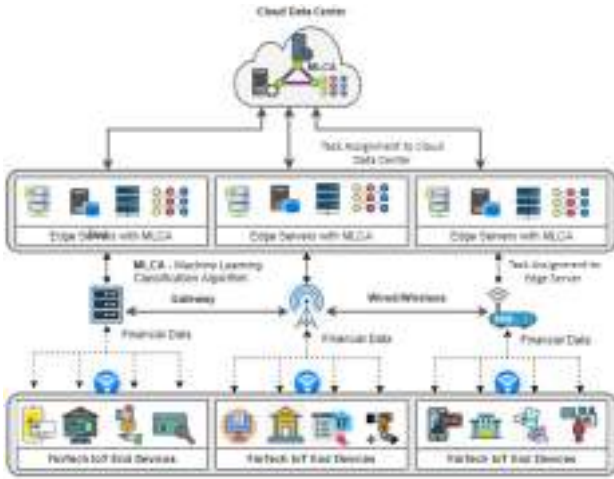


Fig. 1. Edge framework for predictive analysis of FinTech applications.

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 $\gamma_{ij}^{\text{up}} = \mathcal{W}_{ij}^{\text{in}} \log(1 + P_j^{\text{up}} \times [\delta_i^{\text{power}} / \alpha_i^2])$. Here, $\mathcal{W}_{ij}^{\text{in}}$ indicates the channel utilization factor between the i th IoT device and the j th computing device. α_i^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:

$$\text{minimize } \sum_{i=1}^n P_{ij}^{\text{total}}(t) \quad (13a)$$

$$\text{subject to } P_{ij}^{\text{total}}(t) \leq \eta_j^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13b)$$

$$l_{ij}(t) \leq l_j^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13c)$$

$$\tau_i^{\text{CPU}}(t) \leq \tau_i^{\text{max}}, j \in (\mathcal{S} \cup \mathcal{R}) \quad (13d)$$

$$\sum_{i=1}^{(|\mathcal{T}|)} \sum_{j=1}^{(|\mathcal{SR}|)} X(i, j) \leq |\mathcal{S} \cup \mathcal{R}| \quad (13e)$$

$$\sum_{j=1}^{(|\mathcal{T}|)} X(i, j) = 1. \quad (13f)$$

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_j^{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

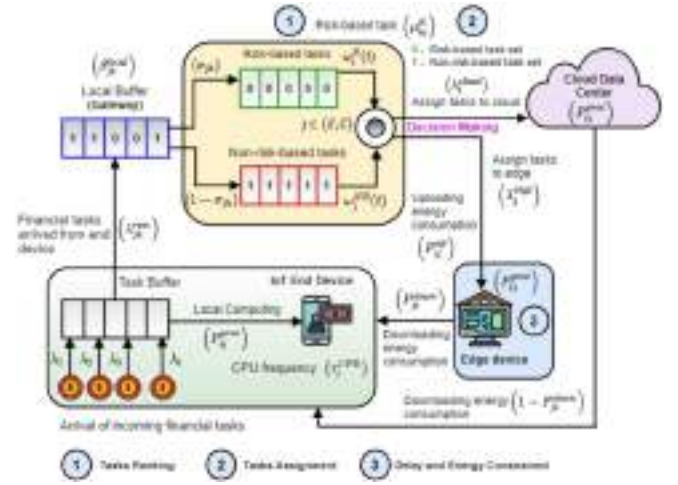


Fig. 2. Workflow of ranked-based task classification.

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

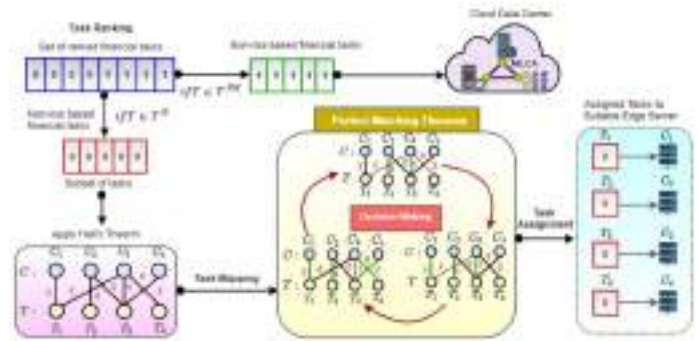


Fig. 3. Service deployment with perfect matching theorem.

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^{NR} 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

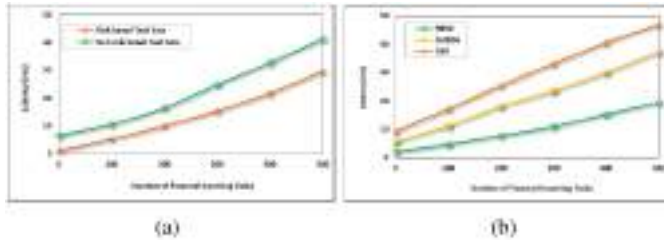


Fig. 4. Impact of task assignment over delay. (a) Various financial tasks. (b) Comparative analysis with baseline schemes.

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.

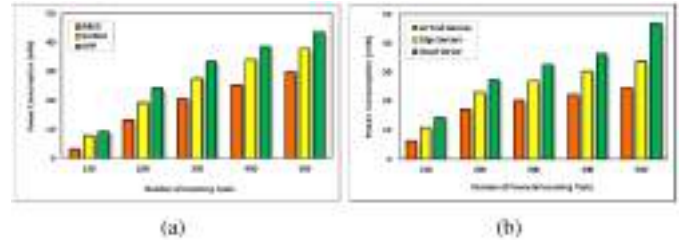


Fig. 5. Impact of task assignment over power consumption. (a) Various financial tasks. (b) Comparative analysis with baseline schemes.

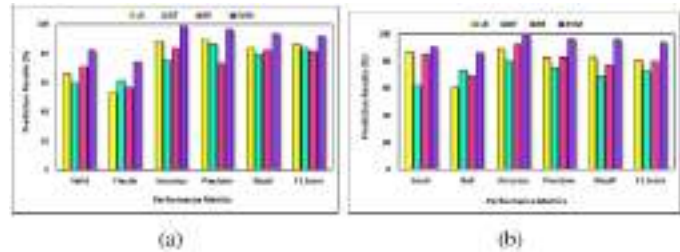


Fig. 6. Edge-level analysis using MLCAs. (a) Prediction results of D1. (b) Prediction results of D2.

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

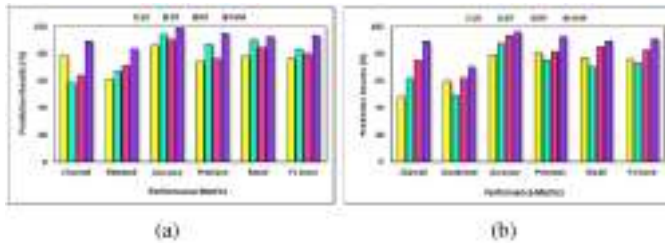


Fig. 7. Cloud-level analysis using MLCAs. (a) Prediction results of D3. (b) Prediction results of D4.

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.

Feature Article: Security, Privacy, Content Protection,
and Digital Rights Management

Security in Edge-Centric Intelligent Internet of Vehicles: Issues and Remedies

Mainak Adhikari

Indian Institute of Information Technology
Lucknow

Ambigavathi Munusamy

Indian Institute of Information Technology Una

Abhishek Hazra

Indian Institute of Technology (ISM) Dhanbad

Varun G Menon

SCMS School of Engineering and Technology

Vijay Anavangot

Indian Institute of Technology Bombay

Deepak Puthal

Newcastle University

Abstract—Propelled by the growth of automotive industry, and the ubiquity of smart sensors, intelligent transport systems such as the Internet of Vehicles (IoV) have seen significant research interest in recent times. The emerging distributed IoV networks support real-time vehicular applications through on-device computing, communication-efficient data processing, edge computing, and cloud aggregation. While enriching the user experience by minimizing the end-to-end latency through efficient energy management, IoV deployments face the fundamental challenge of security attacks. In this article, we discuss various security attack modes in an edge-centric intelligent IoV framework, consisting of distributed smart vehicles, and remote processing units. We highlight various attack detection and mitigation mechanism for the proposed IoV framework, to address the security challenges. Finally, we shed light on several future research directions to ensure security of sensory data in edge-centric IoV systems.

Digital Object Identifier 10.1109/MCE.2021.3116415

Date of publication 30 September 2021; date of current

version 6 October 2022.

■ **THE INTERNET OF Things (IoT)** facilitates the omnipresent sharing of information and knowledge between connected devices with limited human interference, which is a crucial instigator for different applications like green infrastructure, smart transportation networks, etc.¹ Within this paradigm, the term Internet of Vehicles (IoV) is often encountered in the smart transportation system, referring to the vehicular subset of the IoT. IoV is modern technology, integrating smart devices, communication technologies (e.g., 5G and 6G) and intelligent vehicles, sophisticated business analytics, and human-machine cooperation to improve efficiency, performance, and reliability.

Edge computing is a promising technology that can improve the processing of traffic monitoring data and enhance the performance of smart transportation systems by combining knowledge and technologies such as artificial intelligence and 6G communication.² 6G technology helps deliver physical level security, new security protocol, and state-of-the-art security standardization, and interoperability in the vehicular network while managing mobility and high dynamicity among the smart IoV devices. With the emergence of edge-centric IoV systems, security becomes a critical challenge, as attackers can imitate legitimate users to access the IoV infrastructure. In most scenarios, attackers can alert the availability, integrity, and confidentiality of the remote computing devices in the edge networks.³ Thus, this work aims to highlight several burning IoV challenges for security and possible solutions using edge computing in an intelligent environment.

Accordingly, we introduce edge-centric IoV concepts in Sections “Edge-Centric IoV Framework” and “Attacks in Edge-Centric IoV.” Security issues and detection strategies for IoV networks are presented in Sections “Security Requirements for Edge-Centric IoV Network” and “Attack Detection Strategies for Edge-Centric IoV Network,” respectively. Future research directions and conclusion of the work are presented in Sections “Future Research Directions” and “Conclusion,” respectively.

EDGE-CENTRIC IOV FRAMEWORK

The IoV is a shared system that promotes the use of data, generated by smart devices and vehicular ad hoc networks in a distributed manner.⁴ A fundamental aim of the IoV is to enable

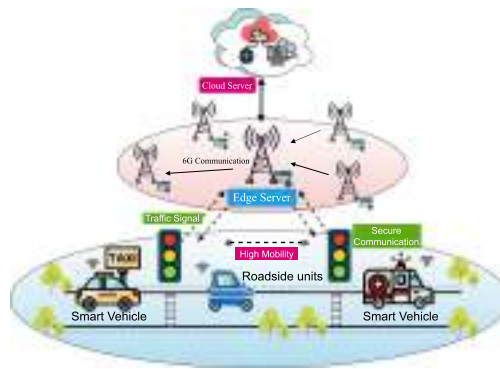


Figure 1. Edge-centric IoV networks for transportation system.

vehicles to interact in real-time with their drivers, other vehicles, roadside edge servers, and pedestrians for making an intelligent transportation system. Several initiatives throughout the world, for example, Japan, South Korea, Singapore, United States, and Australia have admitted to developing a fully automated and intelligent transportation system.

The IoV system mainly supports three types of level-wise communication among various components, where the infrastructure layer connects with smart vehicles, devices, and traffic control units as presented in Figure 1. The edge layer connects infrastructure devices with the roadside processing units, and finally, the cloud layer gathers the IoV data for data analytics. A summary of the three-level IoV-communications is explained as follows.

- *Vehicle-to-Infrastructure (V2I)*: V2I system is a communication framework with fixed infrastructure, such as roadside units (RSU), sensor technology, and network infrastructure, to support the wireless exchange of information. Technologies used for V2I communication are RFID Tags, Wi-Fi, and Bluetooth.

- *Vehicle-to-Edge (V2E)*: Edge server-enabled RSU ensure seamless communication between the moving vehicles, where the smart cars process their generated data, share experience, road weather conditions, and traffic signaling information on a real-time basis. Devices used in V2E communication are street lights, base stations, cameras, and cell towers.

- *Vehicle-to-Cloud (V2C)*: Cloud computing with mature storage and processing technology is used in the IoV network for long-term data

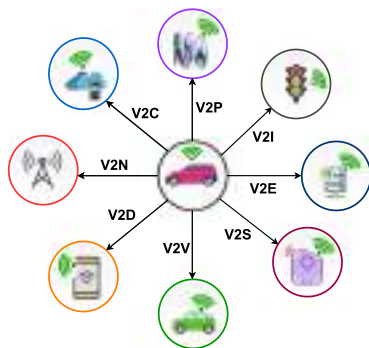


Figure 2. Types of communications links in IoV network.

analysis and traffic management. The main goal of V2C communication is to increase intelligent decision making, monitor vehicles remotely, optimize traffic and traveling costs, and improve safety for participants.

The abovementioned communication strategies can be achieved through a secured and reliable IoV system for data transmission and storage.

ATTACKS IN EDGE-CENTRIC IOV

The security attack in the edge-centric IoV system is one of the most crucial issues due to frequent topology changes, distribution of vehicles, a large volume of traffic data, limited transmission power, etc. This IoV system supports various transmission modes including

vehicle-to-infrastructure (V2I), vehicle-to-sensor (V2S), vehicle-to-network (V2N), vehicle-to-pedestrians (V2P), vehicle-to-device (V2D), vehicle-to-edge (V2E), vehicle-to-cloud (V2C), and vehicle-to-vehicle (V2V) to improve the safety on the road and provide intelligent traffic management and more convenience to the drivers, as shown in Figure 2. However, if there is any network intrusion, the vehicles can be controlled by attackers, which leads to severe accidents, deaths, and also affects the people on the roadside. There are several types of security attacks,⁵ which can affect the overall services and operations of the intelligent transport system, as depicted in Table 1.

Denial of Service Attack In the Denial-of-Service (DoS) attack, the attacker intends to disrupt the normal traffic and reduces the performance of the network by flooding the target with irrelevant messages. The primary target of the attacker would be the RSU, which acts as a core component to authenticate, manage, and update the vehicle's information.

Blackhole Attack In IoV, a malicious node sends false route information to other participating vehicle nodes. This introduces itself as an optimum route and causes other vehicles to route the traffic information via malicious ones, i.e., no traffic can move through the network to

Table 1. Comparison of security attacks with different properties.

Type of attack	Type of attacker	Security requirements	Level of attacks	Types of communication
Denial of Service	Malicious, Active, and Insider	Availability	High	V2I/V2E/V2C
Black hole	Passive and outsider	Availability	Moderate	V2V
Timing	Malicious and Insider	Integrity and Authentication	High	V2V/V2I
Warm hole	Malicious and Outsider	Availability and Confidentiality	Moderate	V2V
Sybil	Insider and Network Attack	Authentication and Availability	High	V2V
Gray hole	Passive and Outsider	Availability	Moderate	V2V
Illusion	Insider and Outsider	Authentication and Integrity	Low	V2V/V2I
Man-in-the-middle	Insider and Monitoring Attack	Integrity and Confidentiality	Moderate	V2V
Replay	Insider and monitoring Attack	Authentication and Integrity	High	V2V/V2C
Malware injection	Outsider	Availability and Integrity	Moderate	V2V/V2I/V2E/V2C

other vehicles. Thus, the malicious node drops all the incoming traffic instead of forwarding them to a specific RSU.

Timing Attack A timing attack in intelligent connected vehicles is a serious attack, where a malicious vehicle receives time-critical updates and traffic information and it does not forward the message to the neighboring vehicles at the right time instead it includes some false time slots to the original message to make further delay.

Warmhole Attack In a wormhole attack, two or more malicious vehicles create a tunnel to hide the true distance and entice other normal vehicles to transmit their traffic information across the malicious vehicles. Also, they start to absorb the normal flow of messages and cause traffic conjunction or collaborate with other malicious nodes.

Sybil Attack In this attack, an attacker or malicious vehicle generates multiple identities to imitate multiple vehicles in the IoV network at the same time. Due to these wrong identities, a driver or normal vehicle cannot identify the false position information, transmitted by the malicious vehicle. Thus, the malicious vehicle creates chaos among normal vehicles and increases the huge security risks in the IoV system.

Grayhole Attack In a gray hole attack, the attacker or malicious vehicle advertises an optimum route and selectively drops the traffic route information from a specific set of vehicle nodes or drops traffic information probabilistically and forwards all other traffic to a specific RSU.

Illusion Attack In an illusion attack, attackers collect and alter the readings from the sensors and RSUs of the IoV system. They generate an illusion by disseminating false traffic warning alerts to other neighborhood vehicles based on the current road conditions. Thus, spreading more illusion can increase the possibility of traffic jams, accidents, and reduce the network performance.

Man-in-the Middle Attack In a man-in-the-middle attack, an attacker or malicious vehicle intercepts the communication process between the vehicle and RSU. The malicious vehicle

modifies the sensor readings, secretly eavesdrops, and steals personal information. Thus, the vehicle can inject false information and secretly relays between two vehicles.

Replay Attack In a replay attack, an attacker captures the data, sent over the IoV network. This attacker acts as an original sender and deceptively delays or repeats the valid transmissions to misdirect the RSU. Thus, the replay attack affects the secure communication between a vehicle and RSU.

Malware Injection Attacks This type of attack is mostly executed by the attackers to take control of the driver's or vehicle's information, stored in an edge server. For this purpose, the malware software or codes are injected into the running process to perform malicious activities such as manipulating or stealing the traffic information.

SECURITY REQUIREMENTS FOR EDGE-CENTRIC IOV NETWORK

Security are the two significant key requirements for an edge-centric IoV network due to sharing critical safety data and computing resources on the local edge server for further analysis. The layered architecture components of edge-centric IoV are shown in Figure 3, which comprises three distinct layers: 1) vehicle-to-infrastructure (V2I), 2) vehicle-to-edge (V2E), and 3) vehicle-to-cloud (V2C), as depicted in Figure 2. Every layer constitutes a potential threat and security attack that can reduce the overall performance of secure data communication. Different security requirements of each layer of the IoV system are described as follows.

Vehicle-to-Infrastructure (V2I) Layer

The first layer consists of various onboard and traffic monitoring sensors that are connected to the RSUs or moving autonomous vehicles. In this layer, the traffic data collected from the vehicle's sensors and the data received by the RSUs or other vehicles should be authenticated in terms of origin, content, and time. Thus, this layer must satisfy the following security requirements:

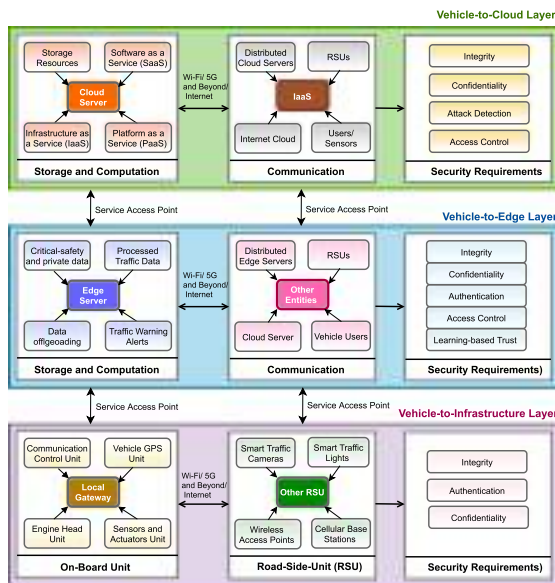


Figure 3. Layer-based architectural components for IoV network.

Integrity and Authenticity of In-Vehicle Communication The valid or unauthorized traffic data transmitted from the moving vehicles are verified and detected by the RSUs.

Integrity and Authenticity of Other Vehicles The data received from vehicles and RSUs in a network are verified and guaranteed to provide authenticity to other connecting vehicles.

Confidentiality of Communication and Data The vehicle broadcasts the captured data to nearby RSU or another vehicle to update the current road conditions. Thus, the broadcasted data must be reliable and confidential to the authorized users.

Vehicle-to-Edge (V2E) Layer

This layer is primarily responsible for storing and processing the data from the vehicle's sensors and sending traffic warning alerts to the right users or vehicles. In this layer, the stored or processed data should be authenticated and protected from malware injection attackers. Thus, the security requirements of V2E are summarized as follows.

Confidentiality Confidentiality requirement provides a guarantee to the stored and processed data. These data should be encrypted

accurately, shareable, and accessible only to authorized users.

Integrity The stored and exchanged data must be validated and verified in an edge server. Thus, exchanged critical safety data should be protected from unauthorized modifications or malware injections. The received data must be authenticated in terms of source, location, and content. In this way, the data should be protected from various attacks and also ensured the integrity of the received data.

Access Control Most of the data, collected from moving vehicles are important for an intelligent traffic management system. With the help of an access control mechanism, only authorized users can be able to access the data, and resources, and private information.

Learning-Based Trust Management The modern IoV system must adopt the traditional attack detection strategy by developing a novel machine or deep learning-based trust management algorithm for connected vehicles. Thus, the vehicles can estimate their future trust values based on the present and past trust values.

Vehicle-to-Cloud (V2C) Layer

The third layer can be connected with the internet to provide on-demand services or additional services to authorized users. A single or multiple vehicular cloud servers are deployed in different locations to store and process the large volume of traffic data. This layer must consider the following security requirements for enhancing the accessibility of their services.

Attack Detection This security requirement enables the cloud server to effectively identify, store, and alerts various internal and external attacks. Thus, various attacks should be detected on the cloud resources from unauthorized access.

Integrity This requirement ensures the integrity of the vehicular cloud server and other tampering events.

Access Control This requirement provides limited and controlled access to authorized cloud users and other virtual cloud servers.

Table 2. Summary of various attacks detection mechanisms in vehicular networks.

Type of attack	Key idea	Mode of routing	Mode of transmission	Limitations
Denial of Service	To mitigate jammer and improve valid transmissions	Network-centric	V2I	Not able to handle jammers from multiple malicious vehicles
Blackhole	To reduce malicious nodes and improve reliable transmission	Node-centric	V2V	No critical path detection mechanism used between two-hop transmissions
Timing	To study the safety-critical vehicular communication	Data-centric	V2V/V2I	Not considered any security detection mechanisms
Sybil	To detect Sybil nodes and reduce false identities	Node-centric	V2V	Not scalable when the number of Sybil nodes in the IoV network are very high
Man-in-the-middle	To identify the malicious nodes from sharing content	Node-centric	V2V	No intelligent detection mechanism used to identify the malicious vehicles
Gray hole	To find the presence of malicious nodes	Node-centric	V2V	Difficult to identify similar types of malicious nodes in an edge-centric IoV
Warm hole	To detect in-band and out-of-band wormholes	Node-centric	V2V	Not suitable for large number of malicious vehicle nodes, present in the IoV network.
Malware injection	Malware injection mechanisms using an edge-centric framework	Edge-centric	V2I	No mechanisms developed to reduce the device or server-side malware injections

Confidentiality This type of security requirement must prevent the critical safety data and sensitive information of the users from unauthorized users. Thus, the stored data should be accessed and shared only with legitimate entities.

ATTACK DETECTION STRATEGIES FOR EDGE-CENTRIC IOV NETWORK

There are plenty of research works available in the literature to mitigate the security issues in IoV networks. In such networks, all the participating vehicles on the road need to exchange traffic information at any time, any place, any path, and to any relevant network. Despite benefits, this network faces a lot of security challenges due to high mobility and frequent changes in topology. Therefore, it is very important to inspect various detection mechanisms to minimize the attacks and provide more secure communication in edge-centric IoV networks. Table 2 describes the importance and limitation of some key attack detection mechanisms in the vehicular networks.

Boche *et al.*⁶, have designed a suitable security detection mechanism using Turing machines to mitigate the jammer in between senders and receiver node. However, this mechanism failed to focus on detecting multiple jammers within the same network. To reduce the packet drop ratio

and detect the malicious nodes, Hassan *et al.*⁷ have presented an intelligent black hole attack detection scheme. Based on the recent real-world security attacks on automotive systems, Hasan *et al.*⁸ have extensively studied the potential threats and security detection mechanisms in the vehicle-to-everything (V2X) communication framework. Further, a novel mechanism has been proposed in Quevedo *et al.*'s work⁹ to detect Sybil nodes and false identities in the vehicular networks. However, this technique is not scalable when the number of Sybil nodes is very high.

To identify the malicious vehicle nodes, a man-in-the-middle attack resistance trust model has been developed for connected vehicles in Ahmad *et al.*'s work.¹⁰ However, this model contains no intelligent detection mechanism to automatically predict the opposing vehicles. An software-defined networking based EC-SDIoV approach has been considered to quickly identify the presence of malicious vehicle nodes.¹¹ However, this approach cannot able to support a large number of malicious vehicular nodes at the edge-centric IoV network. In the work of Tahboush and Agoyi,¹² a hybrid wormhole attack detection algorithm has been proposed to detect in-band and out-of-band wormhole attack between two successive vehicle nodes based on the hop count. Based on this algorithm, only the detection of wormhole attack between two

successive vehicle nodes is possible, however, it is not much effective mechanism when the number of vehicles is very high at the edge-centric network. Similarly, Chavhan *et al.*¹³ and Wang *et al.*¹⁴ have studied various types of security attacks for fog/edge-centric IoV network.

The aforementioned studies in existing literature are leaned toward security attack challenges and detection mechanisms in traditional vehicular networks. However, these mechanisms are unable to solve the issues and satisfy all the requirements of modern edge-centric IoV networks. Thus, the modern edge-centric IoV network requires a novel and hybrid attack detection mechanism to automatically mitigate, identify, and predict a large number of malicious vehicle nodes.

FUTURE RESEARCH DIRECTIONS

Although modern IoV and intelligent transportation systems have extensively benefited from the IoT technology, risks, and challenges in safety and reliability have also increased substantially. Among other, few open research challenges are explained as follows.

A. Intelligent and Reliable Autonomous Vehicles: The integration of advanced sensing, AI technology, and edge computing into the IoV system enables intelligent transportation systems that can autonomously navigate the environment and improve the robustness of autonomous vehicles while making them safer on the road. The major challenge of a context prediction approach is in the prediction accuracy at the edge networks while transmitting the contextual information through a reliable communication path. Besides that, introducing advanced AI-enabled technologies with reliable data transmission protocols is another important challenging task for increasing the prediction accuracy at the edge of the network.

B. Scalable Blockchain-Enabled Intelligent IoV System: Blockchain is a constantly evolving peer-to-peer distributed ledger technology in edge networks with characteristics of decentralization, security, interoperability, and trust establishment, and potentially lower the costs of the underpinning infrastructure. The crucial objective of blockchain technology is to provide security of the IoV data for future data analysis.¹⁵ Further, blockchain technology is transforming IoV data to the local edge

devices or centralized cloud servers by enabling anonymous and trustful transactions in a decentralized and trustless environment. Despite these advances, several challenges remain to be addressed in the edge-centric IoV system, including the poor scalability, heterogeneity, and the impact of integration on network performance.

C. Secure Data Science for Cooperative IoV System: A cooperative intelligent transport system (C-ITS) aims to improve comfort, efficiency, sustainability, and safety over the stand-alone systems by increasing the effectiveness of the secured communication channel and cooperation between the IoV system and edge networks. Therefore, data science techniques with proper security of analyzing the sensed data are the key components of C-ITS at edge networks. However, with the huge amount of data from transportation sensors, the integration of data science techniques presents a real challenge for realizing C-ITS applications in a secured edge network. Although some attempts have been made to explore data science for CITS, there exist various scientific and engineering challenges, including data multi-source heterogeneity, privacy protection, and computational complexity.

D. Reliable Edge Intelligence for IoV System: Intelligent IoV system enables different types of transportation applications, including real-time video analytics, autonomous driving, etc. that bring various challenges to monitor and significantly enhance safety driving, comfort riding, traffic management, etc. However, these powerful applications pose significant challenges by requesting intensive computation services with low latency. To address the challenges, edge intelligence provides promising results in terms of data analysis at the edge of the networks with minimum delay. The AI-enabled technologies at edge networks help to analyze the mission-critical applications efficiently by providing the services to the IoV systems with higher security using reliable communication protocols.

E. Secure Federated Learning Solutions for IoV System: In standard machine learning (ML) a centralized training is followed, i.e., data collected from the IoV sensors are moved to the centralized server for training. However, the centralized model raises security concerns as the raw data can be accessed at these third-party servers. Federated learning (FL), an emerging distributed learning architecture, secures the data at the edge devices

by only sending a focused update of locally learned ML model from the participating edge devices in the network. As the FL technique perform only a local training on the data at the edge and remote computing devices, a more secure distributed learning is possible as compared to the ML techniques.

CONCLUSION

Edge centric IoV is an emerging paradigm of smart vehicular networks, enabled by the advancements in vehicular infrastructure, local RSU, edge servers, and cloud servers. These decentralized transport networks utilize the edge computing resources to reduce processing delays and minimize energy consumption. Security were identified as the topmost challenges in these edge-centric IoV networks. Through this article, we have analyzed different security attack mechanisms, and reviewed the strategies to detect and eliminate such security breaches. Finally, we have highlighted various future research directions for reliable data transmission and protection in edge-centric IoV.

REFERENCES

1. T. Alladi, V. Chamola, B. Sikdar, and K.-K. R. Choo, "Consumer IoT: Security vulnerability case studies and solutions," *IEEE Consum. Electron. Mag.*, vol. 9, no. 2, pp. 17–25, Mar. 2020, doi: [10.1109/MCE.2019.2953740](https://doi.org/10.1109/MCE.2019.2953740).
2. 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, doi: [10.1109/JIOT.2020.3041102](https://doi.org/10.1109/JIOT.2020.3041102).
3. A. Hazra, M. Adhikari, T. Amgoth, and S. N. Srirama, "Collaborative AI-enabled intelligent partial service provisioning in green industrial fog networks," *IEEE Internet Things J.*, to be published, doi: [10.1109/JIOT.2021.3110910](https://doi.org/10.1109/JIOT.2021.3110910).
4. G. Kumar *et al.*, "A privacy-preserving secure framework for electric vehicles in IoT using matching market and signcryption," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7707–7722, Jul. 2020, doi: [10.1109/TVT.2020.2989817](https://doi.org/10.1109/TVT.2020.2989817).
5. M. Yahuza *et al.*, "Systematic review on security and privacy requirements in edge computing: State of the art and future research opportunities," *IEEE Access*, vol. 8, pp. 76 541–76 567, 2020, doi: [10.1109/ACCESS.2020.2989456](https://doi.org/10.1109/ACCESS.2020.2989456).
6. H. Boche, R. F. Schaefer, and H. V. Poor, "Denial-of-service attacks on communication systems: Detectability and jammer knowledge," *IEEE Trans. Signal Process.*, vol. 68, pp. 3754–3768, Jan. 2020, doi: [10.1109/TSP.2020.2993165](https://doi.org/10.1109/TSP.2020.2993165).
7. Z. Hassan, A. Mehmood, C. Maple, M. A. Khan, and A. Aldegheishem, "Intelligent detection of black hole attacks for secure communication in autonomous and connected vehicles," *IEEE Access*, vol. 8, pp. 19618–199628, 2020, doi: [10.1109/ACCESS.2020.3034327](https://doi.org/10.1109/ACCESS.2020.3034327).
8. M. Hasan, S. Mohan, T. Shimizu, and H. Lu, "Securing vehicle-to-everything (V2X) communication platforms," *IEEE Trans. Intell. Veh.*, vol. 5, no. 4, pp. 693–713, Dec. 2020, doi: [10.1109/TIV.2020.2987430](https://doi.org/10.1109/TIV.2020.2987430).
9. C. H. O. O. Quevedo, A. M. B. C. Quevedo, G. A. Campos, R. L. Gomes, J. Celestino, and A. Serhrouchni, "An intelligent mechanism for Sybil attacks detection in VANETs," in *Proc. ICC IEEE Int. Conf. Commun.*, 2020, pp. 1–6, doi: [10.1109/ICC40277.2020.9149371](https://doi.org/10.1109/ICC40277.2020.9149371).
10. F. Ahmad, F. Kurugollu, A. Adnane, R. Hussain, and F. Hussain, "MARINE: Man-in-the-middle attack resistant trust model in connected vehicles," *IEEE Internet Things J.*, vol. 7, no. 4, pp. 3310–3322, Apr. 2020, doi: [10.1109/JIOT.2020.2967568](https://doi.org/10.1109/JIOT.2020.2967568).
11. M. Wazid, P. Bagga, A. K. Das, S. Shetty, J. J. P. C. Rodrigues, and Y. Park, "AKM-IoV: Authenticated key management protocol in fog computing-based Internet of Vehicles deployment," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8804–8817, Oct. 2019, doi: [10.1109/JIOT.2019.2923611](https://doi.org/10.1109/JIOT.2019.2923611).
12. M. Tahboush and M. Agoyi, "A hybrid wormhole attack detection in mobile ad-hoc network (MANET)," *IEEE Access*, vol. 9, pp. 11 872–11 883, 2021, doi: [10.1109/ACCESS.2021.3051491](https://doi.org/10.1109/ACCESS.2021.3051491).
13. S. Chavhan, D. Gupta, C. B. N., A. Khanna, and J. J. P. C. Rodrigues, "Agent pseudonymous authentication-based conditional privacy preservation: An emergent intelligence technique," *IEEE Syst. J.*, vol. 14, no. 4, pp. 5233–5244, Dec. 2020, doi: [10.1109/JSYST.2020.2994631](https://doi.org/10.1109/JSYST.2020.2994631).
14. J. Wang, C. Jiang, K. Zhang, T. Q. S. Quek, Y. Ren, and L. Hanzo, "Vehicular sensing networks in a smart city: Principles, technologies and applications," *IEEE Wireless Commun.*, vol. 25, no. 1, pp. 122–132, Feb. 2018, doi: [10.1109/MWC.2017.1600275](https://doi.org/10.1109/MWC.2017.1600275).
15. M. R. Jabbarpour, M. Sookhak, S. SeyedFarshi, and A. Zomaya, "Proposing a secure self-financing vehicle using blockchain and vehicular edge computing," *IEEE Consum. Electron. Mag.*, to be published, doi: [10.1109/MCE.2020.3038029](https://doi.org/10.1109/MCE.2020.3038029).

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.rafi@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.

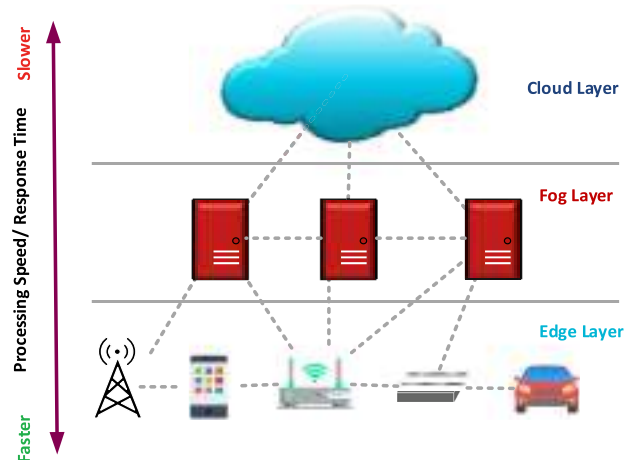


Fig. 1. IoV data processing layer stack.

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
$c_{io}(t)$	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

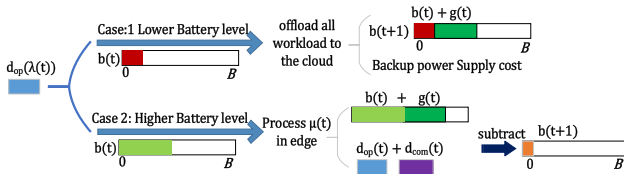


Fig. 2. Two modes of battery status.

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,

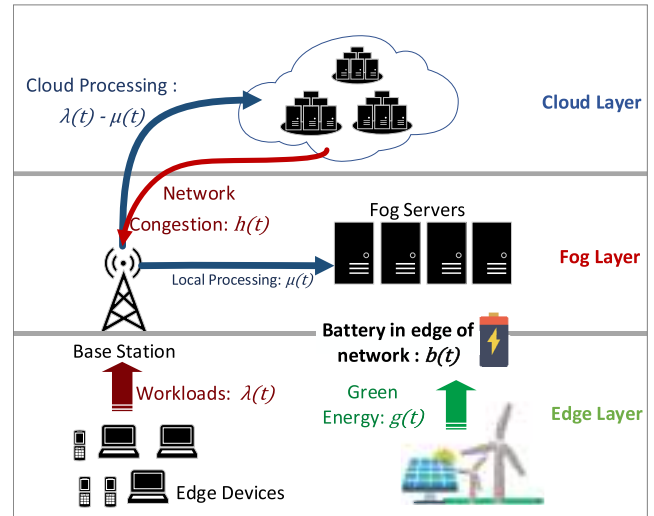


Fig. 3. The cloud-fog architecture.

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:          |      $\mu(t) \leftarrow best\_chromosom(P)$ 
27:          |     while
28:            |        $(battery + green < PowerConsumption(\mu(t)))$ 
29:              |          $\mu(t) \leftarrow next\_best\_chromosome(P)$ 
30:            |     end while
31:          |   end while
32:        |   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

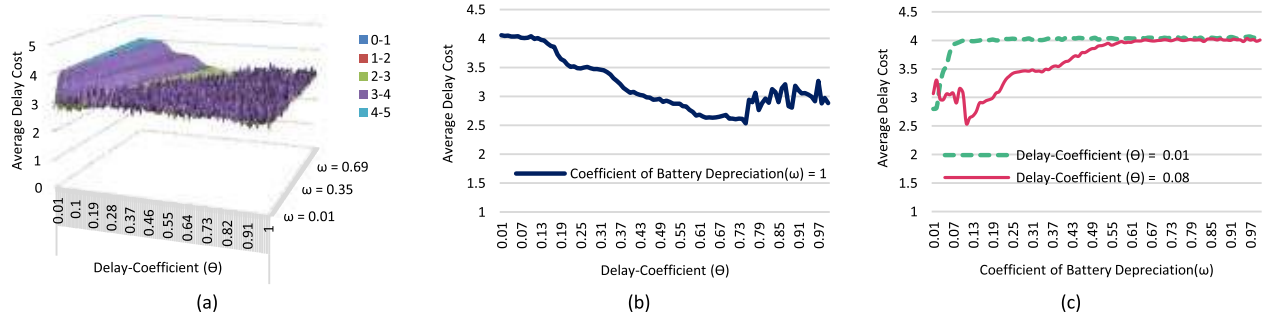


Fig. 4. The average delay cost based on the delay coefficient (θ) and the coefficient of the battery depreciation (ω). (a) The average delay cost by changing the coefficients ω and θ . (b) The effect of delay coefficient (θ) on delay cost. (c) The effect of the coefficient of battery depreciation (ω) on delay cost.

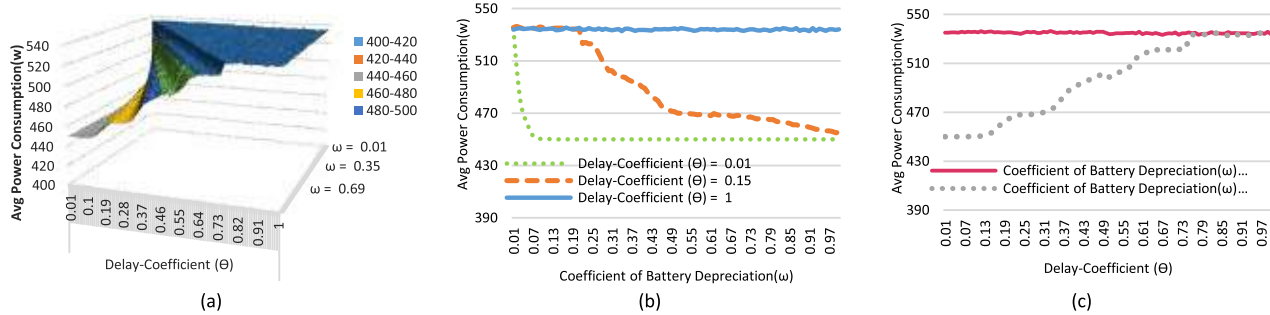


Fig. 5. The average power consumption based on the delay coefficient (θ) and coefficient of battery depreciation (ω). (a) The average power consumption by changing ω and θ . (b) The effect of the coefficient of battery depreciation (ω) on power consumption. (c) The effect of delay coefficient (θ) on 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.

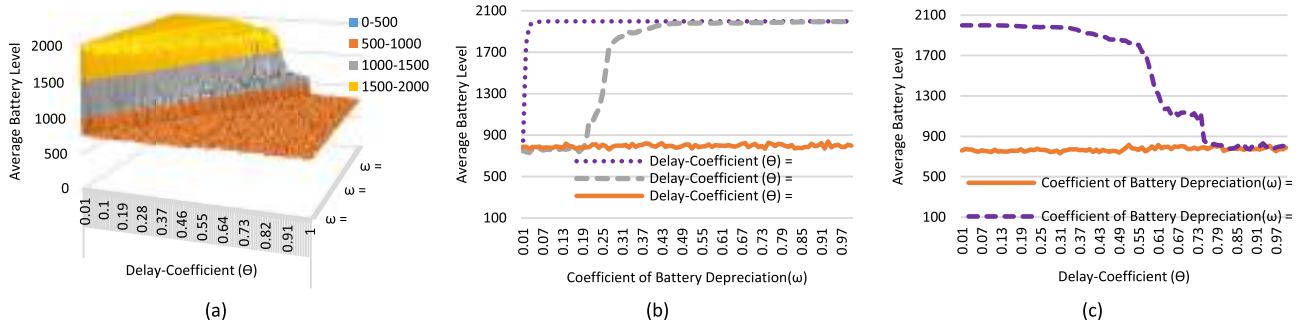


Fig. 6. The average battery level based on the delay coefficient (θ) and the coefficient of battery depreciation (ω). (a) The average battery level by changing ω and θ . (b) The effect of the coefficient of battery depreciation (ω) on battery level. (c) The effect of delay coefficient (θ) on the battery level.

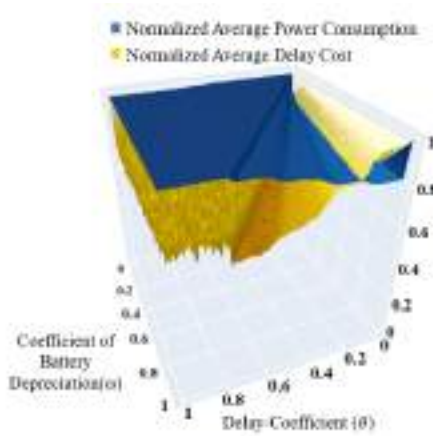


Fig. 7. Normalized values of delay cost and average power consumption.

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 the two normalized levels are the points of balance. The intersection 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

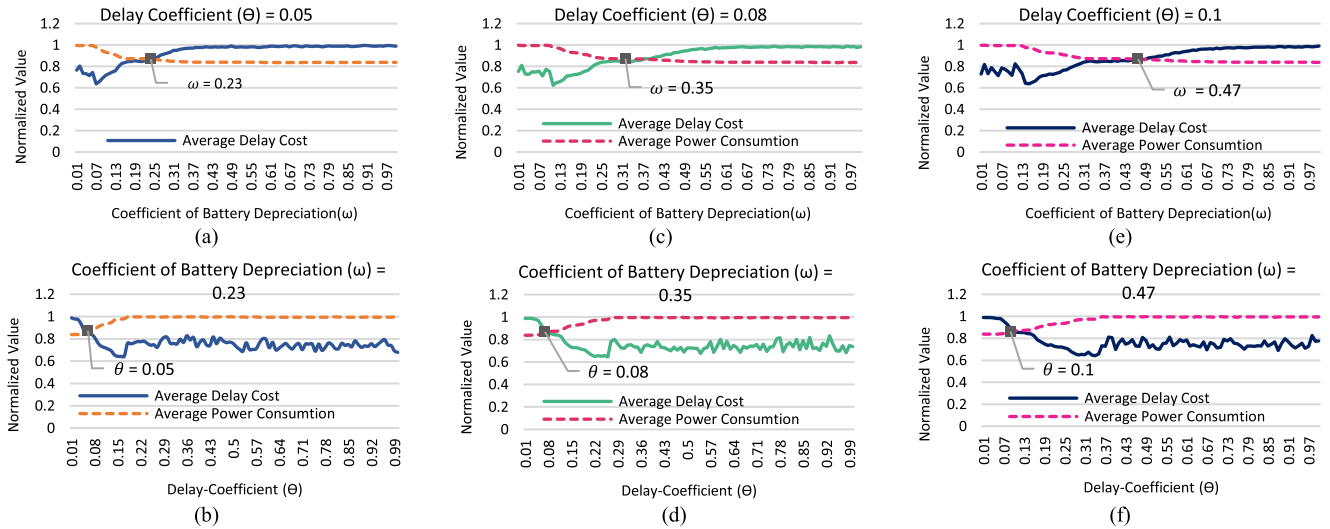


Fig. 8. Normalized values of average delay cost and power consumption for the optimum points. (a) Variations of the coefficient of battery depreciation (ω) for the first point. (b) Variations of delay cost (θ) for the first point. (c) Variations of the coefficient of battery depreciation (ω) for the 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

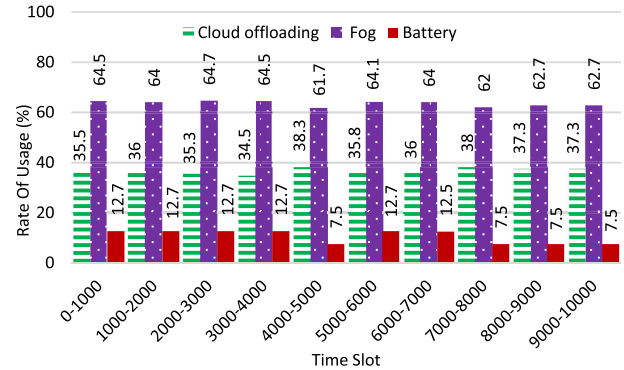


Fig. 9. The rate of usage of cloud and fog resources and power supply over time.

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

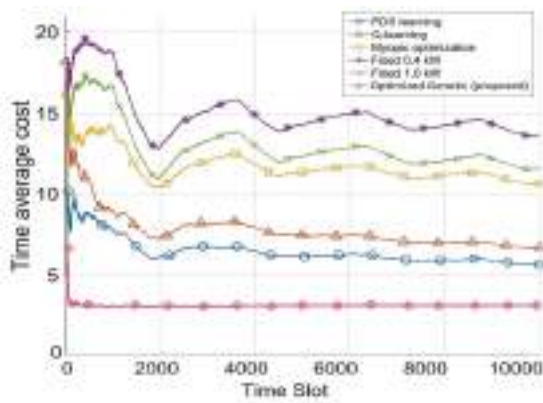


Fig. 10. The average delay cost.

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 further. 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).

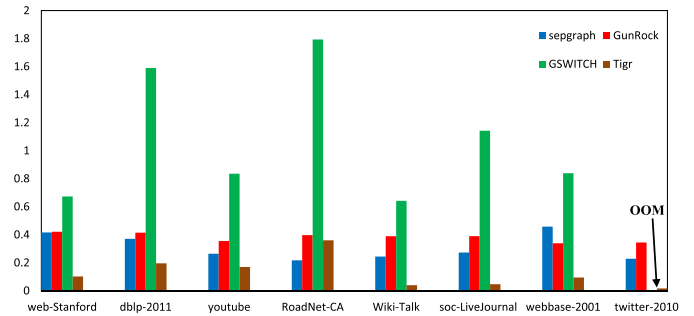


Fig. 1. Average IPC of PR on different data sets with different implementations.

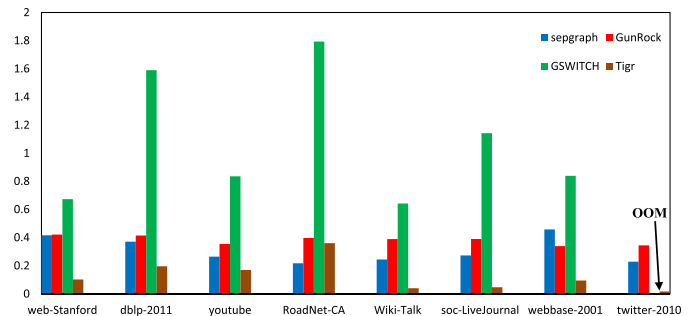


Fig. 2. Average IPC of BFS on different data sets with different implementations.

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.

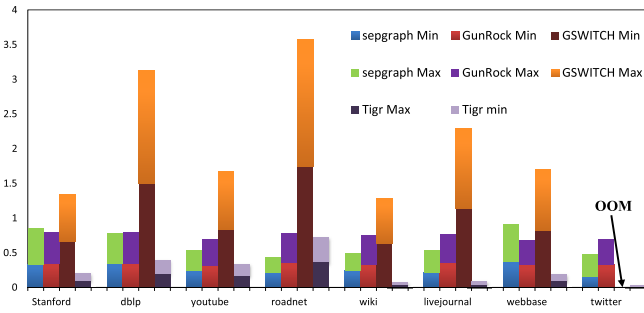


Fig. 3. Max/Min IPC of PR on different data sets with different implementations.

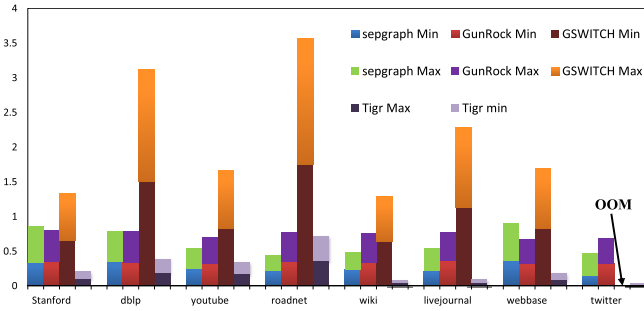


Fig. 4. Max/Min IPC of BFS on different data sets with different implementations.

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.

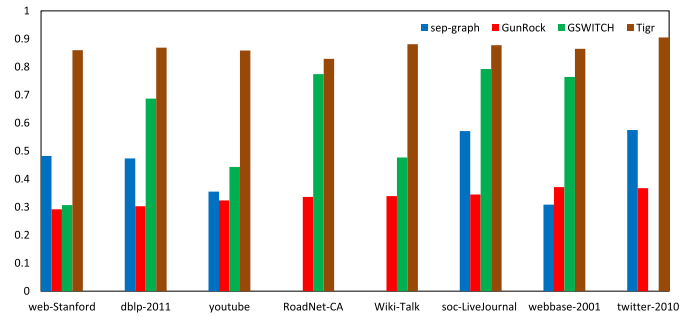


Fig. 5. Achieved occupancy of PR on different data sets with different implementations.

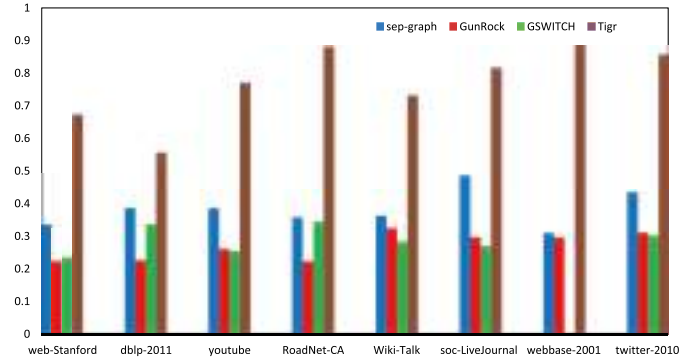


Fig. 6. Achieved occupancy of BFS on different data sets with different implementations.

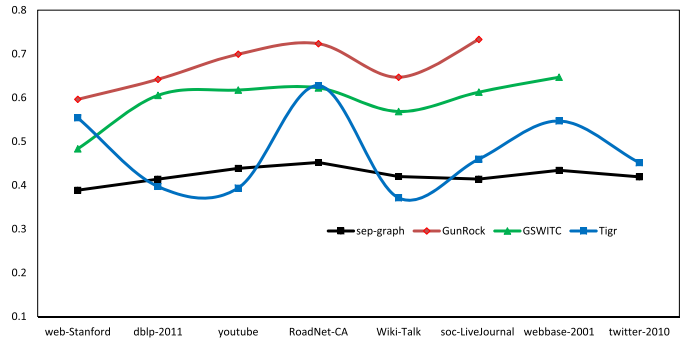


Fig. 7. Global memory efficiency for BFS on different data sets with different implementations.

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

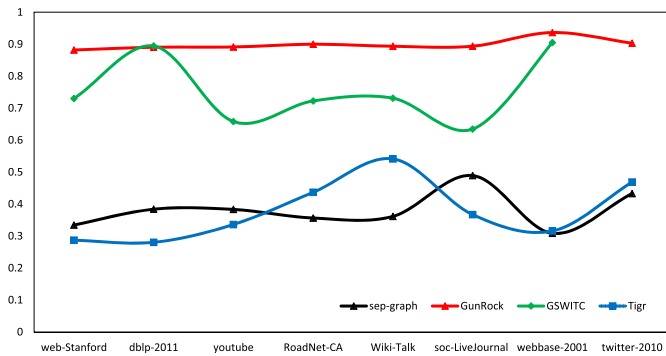


Fig. 8. Global memory efficiency for PR on different data sets with different implementations.

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

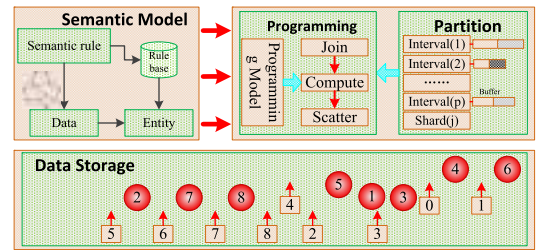


Fig. 9. Illustration of the proposed HITL CPS data processing framework.

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

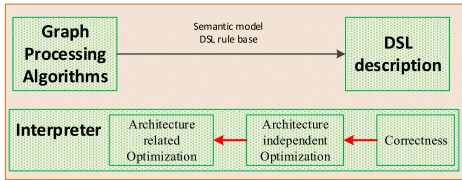


Fig. 10. Graph computing semantic model adapted to new architecture.

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

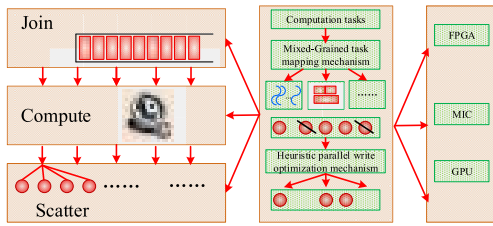


Fig. 11. Graph computing programming model adapted to multiple hardware characteristics.

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).

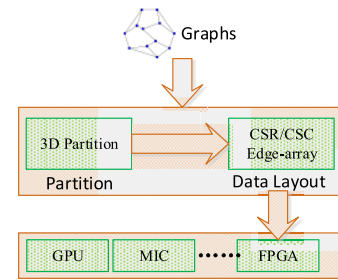


Fig. 12. Data partitioning and placement strategies for data-aware and structure-aware.

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.
- [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.
- [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

impacts of the important system parameters, such as normalized Doppler frequency and the correlation factor are evaluated in different setup scenarios. Simulation results suggest that, compared with the classical maximum likelihood detector, the proposed learning based detectors shows considerable gains.

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.](#)

[Learn more about Institutional subscriptions](#)

References

1. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, Ghemawat S, Goodfellow I, Harp A, Irving G, Isard M, Jia Y, Jozefowicz R, Kaiser L, Kudlur M, Levenberg J, Mané D., Monga R, Moore S, Murray D, Olah C, Schuster M, Shlens J, Steiner B, Sutskever I, Talwar K, Tucker P, Vanhoucke V, Vasudevan V, Viégas F., Vinyals O, Warden P, Wattenberg M, Wicke M, Yu Y, Zheng X (2015) TensorFlow: Large-scale machine learning on heterogeneous systems. <http://tensorflow.org/>. Software available from tensorflow.org

2. Aquilina P, Ratnarajah T (2015) Performance analysis of ia techniques in the mimo ibc with imperfect csi. IEEE Trans Commun 63(4):1259–1270. <https://doi.org/10.1109/TCOMM.2015.2408336>

3. Chen C, Zhong W, Yang H, Du P (2018) On the performance of mimo-noma-based visible light communication systems. IEEE Photon Technol Lett 30(4):307–310. <https://doi.org/10.1109/LPT.2017.2785964>

4. Chen J, Si J, Li Z, Huang H (2012) On the performance of spectrum sharing cognitive relay networks with imperfect csi. IEEE Commun Lett 16(7):1002–1005.

<https://doi.org/10.1109/LCOMM.2012.042512.120100>

5. Chen Z, Sohrabi F, Yu W (2019) Multi-cell sparse activity detection for massive random access: Massive mimo versus cooperative mimo. *IEEE Trans Wirel Commun* 18(8):4060–4074. <https://doi.org/10.1109/TWC.2019.2920823>
-

6. Deng D, Fan L, Lei X, Tan W, Xie D (2017) Joint user and relay selection for cooperative noma networks. *IEEE Access* 5:20220–20227
-

7. Deng D, Fan L, Zhao R, Hu RQ (2016) Secure communications in multiple amplify-and-forward relay networks with outdated channel state information. *Trans Emerg Telecommun Technol* 27(4):494–503
-

8. Deng D, Li X, Zhao M, Rabie KM, Kharel R (2020) Deep learning-based secure MIMO communications with imperfect CSI for heterogeneous networks. *Sensors* 20(6):1730. <https://doi.org/10.3390/s20061730>
-

9. Fan L, Lei X, Fan P, Hu R (2012) Outage probability analysis and power allocation for two-way relay networks with user selection and

outdated channel state information.

Communications Letters, IEEE 16(5):638–641

10. Farrag S, Alexan W (2020) Secure 3d data hiding technique based on a mesh traversal algorithm. Multimedia Tools and Applications pp(99):1–1. <https://doi.org/10.1007/s11042-020-09437-w>

11. Feng C, Jing Y, Jin S (2016) Interference and outage probability analysis for massive mimo downlink with mf precoding. IEEE Signal Process Lett 23 (3):366–370. <https://doi.org/10.1109/LSP.2015.2511630>

12. Gao H, Su Y, Zhang S, Diao M (2019) Antenna selection and power allocation design for 5g massive mimo uplink networks. China Communications 16 (4):1–15

13. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge. <http://www.deeplearningbook.org>

14. Gradshteyn I, Ryzhik I (2007) Table of integrals, series, and products, 7 edn. Academic Press, Elsevier Inc, San Diego, California 92101-4495, USA

15. He Q, Wang Z, Hu J, Blum RS (2019) Performance gains from cooperative mimo radar and mimo communication systems. IEEE Signal Process Lett 26(1):194–198.
<https://doi.org/10.1109/LSP.2018.2880836>

16. Lee H, Lee SH, Quek TQS (2019) Deep learning for distributed optimization: Applications to wireless resource management. IEEE J Sel Areas Commun 37(10):2251–2266.
<https://doi.org/10.1109/JSAC.2019.2933890>

17. Li L, et al. (2020) A unified framework for hs-uav noma networks: Performance analysis and location optimization. IEEE Wireless Commun Lett 99(1).
<https://doi.org/10.1109/ACCESS.2020.2964730>

18. Li L, Li J, Li L (2020) Performance analysis of impaired swipt noma relaying networks over imperfect weibull channels. IEEE Syst J 99(1):680–695.
<https://doi.org/10.1109/JSYST.2019.2919654>

19. Li L, Li J, Liu Y, Ding Z, Nallanathan A (2020) Residual transceiver hardware impairments on cooperative noma networks. IEEE Trans Wirel Commun 19(1):680–695

20. Li L, Liu M, Deng C (2020) Full-duplex cooperative noma relaying systems with i/q imbalance and imperfect sic. *IEEE Wirel Commun Lett* 9(1):17–20.
<https://doi.org/10.1109/JSYST.2019.2919654>

21. Li Y, Hu X, Zhuang Y, Gao Z, Zhang P, El-Sheimy N (2019) Deep reinforcement learning (drl): Another perspective for unsupervised wireless localization. *IEEE Internet of Things J*, 1–1.
<https://doi.org/10.1109/JIOT.2019.2957778>

22. Liu S, Liu Z, Sun D, Yang H, Yi K, Wang K (2018) On the performance of wireless-powered cooperative df relaying networks with imperfect csi. *China Communications* 15(11):79–92.
<https://doi.org/10.1109/CC.2018.8543051>

23. jian Luo T (2020) High-resolution sar images segmentation using nsct denoising and qiga based parameters selection of pcnn model. *Multimed Tools Appl* pp(99):1–1.
<https://doi.org/10.1007/s11042-020-09536-8>

24. Mao Q, Hu F, Hao Q (2018) Deep learning for intelligent wireless networks: A

comprehensive survey. IEEE Communications Surveys Tutorials 20 (4):2595–2621.

<https://doi.org/10.1109/COMST.2018.2846401>

-
25. Michalopoulos DS, Suraweera HA, Karagiannidis GK, Schober R (2012) Amplify-and-forward relay selection with outdated channel estimates. IEEE Trans Commun 60(5):1278–1290
-
26. Ming Z, Zhou S, Zhou W, Zhu J (2017) An improved uplink sparse coded multiple access. IEEE Commun Lett 21(1):176–179
-
27. O’Shea TJ, Corgan J, Clancy TC Jayne C, Iliadis L (eds) (2016) Convolutional radio modulation recognition networks. Springer International Publishing, Cham
-
28. Otoum S, Kantarci B, Mouftah HT (2019) On the feasibility of deep learning in sensor network intrusion detection. IEEE Netw Lett 1(2):68–71.
<https://doi.org/10.1109/LNET.2019.2901792>
-
29. Pan C, Ren H, Wang K, Xu W, ElKashlan M, Nallanathan A, Hanzo L (2020) Multicell mimo communications relying on intelligent

reflecting surfaces. IEEE Trans Wirel
Commun pp(pp):1–12

30. Qiu S, Chen D, Qu D, Luo K, Jiang T (2018) Downlink precoding with mixed statistical and imperfect instantaneous CSI for massive MIMO systems. IEEE Trans Veh Technol 67(4):3028–3041.
<https://doi.org/10.1109/TVT.2017.2774836>
-

31. Simon MK, Alouini MS (2005) Digital communication over fading channels, 2nd edn. Wiley, Hoboken
-

32. Su Y, Lu X, Zhao Y, Huang L, Du X (2019) Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks. IEEE Sensors J 19(20):9561–9569.
<https://doi.org/10.1109/JSEN.2019.2925719>
-

33. Van Luong T, Ko Y (2017) A tight bound on BER of MC-OFDM with greedy detection and imperfect CSI. IEEE Commun Lett 21(12):2594–2597.
<https://doi.org/10.1109/LCOMM.2017.2747549>
-

34. Wang Y, Zhao M, Deng D, Zhou S, Zhou W (2018) Fractional sparse code multiple access and its optimization. *IEEE Wirel Commun Lett* 7 (6):990–993

35. Ye H, Li GY, Juang B (2018) Power of deep learning for channel estimation and signal detection in ofdm systems. *IEEE Wirel Commun Lett* 7(1):114–117.
<https://doi.org/10.1109/LWC.2017.2757490>

36. Zhu J, Gong C, Zhang S, Zhao M, Zhou W (2018) Foundation study on wireless big data: Concept, mining, learning and practices. *China Ccommunications* 15(12):1–15

Acknowledgments

This work was partly supported by Natural Science Foundation of Guangdong Province with grant number 2018A030313736, Scientific Research Project of Education Department of Guangdong with grant number 2019GZDXM002, Application Technology Collaborative Innovation Center of GZPYP with grant number 2020ZX01, Yangcheng scholar, scientific research project of Guangzhou Education Bureau with grant number 202032761, Project of Technology Development Foundation of Guangdong with grant number 706049150203, the National Natural Science Foundation of China Grant 62001320, Key Scientific Research Projects

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 flexibly 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 sent 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

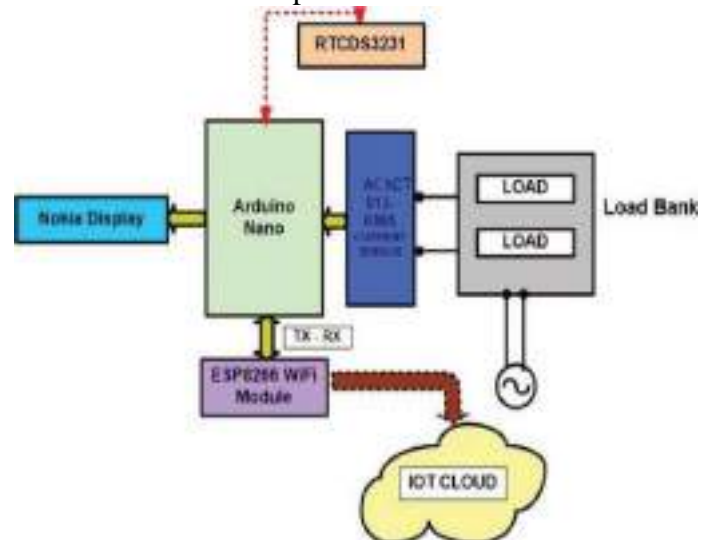


Fig. 2: Power Analyzer Architecture

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

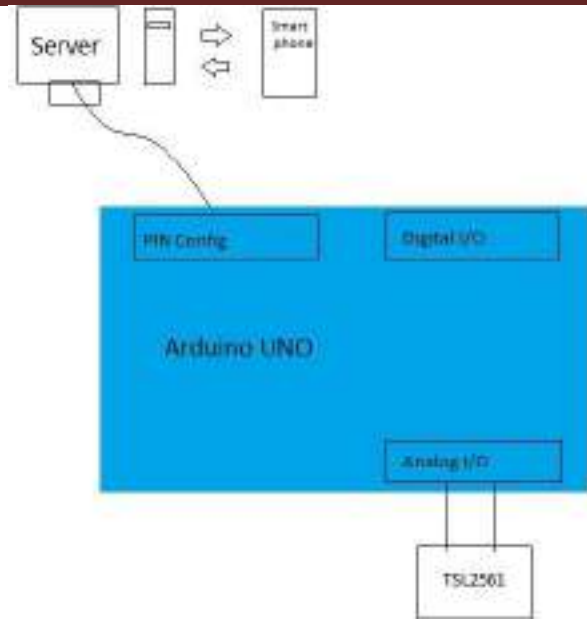


Fig. 3: Sensor Module outline

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.

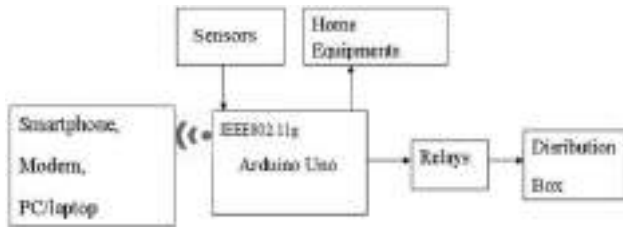


Fig. 4: Home Automation Architecture

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.

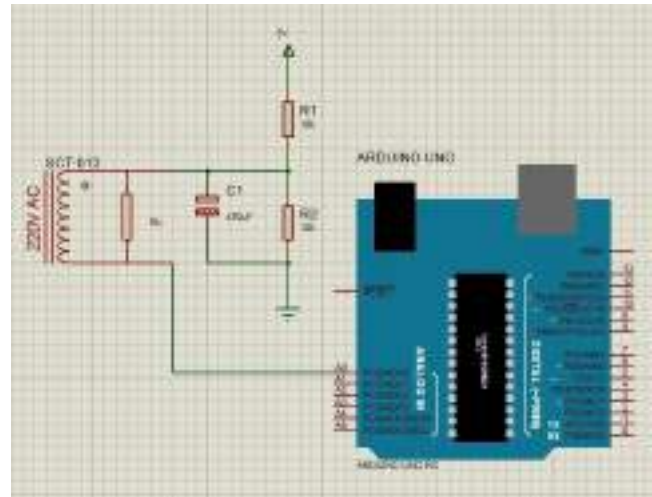


Fig. 5: (CT sensor interfaced with arduino)

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.



EARN YOUR MBA

WWW.IIMPS.IN



Accreditation & Ranking



UGC / NCTE Approved.

INFO@IIMPS.IN

☎ 011-41005174

R
E
S
E
A
R
C
H
G
A
T
E
W
A
Y

STOP PLAGIARISM



Arogyam Ayurveda
Holistic Healing through herbs



A
R
O
G
Y
A
M
O
N
L
I
N
E

PARIVARTAN PSYCHOLOGY CENTER

परिवर्तन

COLOR PSYCHOLOGY : HOW COLOR AFFECT YOUR CHILD



- BLUE** Calms your Child's Mind & Body
- YELLOW** Promotes Concentration, Stimulates the Memory
- PINK** Evokes Empathy, makes your Child Calm
- RED** Excites and energizes your Child's body
- GREEN** Improves Reading speed and Comprehension

www.parivartan4u.com



Confuse about your children's future?

परिवर्तन

भारतीय भाषा, शिक्षा, साहित्य एवं शोध

ISSN 2321 – 9726

WWW.BHARTIYASHODH.COM



**INTERNATIONAL RESEARCH JOURNAL OF
MANAGEMENT SCIENCE & TECHNOLOGY**

ISSN – 2250 – 1959 (O) 2348 – 9367 (P)

WWW.IRJMSST.COM



**INTERNATIONAL RESEARCH JOURNAL OF
COMMERCE, ARTS AND SCIENCE**

ISSN 2319 – 9202

WWW.CASIRJ.COM



**INTERNATIONAL RESEARCH JOURNAL OF
MANAGEMENT SOCIOLOGY & HUMANITIES**

ISSN 2277 – 9809 (O) 2348 - 9359 (P)

WWW.IRJMSH.COM



**INTERNATIONAL RESEARCH JOURNAL OF SCIENCE
ENGINEERING AND TECHNOLOGY**

ISSN 2454-3195 (online)

WWW.RJSET.COM



**INTEGRATED RESEARCH JOURNAL OF
MANAGEMENT, SCIENCE AND INNOVATION**

ISSN 2582-5445

WWW.IRJMSI.COM



**JOURNAL OF LEGAL STUDIES, POLITICS
AND ECONOMICS RESEARCH**

WWW.JLPER.COM

JLPE



IMAGE SPLICING DETECTION - COMPARISON OF DMAC AND DMVN NETWORKS

Roshan Prasad
Student, Dept of CSE
SCMS School of Engg.
and Technology Ernakulam
roshanp60@gmail.com

Sarath J
Student, Dept of CSE
SCMS School of Engg
and Technology Ernakulam
jsarathkaralmanna@gmail.com

Sten Benny
Student, Dept of CSE
SCMS School of Engg
and Technology Ernakulam
stenbenny02@gmail.com

Vinay Stephen
Student, Dept of CSE
SCMS School of Engg.
and Technology Ernakulam
vinaystephen1@gmail.com

Litty Koshy
Asst Professor, Dept of CSE
SCMS School of Engg. and Technology, Ernakulam
littykoshy@scmsgroup.org

Abstract---Constrained image splicing detection and localization (CISDL) is a difficult task for image forensics that examines two input suspicious pictures and determines whether one contains suspected portions copied from the other. Here a unique adverse learning approach for training the deep matching network for CISDL is presented. The goal of the deep matching network based on atrous convolution (DMAC) is to create two high-quality candidate masks that show the suspicious regions of the two input pictures. The correlation layer based on the skip architecture is proposed to capture hierarchical features in DMAC, and Atrous spatial pyramid pooling is used to extract features with rich spatial information. Another model called DMVN uses the same process as DMAC but it is not use atrous convolution. A comparative study of both models was done, in which the DMAC model is better because it gives high resolution fined grained mask.

Keywords: Atrous convolution, DMAC, DMVN, CISDL

I. INTRODUCTION

Malicious image forgery is becoming a global epidemic in recent years, due to the rapidly declining cost of digital cameras and quick development of sophisticated image editing tools. Forgers may use forged images to produce fake news, spread rumors or give false testimony, which result in negative social impacts. Image forensics, which seeks to distinguish forged images and prevent forgers from using forged images for unscrupulous business or political purposes, has attracted great attention in research and industrial communities. A variety of image forensics

methods investigate an individual image and detect its high-level or low-level inconsistencies caused by image manipulation. However, it is still a challenging task to accurately distinguish forged images, due to advanced image manipulation techniques and limited information provided by a single image. Moreover, these image forensics methods identify forged images or regions without providing the source of forged regions or specific tampering process, but these auxiliary evidences can provide more clues and make results more convincing in real applications. Constrained image splicing detection and localization (CISDL) is newly formulated in the Media Forensics Challenge. Different from “conventional” splicing detection, “constrained” means that the inputs are two images: one is a probe image and the other is a potential donor image. In CISDL given a probe image P and a potential donor image D , CISDL aims to detect if a region of D has been spliced into P , and consequently provide mask images P_m and D_m indicating the regions of P were spliced from D . DMVN generates correlation maps by comparing high-level low-resolution feature maps of VGG, and constructs an inception-based mask convolution module to locate suspected regions. However, low-resolution feature maps restrict DMVN's ability to detect accurate boundaries and small suspected regions. Here proposed a deep matching



network based on atrous convolution (DMAC) to generate high-quality candidate masks from high-resolution feature maps.

The basic DMAC architecture achieves significant improvements over DMVN.

This work, proposes a DMAC network which takes two images as inputs. These input images are fed into a atrous convolution network for feature map extraction. The extracted feature maps fed into the correlation layer and atrous spatial pyramid pooling for feature maps comparisons.



Fig 1. a) Donor Image b) Donor Mask c) Probe Image d) Probe mask

II. LITERATURE SURVEY

In [1] Proposes a new convolutional layer that suppresses image content and learns forgery detection. In [1] they proposed a CNN to learn manipulation detection features directly from data and it is used in image forensics. In [2] DMAC is combined with adversarial learning for effective image forgery detection. In [3] propose an optimized 3D lighting estimation method by incorporating a more general surface reflection model. In [4] propose a framework to improve the performance of forgery localization via integrating tampering possibility maps. In [5] proposed algorithm automatically computes a

likelihood map indicating the probability for each 8×8 discrete cosine transform block of being doubly compressed.

III. PROPOSED METHODOLOGY

In this section, explains the proposed framework, as shown in fig 2. In the DMAC model there are three modules namely feature extraction module, correlation module and ASPP (Atrous Spatial Pyramid Pooling). In the feature extraction module atrous convolution is adopted to enrich the spatial information of convolutional features.

In the correlation module, the skip architecture is designed for hierarchical features comparisons and in the ASPP module is used to capture the information of different scales, Atrous Spatial Pyramid Pooling is constructed to generate the final mask. ASPP contains multiple parallel atrous convolutional layers with different sampling rates.

In DMAC Model, two images as inputs probe and donor. In which donor is the original image which is captured by camera also called as authentic image and probe image is the image containing spliced portion of donor image also called as tampered image. This two image is given as input to the model and it produce high resolution fine grained mask as output. The DMAC model is using atrous convolution which is used to give high resolution mask. At last creating another model called DMVN, in this which is not using atrous convolution and compare with DMAC model for accuracy.

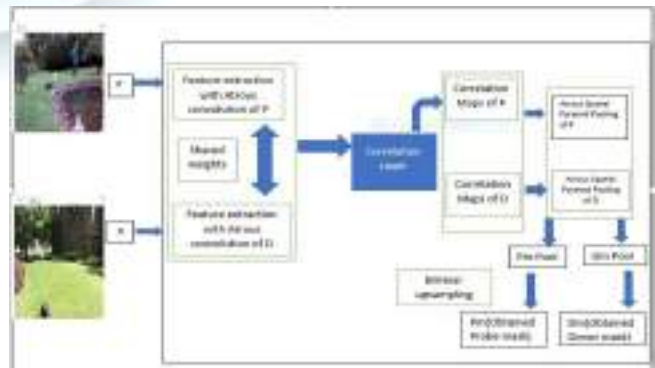


Fig 2. Proposed System Architecture

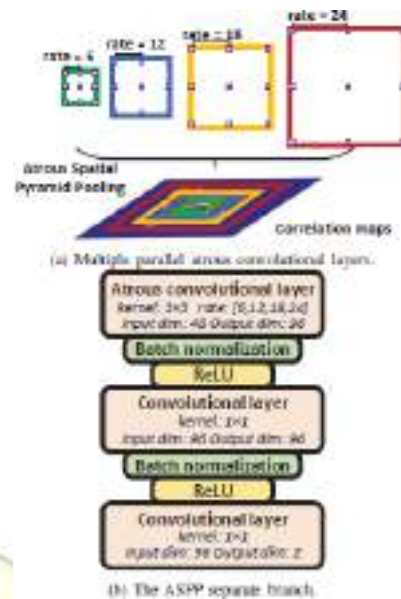


The DMAC network is a unique adversarial network in which feature extraction modules employing atrous convolution, the correlation layer with skip architecture, and ASPP are designed to enrich geographical information. In DMAC atrous convolution, the correlation layer and ASPP are used to capture hierarchical properties and localise impacted regions at many scales, respectively. The detection network and discriminative network, which act as losses with supplementary parameters, monitor DMAC's adversarial training.

A. Feature Extraction with Atrous Convolution

CNNs' pooling or downsampling techniques necessarily degrade the spatial resolution of the output feature maps. As a result, in this research, atrous convolution is employed to create high-resolution feature maps. Atrous convolution allows us to vary the field-of-view of filters by adjusting the rate value without adding any more parameters. This module alters the image by adjusting the colour, contrast, and light intensity, which aids in the creation of a high-resolution mask.

Assume the input feature maps are scaled down by a factor of two before being convolved with standard convolution filters. The created feature maps are just a fourth the size of the original feature maps, and traditional filters only acquire answers from a quarter of the image locations. If we eliminate the downsampling layer and directly convolve the input feature maps, the filters will have a smaller field of vision. Fortunately, we may keep the original field-of-view by employing atrous convolution with rate $r = 2$. Using atrous convolution techniques, we can create high-resolution feature maps, get all answers from the input feature maps, and don't need any additional parameters or calculation. Despite the fact that the effective filter size increases, we only need to examine non-zero filter values, resulting in a constant number of filter parameters and operations per site.



B. Correlation Computation Module

To build dense high-resolution feature maps, atrous convolution is used as the basic process. One of the most difficult difficulties in deep matching tasks is deep feature comparison. For other tasks, only the neighbouring fields are compared, allowing for the creation of complex correlation layers. They usually compute the scalar product of a pair of individual descriptors at each place for long-range correlation computing tasks. We can denote correlation computation procedure as a function. The skip architecture is proposed to effectively organize the atrous convolution and hierarchical convolution features. It makes full use of the feature extraction module's wealth of information. Three sets of feature maps are produced by the atrous convolution layer: f_3 , f_4 , and f_5 . As a result, three sets of correlation feature maps can be created using the feature maps f_3 , f_4 , and f_5 , with no upsampling or mapping functions required. The computation procedure of the proposed correlation layer based on the skip architecture can be summarized as Algorithm: The skip architecture is used to compute the correlation layer.

C. Atrous Spatial Pyramid Pooling

The fact that the tampered portions are on different scales is a problem in the image splicing detection technique. The final masks are generated using Atrous Spatial Pyramid Pooling (ASPP), which captures the information



on multiple scales provided by the correlation maps. Simply ASPP is a discriminative network that drives the DMAC network to produce masks that are hard to distinguish from ground truth ones. Multiple atrous convolutional layers with varying sampling rates are present in ASPP. As a result, those obscene convolution filters have varying field-of-views and can focus on altered parts of various scales. A separate branch of convolutional layers, batch normalisation, and ReLU layers follow each atrous convolutional layer with one sample rate. The individual branches are then merged to create the finished masks. During mask formation, there is no upsampling operation with learnable parameters, thus we just use bilinear upsampling during test time.

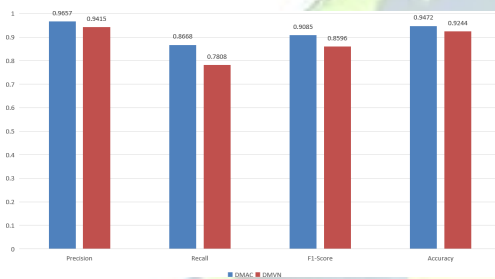
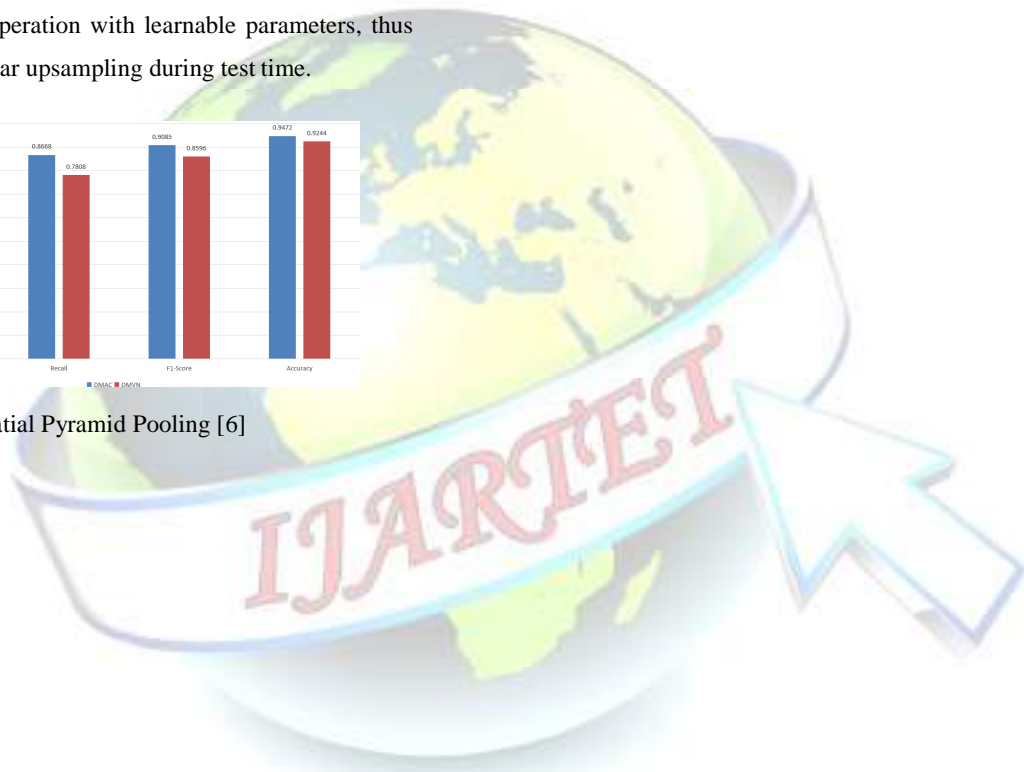


Fig 3: Atrous Spatial Pyramid Pooling [6]





IV. RESULT AND DISCUSSION

```
Precision: 0.8541114489666498  
Recall: 0.9286295170072838  
F1 Score: 0.8898130659054911  
Accuracy: 0.9384002685546875
```

Fig 4- Result of DMAC Model

Fig 5- Comparison Between DMAC and DMVN

V. CONCLUSION

This work provides a simple but effective framework for detecting image splices. A unique adversarial learning framework is proposed to deal with the CISDL task. To improve the DMAC network's ability to detect small matching regions and multi-scale regions, atrous convolution, the skip architecture, and ASPP are used. A lot of experiments are conducted on all generated datasets and also all publicly available datasets. The experimental results demonstrate the appealing performance of the proposed adversarial learning framework and the DMAC network. The use of atrous convolution and ASPP has clearly increased the effectiveness of the algorithm compared to the existing ones. Although the techniques to detect small tampered regions and regions under huge changes still need further research.

REFERENCES

- [1] Belal Ahmed, T. Aaron Gulliver, Saif alZahir (2020) "image splicing detection using mask-RCNN", "Verlag London Ltd., part of Springer Nature 2020".
- [2] Snigdha K. Mankar, Prof. Dr. Ajay A. Gurjar (2015) image splicing detection based on SVM (Support Vector Machines) classifier at IETE Amravati Center Volume 5.
- [3] A. Alahmadi, M. Hussain, H. Aboalsamh, G. Muhammad, G. Bebis, and others, "Splicing image forgery detection based on DCT and Local Binary Pattern," in Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE, pp. 253–256, 2013.
- [4] Hsu Y-F, Chang S-F. (2007). Image splicing detection using

camera response function consistency and automatic segmentation. In: Multimedia and Expo, IEEE International Conference on, p. 28–31.

[5] Chi-Man Pun, Bo Liu, Xiao-Chen Yuan, "Multi scale noise estimation for image splicing forgery detection" in Journal of visual communication and image representation, pp. 195–206, 2016.

[6] Y. Liu, X. Zhu, X. Zhao and Y. Cao, "Adversarial Learning for Constrained Image Splicing Detection and Localization Based on Atrous Convolution," in IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2551–2566, Oct. 2019, doi: 10.1109/TIFS.2019.2902826.

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-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

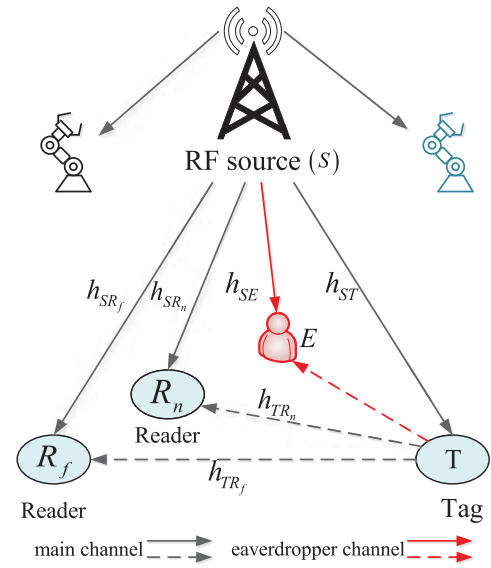


Fig. 1. Ambient backscatter NOMA system model.

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 $\gamma_{R_f}^{x_2}$.

$$\sigma_{e_{ST}}^2 + \sigma_{e_{S R_n}}^2 (1 + \varpi \varphi_J + \kappa_{S R_n}^2), \quad O_{R_n} = \varepsilon a_2 + \varpi \varphi_J + \kappa_{S R_n}^2, \quad m_{R_n} = \beta^2 (\kappa_{S R_n}^2 + \varpi \varphi_J), \quad \xi_{R_n} = \varepsilon + \varpi \varphi_J + \kappa_{S R_n}^2; \\ B_E = \beta^2 (1 + \varphi_J + \kappa_{S E}^2), \quad C_E = B_E \sigma_{e_{ST}}^2, \quad M_E = B_E \sigma_{e_{TE}}^2, \quad Q_E = a_1 + \varphi_J + \kappa_{S E}^2, \quad \psi_E = B_E \sigma_{e_{TE}}^2 \sigma_{e_{ST}}^2 + \sigma_{e_{SE}}^2 (1 + \varphi_J + \kappa_{S E}^2), \quad O_E = \varepsilon a_2 + \varphi_J + \kappa_{S E}^2, \quad m_E = \beta^2 (\kappa_{S E}^2 + \varphi_J), \quad \xi_E = \varepsilon + \varphi_J + \kappa_{S E}^2.$$

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 $\Delta_1^i = \frac{\lambda_{Si} (a_2 - Q_i \gamma_{th2}^i)}{\lambda_{ST} \lambda_{Ti} B_i \gamma_{th2}^i}$, ($i \in \{R_n, R_f, E\}$), 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..

$$\begin{aligned} \text{In (28), } \Delta_6 &= \frac{\lambda_{TR_n} \varsigma_{R_n} C_{R_n} \gamma_{thc}^{R_n} (B_{R_n} \gamma_{thc}^{R_n} + \Delta_5) + \lambda_{SR_n} \gamma_{thc}^{R_n} \Delta_5^{R_n}}{\Delta_5^{R_n} \gamma}, \\ \Delta_7^i &= (\lambda_{Si} \xi_i - \lambda_{Ti} C_i) \gamma_{thc}^i, \quad \Delta_8^i = (\lambda_{Ti} C_i \gamma_{thc}^i + \Delta_7^i) \gamma, \\ A_1^i &= \frac{-1}{\lambda_{Si} \xi_i \lambda_{ST} \lambda_{TE} \Delta_5^i \gamma^2}, \quad A_2^i = -\left(\frac{C_i \gamma_{thc}^i}{\lambda_{ST} \Delta_5^i} + \frac{M_i \gamma_{thc}^i}{\lambda_{TE} \Delta_5^i} \right), \\ A_3^i &= \frac{\lambda_{Ti} M_i (C_i \gamma_{thc}^i)^2 + (\lambda_{Ti} \psi_i \Delta_5^i + \Delta_7^i M_i) C_i \gamma_{thc}^i}{\Delta_5^i} + \Delta_7^i \psi_i \gamma^2, \\ A_4^i &= \lambda_{Si}^2 \xi_i^2 \gamma_{thc}^i, \quad B_1 = \frac{\lambda_{SR_n} + \lambda_{ST} \varsigma_{R_n} M_{R_n}}{\lambda_{SR_n}^2 \lambda_{ST} \lambda_{TE} R_n \gamma \Delta_5^{R_n}} + \frac{\varsigma_{R_n} C_{R_n} B_{R_n} \gamma_{thc}^{R_n}}{\lambda_{SR_n}^2 \lambda_{TE} R_n (\Delta_5^{R_n})^2}, \\ (i \in \{R_n, R_f, E\}), \quad B_5 &= \frac{(\lambda_{ST} \varsigma_{R_n} C_{R_n} + \lambda_{SR_n}) C_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{SR_n} \Delta_5^{R_n}} + B_2 + \frac{\varsigma_{R_n} \psi_{R_n}}{\lambda_{SR_n}}, \end{aligned}$$

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 $\Delta_{17} = \left(\frac{M_E \gamma_{th2}}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)} + \frac{1}{\lambda_{ST}} \right) \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{TE} B_E \gamma_{th2}^E}$,

$$\Delta_{16} = \frac{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}{\lambda_{ST} \lambda_{TE} B_E \gamma_{th2}^E}, \quad \text{and} \quad \Delta_{18} = \frac{\psi_E \gamma_{th2}^E}{\lambda_{SE} (a_1 - O_E \gamma_{th2}^E)}.$$

For non-ideal conditions, the analytical expression for the IP of T in (37), as shown at the bottom of the page.

$$\text{In (27), } \Delta_{13} = 1 / (\lambda_{ST} \lambda_{TE} \Delta_5^E), \quad \Delta_{14} = \frac{C_E \gamma_{thc}^E (\lambda_{ST} + \lambda_{TE})}{\lambda_{TE} \lambda_{ST} \Delta_5^E}, \quad \text{and} \quad \Delta_{15} = \frac{C_E^2 (\gamma_{thc}^E)^2}{(\Delta_5^E)^2 \lambda_{TE}} + \left(\psi_E + \frac{1}{\gamma} \right) \gamma_{thc}^E.$$

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

$$P_{int, \infty}^{R_f, ni} = -\Delta_{1'}^E e^{\Delta_{2'}^E - \Delta_{3'}^E - \frac{\gamma_{th2}^E}{\lambda_{SE} \gamma (a_2 - Q_E \gamma_{th2}^E)}} (1 + b_1') \text{Ei}(-(\Delta_{2'}^E + b_1')), \quad (38)$$

where $\Delta_{2'}^E = \frac{M_E}{\lambda_{TE} B_E} + \frac{M_E C_E \gamma_{th2}^E}{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E) B_E}$, $\Delta_{3'}^E = \frac{\psi_E \gamma_{th2}^E}{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E)}$,

$$\text{and } b_1' = \frac{\lambda_{SE} (a_2 - Q_E \gamma_{th2}^E) + \lambda_{TE} C_E \gamma_{th2}^E}{\lambda_{me} \lambda_{TE}^2 B_E \gamma_{th2}^E}.$$

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.

$$\begin{aligned} P_{out, \infty}^{T, ni} &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TE} \Delta_5^{R_n}} e^{A_2^{R_n}} \left(\frac{\pi}{N} \sum_{k=0}^N e^{-\left(\frac{2A_3^{R_n}}{A_4^{R_n} (\vartheta_k + 1)} - \frac{A_1^{R_n} A_4^{R_n} (\vartheta_k + 1)}{2} \right)} \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} + \frac{1}{\vartheta_k + 1} \right) - 2K_0 \left(2\sqrt{-A_1^{R_n} A_3^{R_n}} \right) \right) \\ &\quad + \frac{2\lambda_{SR_n}}{\lambda_{TR_n} \lambda_{ST} \varsigma_{R_n} B_{R_n}} e^{-B_5} \sum_{v=1}^{\infty} (-1)^v \frac{1}{B_4^v} \left(\frac{B_3}{B_1} \right)^{\frac{v}{2}} K_v \left(2\sqrt{B_3 B_1} \right). \quad (32) \end{aligned}$$

$$\begin{aligned} P_{int}^{T, ni} &= \frac{\lambda_{SE} \xi_E \gamma_{thc}^E}{\lambda_{ST} \lambda_{TE} \Delta_5^E} e^{A_2^E} \left(\frac{\pi}{N} \sum_{k=0}^N e^{-\left(\frac{2(A_3^E + \Delta_8^E)}{A_4^E (\vartheta_k + 1)} - \frac{A_1^E A_4^E (\vartheta_k + 1)}{2} \right)} \sqrt{1 - \vartheta_k^2} \left(\frac{1}{\vartheta_k + 3} - \frac{1}{\vartheta_k + 1} \right) + 2K_0 \left(2\sqrt{-A_1^E (A_3^E + \Delta_8^E)} \right) \right) \\ &\quad + 2\sqrt{\Delta_{15} \Delta_{13}} e^{-\Delta_{14}} K_1 \left(2\sqrt{\Delta_{13} \Delta_{15}} \right). \quad (37) \end{aligned}$$

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)	$\{\gamma_{th1}^{R_n}, \gamma_{th2}^{R_n} = \gamma_{th2}^{R_f}, \gamma_{thc}^{R_n}\} = \{1.2, 1, 0.001\}$
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. $l = 1, 2, 3$.

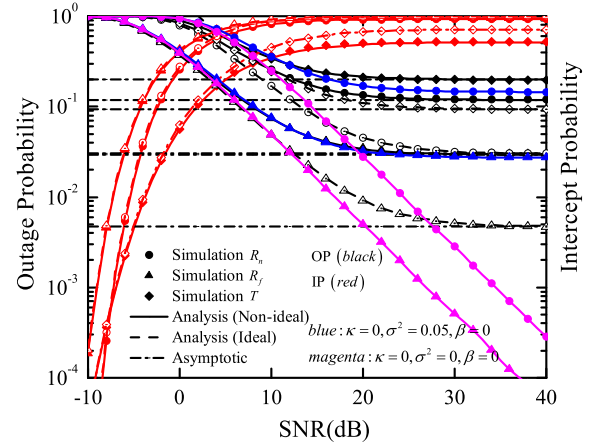


Fig. 2. OP and IP versus the transmit SNR.

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)

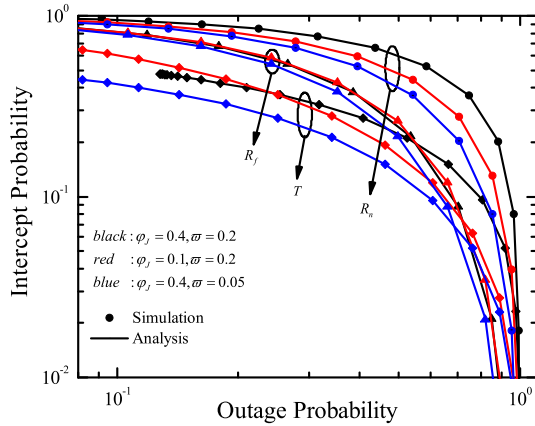
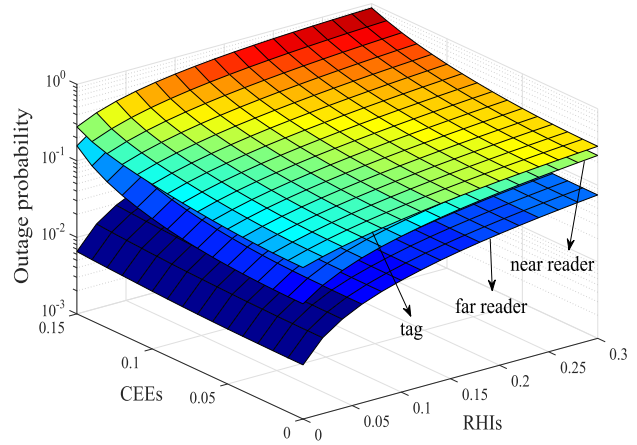


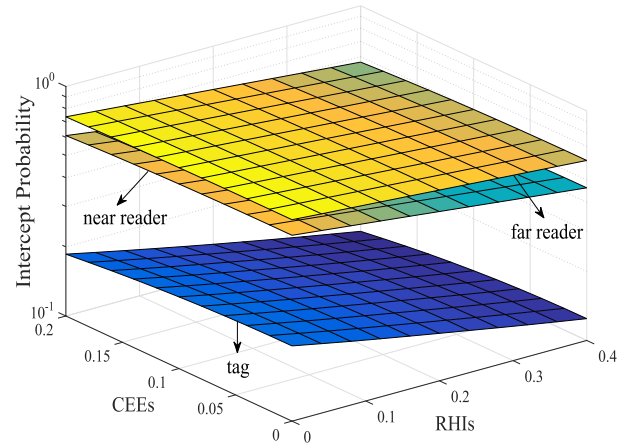
Fig. 3. IP versus OP for different power coefficient of artificial noise φ_J .



(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.

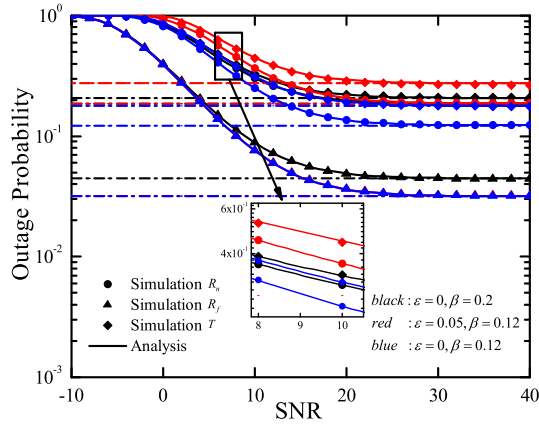
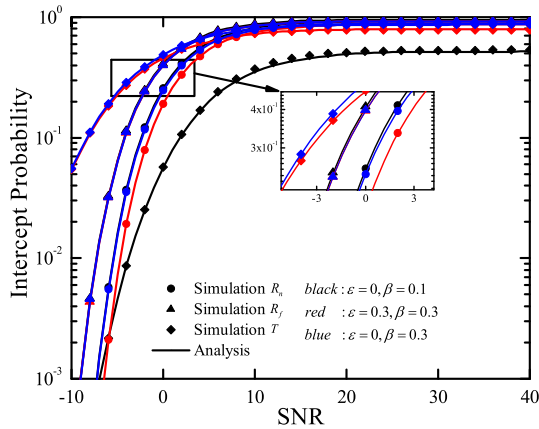
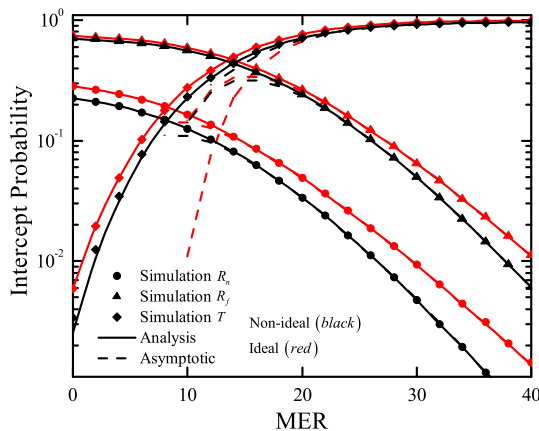
(a) OP versus the transmit SNR for different ε and β .(b) IP versus the transmit SNR for different ε and β Fig. 5. OP and IP versus the transmit SNR for different ε and β .

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

$$P_{out}^{R_f} = 1 - \underbrace{\text{Pr}\left(\gamma_{R_f}^{x_2} > \gamma_{th2}^{R_f}\right)}_{I_1}, \quad (\text{A.1})$$

where I_1 is calculated as follows:

$$\begin{aligned} I_1 &= \text{Pr}\left(\gamma_{R_f}^{x_2} > \gamma_{th2}^{R_f}\right) \\ &= \int_{\alpha_1}^{\infty} \frac{1}{\lambda_{SR_f}} e^{-\frac{x}{\lambda_{SD_f}}} \frac{1}{\lambda_{TR_f}} e^{-\frac{y}{\lambda_{TR_f}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dx dy dz \\ &\stackrel{u=z+\alpha}{=} \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f} - \frac{\gamma_{th2}^{R_f}}{\lambda_{SR_f} \gamma_{(a_2 - Q_{R_f} \gamma_{th2}^{R_f})}}} \int_{\alpha}^{\infty} e^{-\alpha_3 u} \frac{1}{u} du \\ &\stackrel{l_1}{=} 1 + \Delta_1^{R_f} e^{\Delta_2^{R_f} - \Delta_3^{R_f} - \frac{\gamma_{th2}^{R_f}}{\lambda_{SR_f} \gamma_{(a_2 - Q_{R_f} \gamma_{th2}^{R_f})}}} \text{Ei}\left(-\Delta_2^{R_f}\right), \end{aligned} \quad (\text{A.2})$$

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

$$P_{out}^T = 1 - \underbrace{\Pr \left(\gamma_{R_n}^{x_2} > \gamma_{th2}^{R_n}, \gamma_{R_n}^{x_1} > \gamma_{th1}^{R_n}, \gamma_{R_n}^{c(t)} > \gamma_{thc}^{R_n} \right)}_{I_2}, \quad (\text{B.1})$$

• 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 $u = \lambda_{SR_n} \lambda_{TR_n} \Delta_5^{R_n} z - \lambda_{SR_n} \lambda_{TR_n} C_{R_n} \gamma_{thc}^{R_n}$, $\alpha_4 = \frac{[\lambda_{TR_n} \varsigma_{R_n} (B_{R_n} z + C_{R_n}) + \lambda_{SR_n} (\psi_{R_n} + \frac{1}{\gamma}) \gamma_{thc}^{R_n} + M_{R_n} \gamma_{thc}^{R_n} z]}{\lambda_{SR_n} \lambda_{TR_n} (\Delta_5^{R_n} z - C_{R_n} \gamma_{thc}^{R_n})}$, $\alpha_5 = \frac{\lambda_{SR_n}}{\lambda_{TR_n} \varsigma_{R_n} (B_{R_n} z + C_{R_n}) + \lambda_{SR_n}}$, $\alpha_6 = B_5 + \frac{\lambda_{TR_n} \varsigma_{R_n} B_{R_n} \gamma_{thc}^{R_n}}{\lambda_{SR_n} \lambda_{TR_n} \gamma_{thc}^{R_n} \Delta_5^{R_n}}$, and l_2 is obtained by utilizing [15, eq. (3.471)].

$$\begin{aligned} I_{22} &= \int_{\frac{C_{R_n} \gamma_{thc}^{R_n}}{\Delta_5^{R_n}}}^{\infty} e^{\frac{\psi_{R_n} + \frac{1}{\gamma}}{\lambda_{SR_n} \xi_{R_n}} + \left(\frac{M_{R_n}}{\lambda_{SR_n} \xi_{R_n}} - \frac{1}{\lambda_{ST}} \right) z - \alpha_7} \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} (\lambda_{TR_n} \Delta_5^{R_n} z + \Delta_7^{R_n})} dz \\ &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TR_n} \Delta_5^{R_n}} e^{A_2} \int_0^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u + A_4^{R_n}} du \\ &= \frac{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}{\lambda_{ST} \lambda_{TR_n} \Delta_5^{R_n}} e^{A_2} \int_0^{\infty} e^{-\left((-A_5^{R_n}) u + \frac{A_3}{u} \right)} \frac{1}{u + A_4^{R_n}} du \\ &\quad + \int_{A_4^{R_n}}^{\infty} e^{-\left((-A_1^{R_n}) u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u + A_4^{R_n}} du, \quad (\text{B.4}) \end{aligned}$$

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 &= \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 \\ &= \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 \\ &\quad - \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_{R_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. \quad (\text{B.2}) \end{aligned}$$

$$\begin{aligned}
P_{out}^{T,id} &= 1 - \int_0^\infty \frac{(\psi_{R_n} + 1/\gamma)^{\gamma_{thc} R_n}}{\Delta_5^{R_n}} \left(e^{-\frac{\varsigma_{R_n}(B_{R_n} y + \psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n} y - (\psi_{R_n} + 1/\gamma)^{\gamma_{thc} R_n}}{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST} \lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST} \lambda_{TR_n}}} \right) dy \\
&= 1 - \underbrace{\int_0^\infty \left(e^{-\frac{\varsigma_{R_n}(B_{R_n} y + \psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n} y - (\psi_{R_n} + 1/\gamma)^{\gamma_{thc} R_n}}{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST} \lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST} \lambda_{TR_n}}} \right) dy}_{I_{31}} \\
&\quad + \underbrace{\int_0^\infty \frac{(\psi_{R_n} + 1/\gamma)^{\gamma_{thc} R_n}}{\Delta_5^{R_n}} \left(e^{-\frac{\varsigma_{R_n}(B_{R_n} y + \psi_{R_n})}{\lambda_{SR_n}}} - e^{-\frac{\Delta_5^{R_n} y - (\psi_{R_n} + 1/\gamma)^{\gamma_{thc} R_n}}{\lambda_{SR_n} \xi_{R_n} \gamma_{thc}^{R_n}}} \right) \frac{2}{\lambda_{ST} \lambda_{TR_n}} K_0 \left(2\sqrt{\frac{y}{\lambda_{ST} \lambda_{TR_n}}} \right) dy}_{I_{32}}. \quad (B.6)
\end{aligned}$$

$$\begin{aligned}
& - \int_0^{A_4^{R_n}} e^{-\left((-A_1^{R_n})u + \frac{A_3^{R_n} + \Delta_8^{R_n}}{u} \right)} \frac{1}{u} du \\
& = 2K_0 \left(2\sqrt{-A_1^{R_n} \left(A_3^{R_n} + \Delta_8^{R_n} \right)} \right) \\
& \quad - \frac{\pi}{N} \sum_{k=0}^N \frac{1}{\vartheta_{k+1}} e^{-\left(\frac{2(A_3^{R_n} + \Delta_8^{R_n})}{A_4^{R_n} (\vartheta_{k+1})} - \frac{A_1^{R_n} A_4^{R_n} (\vartheta_{k+1})}{2} \right)} \sqrt{1 - \vartheta_k^2}. \quad (B.5)
\end{aligned}$$

$$\begin{aligned}
& \times \int_0^\infty \frac{(\Delta_5^E z - C_E \gamma_{thc}^E) y - M_E \gamma_{thc}^E z - (\psi_E + 1/\gamma)^{\gamma_{thc}^E}}{\Delta_5^E z - C_E \gamma_{thc}^E} e^{-\frac{(\Delta_5^E z - C_E \gamma_{thc}^E) y - M_E \gamma_{thc}^E z - (\psi_E + 1/\gamma)^{\gamma_{thc}^E}}{\lambda_{SE} \xi_E \gamma_{thc}^E}} \\
& \times \frac{1}{\lambda_{TE}} e^{-\frac{y}{\lambda_{TE}}} \frac{1}{\lambda_{ST}} e^{-\frac{z}{\lambda_{ST}}} dy dz. \quad (C.1)
\end{aligned}$$

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

$$I_{31} = \Delta_{11} e^{\Delta_{11} + \frac{1}{\lambda_{SR_n} \gamma_{thc}^{R_n}} \text{Ei}(-\Delta_{11}) - \Delta_9} e^{\Delta_9 - \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma_{thc}^{R_n}} \text{Ei}(-\Delta_9)}, \quad (B.7)$$

$$\begin{aligned}
I_{32} &= \frac{\gamma_{thc}^{R_n} \pi}{N \lambda_{TR_n} \lambda_{ST} \gamma_{\Delta_5^{R_n}}} \sum_{k=0}^N K_0 \left(2\sqrt{\Delta_{10}} \right) \sqrt{1 - \vartheta_k^2} \\
& \quad \times \left[e^{-\left(\varsigma_{R_n} B_{R_n} \Delta_{10} + \frac{\varsigma_{R_n}}{\lambda_{SR_n} \gamma_{thc}^{R_n}} \right)} - e^{-\frac{1}{\lambda_{SR_n} \gamma_{thc}^{R_n}} - \frac{\vartheta_{k+1}}{2 \lambda_{SR_n} \gamma_{thc}^{R_n}}} \right]. \quad (B.8)
\end{aligned}$$

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

$$\begin{aligned}
P_{int}^{T,ni} &= \int_0^\infty \int_{\frac{C_E \gamma_{thc}^E}{\Delta_5^E}}^\infty \int_{\frac{M_E \gamma_{thc}^E z + (\psi_E + 1/\gamma)^{\gamma_{thc}^E}}{\Delta_5^E z - C_E \gamma_{thc}^E}}^\infty \frac{1}{\lambda_{TE} \lambda_{ST}} e^{-\left(\frac{y}{\lambda_{TE}} + \frac{z}{\lambda_{ST}} \right)} dy dz \\
& \quad - \int_0^\infty \frac{C_E \gamma_{thc}^E}{\Delta_5^E}
\end{aligned}$$

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.



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, Forsell 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.





Evaluation of adversarial machine learning tools for securing AI systems

S. Asha¹ · P. Vinod²

Received: 15 November 2020 / Revised: 22 July 2021 / Accepted: 15 September 2021 / Published online: 29 September 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

Artificial intelligence aims to build intelligent systems capable of performing tasks that need human intelligence. Research works in recent years have revealed many potential vulnerabilities in machine learning algorithms. Precisely to exploit these vulnerabilities, an attacker may attempt to design an adversarial input to be incorrectly processed by machine learning algorithms. This paper focuses on methods of generating adversarial samples and discusses possible counter-measures. Our proposed method effectively checks the robustness of different machine learning models using six available hostile robustness tools and summarizes their current development state. The work compares the features of these tools, highlights similarities, and differences among the tools, their strength and weaknesses and trace connections among theoretical methods and their implementations. This paper will provide more insight for researchers and scientists to develop robust solutions and accelerate their experimentation.

Keywords Adversarial machine learning · Evasion attack · Poisoning attack · Adversarial robustness tools · Machine learning models

1 Introduction

Recent years have witnessed tremendous advances in the development of artificial intelligence (AI). These systems learn from experience concerning specific tasks. Based on the outcome, we can improve the performance of the learning. The advancements in machine learning and deep learning domains have tremendously accelerated the reliability of AI-powered systems. AI systems are used widely in many application domains; it has set itself as a prominent tool by achieving higher performance in forensic

analysis, anomaly detection, computer vision applications primarily in autonomous vehicles, etc.

As the machine learning models become more widely deployed, the spur for defeating them also increases. Research on machine learning models has reported that the most popular machine learning models are vulnerable to adversarial situations. AI systems can be confused by creating virtually undetectable alterations of an image, video, speech, and other data. Generally, an attacker causes the machine learning algorithms to lack the power to distinguish between “good” and “bad” under adversarial settings. The existence of adversaries has spawned several kinds of research in the field of adversarial machine learning (AML) on identifying the vulnerabilities in popular machine learning algorithms.

Adversarial machine learning is the area of study that focuses on identifying the vulnerabilities in machine learning models. An adversary injects slight perturbation into the input sample to increase the misclassification rate of the models. Such tainted samples are known as adversarial examples. The goal of an attacker is to poison the training data for compromising system security. Adversarial attacks can be classified broadly as *evasion attacks*

✉ P. Vinod
vinod.p@cusat.ac.in

S. Asha
ashas@scmsgroup.org

¹ Department of Computer Science & Engineering, SCMS School of Engineering & Technology (Affiliated to APJ Abdul Kalam Technological University, Trivandrum, India), Ernakulam, India

² Department of Computer Applications, Cochin University of Science & Technology, Cochin, India

and *poisoning attacks*. In the former, the attacker changes the behavior of machine learning models during the testing process, while in the latter, attacks are crafted by corrupting the training data.

This paper examines the various adversaries that affect the classifier performance and discusses the taxonomy of adversarial attacks. The article also aims to identify the tools which are robust enough to prevent attacks. Further, our work conducts comprehensive experiments using several open-source adversarial tools developed in Python. These libraries offer reference implementations of several standardized, state-of-the-art adversarial attacks, defenses and detections, robustness certifications, metrics, and formal verifications. The tools we have taken into consideration are *CleverHans* [1], *Adversarial Robustness Toolbox* [2], *Foolbox* [3], *AdvBox* [4], *DEEPSEC* [5] and *AdverTorch* [6].

The contributions of the article are listed below:

- To provide a comprehensive analysis of adversary attack strategies and their effectiveness in various machine learning models.
- To identify possible defenses against the adversarial samples. We aim to introduce defenses that would be simple and at least partially effective.
- To evaluate the characteristics of various AML tools on the benchmark datasets in the literature.

This paper presents an insight into robust AML tools. Section 2 presents a summary of prior research works conducted using AML tools. Section 3 gives a brief introduction to AML and the taxonomy of various adversarial attacks. Section 4 discusses the possible methods of adversarial attacks, and Sect. 5 elaborates the countermeasures. Section 6 brief on the several existing machine learning models used in literature. Section 7 provides a comprehensive analysis of AML tools. The next Sect. 8 discusses the attack effectiveness and attack perturbation rate of each AML tool on popular data sets. Finally, we conclude our study on various AML tools and sketches areas for future work in Sect. 9.

2 Background of AML tools

In this section, we introduce different tools used for generating adversarial samples. Here, we brief the state-of-the-art AML tools used in literature [1–6]. This section also explains the high-level features, architectural choices, implementation details, and potential applications of these AML tools.

Carlini et al. [7] critiques *DEEPSEC* and its official report [5]. It asserts that several attacks and defense implementations in this model are incorrect and the

reported results are flawed and misleading. A more detailed description of common flaws, and recommendations on how best to avoid them while evaluating adversarial robustness, is provided in [8]. It discusses the methodological foundations, reviews, and commonly accepted best practices. In addition, it suggests new methods for evaluating defenses to adversarial examples.

Various papers emphasize the support of robustness tools, but very few of them used more than one of such tools in their system. Brendel et al. [9] leverages *Foolbox* to evaluate the effectiveness of a novel class of gradient-based attacks and compare them to the state-of-the-art methods. Duesterwald et al. presents a practical approach in [10] that leverages hyperparameter optimization techniques for tuning adversarial training to maximize robustness while keeping the loss within a defined budget. They use *ART* in their experimentation. Machado et al. [11] provides an adversarial image detection method that arranges ensembles at runtime by randomly choosing multiple defense components. For their experimentations, they take advantage of *ART*. Chen et al. implemented and released a clustering method [12] through *ART* for detecting tainted samples.

The system involving multiple robustness tools are discussed below. Petrov et al. [13] employs *ART* and *Foolbox* to evaluate the transferability of adversarial examples. Fidel et al. leverage *ART* and *Foolbox* to investigate a connection between model explainability and adversarial examples in [14]. Serban et al. in [15] provide a characterisation of the phenomenon of adversarial examples, using *ART*, *CleverHans* and *Foolbox*. Duan et al. in [16] discusses the effectiveness of various robustness tools. It considers many methods and open source tools for crafting adversarial examples. However, it focuses on the discretization problem, which is less general compared to the work discussed in this system.

3 Adversarial machine learning

Adversarial machine learning focuses on the analysis of existing vulnerabilities in machine learning algorithms. Most of the state-of-the-art machine learning algorithms are profoundly unsafe to distorted input [13, 17–19]. An adversarial example is created by perturbing a sample input data causing a machine-learning algorithm to process it incorrectly. AML is an area of research that focuses on practical machine learning techniques against attacker tactics. In most cases, the attacks are targeted ones; the attacker has a specific reason to assault the system. The opponent can violate the integrity and availability to evade the system. In the former, malicious samples are wrongly

labeled as genuine, whereas, in the latter, the opponent tries to make the classification system unusable.

There exist two approaches to cause integrity and availability violation; one by introducing vulnerabilities through manipulating the training examples and the other by fabricating the test data so that the modified test data are misclassified as legitimate. Adversarial examples exhibit the property of *Transferability* [20], where the samples trained for one model often mislead the performances of another model, even if the models differ in architecture and training dataset.

AML can be mathematically formulated as: Given a training set (X, Y) , where X denotes a set of n dimensional input feature vector X_i ; and Y denotes the corresponding output class label. Each input vector is of the form $[x_1, x_2, \dots, x_n]$. For each X_i presented to the machine learning model C , the system has to correctly classify it to a class label Y_i , s.t. $C(X) : X_i \in X \rightarrow Y_i \in Y$, that correctly classify the given input X_i into its corresponding class label Y_i , i.e. $C(X_i) = Y_i$. Adversary model creates sample modified feature vectors X^* , which when subjected to the trained classifier, will get misclassified as $C(X_i^*) = Y_i; Y_i \in Y$ for some or all X_i^* .

Figure 1 illustrates the AML scenario. Typically, machine learning models have two phases of operations; the training phase and the testing phase. In the training phase, the given classifier learns to output the target class label. Once the training phase is over, an attacker create adversarial samples and evaluate the performance of the model. The attack is successful if the classifier misclassify the given perturbed data. Contrary to the above scenario, if the classifier correctly classify the adversarial input, then attacker craft an attack to the system by increasing the noise.

Adversarial attacks can be categorised based on the type of target class, the adversary desires:

- *Non-targeted attack* [21], causes the classifier to predict any incorrect label. This type of attack enables a machine learning model $C(X_i) = Y_i$ to generate samples X_i^* , such that $C(X_i^*)$ gives some $Y_j \in Y$, where $i \neq j$. Optimization-based approaches and fast gradient based approaches can be used to generate non targeted attack samples.
- *Targeted attack* [21], aims to increase the classifier's prediction to a targetted output label Y_i . Here the attacker generate samples $X_i^* \in X^*$, that makes the classifier to generate a particular class label such that $C(X_i^*)$ gives $Y_i \in Y$. Optimization-based approaches and gradient based approaches can be used to perform targeted attack.

Adversarial scenarios can be classified based on the amount of the information the opponent has about the classifier. Table 1 provides an overview of these attacks.

- *White-box* Here the adversary has got internal knowledge about the classifier network, including the type, architecture, weights associated with the connection and values of all parameters. E.g: Optimization-based methods [18], Fast gradient (sign) method (FGSM) [19], Iterative FGSM [26], Jacobian-based saliency map attack (JSMA) [27].
- *Black-box with probing* The attacker is completely ignorant about the system, but he can infer the model parameters. Sometimes the adversary may possess knowledge about the architecture but not of the network weights. Alternatively, the adversary may be entirely ignorant about the architecture. Also in certain cases

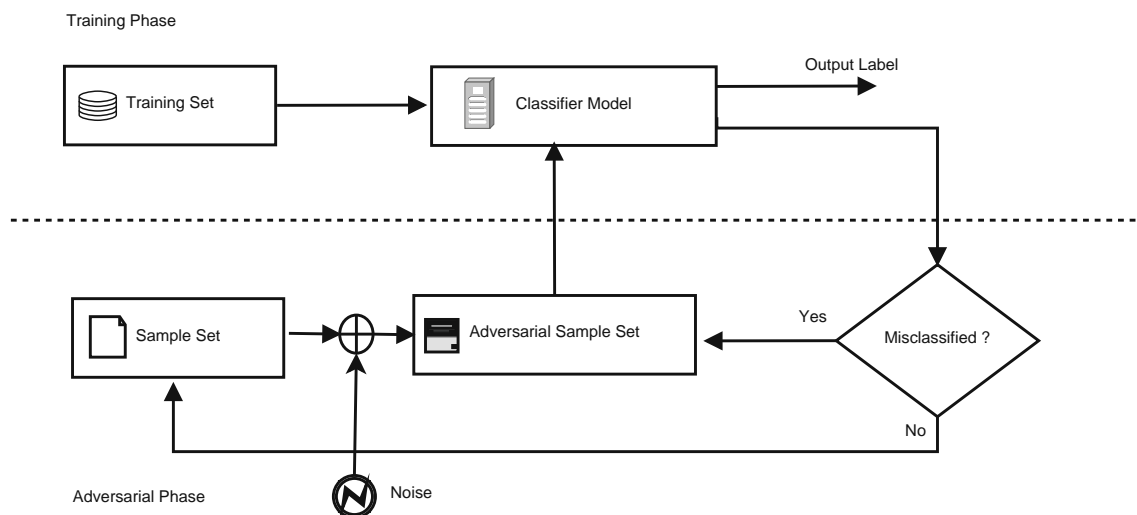


Fig. 1 Basic architecture of AML

Table 1 Comparison of various attacks

Attack category	Attack definition	Attack example	Defenses
White box attack	The adversary has got internal knowledge about the classifier network	Gradient based attacks [Sect. 4.1]	Defensive distillation [22], Ensemble based methods [11], Adversarial retraining [19]
Black box attack	No knowledge about classifier	Score based attacks [Sect. 4.2], Transfer based attacks [Sect. 4.3], Decision based attacks [Sect. 4.3],	Privacy preserving methods [23]
Grey box attack	Some knowledge about the classifier	Generative models [24]	Defense GANs [25]

the hacker can inspect the output probabilities of each class or the most likely class.

- *Black-box without probing* The opponent’s knowledge is limited, or sometimes the attacker doesn’t know anything about the model. Here, the attacker is not allowed to examine or question the model. In this case, the adversary needs to create general adversarial patterns that fool the machine learning network.

Based on the type of vulnerabilities imposed on the training data attack scenarios are classified as

- *Evasion attack* Also known as *exploratory attack* raises security breach by exploiting the available classifier knowledge. They aim is to exploit vulnerabilities during the classification phase and subsequently evade the security of classification. Studies carried out in [17, 28] crafts evasion attacks by manipulating the malicious sample during the testing phase to make them misclassified as benign by an already trained model, without altering the training samples. The attacker aims to violate the system’s integrity, either through a *targeted* or an *untargeted* attack.
- *Poisoning attack* This category of attack causes manipulation of training samples. Poisoning refers to a *causative attack*, here the training samples are manipulated by inserting specially crafted attack attributes thereby influencing the classifier decision boundary. Adversaries at the training stage can exert a long-lasting impact on learning. For example, essential data points may get poisoned, which makes the learning process highly complex.

Adversarial examples are proven to significantly impact a variety of real-world applications [29] that employ deep-learning algorithms. The models that get tricked include trained classifiers deployed in malware detection, speech recognition, facial recognition, semantic segmentation, and video processing. Adversarial examples can be blurry or over-exposed (or under-exposed) images from an image

acquisition device (camera or an image scanner) [30] or a mimicked voice [31] of a person by a third party (a person or a device). It is naturally possible for these inputs to become adversarial and harm the use case. For example, many artists around the world can mimic the sound of celebrities in a near-perfect way. Similarly, in the case of a driver-less or auto-pilot car, natural phenomena like rain, dust, snow, or even sun can cause the system to read the signboards differently and cause a misclassification. Placing carefully created stickers on signboard can also fool the sign detection network. In order to make the deep learning models re-silent to adversarial conditions is suggested to retrain the models by augmenting the training set with adversarial examples.

4 Adversarial attack methods

Adversarial attack methods can be categorized as *Gradient based*, *Score based*, *Decision based*, and *Neural model generated* attacks. A comparison of various attacks are given in Table 1. The meaning of various notations used in this article is provided in [Appendix](#).

4.1 Gradient-based attacks

Gradient-based attacks finds the linear loss L (e.g. cross-entropy), that can be applied on an input \mathbf{x} to find directions ϵ which makes the output of the network being sensitive for class y ,

$$\mathcal{L}(\mathbf{x} + \epsilon, y) \approx \mathcal{L}(\mathbf{x}, y) + \epsilon^T \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y) \quad (1)$$

where $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y)$ is the gradient of the loss w.r.t. to the input \mathbf{x} . The various attacks based on gradients are described below.

- *Fast gradient (sign) method (FG(S)M)* [19] computes the gradient $g(\mathbf{x}) = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, y)$ and obtains the minimum perturbation ϵ such that $\mathbf{x} + \epsilon^T g(\mathbf{x})$ is adversarial

- (FGM case), or $x + \epsilon^T \text{sign}(g(x))$ is adversarial (FGSM case).
- *Basic iterative method (BIM)* [26], iteratively obtains the direction of the sign of gradient. During each iteration, the system estimates a step size of α . Subsequently, the approach adds the direction of the gradient using α to the ϵ .
 - *Projected gradient descent (PGD)* [32] iteratively applies FGM or FGSM multiple times. Each iteration obtains a gradient score based on ϵ -norm ball on the original image and then projects back to the original ϵ .
 - *DeepFool* [33] is an untargeted adversarial methodology that uses the ℓ_2 norm. The approach consider an image x and a classifier C . The image x is iterated at each step and the perturbation γ_i at each step i is obtained. This γ_i is used to obtain the next image x_{i+1} . The process stops when the sign of x_{i+1} and x_i changes. Finally the total perturbations γ_i is computed as the final perturbation value γ^* to create the adversarial sample x^* as $x^* = x + \gamma^*$.
 - *Limited memory- Broyden-Fletcher-Goldfarb-Shanno (L-BFGS)* Szegedy et al. [18] generated adversarial examples using box-constrained L-BFGS algorithm. Given an input image x with a target class y , the method determines distinct image x^* identical to x under ℓ_2 distance. This x^* will be labeled differently by the classifier C . They model the problem as a second-order optimizer problem. The attacks produce x^* under the constraints $\text{dist} = \min(x - x^*)$ and $C(x + \text{dist}) = y$.
 - *Carlini–Wagner attack* [34] create adversarial examples based on optimization problem. Their approach employs small perturbation ϵ which is applied on the input x to increase misclassification. This method proposed a set of three adversarial attacks in the wake of defensive distillation. These attacks make the disturbances quasi-imperceptible by restricting their ℓ_2 , ℓ_∞ , and ℓ_0 norms. The defensive distillation mechanism for the targeted networks mostly fails against these three attacks.
 - *Jacobian-based saliency map attack (JSMA)* [27] uses the gradient to estimate a *saliency score* for each input feature (e.g. pixel). This saliency score indicates how strongly each feature can be perturbed, thereby deceiving the classifier. JSMA iteratively discovers the input feature with a higher saliency score for perturbation.
 - *Universal Adversarial Perturbations* [30] are untargeted iterative attacks. The approach attempt to find a constant disturbance (universal perturbation) iteratively. During each iteration, the perturbed value successfully increases the misclassification. After each iteration, it adds a constant perturbation to the ℓ_p ball

- with radius ϵ to control the intensity of the attack. The system calculates the universal fooling rate based on the effect of the perturbation on additional input samples.
- *Elastic net method (EAD)* [35] performs elastic-net regularization on the input sample to generate adversarial input.
- *Momentum iterative method (MIM)* [36] proposes another algorithm that accelerates the gradient descent algorithms by accumulating a velocity vector in the gradient direction of the loss function across iterations.
- *NewtonFool* [37] belongs to an untargeted attack category that tries to minimize the likelihood of the correct output label by manipulating the gradient.
- *Virtual adversarial method* [38] is a poisoning attack which intends to cause the local distributional smoothness of the trained model. It constructs a perturbation γ of ℓ_2 unit norm maximizing the *Kullback Leibler (KL)* divergence $KL[C(x)||C(x + \gamma)]$. The attack iteratively observe the perturbation strength using gradient ascent along with finite differences. Subsequently, the system constructs an adversarial sample by adding $\epsilon \cdot \gamma$ to the input x , where ϵ is the perturbation strength parameter.
- *Feature adversaries* [39] attack finds adversaries which disrupt both classification (*label adversaries*) and internal representations. This attack strategy constructs feature adversary that minimizes the euclidean distance between the deep internal representation (at a specified layer).

4.2 Score-based attacks

Score-based attacks use probability or logits values to attack the target system. The various attack to create adversarial samples under this category are:

- *Single pixel attack* [40] measures the robustness of a trained model by modifying a single pixel value. It is an iterative process, where the attacker toggles one pixel to either white or black. This process recurs for every pixel in the image. Su et al. [40] demonstrated that their approach was capable of fooling image classifiers by altering a single pixel.
- *Local search attack* [40] observes the probability of occurrence of the correct class, even after applying extreme perturbations to pixel values. The method identifies the pixel value that makes the model highly sensitive. The process repeats for all other pixels (neighborhood of the most sensitive pixel) until the classifier gets deceived by the image.
- *Simultaneous perturbation stochastic approximation (SPSA)* [41] approximates gradients using finite different estimates in random directions.

4.3 Decision-based attacks

Attacks based on the model's target class decision, rather than on gradients and probabilities, constitute the decision based attacks. The few decision-based attacks identified so far are:

- *Transfer based attacks* rely on information about training samples without the knowledge of the training model. The opponent uses this information to train a substitute model and to synthesize adversarial perturbations [20]. They showed that perturbed samples on an ensemble of substitute models could drop the classifier performance by 100%. An effective defense strategy against this transferability approach is the adversarial training method. This defense approach is proven successful against numerous attacks reported in the Kaggle Competition (during 2017) on Adversarial Attacks.
- *Boundary attack* [21] minimizes the ℓ_2 of adversarial perturbations. The algorithm begins from an adversarial point and then considers a random move to find a boundary point, that lies between the adversarial and the non-adversarial region. The newly identified boundary point must (i) lie in the adversarial region and (ii) have a minimum distance from the original image.
- *Pointwise attack* [42] minimizes the ℓ_0 of adversarial perturbations.
- *Additive uniform noise attack* disturbs the robustness of a model by introducing Independent and identically distributed (i.i.d) uniform noise.
- *Additive Gaussian noise attack* uses i.i.d normal noise.
- *Salt-and-pepper noise attack* probes using i.i.d salt-and-pepper noise.
- *Gaussian blur attack* attacks the model by performing Gaussian blur on the input.
- *Contrast reduction attack* examines the robustness of a model by adjusting the contrast value.

4.4 Neural model generated attacks

Various neural network models can be used to generate adversarial attacks. The most common among them is the generative adversarial networks [24] and recurrent neural network [43] based adversarial attacks.

- *Generative adversarial networks (GAN)* are proposed by Goodfellow et al. for the generation of adversarial image samples. GANs are deployed in numerous application domains like computer vision, natural language processing, and speech synthesis. GAN architecture uses two network models; a generative model and a discriminative model. Both the networks

follow a min-max game. The generator generates realistic samples from the given domain and tries to fool the discriminator model. The discriminator learns to distinguish between fake or real input samples. GANs are capable of generating adversarial samples by learning the distribution of real samples [44]. Xiao et al. [45] proposes the generation of adversarial samples using GANs. Zhao et al. in [46], proposes unsupervised adversarial attacks (UAA-GAN) to generate fake images and thereby fools the deep learning image retrieval system. GANs can generate realistic face samples retaining the identity of the original face called deepfakes that can fool a face recognition system. Nowadays, GANs are widely used to generate sentences (RankGAN [47]) and in audio processing systems for speech, music, and poetry generation (SeqGAN [48])

- *Recurrent neural networks (RNN)* [43] are neural models that are extremely good in handling both sequential and temporal data. This network model is capable of making accurate predictions based on the current input and previous inputs. RNNs are available in multitude types, notably vanilla RNNs, bidirectional RNNs, recursive RNNs, and long short-term memory RNNs (LSTM). RNN cells' intrinsic sequential processing capabilities aid in the creation of successful adversarial samples. Chang et al. in [49] proposes an audio perturbation sample for an Automatic Speaker Recognition system. The system uses LSTM as the RNN cell to craft perturbed audio examples. The system generates an efficient and effective adversarial audio by combining the RNNs pre-training and fine-tuning parameters. The system in [50] proposes a novel LSTM-CGAN method to craft high-quality vulnerable samples for blockchain-based wireless network detection models. A generic adversarial testing framework RNN-Test [51] uses RNN to generate vulnerable inputs. RNN-Test relies on RNN wrapper cells that extract the hidden states and cell states of each RNN cell in the given RNN model without affecting its inherent process. The extracted hidden state parameters help in generating adversarial examples.

5 Defenses and detectors

Methods for adversarial defenses can roughly be categorized as *model hardening*, *data preprocessing*, *adversarial detectors* of adversarial inputs and *Certifications*.

5.1 Model hardening

This defense strategy deals with the techniques to model a new classifier with better robustness when compared with the primary classifier based on given metrics.

- *Adversarial training* [19] enhances the classifier’s robustness metric by augmenting adversarial samples with the training data. This method proactively generates vulnerable inputs with the correct class targets and makes the model trained using these generated samples. Szegedy et al. [18] demonstrated a reasonably inexpensive way to create adversarial examples with the FGSM. Studies in [52] reported that this strategy enhances the robustness of the system to FGSM attacks, and it has multiple downsides: (i) It does not help against more sophisticated white-box attacks like RAND+FGSM. (ii) It does not help against black-box attacks either.
- *Defensive distillation* Nicholas Papernot et al. [22] proposed an efficient defensive mechanism for adversaries on deep neural network. Their solution utilizes an adversarial crafting framework to determine the direction by adding noise to increase misclassification. Based on this direction they applied perturbations to the training set. This work uses two networks an initial network and a distilled network. The initial system generates a probability vector representing predictions for each input. Subsequently, the distilled network takes the prediction probabilities as input and train the classifier. The distilled model mimics a computationally expensive large model. The distilled model is more robust to attacks comparing the FGSM or the JSMA approach.
- *Ensemble adversarial training* [11], the existence of vulnerabilities in an adversarially trained model against black-box attacks leads the authors to propose a highly secure defense strategy called *Ensemble Adversarial Training*. The model is trained using modified adversarial samples from an ensemble of more than two models.
- *Safe reinforcement learning* (Safe RL) is a defense model proposed in [53] to make reinforcement learning models defend the unseen attack scenarios. Reinforcement learning focuses on taking the best possible behavior or path to maximize reward in a given situation. Learned policies should be robust to uncertainty and parameter variation to ensure predictable behavior, which is essential for many practical applications of RL like robotics.

5.2 Data preprocessing

Input data preprocessing techniques transform input data and/or labels at training and/or test time.

- *Feature squeezing* [54] lessens the precision of the components of input samples by encoding that on a reduced number of bits.
- *Label smoothing* [55] alters the class labels during the learning phase. Label smoothing uses the following representation instead of using one-hot encoding :

$$y_i = \begin{cases} y_{max} & \text{if } y = i \\ \frac{(1 - y_{max})}{(K - 1)} & \text{otherwise} \end{cases} \quad (2)$$

- where $y_{max} \in (0, 1)$ and $i \in \{1, \dots, K\}$. The label representation is *smoothed* by reducing the variation between its highest and lowest components and thereby increasing its entropy. This approach helps to subdue the way an adversary could use the gradient to construct adversarial samples.
- *Spatial smoothing* [54] strives to refine out adversarial signals employing local spatial smoothing. It is a defense designed explicitly for images.
- *JPEG compression* [56, 57] divides images into square blocks and from each block removes the high-frequency signal elements. This operation is similar to performing a selective blurring of the image; it also helps to reduce additive perturbations.
- *Thermometer encoding*¹ discretizes the input, encodes each input to a fixed-size binary vector.
- *Gaussian data augmentation* [58] augments the original training dataset with copies of original samples that are perturbed by adding Gaussian noise.

5.3 Adversarial detectors

Methods for adversarial detection detect if the given data is adversarial or not. Adversarial detectors are classified into *poisoning* and *evasion* detectors, according to the stage they are applied, specifically at training/test time and inference time, respectively.

- *Poisoning detectors* [59, 60] check whether a given input provided at training or test time is adversarial or not and filters out such malicious data. The perturbed sample causes to alters the decision boundary of machine learning models, thereby causes an increased rate of misclassification and reduces the overall performance. For example, by augmenting malicious samples into the training set, adversaries can generate

¹ <https://openreview.net/forum?id=S18Su-CW>.

backdoors or *trojans* into the machine learning network [61]. This detection works well on targeted training and testing samples but acts upon specific attacker-chosen inputs. The *activation clustering* [12] defense method is also capable of identifying malicious data crafted to misclassify the trained machine learning models.

- *Evasion detectors* check whether a given input provided at inference time is adversarial or not. The *binary input detector* is a binary classifier whose output class labels indicate whether the given input is adversarial or not. The *binary activation detector* is a binary classifier where the *input* consists of the activation values produced by a layered classifier, and the output *labels* denotes whether the given sample is adversarial or not.

5.4 Robustness certifications

Robustness certifications are defenses provably robust to certain kinds of adversarial perturbations. On the other hand, a model is *certifiably robust* if, for any input x , one can obtain a guarantee that its prediction is constant within some set around x , often an ℓ_2 or an ℓ_∞ ball. An example of such a defense is *Randomized Smoothing*. *Randomized Smoothing* [62] is a provable adversarial defense in ℓ_2 – *norm* which scales to *ImageNet*. Specifically, it provides a tight robustness guarantee in ℓ_2 – *norm* for smoothing with Gaussian noise.

6 Machine learning models

This section gives an overview of popular Deep Learning Architectures: *ResNet*, *InceptionV3*, and *Madry*. The architecture of these models, the pros and cons of these networks are briefly discussed.

6.1 InceptionV3

The Inception net [63] achieved a milestone in convolutional neural network (CNN) classifiers. The Inception network is heavily engineered to push performance, both in terms of speed and accuracy. It is the winner of the ImageNet Large Scale Visual Recognition Competition in 2014.

The computational cost in deep convolutional networks is reduced drastically by introducing a 1×1 convolution. Inception net adds a new 1×1 convolution layer before the 3×3 and 5×5 convolutions. Inception net applies the 1×1 convolution after the max-pooling layer also. Towards the end, a global average pooling replaces the fully connected layers. The global average pooling

calculates the average of every feature map. GoogLeNet Architecture of Inception Network has 22 layers in total.

6.2 ResNet

Residual networks have become quite popular for image recognition and classification tasks as it is capable of solving vanishing and exploding gradients when adding more layers to an already existing deep neural network. Deep Residual Learning network [64] works based on the concept of Residual Learning. Residual learning means each layer in a neural network gets a fine-tuned output from the previous layer. Adding a learned “residual” to the input obtains fine tuned result. A residual network consists of residual units or blocks which have skip connections, also called identity connections. A residual block has a 3×3 convolution layer followed by a batch normalization layer and a ReLU activation function. The skip connection skips both these layers and adds directly before the ReLU activation function. Such residual blocks repeatedly constitute a residual network. The hop or skip could be 1, 2, or even 3. While adding, the dimensions of input may be different than the output due to the convolution process, resulting in a reduction in dimensions. Thus, a 1×1 convolution layer is added to change the dimensions of input.

6.3 Madry

Aleksander Madry in [32] proposed a deep learning neural network with significantly improved strength to a wide range of adversarial attacks. The neural model utilizes the projected gradient descent (PGD) as a universal “first-order adversary”. It uses the local first-order information about the network. They have experimentally identified that for typically trained networks and adversarially trained networks, the local maxima obtained by PGD gives similar loss values. The experimental result also demonstrated that a system is robust against PGD adversaries, also resilient to a wide range of other attacks.

7 Comprehensive analysis of tools

This section gives a detailed view on various attacks and defenses implemented in the six machine learning tools, namely *CleverHans* [1], *ART* [2], *Foolbox* [3], *Adv-Box* [4], *DEEPSEC* [5], and *AdverTorch* [6]. One of the most commonly used AML libraries is *ART*. It is an open-source python library. The *ART* architecture design makes it simple to couple defenses, e.g. adversarial training with data preprocessing and runtime detection of adversarial inputs. Nicolas Papernot and Ian Goodfellow in [1] developed a library package ‘Cleverhans’, which provides

the attackers and defenders with standard implementations of adversary construction and training scenarios.

AdverTorch Library [6] provides a wide variety of dynamic computational graphs. It provides simple and consistent attack and defense APIs. It facilitates faster execution of commands with the help of GPU-powered PyTorch implementations. *Foolbox* [3] is a python library that helps researchers to generate adversarial samples for machine learning models. It provides features that help to analyze the robustness of existing machine learning networks. *DEEPSEC* [7] is a uniform platform for security analysis of deep learning models. This library enables researchers and practitioners to estimate the vulnerability of deep learning models. *AdvBox* [4] is a toolbox to generate adversarial examples that fools neural networks in PyTorch, PaddlePaddle, MxNet, Caffe, Keras, and TensorFlow. Like any other AML tool, *AdvBox* also provides mechanisms to evaluate and benchmark the robustness of various machine learning models. *AdvBox* gives a command-line interface to create perturbed patterns with Zero-Coding.

All six tools under analysis support at least one machine learning library, though four support more than one of such libraries. Table 2 shows the ML libraries supported by the six tools under consideration. *ART* is compatible with 11 ML libraries. *Foolbox* and *AdvBox* supports 6 ML libraries each. *CleverHans* is currently fully compatible with two ML libraries Tensorflow and Keras. In contrast, *AdverTorch* and *DEEPSEC* are based on *PyTorch*, and it seems they have not planned an extension of their compatibility to other ML libraries.

All the above tools offer a considerable amount of attack methods. Table 3 gives an overview of various adversarial attacks incorporated in the six AML tools. *Foolbox* supports 29 attacks including FGSM, C&W, BIM, and DeepFool. *AdvBox* supports the least number of attacks; only 11 adversarial attacks. *ART* supports 19 adversarial attacks; *CleverHans*, *AdverTorch*, *DEEPSEC* implements 16, 14, 13 attacks respectively. It is interesting to highlight that all the six ML robustness tools implement FGSM, C&W, BIM, PGD, JSMA, DeepFool, elastic net, and L-BFGS.

Table 4 gives the details of various defenses and detectors supported by *ART*, *DEEPSEC*, *AdverTorch*, *AdvBox*, *CleverHans* and *Foolbox*. From Table 4, it can be inferred that five tools out of six implement at least one form of defense in opposition to adversarial attacks. *ART* provides nine defense methods and six detector methods supporting a total of 15. *Foolbox* supports no defense mechanisms. *CleverHans* supports only *Adversarial Training* strategy. *DEEPSEC* implements 10, *Adver Torch* provides 7, and *AdvBox* implements 5 defense strategies. The only defense strategy that seems to work well in literature is *Adversarial Training*.

Figure 2 gives a visual comparison of the features considered for all the six machine learning tools. For each tool, it plots the number of supported attacks, number of implemented defenses, number of ML libraries supported, and the metric value. Figure 2, depicts that *ART* has the broader compatibility to ML libraries, and it supports almost double the number for every parameter that is taken into consideration. It is the only tool that provides inbuilt

Table 2 Compatibility of AML robustness tools

Libraries	Tools					
	ART v1.0	Fool box v2.0	Adv Box	CleverHans v3.0.1	AdverTorch v0.1	DeepSec
Tensorflow 1	✓	✓	✓	✓		
Tensorflow 2	✓					
Keras	✓	✓	✓	✓		
PyTorch	✓	✓	✓		✓	✓
Theano		✓				
Caffe2			✓			
MXNet	✓	✓	✓			
Lasagne		✓				
Scikit-learn	✓					
XGBoost	✓					
LightGBM	✓					
CatBoost	✓					
PaddlePaddle			✓			
GPy	✓					
Tesseract	✓					
Total	11	6	6	2	1	1

Table 3 Adversarial ML attacks available in robustness tools

Attacks	Tools					
	Foolbox	ART	CleverHans	AdverTorch	DeepSec	AdvBox
Fast gradient method (FGM)	✓	✓	✓	✓	✓	✓
Carlini–Wagner (C&W)	✓	✓	✓	✓	✓	✓
Basic iterative method (BIM)	✓	✓	✓	✓	✓	✓
Projected gradient descent (PGD)	✓	✓	✓	✓	✓	✓
Jacobian saliency map (JSMA)	✓	✓	✓	✓	✓	✓
DeepFool	✓	✓	✓	✓	✓	✓
Elastic net	✓	✓	✓	✓	✓	✓
L-BFGS	✓		✓	✓	✓	✓
Momentum iterative method	✓		✓	✓	✓	✓
Feature adversaries			✓	✓		
Simultaneous perturbation stochastic approximation (SPSA)			✓			
Universal adversarial perturbation		✓			✓	
NewtonFool	✓	✓				
Virtual adversarial method		✓	✓			
Query-efficient black-box		✓				
Zeroth-order optimization (ZOO)		✓				
Adversarial patch		✓				
Semantic			✓			
Max confidence			✓			
ADef	✓					
SLSQ	✓					
Decoupled direction and rm	✓			✓		
Sparse Fool	✓					
Sparse descent				✓		
Least-likely class method (LLCM)					✓	
Iterative LLCM					✓	✓
Single pixel	✓			✓		✓
LocalSearch	✓			✓		✓
Approximate LBFGS	✓					
Boundary	✓	✓				
OptMargin					✓	
Spatial transform	✓	✓	✓	✓		
Poinwise	✓					
Noise			✓			
Gaussian blur	✓					
Contrast reduction	✓					
Additive uniform noise	✓					
Additive Gaussian noise	✓					
Salt and pepper noise	✓					
Blended Uniform Noise	✓					
HopSkipJump	✓	✓				
Decision tree attack		✓				
Poisoning attack						✓
Poisoning attack on SVM		✓				
Binarization refinement	✓					

Table 3 (continued)

Attacks	Tools					
	Foolbox	ART	CleverHans	AdverTorch	DeepSec	AdvBox
Precomputed images	✓					
High confidence low uncertainty		✓				
Total	29	19	16	14	13	11

robustness metrics and verification. *ART* implements precisely three robustness metrics; empirical robustness (EC) metric [33], loss sensitivity (LS)metric [65], Cross Lipschitz Extreme Value for nEetwork Robustness (CLEVER) [66] and one verification, Clique Method Robustness Verification [67].

It is essential to point out that nevertheless, *DEEPSEC* supplies ten defenses, several of them, like *Defensive Distillation*, have already been proved as not robust in the literature. Also, some of its defense implementations, like *Adversarial Training*, have been proven to be effective only against simple cases like *Single Pixel Attack*. In contrast, *CleverHans* provides only one type of defense, *Adversarial Training*, but its applicability is very general and flexible. It is effective against vulnerable examples generated with any attack methodologies. The *Foolbox* does not provide any form of protection against adversarial attacks. It is designed with the unique purpose to provide a wide range of methods to test the model's robustness.

It is observed that *ART* has reasonable value for all the four metrics under consideration. Its divergence from the other tools is probably because *ART* belongs to a more significant project for AI trusting. Figure 3 shows the usage statistics of the six AML tools in research studies. The value is obtained based on the citation count of the respective research paper². It shows that, among the six tools, the most commonly used tool for research studies are *ART*(21.1%), *Foolbox*(32.5%) and *CleverHans*(26.3%).

8 Results and discussions

This section describes the benchmark datasets used in the experiments and the metrics used to evaluate the attack effectiveness and perturbations. It compares various AML tools used in the literature and summarizes the experimental outcomes.

8.1 Datasets

For evaluating the robustness of AML tools, the experiment uses three datasets; *ImageNet* [68], *Cifar-10* [69] and *Modified National Institute of Standards and Technology database(MNIST)*³. *ImageNet* contains a total of 15 million labeled high-resolution images categorized with around 20,000 labels. In this system, the *Inception V3* machine learning model gets evaluated using the *ImageNet* dataset. *Cifar-10* consists of ten image categories with 6000 photos each, making a total of 60,000 images. *ResNet* Classifier model is assessed using the *Cifar-10* dataset. *MNIST* is an image data store of handwritten digits that consists of 70,000 28 x 28 grayscale images of numbers 0–9. For the *Madry* machine learning network we use the *MNIST* dataset. From each dataset, the system chooses 100 samples at random and ensures that the same sample set is used for all the tests performed. To evaluate the effectiveness of attacks, for each dataset, two state-of-the-art models were selected; one adversarially trained and another traditionally trained. In particular, for the *ImageNet* dataset, two *Inception V3* models were used, two *ResNet* models for *Cifar-10* dataset and two *Madry* models for *MNIST*.

From Table 5, considering *ResNet* machine learning model, *ART* and *Foolbox* implements six attack strategies each, taking a sample size of 600. Since *CleverHans* does not support C&W attack, only five attack strategies are employed. Hence, *CleverHans* takes only 500 samples. Similarly, in *ImageNet* dataset, for *Inception V3* evaluation, *ART* and *Foolbox* employs six attack strategies, taking a total sample size of 600. Since *CleverHans* does not support C&W and *Saliency Map* attack on *ImageNet*, only four attack strategies are used, taking a total of 400 samples. For *Madry*, *ART* and *Foolbox* employ six attack strategies, taking a total sample size of 600. *CleverHans* employs only five attack strategies, so the sample size is 500.

² <https://www.scholar.google.com/>.

³ <http://yann.lecun.com/exdb/mnist/>.

Table 4 Adversarial ML defenses and detectors available in robustness tools

Defenses	Tools					
	ART	DEEPSEC	AdverTorch	AdvBox	CleverHans	Foolbox
Adversarial training	✓				✓	
Naive adversarial training		✓				
Ensemble adversarial training		✓				
PGD-based adversarial training		✓				
Virtual adversarial training	✓					
Bit squeezing			✓			
Conv smoothing 2D			✓			
Average smoothing 2D			✓			
Gaussian smoothing 2D			✓			
Median smoothing 2D			✓			
Label smoothing	✓			✓		
Gaussian data augmentation	✓			✓		
Thermometer encoding	✓	✓		✓		
Total variance minimization	✓					
JPEG compression	✓					
JPEG filter			✓			
Pixel defend	✓	✓				
Binary filter			✓			
Region-based classification		✓				
Random transformations based defense		✓				
Ensemble input transformation		✓				
Defensive Distillation		✓				
Input gradient regularization		✓				
Randomized smoothing	✓					
Based on inputs	✓					
Trained on the activations of a specific layer	✓					
Fast generalized subset scan	✓					
Local intrinsic dimensionality based detector						
Spatial smoothing	✓			✓		
Feature squeezing detector	✓			✓		
Based on activations analysis	✓					
Total	15	10	7	5	1	0

8.2 Robustness evaluation metrics

Robustness metrics evaluate the effectiveness of machine learning models. It measures how weak a classifier behaves under various attack scenarios and its efficacy to different defensive mechanisms. They also quantify the required perturbation rate that causes misclassification. It evaluates the model's sensitivity rate concerning changes in their inputs. Possible such metrics are *EC*, *LS* and *CLEVER*. The metrics suggested below considers $C(x)$ as the classifier, $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ as the test set, and y as the target class.

8.2.1 Empirical robustness

The *EC* metric [33] measures the average minimum perturbation that causes a misclassification. The robustness of a classifier is evaluated based on a specific attack on a test dataset. *EC* is formally defined as

$$EC(C, \rho, \mathbf{X}) = \frac{1}{|I|} \sum_{i \in I} \frac{\|\rho(\mathbf{x}_i) - \mathbf{x}_i\|_p}{\|\mathbf{x}_i\|_p}, \quad (3)$$

$$\rho(x) = \frac{\Delta(x)}{\|x\|_p} \quad (4)$$

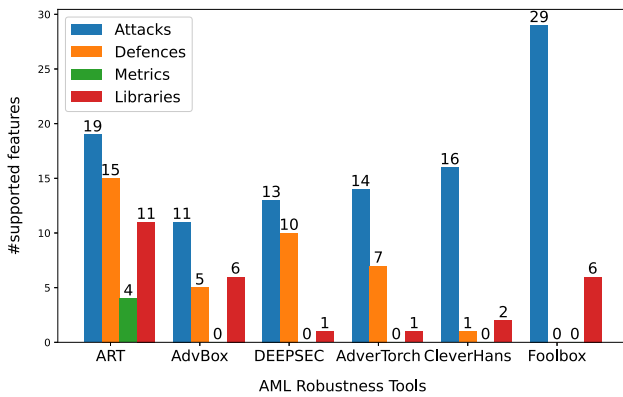


Fig. 2 Robustness evaluation of ML tools based on different metric such as attacks,defences,robustness evaluation metrics, and ML libraries

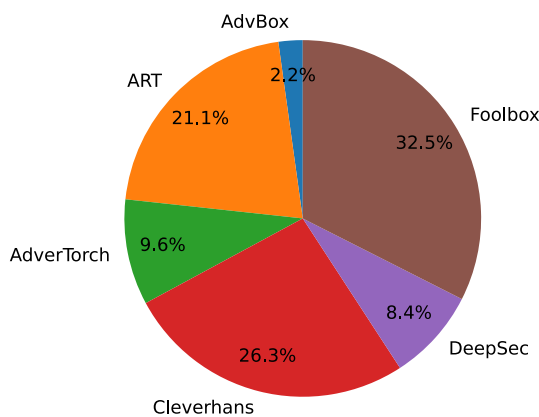


Fig. 3 AML tools usage statistics

where $C(\rho(\mathbf{x})) \neq C(\mathbf{x})$, p is the norm used in the creation of the adversarial samples.

8.2.2 Loss sensitivity

The *LS* metric [65] evaluates the sensitivity of a classifier model based on gradients. This metric is based on the properties of the model rather than any specific attacks. It measures the perturbation by making variations in the test data. It quantifies the model’s robustness by using *Lipschitz Constant*.

Loss sensitivity is formally defined as

$$LS(C, \mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \|\nabla \mathcal{L}(x_i, y_i)\|_2, \tag{5}$$

where $\nabla \mathcal{L}(\mathbf{x}, \mathbf{y})$ is the gradient of the loss w.r.t. to the input

8.2.3 CLEVER

The *CLEVER* metric [66] estimates the minimal perturbation γ that is required to change the classification of an

input \mathbf{x} , i.e. $\|\mathbf{x} - \mathbf{x}'\|_p < \gamma$ implies $C(\mathbf{x}) = C(\mathbf{x}')$. The perturbation factor γ is obtained based on the *Lipschitz Constant*. This metric is attack-agnostic and applicable to any neural network.

The *CLEVER* score for *targeted* attacks is formulated as

$$CLEVER(C, \mathbf{x}, y, N_b, N_s, p, R) = \min \frac{C(\mathbf{x}) - C_y(\mathbf{x})}{MLE_{Weibull}(S)}, \tag{6}$$

where $C(\mathbf{x})$ is a K -class classifier, $MLE_{Weibull}$ is the maximum likelihood estimate and

$$S = \bigcup_{i=1}^{N_b} \max_{j=1, \dots, N_s} \|\nabla g(\mathbf{x}^{(i,j)})\|_q, \tag{7}$$

where $\mathbf{x}^{(i,j)} \in B_p(\mathbf{x}_0, R)$ is randomly selected, N_b and N_s are

respectively the number of batches and samples per batch, R is the maximum perturbation, y is the target class and $q = \frac{p}{p-1}$ where p is a perturbation norm. On the other hand, the least *CLEVER* score for targeted attacks over all classes $y \in Y$ with $y \neq C(\mathbf{x})$ gives the *CLEVER* score for *untargeted* attacks.

8.3 Experimental outcome

For our experimental setup, we choose three AML Tools *ART*, *CleverHans*, and *Foolbox*. We considered the number of attacks supported by each tool as the criteria for evaluating the robustness of classifier models. From Fig. 2, *Foolbox* tops in the list with 29 attacks followed by *ART* and *CleverHans* with 19 and 16 attacks respectively. Among them, *ART* supports a reasonable number of defense strategies, libraries, and evaluation metrics as listed in Tables 3 and 4. The effectiveness of these three tools in generating adversarial samples is evident in their usage statistics as depicted in Fig. 3.

We analyzed the performance of common attacks on the selected AML tools using popular datasets *ImageNet*, *Cifar-10*, and *MNIST*. For each experiment, three relevant robustness tools: v1.0.0 of *ART*, v2.0.0 of *Foolbox*, and v3.0.1 of *CleverHans* were selected. For each tool, the system tested the following six attacks: *FGSM*, *DeepFool*, *C&W*, *PGD*, *JSMA*, and *BIM*. The system performed only untargeted attacks.

The experiment to perform the analysis of machine learning tools on different datasets is structured as follows. For each experiment,

- Select the data set, the attack, the model, and the AML tool to be evaluated.
- Perturb the images from the selected dataset with the selected attack for the selected tool.
- Classify the perturbed images using the selected model.

Table 5 Adversarial ML robustness tool effectiveness

Dataset	Model	Adversarial model	Tool	#Attacks	#Samples	Attack effectiveness
CIFAR10	ResNet	-	ART	6	600	0.87
CIFAR10	ResNet	-	CleverHans	5	500	0.92
CIFAR10	ResNet	-	Foolbox	6	600	0.69
CIFAR10	ResNet	✓	ART	6	600	0.15
CIFAR10	ResNet	✓	CleverHans	5	500	0.46
CIFAR10	ResNet	✓	Foolbox	6	600	0.13
ImageNet	Inception V3	-	ART	6	600	0.76
		-	CleverHans	4	400	0.95
ImageNet	Inception V3	-	Foolbox	6	600	0.85
ImageNet	Inception V3	✓	ART	6	600	0.23
ImageNet	Inception V3	✓	CleverHans	4	400	0.75
ImageNet	Inception V3	✓	Foolbox	6	600	0.09
MNIST	Madry	-	ART	6	600	0.62
MNIST	Madry	-	CleverHans	5	500	0.71
MNIST	Madry	-	Foolbox	6	600	0.50
MNIST	Madry	✓	ART	6	600	0.82
MNIST	Madry	✓	CleverHans	5	500	0.71
MNIST	Madry	✓	Foolbox	6	600	0.82

The overall performance analysis is carried out for every combination of tools, attacks, datasets, and models. On each dataset, six attack strategies were applied to original images and adversarial images to calculate the attack perturbation and attack effectiveness for each combination.

Table 6 details the attack perturbation and attack effectiveness of each tool under adversarial and non-adversarial training on various datasets for six attack strategies. In particular, two *Inception V3* models were chosen for *ImageNet* dataset, two *ResNet* models selected for *Cifar-10*, and two *Madry* models for *MNIST*.

Table 5 aggregates the average value of attack perturbation and attack effectiveness on *ART*, *CleverHans* and *Foolbox* on both adversarial and non adversarial environment. For e.g. the *ART* on *MNIST* dataset under normal training obtains an attack effectiveness of 0.62 as the average of 0.86, 0.76, 0.27, 0.41, 0.86 and 0.55. The table also highlights the number of samples taken and the number of attacks implemented.

On *Cifar-10* dataset, *CleverHans* gives the best attack effectiveness on both normal and adversarial classification. Similarly, for *ImageNet* *CleverHans* gives the best attack effectiveness of 0.95 on normal classification and 0.75 on adversarial classification. On the other hand, in the case of *MNIST* dataset, *CleverHans* gives the best attack effectiveness of 0.71 on normal classification whereas *ART* and *Foolbox* outperform it in the adversarial setting by each exhibiting attack effectiveness of 0.82 against 0.71

obtained by *CleverHans*. The attack accuracy of an AML tool in a specific dataset is calculated by averaging the attack effectiveness of the AML tool under both adversarial and non-adversarial training in all the six attack scenarios which are consolidated in Table 5. E.g., the attack accuracy for *ART* in *MNIST* is calculated as the average of 0.62 and 0.82 and is 0.72.

Figure 4 shows the attack accuracy per dataset of all the AML tools. It depicts that, for *MNIST* the attack accuracy of all tools is similar. This congruity could be because the *MNIST* dataset is relatively simpler than the others as it constitutes only grayscale images. *CleverHans* tops in the attack accuracy and exhibits consistent performance on all three datasets. It performs the best on the *ImageNet* dataset. *Cifar-10* accuracy is the least observed attack effectiveness on all machine learning tools. The performance across all the datasets is aggregated to obtain the overall attack accuracy of an AML tool. Figure 5 shows the aggregated value of *CleverHans*, *ART* and *Foolbox*. From Fig. 5, it can be observed that the average attack accuracy of *CleverHans* is higher than *ART* and *Foolbox*.

One of the reasons for the higher attack effectiveness of *CleverHans* is its higher attack perturbation rate. Figure 6 show the average attack perturbation rate of each AML Tool. The ratio of the number of altered pixels to the total number of pixels in the given image defines the perturbation rate. On average, attacks performed by *CleverHans* produce higher perturbation when compared to the attacks

Table 6 Adversarial ML robustness tool attack effectiveness

Attack	Adversarial model	Tool	Resnet(CIFAR10)		InceptionV3(ImageNet)		Madry(MNIST)	
			Attack perturbation	Attack effectiveness	Attack perturbation	Attack effectiveness	Attack perturbation	Attack effectiveness
BIM	-	ART	1e-2	0.96	1e-2	0.98	1e-2	0.86
BIM	-	CleverHans	4e-2	0.96	5e-2	0.97	3e-1	0.68
BIM	-	Foolbox	2e-5	0.36	8e-6	0.91	2e-3	0.52
BIM	✓	ART	1e-2	0.23	1e-2	0.47	1e-2	0.80
BIM	✓	CleverHans	4e-2	0.45	5e-2	0.75	3e-1	0.94
BIM	✓	Foolbox	2e-5	0.11	8e-6	0.08	2e-3	0.82
C&W	-	ART	2e-5	0.96	8e-6	0.94	4e-3	0.76
C&W	-	Foolbox	1e-5	0.83	8e-6	0.94	1e-3	0.30
C&W	✓	ART	2e-5	0.11	8e-6	0.08	4e-3	0.83
C&W	✓	Foolbox	1e-5	0.11	8e-6	0.06	1e-3	0.82
DeepFool	-	ART	2e-5	0.80	8e-6	12	2e-3	0.27
DeepFool	-	CleverHans	3e-2	0.83	5e-2	0.97	2e-1	0.70
DeepFool	-	Foolbox	2e-5	0.87	8e-6	0.94	2e-3	0.40
DeepFool	✓	ART	2e-5	0.11	8e-6	0.07	2e-3	0.82
DeepFool	✓	CleverHans	3e-2	0.29	5e-2	0.36	2e-1	0.03
DeepFool	✓	Foolbox	2e-5	0.11	8e-6	0.08	2e-3	0.82
FGM	-	ART	2e-3	0.90	2e-3	0.73	1e-3	0.41
FGM	-	CleverHans	9e-2	0.90	1e-1	0.88	4e-1	0.79
FGM	-	Foolbox	3e-2	0.84	4e-3	0.91	4e-2	0.64
FGM	✓	ART	2e-3	0.13	2e-3	0.23	1e-3	0.82
FGM	✓	CleverHans	9e-2	0.84	1e-1	0.99	4e-1	0.93
FGM	✓	Foolbox	3e-2	0.21	4e-3	0.18	4e-2	0.83
PGD	-	ART	1e-2	0.96	1e-2	0.98	1e-2	0.86
PGD	-	CleverHans	5e-2	0.98	7e-2	0.96	4e-1	0.73
PGD	-	Foolbox	2e-5	0.37	8e-6	0.90	2e-3	0.43
PGD	✓	ART	1e-2	0.23	1e-2	0.46	1e-2	0.80
PGD	✓	CleverHans	5e-2	0.52	7e-2	0.88	4e-1	0.85
PGD	✓	Foolbox	2e-5	0.11	8e-6	0.07	2e-3	0.82
Saliency Map	-	ART	2e-4	0.66	1e-4	0.41	1e-2	0.55
Saliency Map	-	CleverHans	3e-3	0.94	-	-	1e-2	0.66
Saliency Map	-	Foolbox	2e-4	0.89	1e-5	0.51	5e-3	0.71
Saliency Map	✓	ART	2e-4	0.11	1e-4	0.09	1e-2	0.85
Saliency Map	✓	CleverHans	3e-3	0.19	-	-	1e-2	0.82
Saliency Map	✓	Foolbox	2e-4	0.11	1e-5	0.05	5e-3	0.83

generated by the other tools. *ART* and *Foolbox* employ a batch processing strategy for crafting the adversarial sample. Consequently, only a portion of the pixels in the original input is modified, resulting in a low perturbation rate. While in *CleverHans*, the obtained sign of the gradient is applied to the entire image causing a higher perturbation

rate, which finally yields a higher attack accuracy. Specifically, the average perturbation of *CleverHans* attacks is two orders of magnitude higher than the average perturbation of the other two tools.

We conducted a subjective assessment (manual analysis) of the perturbed images generated by these three

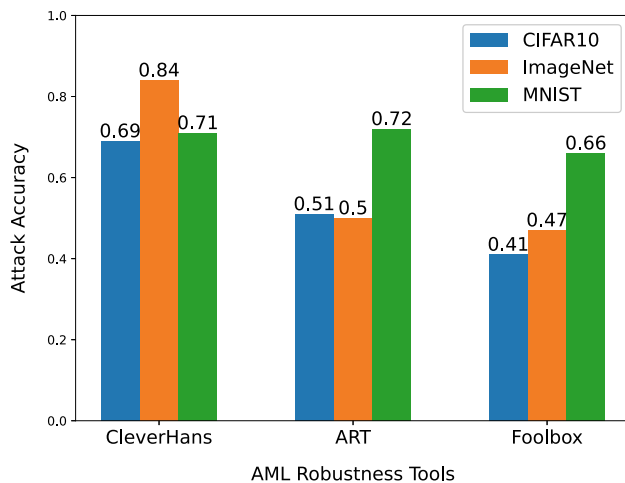


Fig. 4 Performance analysis of AML tools on *Cifar-10*, *ImageNet*, and *MNIST* datasets

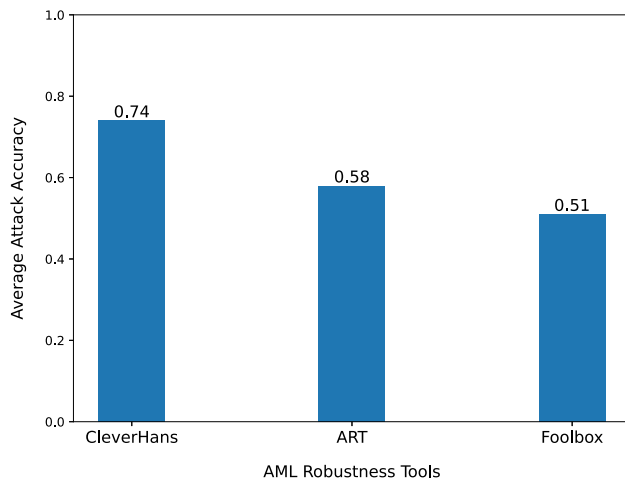


Fig. 5 Overall attack accuracy of AML tools

AML tools. It measures the effectiveness in performing the attack keeping the perturbed image visually closer to the original input. Subjective classification of the attacked image is one of the qualities exhibited by an adversarial attack. Figure 7 shows the perturbed images generated using FGSM attack with an epsilon value of 0.1 in *CleverHans*, *ART*, and *Foolbox* tools. For *ImageNet* and *MNIST*, the perturbed images are perceptually similar to the original image. In the *Cifar-10* dataset, all the three tools generated distorted images when compared to the original image. Nevertheless, the image crafted by *CleverHans* exhibits a better visual perception compared to *Foolbox* and *ART*.

Several objective evaluation methods exist to assess the visual quality of images. We compared the visual quality of perturbed images with the original image using three metrics: Peak Signal to Noise Ratio (PSNR), Mean Squared Error (MSE), and Structural Similarity Index

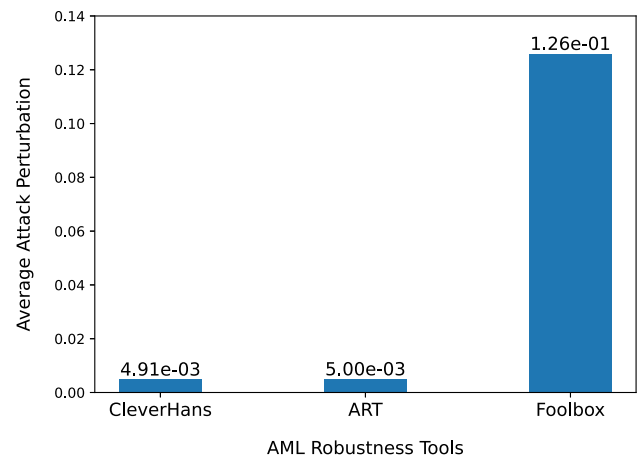


Fig. 6 Overall attack perturbation of AML tools

Measure(SSIM). Figures 7 and 8 summarizes the average MSE, PSNR, and SSIM value obtained by the adversarial samples generated by *CleverHans*, *Foolbox*, and *ART* using images from *ImageNet*, *Cifar-10*, and *MNIST* datasets. Figures 7 and 8 depicts that *CleverHans* achieves a higher PSNR value in both *ImageNet* and *Cifar-10*. In the *MNIST* dataset, all the toolboxes produced images with comparable PSNR rates, with *ART* having a value of 70.62dB. The MSE value is almost null for *CleverHans* in *ImageNet* dataset. From Figs. 7 and 8, the MSE values of *CleverHans* is less compared to the MSE values of other two tools. Taking the SSIM metric as in Fig. 8, *CleverHans* outperform other tools with a higher similarity index of 0.703.

Table 7 summarizes all the three metrics of AML tools across all the three datasets. From Table 7, it is clear that *CleverHans* gives a higher PSNR value of 78.84dB, which indicates the signal-to-noise ratio of perturbed images is higher in *CleverHans*. The squared error value is also on the lower side for *CleverHans*. The similarity of perturbed image and original image has an overall average of 0.703 in *CleverHans*, which is reasonably higher than the other two AML tools. From this study, we infer that even though the attack perturbation is high in *CleverHans*, it performs the best in subjective and objective evaluation. *CleverHans* produce perturbed images that are visually similar to the original input and at the same time give a higher attack accuracy during classification.

9 Conclusion

This paper analyzes and compares three major AML tools *ART*, *CleverHans*, and *Foolbox* from various perspectives. We observe the attack effectiveness of these tools using

Fig. 7 Comparison of original image from *ImageNet*(first row), *MNIST*(middle row) and *Cifar-10*(last row) against FGSM attack on *CleverHans*, *ART* and *Foolbox* Tools

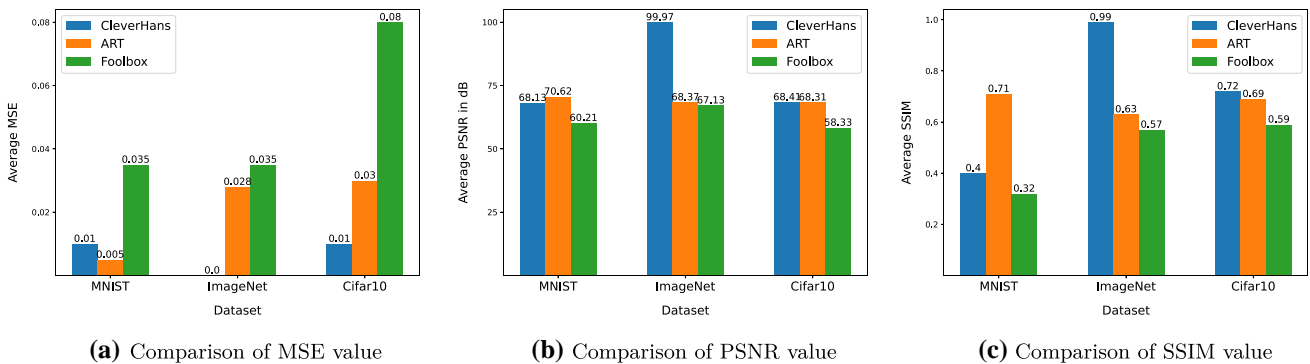


Fig. 8 Comparison of AML tools using MSE, PSNR and SSIM metrics across different datasets

multiple adversarial samples. Each of these tools is evaluated with six adversarial attacks under different datasets, and different machine learning models were evaluated for effectiveness under attack. The machine learning models used are *ResNet*, *InceptionV3* and *Madry*. For *ResNet*, the *Cifar-10* dataset is utilized. For *InceptionV3*, the *ImageNet* dataset is employed, and for *Madry*, the *MNIST* dataset is used. *Foolbox* provides the maximum attack

implementation by supporting 29 attack scenarios. *ART* gives the highest defense strategy implementation, 15 defense mechanisms are supported in *ART*. The attack effectiveness seems to be higher for *CleverHans* tool under all attack conditions. *CleverHans* with *InceptionV3* model gives the highest efficiency of 0.95. The overall attack accuracy and perturbation rate of *CleverHans* are 0.74 and 0.126 respectively. The perturbed image produced by

Table 7 Comparison of AML tools based on average PSNR, MSE and SSIM values

Metric	CleverHans	ART	Foolbox
PSNR	78.84dB	69.1dB	61.89dB
MSE	0.007	0.021	0.05
SSIM	0.703	0.68	0.493

CleverHans gives a better subjective perception when compared to the ones from the other tools. The perturbed output by *CleverHans* gives an average PSNR rate of 78.84 dB, average MSE rate of 0.007, and average SSIM value of 0.703.

Through this work, we feel that ML security researchers/practitioners will be benefited, as this paper is complete by comprehensively discussing the diversity of adversarial attacks, defenses, and metrics, which will help the target group to build security-aware ML models. The paper also gives a detailed taxonomy of different types of tools. We feel that this work can be extended in the future by examining other AML tools also discussing attacks and strategies incorporated in diverse domains such as text, audio, and network traffic.

Appendix

List of notations used in the paper

Symbol	Meaning
X	Training set $[x_1, x_2, \dots, x_n]$ of images
x_i	Input image
Y	The set of output labels $[y_1, y_2, \dots, y_n]$
y_i	Target class label
X^*	Adversarial set
$C(X)$	The classifier takes input X and output Y
x_i^*	Perturbed image
$g(\mathbf{x}) = \nabla_{\mathbf{x}}$	Gradient w.r.t \mathbf{x}
\mathcal{L}	Loss function
ϵ	Scaling factor
ℓ_2	Euclidean norm
$ l $	Absolute value
p	Norm used to create adversarial samples
ρ	A perturbation function
γ	Minimum perturbation value
N_b	Number of batches
N_s	Number of samples per batch
R	Maximum perturbation
$\ x\ $	Normalisation
α	Stepsize

References

- Papernot, N., Faghri, F., Carlini, N., Goodfellow, I., Feinman, R., Kurakin, A., Xie, C., Sharma, Y., Brown, T., Roy, A., Matyasko, A., Behzadan, V., Hambardzumyan, K., Zhang, Z., Juang, Y.-L., Li, Z., Sheatsley, R., Garg, A., Uesato, J., Gierke, W., Dong, Y., Berthelot, D., Hendricks, P., Rauber, J., Long, R.: Technical report on the cleverhans v2.1.0 adversarial examples library. arXiv preprint [arXiv:1610.00768](https://arxiv.org/abs/1610.00768) (2018)
- Nicolae, M.-I., Sinn, M., Tran, M.N., Rawat, A., Wistuba, M., Zantedeschi, V., Molloy, I., Edwards, B.: Adversarial robustness toolbox v0.10.0. *CoRR*, [arXiv:1807.01069](https://arxiv.org/abs/1807.01069) (2018)
- Rauber, J., Brendel, W., Bethge, M.: Foolbox: a python toolbox to benchmark the robustness of machine learning models. arXiv preprint [arXiv:1707.04131](https://arxiv.org/abs/1707.04131) (2018)
- B. X-lab: Adv box: a toolbox to generate adversarial examples that fool neural networks (2019)
- Ling, X., Ji, S., Zou, J., Wang, J., Wu, C., Li, B., Wang, T.: Deepsec: a uniform platform for security analysis of deep learning model. In *IEEE S&P* (2019)
- Ding, G.W., Wang, L., Jin, X.: Advtorch v0. 1: an adversarial robustness toolbox based on pytorch. arXiv preprint [arXiv:1902.07623](https://arxiv.org/abs/1902.07623) (2019)
- Carlini, N.: A critique of the deepsec platform for security analysis of deep learning models. arXiv preprint [arXiv:1905.07112](https://arxiv.org/abs/1905.07112) (2019)
- Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A.: On evaluating adversarial robustness. arXiv preprint [arXiv:1902.06705](https://arxiv.org/abs/1902.06705) (2019)
- Brendel, W., Rauber, J., Kümmerner, M., Ustyuzhaninov, I., Bethge, M.: Accurate, reliable and fast robustness evaluation. arXiv preprint [arXiv:1907.01003](https://arxiv.org/abs/1907.01003) (2019)
- Duesterwald, E., Murthi, A., Venkataraman, G., Sinn, M., Vijaykeerthy, D.: Exploring the hyperparameter landscape of adversarial robustness. arXiv preprint [arXiv:1905.03837](https://arxiv.org/abs/1905.03837) (2019)
- Machado, G.R., Silva, E., Goldschmidt, R.R.: A non-deterministic method to construct ensemble-based classifiers to protect decision support systems against adversarial images: a case study. In: *Proceedings of the XV Brazilian Symposium on Information Systems*, p. 72, ACM (2019)
- Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., Srivastava, B.: Detecting backdoor attacks on deep neural networks by activation clustering. arXiv preprint [arXiv:1811.03728](https://arxiv.org/abs/1811.03728) (2018)
- Petrov, D., Hospedales, T.M.: Measuring the transferability of adversarial examples. arXiv preprint [arXiv:1907.06291](https://arxiv.org/abs/1907.06291) (2019)
- Fidel, G., Bitton, R., Shabtai, A.: When explainability meets adversarial learning: detecting adversarial examples using shap signatures. arXiv preprint [arXiv:1909.03418](https://arxiv.org/abs/1909.03418) (2019)
- Serban, A.C., Poll, E.: Adversarial examples—a complete characterisation of the phenomenon. arXiv preprint [arXiv:1810.01185](https://arxiv.org/abs/1810.01185) (2018)
- Duan, Y., Zhao, Z., Bu, L., Song, F.: Things you may not know about adversarial example: a black-box adversarial image attack. arXiv preprint [arXiv:1905.07672](https://arxiv.org/abs/1905.07672) (2019)

17. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 387–402. Springer, Berlin (2013)
18. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint [arXiv:1312.6199](https://arxiv.org/abs/1312.6199) (2013)
19. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint [arXiv:1412.6572](https://arxiv.org/abs/1412.6572) (2014)
20. Papernot, N., Patrick, M., Goodfellow, I.: Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint [arXiv:1605.07277](https://arxiv.org/abs/1605.07277) (2016)
21. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint [arXiv:1712.04248](https://arxiv.org/abs/1712.04248) (2017)
22. Nicholas Papernot, P.: Distillation as a defence to adversarial perturbations against deep neural networks. In: IEEE Symposium on Security and Privacy, pp. 582–597. IEEE (2016)
23. Rubinstein, B.I., Bartlett, P.L., Huang, L., Taft, N.: Learning in a large function space: privacy-preserving mechanisms for svm learning. arXiv preprint [arXiv:0911.5708](https://arxiv.org/abs/0911.5708) (2009)
24. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint [arXiv:1406.2661](https://arxiv.org/abs/1406.2661) (2014)
25. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: protecting classifiers against adversarial attacks using generative models. arXiv preprint [arXiv:1805.06605](https://arxiv.org/abs/1805.06605) (2018)
26. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint [arXiv:1611.01236](https://arxiv.org/abs/1611.01236) (2016)
27. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, B.Z., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387. IEEE (2016)
28. Demontis, A., Melis, M., Biggio, B., Maiorca, D., Arp, D., Rieck, K., Corona, I., Giacinto, G., Roli, F.: Yes, machine learning can be more secure! A case study on android malware detection. *IEEE Trans. Dependable Secur. Comput.* **16**(4), 711–724 (2019)
29. Kurakin, A., Goodfellow, I., Bengio, S., et al.: Adversarial examples in the physical world (2016)
30. Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773 (2017)
31. Yakura, H., Sakuma, J.: Robust audio adversarial example for a physical attack. arXiv preprint [arXiv:1810.11793](https://arxiv.org/abs/1810.11793) (2018)
32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint [arXiv:1706.06083](https://arxiv.org/abs/1706.06083) (2017)
33. Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582 (2016)
34. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57. IEEE (2017)
35. Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., Hsieh, C.-J.: Ead: elastic-net attacks to deep neural networks via adversarial examples. In: Thirty-second AAAI conference on artificial intelligence (2018)
36. 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)
37. Jang, U., Wu, X., Jha, S.: Objective metrics and gradient descent algorithms for adversarial examples in machine learning. In: Proceedings of the 33rd Annual Computer Security Applications Conference, pp. 262–277. ACM (2017)
38. Miyato, T., Maeda, S.-I., Koyama, M., Nakae, K., Ishii, S.: Distributional smoothing with virtual adversarial training. arXiv preprint [arXiv:1507.00677](https://arxiv.org/abs/1507.00677) (2015)
39. Sabour, S., Cao, Y., Faghri, F., Fleet, D.J.: Adversarial manipulation of deep representations. arXiv preprint [arXiv:1511.05122](https://arxiv.org/abs/1511.05122) (2015)
40. Narodytska, N., Kasiviswanathan, S.P.: Simple black-box adversarial perturbations for deep networks. arXiv preprint [arXiv:1612.06299](https://arxiv.org/abs/1612.06299) (2016)
41. Uesato, J., O’Donoghue, B., Oord, A.V.D., Kohli, P.: Adversarial risk and the dangers of evaluating against weak attacks. arXiv preprint [arXiv:1802.05666](https://arxiv.org/abs/1802.05666) (2018)
42. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Education, London (2001)
43. Zaremba, W., Sutskever, I., Vinyals, O.: Recurrent neural network regularization. arXiv preprint [arXiv:1409.2329](https://arxiv.org/abs/1409.2329) (2014)
44. Pan, Z., Yu, W., Yi, X., Khan, A., Yuan, F., Zheng, Y.: Recent progress on generative adversarial networks (gans): a survey. *IEEE Access* **7**, 36322–36333 (2019)
45. Xiao, C., Li, B., Zhu, J.-Y., He, W., Liu, M., Song, D.: Generating adversarial examples with adversarial networks. arXiv preprint [arXiv:1801.02610](https://arxiv.org/abs/1801.02610) (2018)
46. Zhao, G., Zhang, M., Liu, J., Wen, J.-R.: Unsupervised adversarial attacks on deep feature-based retrieval with gan. arXiv preprint [arXiv:1907.05793](https://arxiv.org/abs/1907.05793) (2019)
47. Lin, K., Li, D., He, X., Zhang, Z., Sun, M.-T.: Adversarial ranking for language generation. arXiv preprint [arXiv:1705.11001](https://arxiv.org/abs/1705.11001) (2017)
48. Yu, L., Zhang, W., Wang, J., Yu, Y.: Seqgan: sequence generative adversarial nets with policy gradient. In: Proceedings of the AAAI conference on artificial intelligence, vol. 31 (2017)
49. Chang, K.-H., Huang, P.-H., Yu, H., Jin, Y., Wang, T.-C.: Audio adversarial examples generation with recurrent neural networks. In: 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC), pp. 488–493. IEEE (2020)
50. Liu, Z., Yin, X.: Lstm-cgan: Towards generating low-rate dds adversarial samples for blockchain-based wireless network detection models. *IEEE Access*, vol. 9, pp. 22616–22625 (2021)
51. Guo, J., Zhao, Y., Han, X., Jiang, Y., Sun, J.: Rnn-test: Adversarial testing framework for recurrent neural network systems. arXiv preprint [arXiv:1911.06155](https://arxiv.org/abs/1911.06155) (2019)
52. Alireza Makhzani, N.J., Jonathon, S., Goodfellow, I.: Adversarial autoencoders. arXiv preprint [arXiv:1511.05644](https://arxiv.org/abs/1511.05644) (2016)
53. Garcia, J., Fernández, F.: A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **16**(1), 1437–1480 (2015)
54. Xu, W., Evans, D., Qi, Y.: Feature squeezing: detecting adversarial examples in deep neural networks. arXiv preprint [arXiv:1704.01155](https://arxiv.org/abs/1704.01155) (2017)
55. Warde-Farley, D., Goodfellow, I.: 11 adversarial perturbations of deep neural networks. *Perturbations Optim. Stat.* (2016). <https://doi.org/10.7551/mitpress/10761.003.0012>
56. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint [arXiv:1608.00853](https://arxiv.org/abs/1608.00853) (2016)
57. Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M. E., Chau, D.H.: Keeping the bad guys out: protecting and vaccinating deep learning with jpeg compression. arXiv preprint [arXiv:1705.02900](https://arxiv.org/abs/1705.02900) (2017)
58. Zantedeschi, V., Nicolae, M.-I., Rawat, A.: Efficient defenses against adversarial attacks. In: Proceedings of the 10th ACM

- workshop on artificial intelligence and security, pp. 39–49. ACM (2017)
59. Barreno, M., Nelson, B., Joseph, A.D., Tygar, J.D.: The security of machine learning. *Mach. Learn.* **81**(2), 121–148 (2010)
 60. Baracaldo, N., Chen, B., Ludwig, H., Safavi, J.A.: Mitigating poisoning attacks on machine learning models: a data provenance based approach. In: *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 103–110. ACM (2017)
 61. Gu, T., Dolan-Gavitt, B., Garg, S.: Badnets: identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733* (2017)
 62. Cohen, J.M., Rosenfeld, E., Kolter, J.Z.: Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918* (2019)
 63. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826. IEEE (2016)
 64. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385,2015* (2015)
 65. Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M.S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al.: A closer look at memorization in deep networks. In: *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 233–242. JMLR (2017)
 66. Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., Daniel, L.: Evaluating the robustness of neural networks: an extreme value theory approach. *arXiv preprint arXiv:1801.10578* (2018)
 67. Chen, H., Zhang, H., Si, S., Li, Y., Boing, D., Hsieh, C.-J.: Robustness verification of tree-based models. *arXiv preprint arXiv:1906.03849* (2019)
 68. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 103–110. IEEE (2009)
 69. Torralba, A., Fergus, R., Freeman, W.T.: 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **30**(11), 1958–1970 (2008)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



S. Asha is presently an Assistant Professor at SCMS School of Engineering and Technology, Ernakulam, Kerala, India. She is pursuing Ph.D. from APJ Abdul Kalam Technological university, Trivandrum, Kerala. Her areas of interests are Adversarial Machine learning, deep learning and Computer vision.



P. Vinod is presently Professor at the Department of Computer Applications at Cochin University of Science & Technology, Cochin, Kerala, India. Prior to this he was Professor in Department of Computer Science & Engg. at SCMS School of Engineering & Technology, Kerala, India. He holds a Ph.D from Department of Computer Engineering, Malaviya National Institute of Technology, Jaipur, India. He was Post Doc at Department of Mathematics, University of Padua, Italy and a Post Doctoral researcher at Malaviya National Institute of Technology, Jaipur, India. He has more than 70 research articles published in peer reviewed Journals and International Conferences. He is a reviewer of a number of security journals, and has also served as programme committee member in the International Conferences related to Computer and Information security. His current research is involved in the development of malware scanners for mobile applications using machine learning techniques. Vinod's area of interest is Adversarial Machine Learning, Malware Analysis, Context aware privacy preserving Data Mining, and Natural Language Processing.

[\[BACK\]](#)*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.

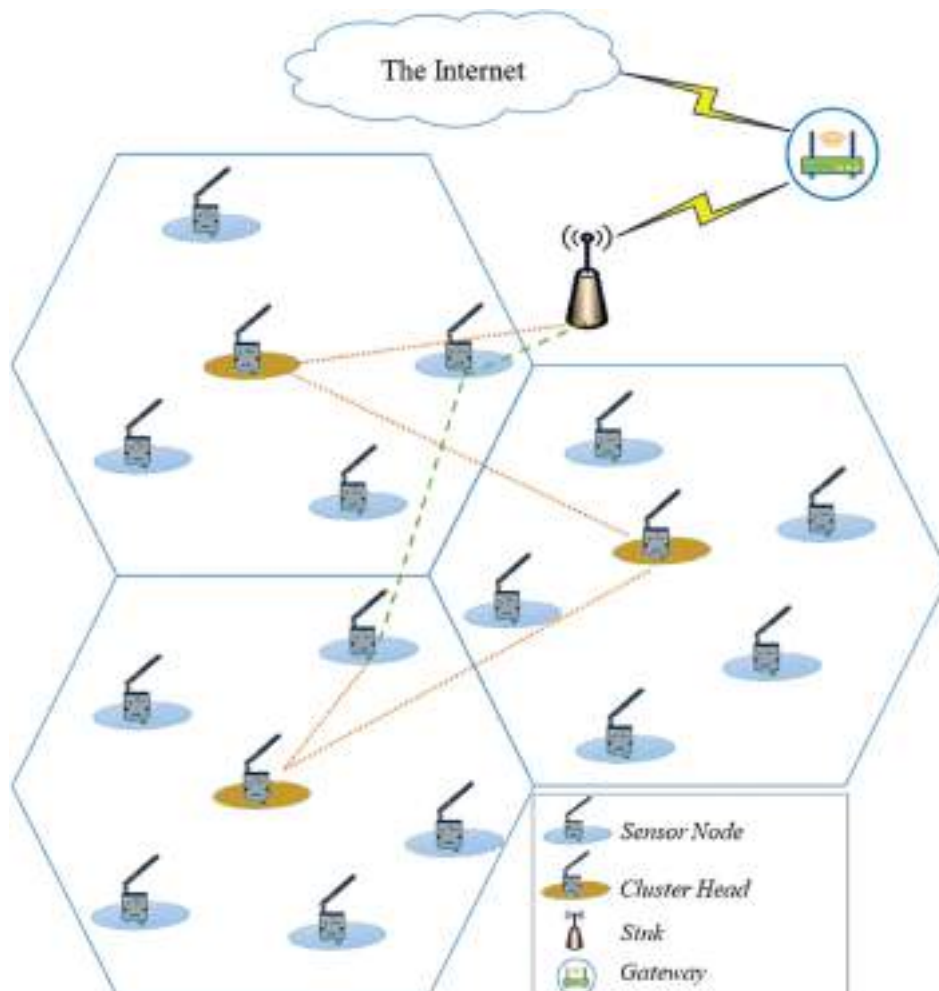


Figure 1: Routing procedure in WSNs

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

$$E_t = E^T + P_i \times d_{ij} + W \times d^m + \sum_{k=1}^{n-1} D_k \tag{11}$$

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^T	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^T . 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

$$\max \sum_{i=1}^n d_i \tag{12}$$

$$\text{s.t. } \sum_{i=1}^n d_i + \sum_{i=1}^n d_i \times E_i \tag{13}$$

$$\sum_{i=1}^n E_i \times P_i \tag{14}$$

$$\sum_{i=1}^n E_i \times P_i \tag{15}$$

$$E_i + P_i = E^T \tag{16}$$

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:

$$\begin{aligned} \text{max } w &= 50x_1 + 100x_2 + 70x_3 && \text{---(1)} \\ \text{s.t. } x_1 + x_2 + x_3 &\leq 100 && \text{---(2)} \\ -2x_1 + 5x_2 + 4x_3 &\leq 100 && \text{---(3)} \\ 4x_1 - 2x_2 - x_3 &\leq 70 && \text{---(4)} \\ x_1, x_2, x_3 &\geq 0 && \text{---(5)} \end{aligned}$$

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:

$$\begin{aligned} \text{min } w &= 100y_1 + 100y_2 + 70y_3 && \text{---(1)} \\ \text{s.t. } y_1 + y_2 + y_3 &\geq 50 && \text{---(2)} \\ -2y_1 + 5y_2 + 4y_3 &\geq 100 && \text{---(3)} \\ 4y_1 - 2y_2 - y_3 &\geq 70 && \text{---(4)} \\ y_1, y_2, y_3 &\geq 0 && \text{---(5)} \end{aligned}$$

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

$$\begin{aligned} \text{min } w &= 100y_1 + 100y_2 + 70y_3 + 0s_1 + 0s_2 + 0s_3 && \text{---(1)} \\ \text{s.t. } y_1 + y_2 + y_3 + s_1 &= 50 && \text{---(2)} \\ -2y_1 + 5y_2 + 4y_3 + s_2 &= 100 && \text{---(3)} \\ 4y_1 - 2y_2 - y_3 + s_3 &= 70 && \text{---(4)} \\ y_1, y_2, y_3, s_1, s_2, s_3 &\geq 0 && \text{---(5)} \end{aligned}$$

$$4x_1 - 2x_2 - x_3 - a_1 = 1 \tag{10}$$

$$x_1, x_2, x_3, a_1, a_2, a_3 \in \mathbb{R} \tag{11}$$

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:

$$\max: -50y_1 + 100y_2 + 70y_3 + 0y_4 + 0y_5 + 0y_6 - M a_1 - M a_2 - M a_3 \tag{12}$$

$$a_1' - y_1 - 2y_2 + 4y_3 + a_4 = 1 \tag{13}$$

$$-2y_1 + 5y_2 - 2y_3 + a_5 = 1 \tag{14}$$

$$4x_1 - 2x_2 - x_3 + a_6 = 1 \tag{15}$$

$$x_1, x_2, x_3, a_4, a_5, a_6 \in \mathbb{R} \tag{16}$$

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;

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} b_1 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} 16b_1 + 1628 \\ 2b_1 + 240 \\ 9b_1 + 2129 \end{bmatrix} \quad (27)$$

$$\begin{cases} 16b_1 + 1628 \geq 0 \Rightarrow b_1 \geq -101.75 \\ 2b_1 + 240 \geq 0 \Rightarrow b_1 \geq -120 \\ 9b_1 + 2129 \geq 0 \Rightarrow b_1 \geq -236.56 \end{cases} \quad (28)$$

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

$$RHS = B^{-1}b = \begin{bmatrix} 16/35 & 2/7 & 9/35 \\ 2/7 & 3/7 & 2/7 \\ 9/35 & 2/7 & 16/35 \end{bmatrix} \begin{bmatrix} -110 \\ 100 \\ 70 \end{bmatrix} = \begin{bmatrix} -278 \\ 228 \\ 228 \end{bmatrix} \quad (29)$$

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 $C_j - Z_j$.

$$\text{For } x_3 = C_j - Z_j = 0 - \left[\left(1 + \frac{36}{105}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{9}{105}\right) \right] = \frac{-26c_1 - 26}{105} \tag{68}$$

$$\text{For } x_2 = C_j - Z_j = 0 - \left[\left(1 + \frac{2}{7}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{2}{7}\right) \right] = \frac{-2c_1 - 4}{7} \tag{69}$$

$$\text{For } x_1 = C_j - Z_j = 0 - \left[\left(1 + \frac{5}{105}\right) + \left(1 + \frac{2}{7}\right) + \left(c_1 + \frac{16}{105}\right) \right] = \frac{-16c_1 - 16}{105} \tag{70}$$

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

$$\frac{-26c_1 - 26}{105} \leq 0 \Rightarrow c_1 \geq \frac{-26}{9} \tag{71}$$

$$\frac{-2c_1 - 4}{7} \leq 0 \Rightarrow c_1 \leq \frac{-2}{7} \tag{72}$$

$$\frac{-16c_1 - 16}{105} \leq 0 \Rightarrow c_1 \geq \frac{-16}{15} \tag{73}$$

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.

Edge-Centric Secure Service Provisioning in IoT-Enabled Maritime Transportation Systems

Ambigavathi Munusamy^{ID}, *Member, IEEE*, Mainak Adhikari^{ID}, *Member, IEEE*,
 Mohammad Ayoub Khan^{ID}, *Member, IEEE*, Varun G. Menon^{ID}, *Senior Member, IEEE*,
 Satish Narayana Srirama^{ID}, *Senior Member, IEEE*, Linss T. Alex, *Member, IEEE*,
 and Mohammad R. Khosravi^{ID}, *Member, IEEE*

Abstract—With the exponential growth of the Internet of Things (IoT) devices in Maritime Transportation Systems (MTS), the centralized cloud-centric framework can hardly meet the requirements of the applications in terms of low latency and power consumption. By inventing the distributed edge-centric framework, real-time IoT applications can meet the requirements of the MTS by analyzing the tasks at the edge of the networks. However, one of the critical challenges of the edge-centric MTS is to provide security and privacy between local IoT devices and distributed edge nodes. Motivated by that, in this paper, we design a blockchain-enabled edge-centric framework for analyzing the real-time data at the edge of the networks with minimum latency and power consumption while meeting the security and privacy issue of MTS. The introduction of blockchain and smart contract in the edge-centric MTS frameworks help to validate the transactions of each block at edge nodes by estimating the lifetime, belief, and trustfulness, and mitigate various types of security threats. Further, we introduce different classification models to predict the malicious vessels over the real-time maritime dataset at a secured edge-centric MTS framework. Extensive simulation results demonstrate that the superiority of the proposed strategy with baseline approaches under various performance metrics.

Index Terms—IoT, maritime transport networks, service provisioning, blockchain, smart contract, vessel tracking.

Manuscript received 15 April 2021; revised 26 June 2021 and 17 July 2021; accepted 1 August 2021. **Date of publication 12 August 2021**; date of current version 8 February 2023. This work was supported by the UoH-IoE through MHRD, India, under Grant F11/9/2019-U3(A). The Associate Editor for this article was A. K. Bashir. (*Corresponding author: Satish Narayana Srirama.*)

Ambigavathi Munusamy is with the Department of Electronics and Communication Engineering, Anna University, Chennai 600025, India (e-mail: ambigaindhu8@gmail.com).

Mainak Adhikari is with the Institute of Computer Science, University of Tartu, 50090 Tartu, Estonia (e-mail: mainak.ism@gmail.com).

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 SCMS School of Engineering and Technology, Kerala 683576, India (e-mail: varunmenon@scmsgroup.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).

Linss T. Alex is with the Albertian Institute of Science and Technology, Kochi 682022, India (e-mail: linsstalex@aisat.ac.in).

Mohammad R. Khosravi is with the Department of Electrical and Electronic Engineering, Shiraz University of Technology, Shiraz 71557-13876, Iran (e-mail: m.r.khosravi.taut@gmail.com).

Digital Object Identifier 10.1109/TITS.2021.3102957

I. INTRODUCTION

THE rapid emergence of digitalization and data explosion from the Internet of Things (IoT) devices has deeply promoted and altered the threat dynamics in maritime networks [1]. With the integration of IoT and edge computing, the Maritime Transport Systems (MTS) can improve the real-time tracking, safety operations of vessels, cargo handling, and cyber risk assessments [2], [3]. However, it is highly attracted by the threat actors and cyber attackers or cyber-criminals in recent times [4]. For instance, the maritime industry was hit by various cyber-attacks [5]. In the year 2017, Maersk was hit by NotPetya ransomware and affected the overall operational capacity and cost up to \$300 million.¹ Besides that, in the year 2018, MTS was hit by two major cyber-attacks including Port of Barcelona and Port of San Diego, respectively.

Recently, in the year 2020, the Mediterranean Shipping Company² was hit by a cyber-attack, which disrupted the services of the company across the world. Therefore, the transmission of real-time data to the distributed local edge devices instead of the centralized cloud server is less vulnerable, which can reduce the chance to tamper the tasks by the attackers [6], [7]. Besides that, analyzing the real-time tasks at the edge of the networks with minimum latency and power consumption is another critical challenge of the MTS [4].

A. Related Works

This section inspects various existing service deployment strategies, security mechanisms, predictive analysis models in the MTS, and edge networks [8]–[11]. The mobility-aware blockchain-enabled resource provisioning strategy has been designed to offload the tasks under edge-enabled secure vehicular network [12]. A blockchain-enabled federated learning model has been introduced to securely upload the tasks to the fog server and the global updates are stored and computed in the end devices [13]. In [14], a scalable mechanism

¹<https://digitalguardian.com/blog/cost-malware-infection-maersk-300-million>

²<https://www.cybersecurity-insiders.com/mediterranean-shipping-company-msc-hit-by-a-cyber-attack/>

called LiTiChain has been proposed to manage the outdated transactions in each block and remove them based on the end time of the blocks. A multi-vessel offloading strategy has been proposed for maritime edge networks [15]. Based on the task offloading decisions, the local device offloads the tasks to the computing servers without any security mechanism for handling the tasks from the vessel terminals.

A two-stage offloading scheme has been developed for a maritime network using mobile edge computing in [16]. To increase the processing ability of the edge server, the computing tasks requested by the IoT devices are offloaded to the local edge devices. The transmission of delay-sensitive tasks during offloading is more vulnerable. For this reason, blockchain has been considered to ensure the data integrity [17]. The blockchain-enabled and learning-based secure intelligent task offloading strategy has been designed for vehicular networks in [18]. This strategy mitigates various types of attacks and increases the trustfulness of network entities by quantifying the task offloading success events. To minimize the issues of data imbalance, authors in [19]–[21] have analyzed difficulties of the maritime datasets to identify better similarities and enhance the accuracy. Thus, one of the critical and yet unsolved challenges on the MTS is to design a secured edge-centric framework for analyzing the real-time MTS data at the edge of the networks with minimum latency and power consumption while meeting security and privacy issues.

B. Motivations

In recent times, many research efforts have been conducted to solve the data breach issues and detect the security threats and cyber-attacks in maritime networks. Similarly, a lot of service provisioning strategies have been proposed to offload the real-time vessel tasks to the remote computing servers for further processing and analysis [22], [23]. However, modern maritime networks should require an efficient service provisioning strategy that can reduce the latency and power consumption of real-time tasks. The huge volume of sensitive tasks from multiple vessels should be validated and verified at the edge of the networks. Further, all the secured tasks stored at the local edge devices should be managed according to the expiry time of the blocks to reduce the processing overhead. From the above perceptions, the following three major research questions come out: *1. how to classify and securely offload the large volume of vessel's IoT data using an efficient service deployment strategy?*, *2. how to manage the outdated blocks according to the lifetime and arrival of data?*, *3. how to instantly detect malicious entities using a suitable predictive analysis model with higher accuracy?* To the best of our knowledge, none of the existing research works have been focused to jointly design secure service provisioning using smart contracts and blockchain along with a predictive analysis model for edge-centric MTS. Thus, we require an edge-centric secure service provisioning strategy to enhance the network performance of the MTS while meeting security and privacy requirements.

C. Contributions

Motivated by the above-mentioned challenges, in this paper, we design a new secure edge-centric framework for the MTS with an efficient service provisioning strategy and blockchain technique. Extensive simulation results demonstrate the efficiency of the proposed strategy over the baseline algorithms in terms of various performance metrics. The main contributions of the proposed strategy are summarized as follows.

- Design a Task Queuing Priority (TQP) based algorithm for classifying the incoming IoT-enabled tasks from maritime vessels based on the degree of importance and deadlines. In this approach, the incoming vessel's tasks are classified into three types including secured, trusted, and general tasks, respectively and analyzed the tasks at remote edge devices or cloud servers as per their importance.
- Introduce smart contract at the edge-centric framework to store all the blockchain ledger, verify the validation of blocks, and identify the malicious behaviors of vessels in maritime networks. This technique greatly reduces various security threats of the network based on the belief estimation and trust assessment.
- Device a task handling strategy using blockchain to efficiently record all the secured transactions based on the lifetime and expiration time of blocks. This is achieved by introducing a scalable MiTiChain to manage the outdated blocks and reduce the processing overhead of the local edge devices.
- Jointly considering the SVM and CNN learning model in the edge-centric framework for identifying the number of malicious entities in the network. Further, the learning models have been validated with a real-time maritime vessel dataset to improve the prediction accuracy and safety operations of maritime vessels.

The rest of this paper is organized as follows. Section II presents the system model and problem formulation. The proposed secure edge-centric service provisioning strategy for MTS is described in section III. Section IV demonstrates the various numerical and predictive analyses of the proposed strategy. Finally, section V concludes the work.

II. SYSTEM MODEL AND PROBLEM FORMULATION

This section presents the system model for an edge-centric IoT-enabled MTS followed by the problem formulation.

A. System Model

An edge-centric MTS framework for analyzing real-time IoT data is shown in Fig.1. This network consists of a set of vessels with IoT devices $\mathcal{V} = \{v_1, v_2, v_3, \dots, v_m\}$. Let \mathcal{D} be the set of maritime edge servers, expressed as $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$ and \mathcal{C} be the set of centralized cloud servers, denoted as $\mathcal{C} = \{c_1, c_2, c_3, \dots, c_o\}$. During the voyage of a vessel, each monitoring IoT device transmits a set of tasks $\mathcal{T} = \{t_1, t_2, \dots, t_k\}$ for further analysis. The incoming tasks, generated from different v_m , are processed

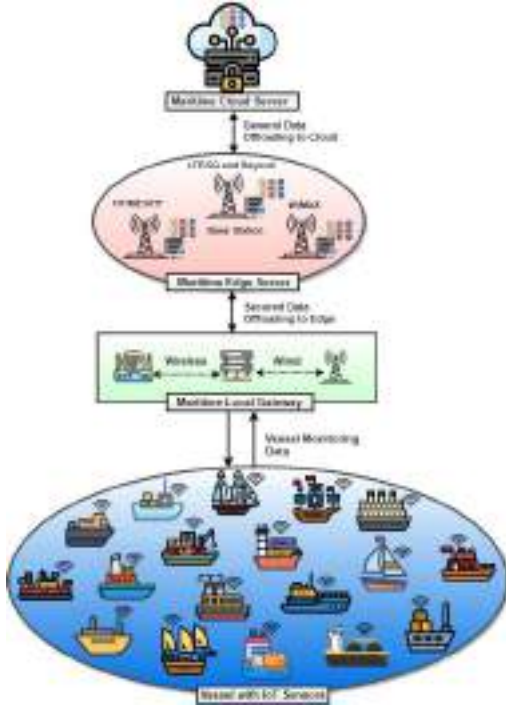


Fig. 1. A hierarchical view of edge-centric maritime transport network.

at local IoT device or offloaded to the remote computing devices d_n or c_o through a set of maritime gateway devices $\mathcal{G} = \{g_1, g_2, \dots, g_l\}$. Further, the set of wireless channels for offloading the vessel's tasks can be represented as $\mathcal{H} = \{h_1, h_2, \dots, h_g\}$. The total workload of a particular vessel, input size of task, and required CPU cycles of the task are expressed as $\mathcal{U}_m = \{\psi_m, \mathcal{I}_m, \sigma_m\}$. The offloading decision χ_{v_m} of an incoming task for further processing is expressed as follows.

$$\chi_{v_m} = \begin{cases} 1 & \text{Offload the vessel's tasks to maritime } d_n \text{ or } c_o \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

1) *Local Computing Offloading Model*: The small-scale real-time tasks, generated from vessel IoT device v_m with limited resource requirement can be processed locally. Let σ_m and $P_{v_m}^{local}$ be the required CPU cycles to process the tasks and local computing power of the vessel IoT device v_m . Thus, the total processing time of the task is formulated as follows.

$$T_{v_m}^{local} = \sigma_m \times (P_{v_m}^{local})^{-1} \quad (2)$$

The vessel IoT device can dynamically adjust the frequency and voltage of the vessel based on the various service requirements and applications running on the same vessel. Thus, the power consumed by the vessel IoT device v_m while processing the task is given as follows.

$$P_{v_m}^{local} = S_m \times (P_{v_m}^{local})^2 \times \sigma_m \quad (3)$$

where S_m denotes the structure of the vessel monitoring unit. The total processing cost $\mathcal{Q}_{v_m}^{local}$ of the tasks in the local vessel IoT device v_m is expressed as follows.

$$\mathcal{Q}_{v_m}^{local} = \mu_{v_m}^{local} \times T_{v_m}^{local} + \omega_{v_m}^{local} \times P_{v_m}^{local} \quad (4)$$

where $\mu_{v_m}^{local}$ and $\omega_{v_m}^{local}$ indicate the time and power weight coefficients for processing the tasks \mathcal{U}_m locally on the vessel. The correlation between these coefficients is defined as follows.

$$\begin{aligned} \mu_{v_m}^{local} + \omega_{v_m}^{local} &= 1 \\ 0 &\leq \mu_{v_m}^{local} \leq 1 \\ 0 &\leq \omega_{v_m}^{local} \leq 1 \end{aligned} \quad (5)$$

Therefore, both $\mu_{v_m}^{local}$ and $\omega_{v_m}^{local}$ can directly impact to process the task \mathcal{U}_m in vessel IoT device v_m . For instance, if the value of $\mu_{v_m}^{local}$ is larger, then the v_m needs to pay more attention to the delay. In contrast, if the value of $\omega_{v_m}^{local}$ is larger, it has to pay more attention to the power consumption. Thus, balancing these two coefficients can minimize the overall delay and power consumption of the tasks.

2) *Edge Computing Offloading Model*: Due to the limited resource capacity of the local v_m , most of the real-time tasks should be offloaded to the distributed edge devices \mathcal{D} in the network. The large number of tasks received from v_m are highly sensitive and must require instant decisions along with secure service provisioning to perform the safety operations. Besides that, it is assumed that there are different interferences between different vessel IoT devices and edge devices during the communication. Therefore, a set of channels \mathcal{H} is considered, where the channel bandwidth B is divided into several subchannels h_g . The bandwidth of each subchannel is denoted as $\frac{B}{h_g}$. By using Shannon's theorem, the uploading rate $r_{d_n}^{ch}$ of vessel tasks via specific channel ch can be expressed as follows.

$$r_{d_n}^{ch} = \phi_{d_n}^{ch} \times \frac{B}{h_g} \log(1 + SNR_{d_n}^{ch}) \quad (6)$$

where $\phi_{d_n}^{ch}$ and $SNR_{d_n}^{ch}$ present the channel assignment decision and the signal-to-noise ratio (SNR), respectively. N_0 represents the background noise of the channel. The transmission power of vessels v_m and edge device d_n on the same channel is considered as $P_{v_m}^{ch}$ and $P_{d_n}^{ch}$, respectively. $\mathcal{G}_{d_n}^{ch}$ is the channel gain between vessels and maritime edge device d_n . $\mathcal{G}_{v_m, d_n}^{ch}$ is the channel gain of v_m and d_n , respectively. Thus, the SNR of data transmission is calculated as follows.

$$SNR_{d_n}^{ch} = \frac{P_{v_m}^{ch} \times \mathcal{G}_{d_n}^{ch}}{N_0 + \sum_{d_n=1, d_n \neq d_n}^S P_{d_n}^{ch} \mathcal{G}_{v_m, d_n}^{ch}} \quad (7)$$

The channel assignment decision $\phi_{d_n}^{ch}$ indicates whether the channel h_g is allocated to edge device d_n or not and it can be written as follows.

$$\phi_{d_n}^{ch} = \begin{cases} 1 & \text{Subchannel is allocated to edge device } d_n \\ 0 & \text{Otherwise} \end{cases} \quad (8)$$

The total transmission rate of uploading the tasks from v_m is obtained as follows.

$$R_{d_n}^{ch} = \sum_{h_m}^H r_{d_v}^{ch} \quad (9)$$

Likewise, the total time required to offload the task from v_m to the edge device is expressed as follows.

$$T_{d_n}^{edge} = \frac{\mathcal{I}_m}{R_{d_n}^{ch}} + T_{d_n}^{wait} + \frac{\sigma^{edge}}{P_{edge}} \quad (10)$$

The total power consumed by the local gateway device while offloading the vessel's task to the maritime edge device is defined as follows.

$$P_{d_n}^{edge} = P_{d_n}^{ch} \left(\frac{\mathcal{I}_m}{R_{d_n}^{ch}} + T_{d_n}^{wait} \right) + P_{d_n}^{addl} + S^{edge} (P_{edge})^2 \sigma^{edge} \quad (11)$$

where $T_{d_n}^{wait}$ and $P_{d_n}^{addl}$ signify the waiting time and tail energy of the edge device d_n . The total processing cost of the task in the maritime edge device is computed as follows.

$$\mathcal{Q}_{d_n}^{edge} = \mu_{d_n}^{edge} \times T_{d_n}^{edge} + \omega_{d_n}^{edge} \times P_{d_n}^{edge} \quad (12)$$

where $\mu_{d_n}^{edge}$ and $\omega_{d_n}^{edge}$ indicate the time and power weight coefficients for processing the vessel's tasks in distributed edge devices.

3) *Cloud Computing Offloading Model*: The remaining computation-intensive tasks, generated by v_m are offloaded to the centralized cloud server c_o through wireless networks. Thus, the total time required to offload the tasks from g_l to the cloud server can be obtained as follows.

$$T_{c_o}^{cloud} = \frac{\mathcal{I}_m}{R_{c_o}^{ch}} + T_{c_o}^{wait} + \frac{\sigma^{cloud}}{P_{cloud}} \quad (13)$$

The total power consumed by the g_l to offload the vessel's tasks to the remote cloud server is formulated as follows.

$$P_{c_o}^{cloud} = P_{c_o}^{ch} \left(\frac{\mathcal{I}_m}{R_{c_o}^{ch}} + T_{c_o}^{wait} \right) + P_{c_o}^{addl} \quad (14)$$

The total processing cost of the computation-intensive vessel tasks can be formulated as follows.

$$\mathcal{Q}_{c_o}^{cloud} = \mu_{c_o}^{cloud} \times T_{c_o}^{cloud} + \omega_{c_o}^{cloud} \times P_{c_o}^{cloud} \quad (15)$$

B. Problem Formulation and Design Goals

Here, we consider an incoming task to the gateway device g_l can be processed by a local edge device d_n or centralized cloud server c_o . Then, the total power consumption of the edge device and centralized cloud server is formulated as follows.

$$\mathcal{Q}_{ec} = \delta_{v_m} \mathcal{Q}_{d_n}^{edge} + (1 - \delta_{v_m}) \mathcal{Q}_{c_o}^{cloud} \quad (16)$$

where δ_{v_m} represents whether the task is processed either on the edge device or remote cloud server. Thus, the offloading decision δ_{v_m} is given as follows.

$$\delta_{v_m} = \begin{cases} 1 & \text{Maritime edge server} \\ 0 & \text{Maritime remote cloud server} \end{cases} \quad (17)$$

Similarly, the total power consumption of local vessel IoT device v_m is measured as follows.

$$\mathcal{Q}_{total} = \chi_{v_m} \mathcal{Q}_{ec} + (1 - \chi_{v_m}) \mathcal{Q}_{v_m}^{local} \quad (18)$$

where χ_{v_m} denote the offloading decision-making parameter at the local vessel IoT device. Let 0 and 1 be the condition of offloading and local processing of the tasks. The offloading decision matrix \mathcal{M}_{off} is represented as follows.

$$\mathcal{M}_{off} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (19)$$

The main objective of this work is to minimize the overall power consumption of the incoming vessel tasks while processing in the local IoT devices or remote edge or cloud servers. Thus, the optimization problem with necessary constraints can be formulated as follows.

$$\text{minimize } \mathcal{Q}_{total} \quad (20a)$$

$$\text{subject to } \chi_{v_m} (\mathcal{N}_{v_m}) + (1 - \chi_{v_m}) T_{v_m}^{local} \leq T_{max} \quad (20b)$$

$$\chi_{v_m} (\mathcal{L}_{v_m}) + (1 - \chi_{v_m}) P_{v_m}^{local} \leq P_{max} \quad (20c)$$

$$0 \leq \sum_{c_m=1}^C \phi_{v_m}^{ch} P_{v_m}^{ch} \leq P_{max} \quad (20d)$$

$$\chi_{v_m} \in 0, 1, \delta_{v_m} \in 0, 1, \phi_{v_m}^{ch} \in 0, 1, \quad \forall v_m \in S \quad (20e)$$

The constraint (20b) presents the maximum delay T_{max} that the v_m can tolerate from the actual delay. Constraint (20c) defines the power consumption should not exceed the maximum value of P_{max} and constraint (20d) specifies the power limitations. The constraint (20e) denotes the offloading decision of v_m to the maritime edge or cloud server. Thus, the total power consumption during offloading the vessel tasks to the remote computing servers should always be less than the total consumption of processing the tasks at v_m .

III. EDGE-CENTRIC SECURE SERVICE PROVISIONING

In this section, we discuss the proposed secure service provisioning strategy in the edge-centric MTS. Here, we propose a secure task handling strategy using smart contracts and blockchain technology to ensure data integrity, fairness, and security in the network. Finally, we apply predictive models to assess the malicious entities in the edge-centric maritime network.

A. Task Queuing Priority (TQP)-Based Algorithm

Initially, the vessel's tasks are prioritized and sorted according to non-decreasing order. If a task T_k is considered as a secured task then its utilization factor $\mathcal{F}(T_k)$ should always be greater than equal to the value of $\frac{1}{2}$. Alternatively, if a task T_k is called a trusted or general task then its utilization factor should always be less than the value of $\frac{1}{2}$. Based on the task ordering, the tasks are placed either into secured task queue τ_k^{Sr} , where $T_k \in \mathcal{T}_k^{Sr}$ or trusted task buffer τ_k^{Gr} . Let $f(t) = \lambda_k^e - \lambda_k^t$ be the arrival rate of maritime vessel's tasks using Poisson process with density function. Then, the arrival rate of the task t_k from v_m for processing locally is derived as follows.

$$\lambda_{t_k}^{local} = (1 - \chi_{v_m}) \times \lambda_{t_k} \quad (21)$$

where λ_{t_k} and χ_{v_m} represent the vessel's task arrival rate and the uploading decision. The set of vessel's tasks (β_{v_m}) that arrive under a secured queue ($\lambda_{t_k}^{Sr}$) of a gateway device is defined as follows.

$$\lambda_{t_k}^{Sr} = \beta_{v_m} \times \lambda_{t_k}^{rem} \quad (22)$$

Similarly, the rest of the vessel's tasks ($\lambda_{t_k}^{rem}$) that arrive under a normal task queue of a gateway device is calculated as follows.

$$\lambda_{t_k}^{rem} = (1 - \beta_{v_m}) \times \lambda_{t_k}^{rem} \quad (23)$$

Let $\gamma_{t_k}^{Sr}$ and $\gamma_{t_k}^{Gr}$ be the secured-tasks and general or trusted tasks, respectively. Thus, the arrival rate of the tasks that are offloaded to the edge device d_n , can be expressed as follows.

$$\begin{aligned} \lambda_{t_k}^{edge} &= \gamma_{t_k}^{Sr} \times \lambda_{t_k}^{Sr} + \gamma_{t_k}^{Gr} \times \lambda_{t_k}^{Gr} \\ &= \gamma_{t_k}^{Sr} \times \beta_{v_m} \times \lambda_{t_k}^{rem} + \gamma_{t_k}^{Gr} \times (1 - \beta_{v_m}) \times \lambda_{t_k}^{rem} \end{aligned} \quad (24)$$

Likewise, the tasks that are offloaded to the remote cloud server c_o can be formulated as follows.

$$\begin{aligned} \lambda_{t_k}^{cloud} &= (1 - \gamma_{t_k}^{Sr}) \times \lambda_{t_k}^{Sr} + (1 - \gamma_{t_k}^{Gr}) \times \lambda_{t_k}^{Gr} \\ &= (1 - \gamma_{t_k}^{Sr}) \times \beta_{v_m} \times \lambda_{t_k}^{rem} + (1 - \gamma_{t_k}^{Gr}) \times (1 - \beta_{v_m}) \times \lambda_{t_k}^{rem} \end{aligned} \quad (25)$$

Besides that, the conditional probability of processing the incoming task t_k at local vessel IoT device v_m can be expressed as follows.

$$\begin{cases} Q_{ec} \leq Q_{v_m}^{local}, & \chi_{v_m} = 1 \\ Q_{ec} \geq Q_{v_m}^{local}, & \chi_{v_m} = 0 \end{cases} \quad (26)$$

The total power consumption of the task t_k while processing in the local vessel IoT device v_m can be obtained as follows.

$$\min \sum_{v_m=1}^S \mu_{v_m}^{local} \left(\sigma_m P_{v_m}^{local-1} \right) + \omega_{v_m}^{local} \left(S_m P_{v_m}^{local^2} \sigma_m \right) \quad (27)$$

where $\sigma_m \left(P_{v_m}^{local-1} \right)$ and $S_m \left(P_{v_m}^{local^2} \right) \sigma_m$ denote the number of CPU cycles required to process the task and structure of the vessel monitoring unit, respectively. Then, $P_{v_m}^{local}$ is the computing power of v_m IoT device. Based on the above conditions, constraints (20b), (20c), and (20d) can be achieved. During task offloading to the remote computing device, i.e., if $\chi_{v_m} = 1$, then vessel IoT device v_m selects a suitable channel to offload the tasks to the computing servers. However, the channel capacity utilized by the g_l should not exceed the maximum channel capacity \mathcal{H} . Thus, v_m choose a suitable subchannel h_g to efficiently offload the tasks, which satisfies the constraint (20e). The conditional task offloading decisions of g_l can be expressed as follows.

$$\begin{cases} Q_{v_m}^{edge} \leq Q_{v_m}^{cloud}, & \text{Maritime edge server} \\ Q_{v_m}^{edge} \geq Q_{v_m}^{cloud}, & \text{Maritime remote cloud server} \end{cases} \quad (28)$$

The first case is applicable when g_l decides to offload the $\gamma_{t_k}^{Sr}$ to the maritime edge server. The second case is used when g_l decides to offload the $\gamma_{t_k}^{Gr}$ to the remote cloud server. Therefore, $Q_{d_n}^{edge} \left(P_{v_m}^{ch} \right)$ is fully related to the $P_{v_m}^{ch}$.

Further, the time interval of $P_{v_m}^{ch}$ is defined as $[P_x, P_y]$. Let $P_x^* = \max P_{min}, P_x$ and $P_y^* = \min P_{max}, P_y$ be the optimal intervals. Then, the total power consumption ($Q_{d_n}^{edge} \left(P_{v_m}^{ch} \right)$) of the tasks while offloading to the local edge devices can be formulated as follows.

$$\left(Q_{d_n}^{edge} \right)^* \left(P_{v_m}^{ch} \right) = \begin{cases} Q_{d_n}^{edge} \left(P_x \right)^*, & \left(P_{v_m}^{ch} \right)^* \leq \left(P_x \right)^* \\ Q_{d_n}^{edge} \left(P_{v_m}^{ch} \right)^*, & \left(P_x \right)^* < \left(P_y \right)^* \\ Q_{d_n}^{edge} \left(P_y \right)^*, & \left(P_{v_m}^{ch} \right)^* \geq \left(P_y \right)^* \end{cases} \quad (29)$$

From the above formulation, $\left(P_{v_m}^{ch} \right)^*$ implies the optimal transmission power of g_l by making $\Delta Q_{d_n}^{edge} \left(P_{v_m}^{ch} \right) | P_{v_m}^{ch} = \left(P_{v_m}^{ch} \right)^* = 0$. Similar to the previous formulation, it can be concluded that the interval range of $P_{v_m}^{ch}$ is $[P_n, P_o]$. Let $P_n^* = \max P_{min}, P_n$ and $P_o^* = \min P_{max}, P_o$ be the optimal intervals while offloading the tasks to the centralized cloud servers. Then, the total consumption ($Q_{c_o}^{cloud} \left(P_{v_m}^{ch} \right)$) of the tasks while offloading to the cloud servers is formulated as follows.

$$\left(Q_{c_o}^{cloud} \right)^* \left(P_{v_m}^{ch} \right) = \begin{cases} Q_{c_o}^{cloud} \left(P_n \right)^*, & \left(P_{v_m}^{ch} \right)^* \leq \left(P_x \right)^* \\ Q_{c_o}^{edge} \left(P_{v_m}^{ch} \right)^*, & \left(P_x \right)^* < \left(P_y \right)^* \\ Q_{c_o}^{edge} \left(P_y \right)^*, & \left(P_{v_m}^{ch} \right)^* \geq \left(P_y \right)^* \end{cases} \quad (30)$$

Based on the above formulations, the proposed strategy satisfies the delay, power consumption, and total consumption constraints of the tasks, refer to (20b)-(20e). The pseudocode of the proposed TQP is depicted in Algorithm 1. However, the vessels remain anonymous during the task offloading, cyber attackers may infer or alter information from anywhere, at any time. Therefore, it is necessary to protect sensitive information at different levels of transmissions, so the privacy information of v_m cannot be inferred by any cyber attacker.

Algorithm 1 TQP Algorithm

INPUT : \mathcal{T} : Set of vessel's tasks, $\gamma_{v_m}^{Sr}$: Secured tasks, $\gamma_{v_m}^{Gr}$: General tasks, δ_{v_m} : Offloading decision
OUTPUT : Prioritized vessel tasks ($\gamma_{v_m}^{Sr}, \gamma_{v_m}^{Gr}$)

```

1 begin
2   for  $i \leftarrow 1$  to  $t_m$  do
3     Compute priority utilization factor  $\mathcal{F}(\mathcal{T}_m) = \frac{S_r}{G_r}$ 
4     Assign priority to each incoming task  $\mathcal{T}_m$ 
5     if  $\mathcal{F}(\mathcal{T}_m) > \frac{1}{2}$  then
6       Classify the task  $\mathcal{T}_m$  as secured-task  $\gamma_{v_m}^{Sr}$ 
7     end
8     if  $\mathcal{F}(\mathcal{T}_m) \leq \frac{1}{2}$  then
9       Classify the task  $\mathcal{T}_m$  as general task  $\gamma_{v_m}^{Gr}$ 
10    end
11    Determine the offloading decisions at  $g_l$ 
12    if  $(\delta_{v_m} == 1)$  then
13      Offload the secured-task  $\gamma_{v_m}^{Sr}$  to  $d_n$ 
14    end
15    if  $(\delta_{v_m} == 0)$  then
16      Offload the general-task  $\gamma_{v_m}^{Gr}$  to  $c_o$ 
17    end
18  end
19 end
```

B. Task Offloading Using Blockchain and Smart Contract

Fig.2 shows the proposed secure task offloading strategy using the smart contract. We consider a scenario with a set

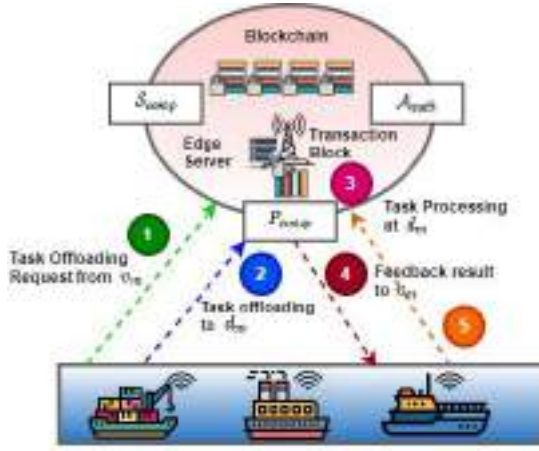


Fig. 2. Smart contract based secure task offloading in maritime network.

of local vessel IoT devices \mathcal{V} continuously offload their tasks to the local edge devices. If the number of incoming tasks is overloaded, then the edge server needs to maintain the transactions through the blockchain. To achieve this, we include three essential components at d_n such as register authority (\mathcal{A}_{auth}), computing component (\mathcal{P}_{comp}), and storage component (\mathcal{S}_{comp}). Here, \mathcal{A}_{auth} is responsible to provide registration and assign authority to each v_m in the form of the digital certificate. Then, \mathcal{P}_{comp} is used to implement the smart contract, and \mathcal{S}_{comp} is used to store all the transactions in the blockchain and verify the validation of blocks. Let $K_{d_n}^{pri}$ and $K_{d_n}^{pub}$ be the public and private key generated by the edge server d_n . This key is mainly used to encrypt and verify the digital signature of the received tasks using asymmetric cryptography. Similarly, v_m and d_n generate their public and private key pairs $(K_{v_m}^{pri}, K_{v_m}^{pub})$ and $(K_{d_n}^{pri}, K_{d_n}^{pub})$, respectively. Similarly, v_m uses its public key for encryption and private key for decryption, respectively. Both v_m and d_n exchange their public keys to \mathcal{A}_{auth} on the d_n . Subsequently, \mathcal{A}_{auth} uses its private key $K_{d_n}^{pri}$ to encrypt the information from v_m and d_n and generate unique signature Sig_{v_m} and Sig_{d_n} , respectively. This unique signature is required when v_m decides to offload the $\gamma_{v_m}^{S_r}$ to the edge server. Finally, \mathcal{A}_{auth} merges the public key $K_{v_m}^{pub}$ of v_m with its Sig_{v_m} , expressed as follows.

$$\begin{aligned} C_{v_m} &= K_{v_m}^{pub} || Sig_{v_m} \\ C_{d_n} &= K_{d_n}^{pub} || Sig_{d_n} \end{aligned} \quad (31)$$

Further, the certificates of both v_m and d_n are stored in \mathcal{S}_{comp} . The detailed step-by-step execution procedures of the smart contract are described as follows.

1) *Task Offloading Request From v_m* : Initially, each v_m chooses nearby d_n for task offloading, so it transmits the offloading request with the maximum delay requirement T_{max} to \mathcal{P}_{comp} , which can be expressed as follows.

$$v_m \rightarrow \mathcal{P}_{comp} : content = Sig_{v_m} || Sig_{d_n} || T_{max} \quad (32)$$

Once \mathcal{P}_{comp} receives a request, it retains the delay requirement T_{max} while executing the smart contract in the blockchain. \mathcal{P}_{comp} sends the certificate of d_n to the v_m and

v_m to the d_n , respectively. In this way, both entities can verify the validity of each other, which is formulated as follows.

$$\begin{aligned} \mathcal{A}_{auth} &\rightarrow v_m : content = C_{d_n} \\ \mathcal{A}_{auth} &\rightarrow d_n : content = C_{v_m} \end{aligned} \quad (33)$$

2) *Task Offloading to d_n* : Once receiving a response from d_n , the local gateway device offloads the tasks $\gamma_{v_m}^{S_r}$ to the edge server and the offloaded tasks are further constructed as a Merkle hash tree at the edge server as leaf nodes. Then, v_m adds the public key of d_n in its certificate for encryption and sends them back to the d_n , which is given as follows.

$$v_m \rightarrow d_n : content = E_{K_{d_n}^{pub}}(t_k) \quad (34)$$

Next, v_m generates the root value of Merkle hash root as \mathcal{R}_{m1} and sends to the \mathcal{P}_{comp} , depicted as follows.

$$v_m \rightarrow \mathcal{P}_{comp} : content = Sig_{v_m} || \mathcal{R}_{m1}. \quad (35)$$

3) *Task Processing in d_n* : In this step, d_n uses the private key for decryption and starts to process the received task $\gamma_{v_m}^{S_r}$. Further, d_n creates a new hash root value \mathcal{R}_{m2} based on the processed result.

4) *Feedback of Processed Results*: In this step, v_m sends \mathcal{R}_{m2} value to the \mathcal{P}_{comp} to verify the results, which is expressed as follows.

$$d_n \rightarrow \mathcal{P}_{comp} : content = Sig || \mathcal{R}_{m2} \quad (36)$$

Further, \mathcal{P}_{comp} compares the value of \mathcal{R}_{m1} and \mathcal{R}_{m2} from v_m and d_n , respectively. Thus, d_n reduces the threats in data computation instead of directly using the offloaded task to generate \mathcal{R}_{m2} . Thus, if the values are equal, then d_n cannot submit \mathcal{R}_{m2} value to the \mathcal{P}_{comp} within the delay requirement. The task offloading failure events are recorded in the block. Otherwise, \mathcal{P}_{comp} assumes that d_n has completed executing its task. Similarly, d_n includes the public key of v_m in its certificate for encrypting the processed result and sends it back to the v_m , which is defined as follows.

$$d_n \rightarrow v_m : content = E_{K_{v_m}^{pub}}(results) \quad (37)$$

5) *Transaction Settlement*: Once the feedback is received, v_m includes its private key $K_{v_m}^{pri}$ for decrypting the processed results and creates a hash root value \mathcal{R}_{m3} to construct a Merkle tree using the value of \mathcal{R}_{m1} and processed results. Further, it sends the \mathcal{R}_{m3} to the \mathcal{P}_{comp} , which is given as follows.

$$v_m \rightarrow \mathcal{P}_{comp} : content = Sig_{v_m} || \mathcal{R}_{m3} \quad (38)$$

After comparing the value of \mathcal{R}_{m2} and \mathcal{R}_{m3} , the transaction is settled by the \mathcal{P}_{comp} . If both values are equal, then it is assumed that v_m has received the processed results while meeting the delay and security requirements. Further, the task offloading success events are recorded in the block.

6) *Trust Assessment of d_n* : The trustfulness of d_n is measured based on the belief and channel uploading rate probability $r_{v_m}^{ch}$. Let $\eta_{v_m,r}$, $\eta'_{v_m,r}$, and \mathcal{U} , represent the belief, disbelief, and uncertainty, respectively. The opinion vector can be represented by $\theta_{v_m,r} = \{\eta_{v_m,r}, \eta'_{v_m,r}, \mathcal{U}_{v_m,r}\}$. Thus, the trust assessment of d_n is measured as follows.

$$trust_{(m,r)} = \eta_{v_m,r} r_{v_m}^{ch} \quad (39)$$

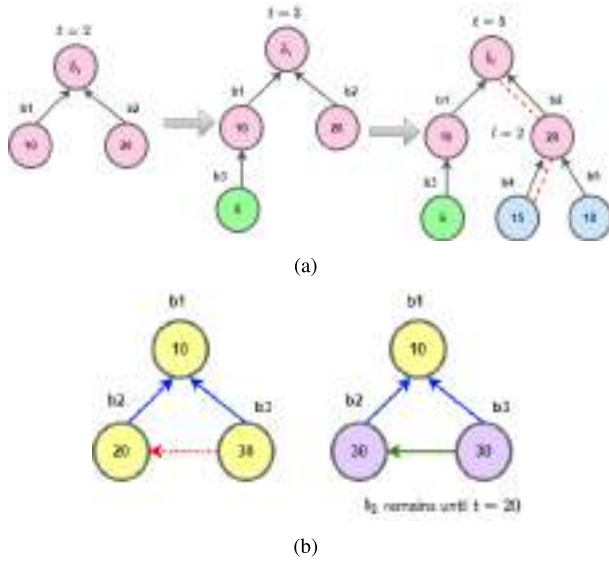


Fig. 3. Blockchain-based task handling (a) Block insertion (b) Block expiration.

7) *Belief Estimation of d_n* : The assessment of trust is a time-varying process at d_n , where the recent $\gamma_{v_m}^{S_r}$ offloading events from g_l creates a larger impact than the past offloading events. Similarly, the failure events of offloading tasks can create high impact than the success events. Therefore, in the proposed strategy, the weights are assigned to recent and past offloading events, which is expressed as \mathcal{E} and $(1 - \mathcal{E})$, respectively, where $\mathcal{E} \in (0.5, 1)$. Similarly, the assigned weights for task failure and success events are defined as $\bar{\vartheta}$ and $(1 - \bar{\vartheta})$, respectively, where $\bar{\vartheta} \in (0.5, 1)$. Let x_n and y_n be the number of success and failure events, respectively. Then, the belief estimation error is computed as follows.

$$\begin{aligned} x_n &= \mathcal{E}(1 - \bar{\vartheta})x_n^{recent} + (1 - \mathcal{E})(1 - \bar{\vartheta})x_n^{past} \\ y_n &= \mathcal{E}\bar{\vartheta}y_n^{recent} + (1 - \mathcal{E})\bar{\vartheta}y_n^{past} \end{aligned} \quad (40)$$

where x_n^{recent} and y_n^{recent} indicate the number of recent success and failure events, respectively. Further, it will be recorded in the blockchain while x_n^{past} and y_n^{past} denote the number of success and failure events, respectively.

C. Task Handling Using Blockchain

Blockchain is a type of distributed or decentralized ledger that stores all the encrypted blocks of data across a peer-to-peer network. In this process, the local edge server d_n uses blockchain to record the information about the tasks in the form of transactions, received from \mathcal{V} . However, it has limited storage compared to the remote cloud server. Therefore, it cannot permanently store the full history of frequently generated tasks from \mathcal{V} . Moreover, in the MTS, the IoT data collected by the v_m can change over time. To overcome this issue, a scalable MiTiChain is introduced to manage the outdated and lifetime of the blocks at d_n . The MiTiChain process contains the following steps:

1) *Block Lifetime*: Each block consists of different transactions with different end times. If the lifetime of any block is expired, it will be deleted from the blockchain. In another

case, if any block is deleted before its end time, it will be recreated again. Otherwise, the block is disconnected from the blockchain. Based on the end time ordering, MiTiChain constructs a graph by merging the hash root value of a new block with an existing block.

2) *Block Insertion*: Let b_i , t_i , and e_i be the block, timestamp at which the block is created, and the end time of a block. The new block is inserted with the existing blocks b_i^* . If the end time of genesis blocks is later than the new block b_i , it will be connected to the block with the earliest end time. In this case, the end time of all genesis blocks is earlier than the new block and it will be connected to the previous block \bar{b}_i with an infinite lifetime. A sample block insertion procedure is shown in Fig.3(a).

3) *Block Height*: Sometimes, the height of b_i can be short or shallow. The height of the chain increases due to the deletion of blocks. In MiTiChain, the height of a chain is increased by merging two sets of edges from end-time ordering (EO) and arrival ordering of blocks. Block b_i contains two directed edges from b_i to b_i^* and \bar{b}_i . Further, if the chain is longer, it is difficult to undo and the blockchain is always considered to be more secure and valid. For instance, as depicted in Fig.3(a), the height of b_4 is 2 using only EO.

4) *Block Expiration*: If the endtime of a block b_i is expired, then it is deleted from the blockchain. However, when there is a block called b_{i+1} , it has arrived before b_i and its end time e_{i+1} is later than e_i . Moreover, if the endtime of b_{i+1} is not yet expired, so b_i can be deleted from the chain and it remains in the chain until e_{i+1} of a block b_{i+1} is expired. As shown in Fig.3(b), the endtime of b_2 is $e_3 = 20$. However, b_2 cannot be deleted at e_2 but it remains in the chain until $t = 30$. Thus, the expiry time of b_2 is extended by 10 time units.

5) *Block Renewal*: The block is renewed by creating a new copy and deleting the old block. If any block needs to be deleted at t , then it immediately checks whether the block contains the indefinite transactions or not. If it is present, a new block is created with lifetime l and inserted into a blockchain. Otherwise, the old block is deleted based on the previous request. The pseudocode of the overall task handling using the MiTiChain strategy is depicted in Algorithm 2.

D. Maritime Data Analysis at Edge Server

Rapid digitalization and the proliferation of IoT devices introduce new types of threats and cyber incidents in the maritime network. With the inclusion of an intelligent predictive model at the edge server, the delay-sensitive information collected from \mathcal{V} can enhance the security and safety operation of maritime vessels that can predict attacks earlier, detect and prevent cyber incidents, prioritize the operations, and provide instant response to the vessel control unit. Many research efforts have been conducted using machine learning classification models in the literature. However, the proposed strategy uses both Support Vector Machine (SVM) and Convolution Neural Network (CNN) as the base model to constructively analyze and detect the malicious vessels and identify the trusted vessels to perform safety operations with higher accuracy. The key idea to select the SVM and CNN classification

Algorithm 2 Task Handling using MiTiChain

```

INPUT :  $b_i$ : Block,  $\bar{b}_i$ : Previous block,  $b_i^*$ : Parent block,  $\tau$ :
          Indefinite lifetime,  $e_i$ : Endtime,  $e_r$ : Expirytime,  $l$ : Lifetime
OUTPUT : Create a block  $b_i$  with endtime  $e_i$ 
1 begin
2   if ( $e_i(b_i) \geq b_i$ ) then
3     | Connect  $b_i$  to the  $\bar{b}_i$  with earliest  $e_i$ 
4   end
5   if ( $e_i(b_i) \leq b_i$ ) then
6     | Connect  $b_i$  to the  $\bar{b}_i$  with infinite  $l$ 
7   end
8   Increase the height of the chain by merging  $b_i \rightarrow \bar{b}_i$  and  $b_i \rightarrow b_i^*$ 
9   if ( $e_i(b_i) == e_r$ ) then
10    | Delete  $b_i$  from the blockchain
11  end
12  if ( $e_i(b_i) \neq e_r$ ) then
13    | Extend the  $e_r$  of  $b_i$  by 10 time units
14  end
15  if  $b_i == \tau$  then
16    | Check the block  $b_i$  contains the transactions with  $\tau$  and not
17    | requested to be deleted
18  end
19  if  $b_i \neq \tau$  then
20    | Delete the old block which are already requested
21  end

```

model is to handle high-dimensional vessel IoT data in the maritime network and increase the prediction accuracy with minimum power consumption.

IV. NUMERICAL ANALYSIS

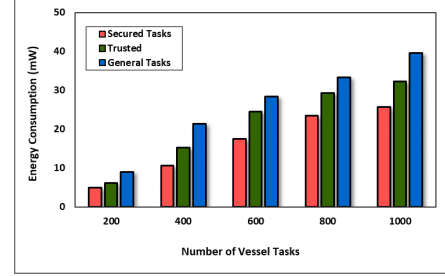
This section describes the numerical evaluation of the proposed strategy and predictive analysis model for IoT-enabled maritime networks. Here, we first evaluate the proposed strategy according to the delay and power consumption of the edge-centric framework of the MTS. The quantitative results are compared with two baseline schemes [15], [16] to show the efficiency of the proposed strategy. Further, we apply SVM and CNN models over maritime vessel dataset at both edge and cloud servers to study the various effect of the predictive analysis model. Besides that, the prediction results are compared with the individual model under different validation metrics such as accuracy, precision, recall, and F1.

A. Simulation Setup and Dataset

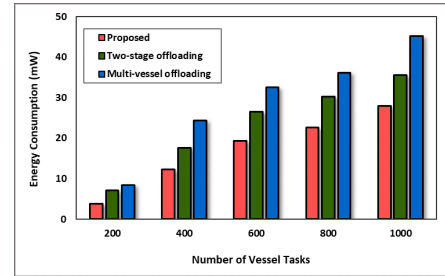
The simulation parameters for the maritime network are listed in Table.I. The proposed maritime network consists of 100 vessel IoT devices and each vessel generates 100 tasks/sec at each time interval. The maximum data rate is set to 7600 *kbps* and the total communication range is 120 *km*. The total size of vessel input tasks \mathcal{T}_m is [200 *kb*-1000 *kb*]. The vessel's task arrival rate to the edge server $\lambda_{v_m}^{edge}$ is 0.120 *ms*. For predictive analysis, we consider the visible maritime image (VMI) dataset [19] for classification, which consists of 3750 images and 25 fine-grained categories of maritime vessels. Additionally, we include 20 malicious vessels to effectively identify and validate the secure edge-centric predictive model. Each category of the vessel contains 150 navigation images and we split the dataset into the ratio of 80% for the training set and 20% for the test set, respectively.

TABLE I
SIMULATION PARAMETERS

Parameters	Values
Number of Vessels with IoT devices (\mathcal{V})	100
Number of Edge servers (\mathcal{D})	5
Number of cloud servers (\mathcal{C})	2
Number of local gateway devices (\mathcal{G})	100
Number of vessel IoT data (λ_{v_m})	100 [tasks/sec]
CPU processing power consumption ($P_{v_m}^{local}$)	0.8 Joules
Transmission power of local gateway devices (T^I)	1 mW



(a)



(b)

Fig. 4. Comparative results of power consumption (a) Different vessel tasks (b) Analysis with baseline schemes.

B. Simulation Results

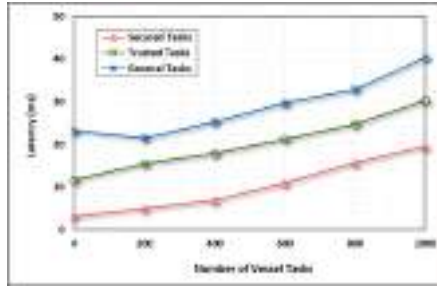
This section examines the simulation results of the proposed service provisioning strategy and predictive analysis model in terms of various performance metrics.

1) *Analysis of Power Consumption*: Fig.4 shows the comparative results of $P_{v_m}^{local}$ with different types of vessel tasks \mathcal{T} and baseline schemes. As shown in Fig.4(a), it is noticed that the total power consumption $P_{v_m}^{local}$ while offloading and processing the secured tasks S_r to edge server is less as compared to the trusted $trust_{(m,r)}$ and general tasks G_r , respectively. The main reason behind that the proposed service provisioning strategy distributes the incoming tasks on the local edge servers with minimum power consumption. Similarly, Fig.4(b) depicts the comparative results of average $P_{v_m}^{local}$ of the proposed strategy with existing baseline schemes. From the simulation results, it can be observed that the total power consumption of the proposed strategy is minimum (28.03 *mW*) as compared to the existing two-stage (35.64 *mW*) and multi-vessel (45.18 *mW*) offloading schemes. The quantitative results show that the proposed strategy reduces the average power consumption by 7.61% and 17.15%, respectively over the baseline schemes.

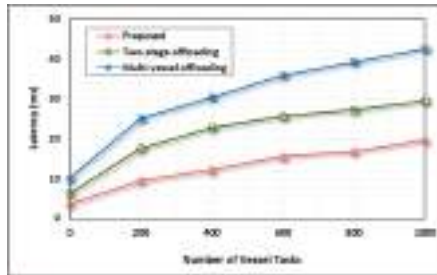
2) *Analysis of Delay*: Fig.5 depicts the comparative analysis results of delay $T_{v_m}^{local}$ with different vessel tasks \mathcal{T} and

TABLE II
ACCURACY OF VARIOUS LEARNING MODELS IN EDGE SERVER

Edge Level Analysis							
Dataset	MLCA Models	Malicious Detection		Accuracy	Precision	Recall	F1 Score
		Trusted	Malicious				
VMI	SVM	0.6921	0.5304	0.8815	0.8730	0.8542	0.8639
	CNN	0.5568	0.4975	0.8243	0.8349	0.8176	0.8485
	CNN+SVM	0.7254	0.8231	0.9567	0.9259	0.8918	0.9176



(a)

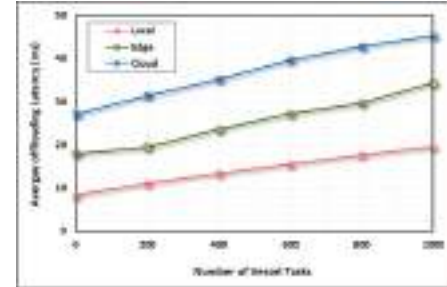


(b)

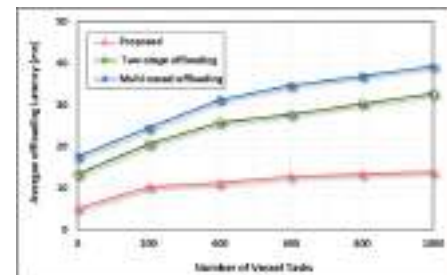
Fig. 5. Comparative results of delay (a) Different vessel tasks (b) Analysis with baseline schemes.

baseline schemes. In Fig.5(a), it is observed that the secured tasks require less time (19.45 ms) than the trusted and general tasks, respectively. The reason behind that proposed service provision strategy distributes the tasks as per their priorities. The average delay of the proposed strategy over the existing baseline schemes is shown in Fig.5(b). From the figure, it is observed that the average delay of two-stage (29.53 ms) and multi-vessel (42.61 ms) strategy increases with the arrival of incoming vessel tasks. However, the average delay of the proposed strategy reduces (19.76 ms) as compared with existing baseline strategies. The proposed strategy has minimized the average delay by 9.77% and 22.85% over two-stage and multi-vessel strategies, respectively.

3) *Analysis of Offloading Delay Under Belief Estimation:* Fig.6 shows the impact of offloading delay based on the belief estimation and the number of incoming vessel tasks. As shown in Fig.6(a), the processing delay of various tasks in v_m is lower than the offloading tasks to the remote computing device d_n and c_o . The multiple successive events \hat{v} or beliefs are estimated based on the recent offloading secured tasks, which reduces the average offloading delay of the v_m . In Fig.6(b), the comparative analysis of average offloading delay is evaluated with different tasks \mathcal{T} . It is noted that the average offloading of the proposed strategy (13.92 ms) is lower than the existing baseline schemes. The average delay of both



(a)



(b)

Fig. 6. Comparative results of offloading delay under belief estimation error (a) Various maritime devices (b) Analysis with baseline schemes.

existing schemes (32.73 ms and 39.25 ms) increases while increasing the task offloading failure. The proposed strategy outperforms the existing offloading schemes by 18.81% and 25.33%, respectively.

4) *Predictive Analysis at Edge Server:* The prediction result of the standard learning model is shown in Table.II. Upon offloading the secured tasks to the edge server, the tasks are placed in the form of transactions to create a blockchain. The group of secured transactions in a block is analyzed based on the end time and lifetime. In Fig.7(a), it is observed that the lifetime of secured tasks is higher than the trusted and general tasks, respectively. However, it may vary according to the end time. The remaining tasks are placed into the remote cloud server c_o for further analysis. As shown in Fig.7(b), the combination of both the SVM and CNN model identifies the number of malicious d_n and provides better accuracy than the other two models, which enhances the accuracy by 82.31% and 72.54 %, respectively. Thus, the proposed model improves 95.67% prediction accuracy, which is higher than the existing models. Hence, the proposed strategy along with the combined learning models at the edge network improves the detection rate of malicious servers and cyber incidents, prevention of various attacks, and enhances the prediction accuracy of the vessel's monitoring operations and services in a secured manner. All these constraints are effectively satisfied

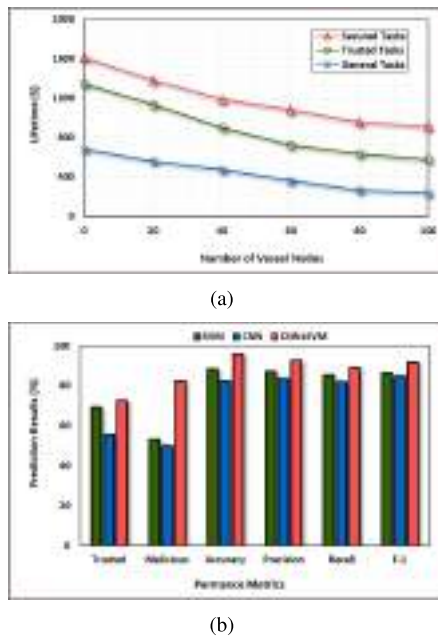


Fig. 7. Edge-level prediction results (a) Lifetime under different vessel tasks (b) Analysis with baseline models.

by considering the secured-tasks offloading, tasks handling using both smart-contract and blockchain, and fine-grained maritime vessel dataset.

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have introduced a secure edge-centric service provisioning strategy for IoT-enabled MTS and a collaborative predictive analysis model at the edge networks. This work is mainly focused on service provisioning of the secured tasks based on the degree of importance from the local gateway device to the maritime edge servers while reducing the delay and power consumption. To achieve this, a task queuing priority-based algorithm is designed to classify the various incoming tasks from the maritime vessels. The secured tasks are efficiently stored, validated, and handled in various blocks by jointly considering the smart contract and blockchain technique. Further, the combined SVM and CNN model is applied to instantly predict the number of malicious entities and improve the prediction accuracy of vessel monitoring units in edge networks. The numerical results show that the average delay and power consumption of the proposed strategy is minimized by 7.61%-17.15% and 9.77%-22.15% over existing baseline schemes. Further, the lifetime of each block is efficiently utilized based on the expiration time and the predictive model enhances the accuracy by 95.67%. In the future, we will extend this work by incorporating blockchain with graph theory to study the various effect of blocks, increase the security of transactions, and predict the cyber risks with higher accuracy at the edge server.

REFERENCES

[1] A. Hazra, M. Adhikari, T. Amgoth, and S. N. Srirama, "Joint computation offloading and scheduling optimization of IoT applications in fog networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 4, pp. 3266–3278, Oct. 2020.

[2] Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, "Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1796–1825, May 2020.

[3] V. F. Arguedas, G. Pallotta, and M. Vespe, "Maritime traffic networks: From historical positioning data to unsupervised maritime traffic monitoring," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 722–732, Mar. 2018.

[4] X. Su, L. Meng, and J. Huang, "Intelligent maritime networking with edge services and computing capability," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13606–13620, Nov. 2020.

[5] J. O. Eichenhofer, E. Heymann, B. P. Miller, and A. Kang, "An in-depth security assessment of maritime container terminal software systems," *IEEE Access*, vol. 8, pp. 128050–128067, 2020.

[6] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "Privacy-preserved task offloading in mobile blockchain with deep reinforcement learning," *IEEE Trans. Netw. Service Manage.*, vol. 17, no. 4, pp. 2536–2549, Dec. 2020.

[7] S.-W. Jo and W.-S. Shim, "LTE-maritime: High-speed maritime wireless communication based on LTE technology," *IEEE Access*, vol. 7, pp. 53172–53181, 2019.

[8] K.-L.-A. Yau, S. Peng, J. Qadir, Y.-C. Low, and M. H. Ling, "Towards smart port infrastructures: Enhancing port activities using information and communications technology," *IEEE Access*, vol. 8, pp. 83387–83404, 2020.

[9] S. Gao, T. Yang, H. Ni, and G. Zhang, "Multi-armed bandits scheme for tasks offloading in MEC-enabled maritime communication networks," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2020, pp. 232–237.

[10] P. S. Sangeerth and K. V. Lakshmy, "Blockchain based smart contracts in automation of shipping ports," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1248–1253.

[11] 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.

[12] A. Lakhani, M. Ahmad, M. Bilal, A. Jolfaei, and R. M. Mehmood, "Mobility aware blockchain enabled offloading and scheduling in vehicular fog cloud computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4212–4223, Jul. 2021.

[13] Y. Qu *et al.*, "Decentralized privacy using blockchain-enabled federated learning in fog computing," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5171–5183, Jun. 2020.

[14] C. K. Pyoung and S. J. Baek, "Blockchain of finite-lifetime blocks with applications to edge-based IoT," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 2102–2116, Mar. 2020.

[15] T. Yang, H. Feng, C. Yang, Y. Wang, J. Dong, and M. Xia, "Multivessel computation offloading in maritime mobile edge computing network," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4063–4073, Jun. 2019.

[16] T. Yang *et al.*, "Two-stage offloading optimization for energy-latency tradeoff with mobile edge computing in maritime Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5954–5963, Jul. 2020.

[17] X. Xu, X. Zhang, H. Gao, Y. Xue, L. Qi, and W. Dou, "BeCome: Blockchain-enabled computation offloading for IoT in mobile edge computing," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4187–4195, Jun. 2020.

[18] H. Liao, Y. Mu, Z. Zhou, M. Sun, Z. Wang, and C. Pan, "Blockchain and learning-based secure and intelligent task offloading for vehicular fog computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 7, pp. 4051–4063, Jul. 2021.

[19] R. Zhao, J. Wang, X. Zheng, J. Wen, L. Rao, and J. Zhao, "Maritime visible image classification based on double transfer method," *IEEE Access*, vol. 8, pp. 166335–166346, 2020.

[20] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "SeaShips: A large-scale precisely annotated dataset for ship detection," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2593–2604, Oct. 2018.

[21] D. Qiao, G. Liu, F. Dong, S.-X. Jiang, and L. Dai, "Marine vessel identification: A large-scale dataset and global-and-local fusion-based discriminative feature learning," *IEEE Access*, vol. 8, pp. 27744–27756, 2020.

[22] M. Adhikari, M. Mukherjee, and S. N. Srirama, "DPTO: A deadline and priority-aware task offloading in fog computing framework leveraging multilevel feedback queueing," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5773–5782, Jul. 2020.

[23] A. Munusamy *et al.*, "Service deployment strategy for predictive analysis of FinTech IoT applications in edge networks," *IEEE Internet Things J.*, early access, May 7, 2021, doi: 10.1109/JIOT.2021.3078148.



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-

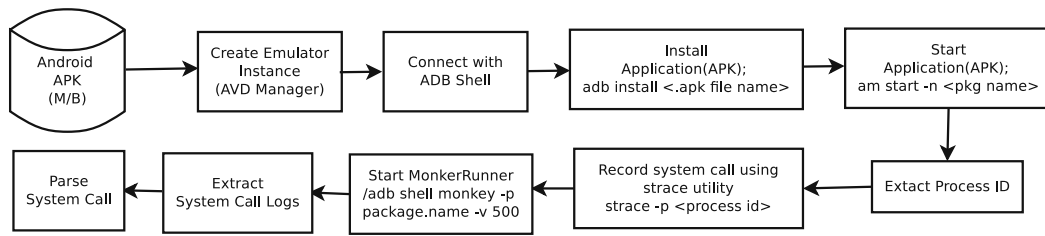


Fig. 1 System call extraction

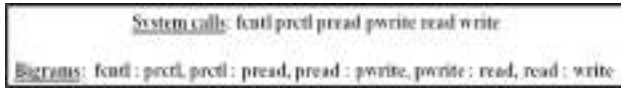


Fig. 2 Sample set of system calls and bigrams

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.

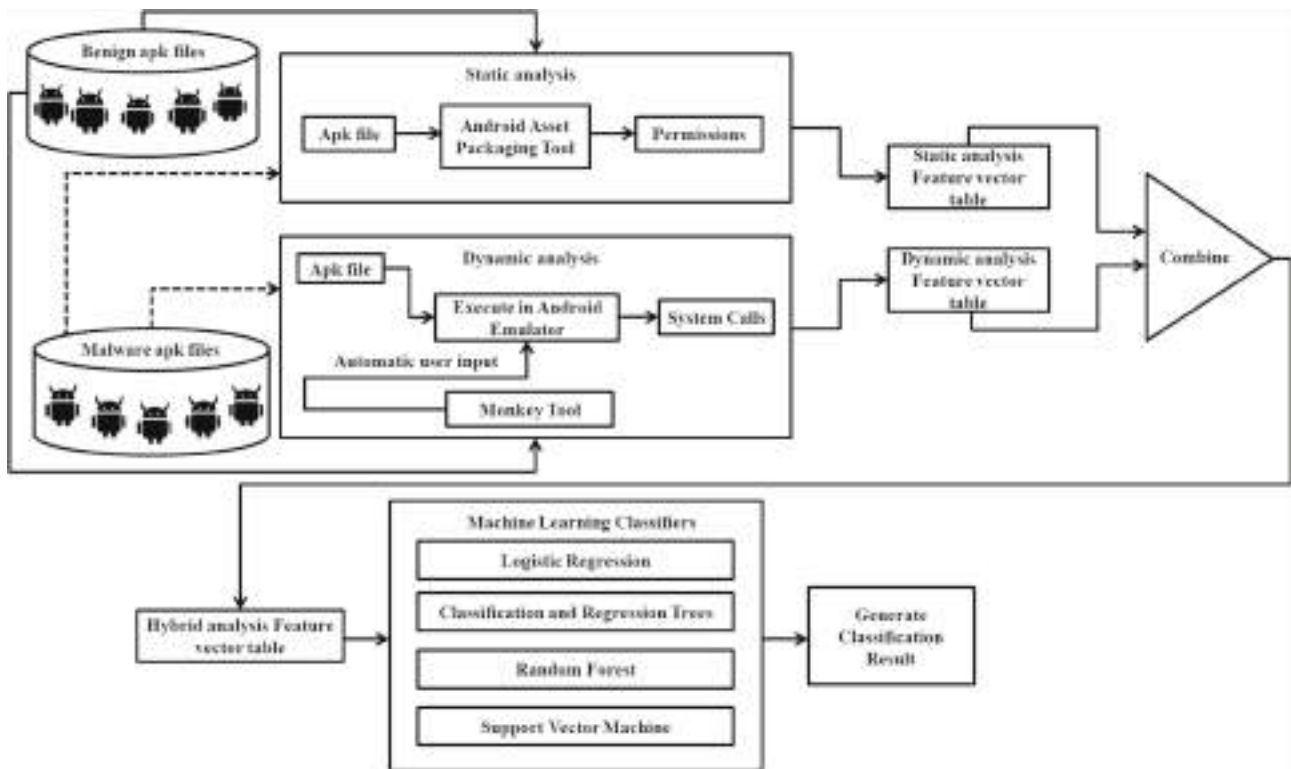


Fig. 3 The architecture of the proposed Hybrid Malware Detector

$F_1, F_2, F_3, \dots, F_p$ 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

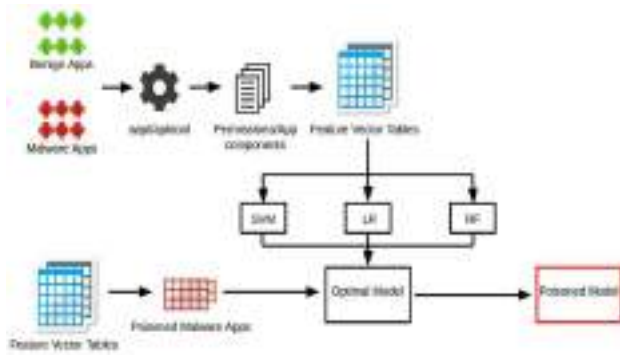


Fig. 4 Adversarial attack

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).

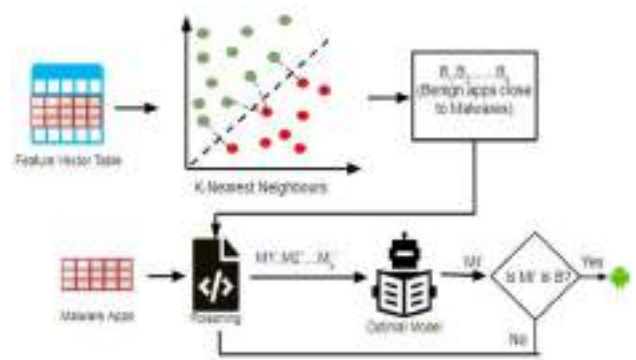


Fig. 5 Evasion Attack based on Hamming Distance

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.

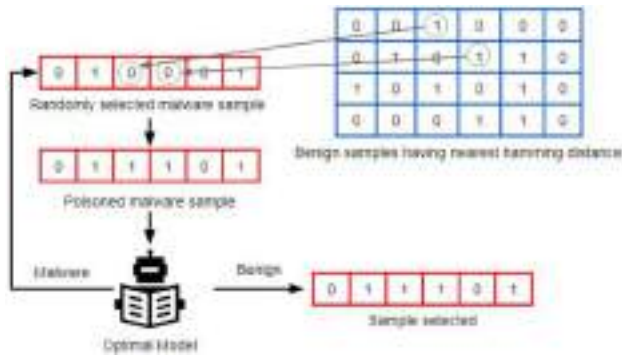


Fig. 6 Evasion attack an example

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.

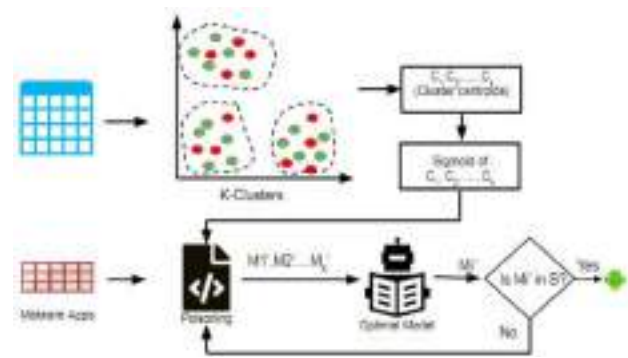


Fig. 7 Poisoning Attack Using KMeans Clustering

- 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

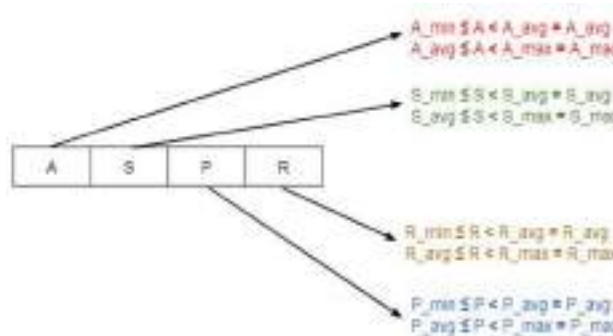


Fig. 8 Modification of app components

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.

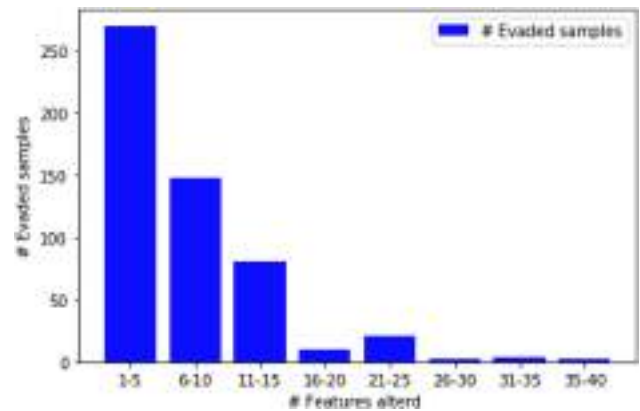


Fig. 9 Number of evaded samples vs number of permissions inserted

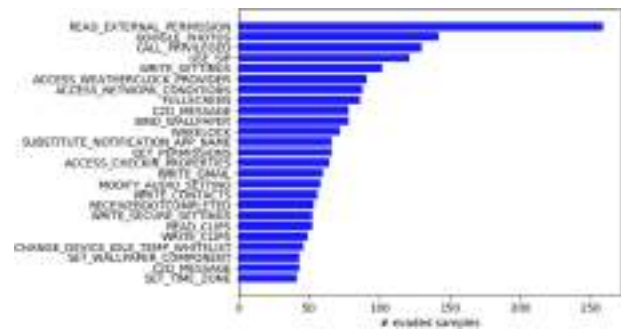


Fig. 10 Inserted permissions in adversarial samples

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

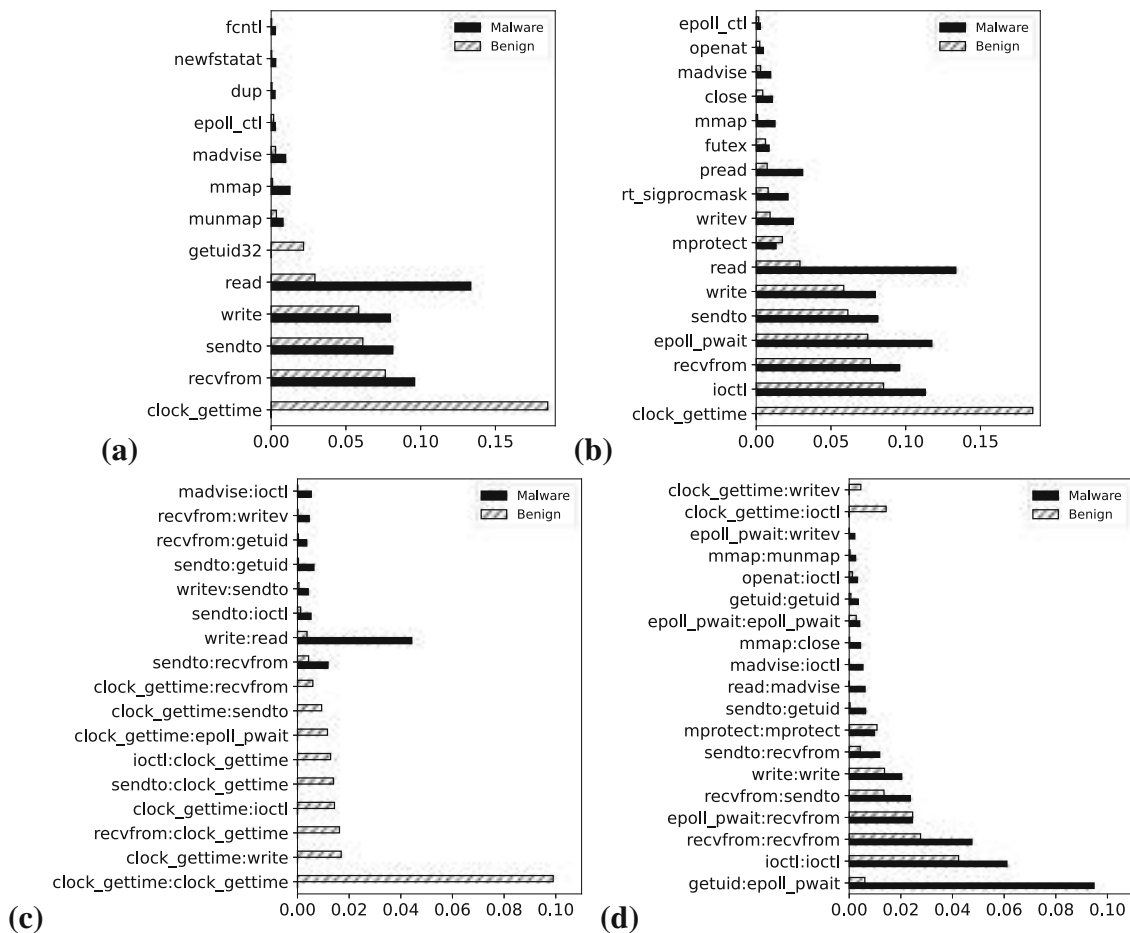


Fig. 11 System call grams **a** uni-gram SelectKBest **b** uni-grams RFE **c** bi-gram SelectKBest and **d** bi-gram RFE

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

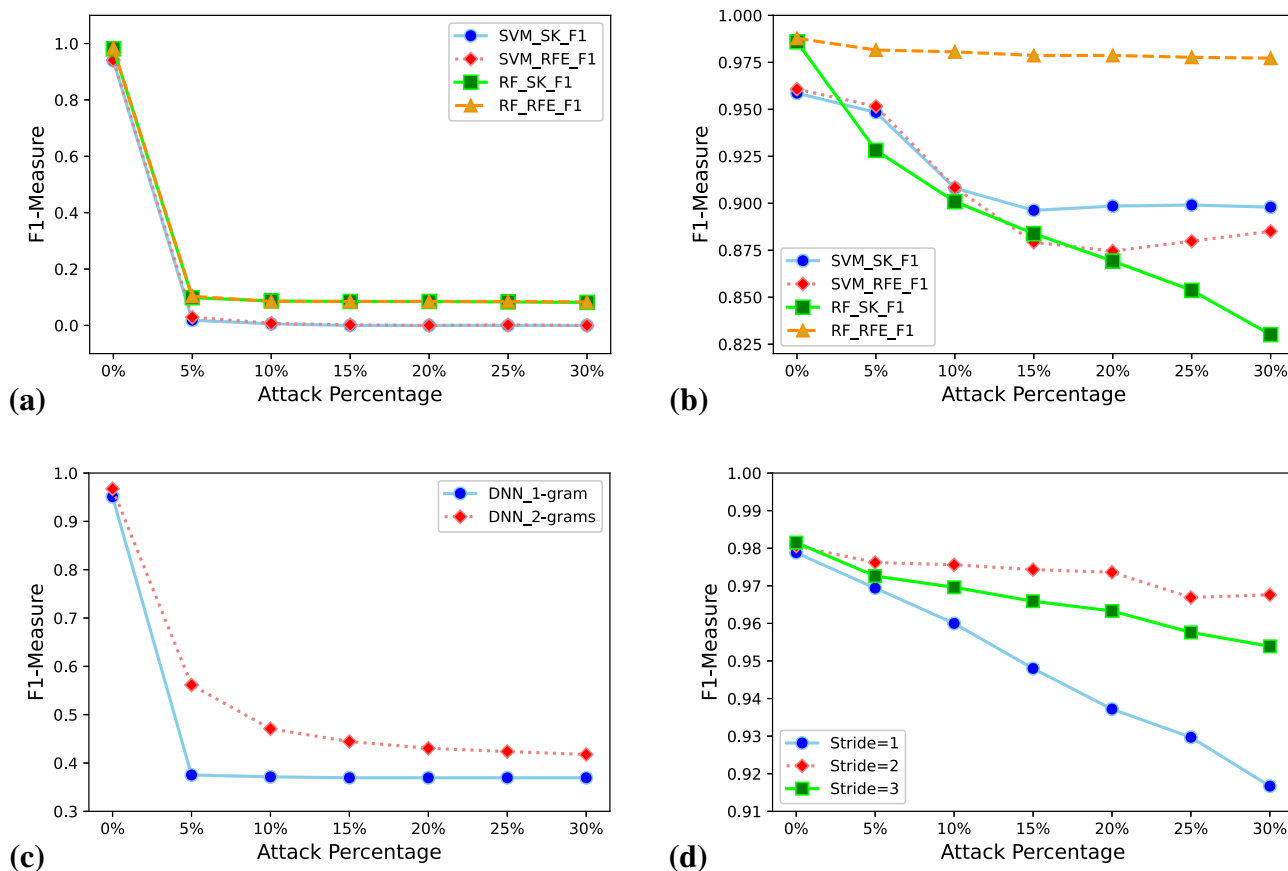


Fig. 12 Evasion attack using a uni-gram SelectKBest b bi-gram RFE c Deep neural network and d 1D-CNN

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

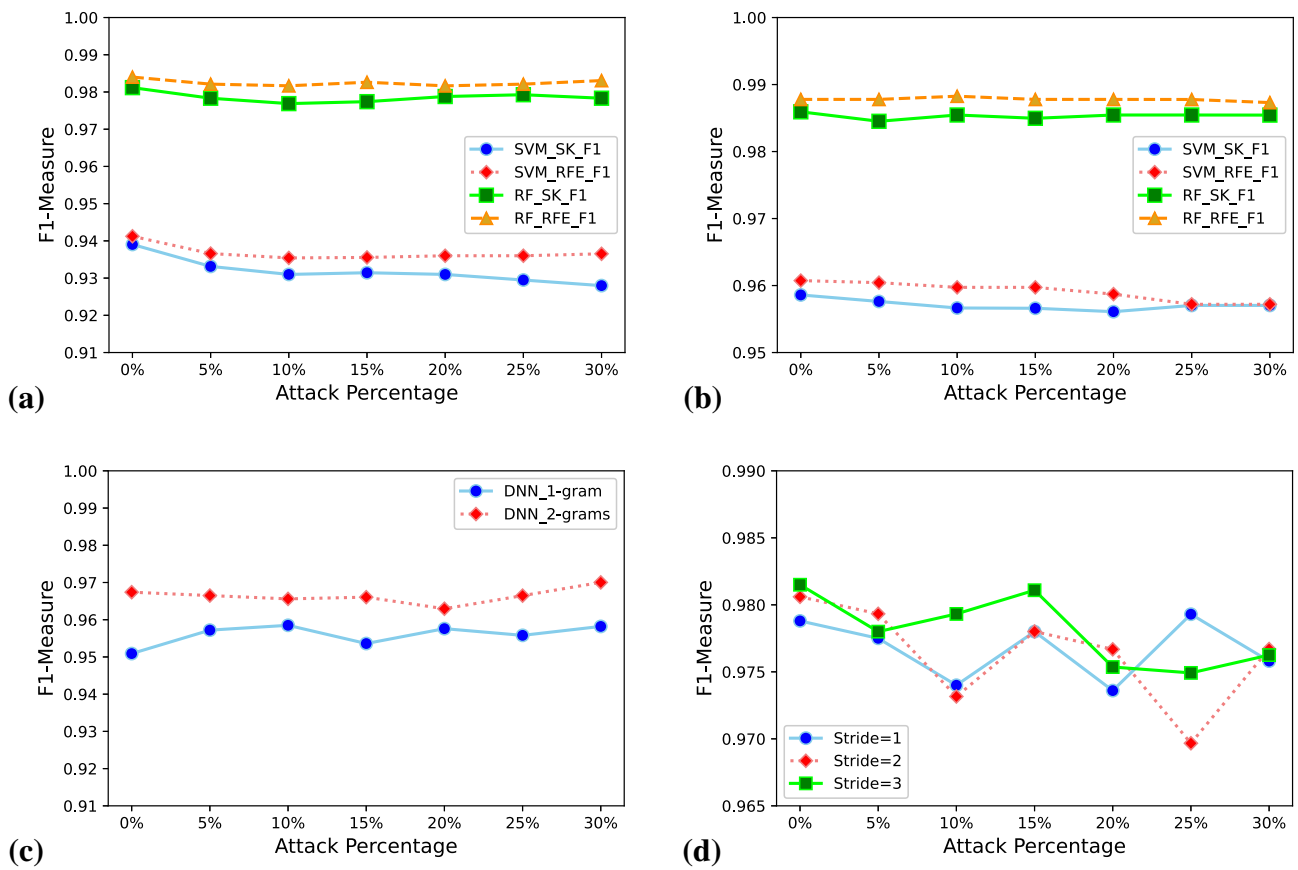


Fig. 13 Poisoning attack employing a uni-gram SelectKBest b bi-gram RFE c Deep neural network and d 1D-CNN

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

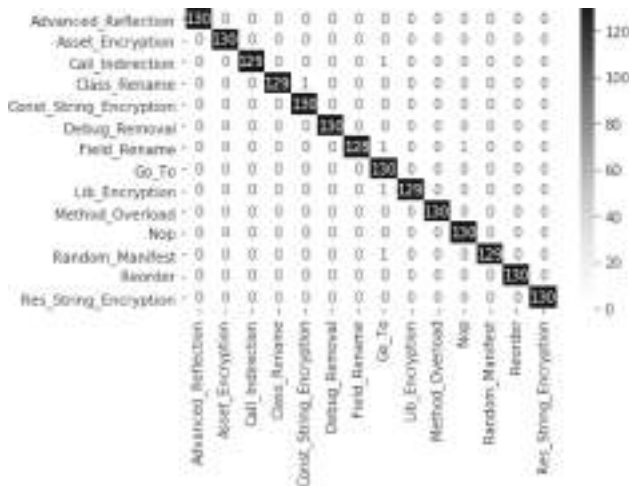


Fig. 14 Classification of obfuscated malware variants

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

- 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)
- 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)
- 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)
- Almin, S.B., Chatterjee, M.: A novel approach to detect android malware. *Procedia Comput. Sci.* **45**, 407–417 (2015)
- Aonzo, S., Georgiu, G.C., Verderame, L., Merlo, A.: Obfuscapck: an open-source black-box obfuscation tool for Android apps. *SoftwareX* **11**, 100403 (2020)
- APKTool: <https://ibotpeaches.github.io/Apktool/install/>
- 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)
- 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)
- 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)
- Biggio, B., Fabio, R.: Wild patterns: ten years after the rise of adversarial machine learning. *Pattern Recognit.* **84**, 317–331 (2018)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- 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)
- Dimjašević, M., Atzeni, S., Ugrina, I., Rakamaric, 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)
- 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)
- Gandotra, E., Bansal, D., Sofat, S.: Malware analysis and classification: a survey. *J. Inf. Secur.* **2014** (2014)
- 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)
- Greengard, S.: Cybersecurity gets smart. *Commun. ACM* **59**(5), 29–31 (2016)
- 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)
- 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)
- 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)
- 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)
- 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)
- Martinelli, F., Marulli, F., Mercaldo, F.: Evaluating convolutional neural network for effective mobile malware detection. *Procedia Comput. Sci.* **112**, 2372–2381 (2017)
- 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)
- MonkeyRunner: <https://developer.android.com/studio/test/monkey>
- Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S.: Malware images: visualization and automatic classification. In: Proceedings

- of the 8th International Symposium on Visualization for Cyber Security, pp. 1–7. (2011)
32. Ni, S., Qian, Q., Zhang, R.: Malware identification using visualization images and deep learning. *Comput. Secur.* **77**, 871–885 (2018)
 33. Patel, K., Buddadev, B.: Detection and mitigation of android malware through hybrid approach. In *International symposium on Security in Computing and Communication*, pp. 455–463. Springer, Cham, (2015)
 34. Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., Cavallaro, L.: TESSERACT: eliminating experimental bias in malware classification across space and time. In: *28th USENIX Security Symposium (USENIX Security 19)*, pp. 729–746. (2019)
 35. Pierazzi, F., Pendlebury, F., Cortellazzi, J., Cavallaro, L.: Intriguing properties of adversarial ML attacks in the problem space. In: *Proceedings of IEEE Symposium on Security and Privacy, 2020*, pp.1332–1349. IEEE (2020)
 36. SL, S.D., Jaidhar, C.D.: Windows malware detector using convolutional neural network based on visualization images. *IEEE Trans. Emerg. Top. Comput.* (2019)
 37. Santos, I., Penya, Y.K., Devesa, J., Bringas, P.G.: N-grams-based file signatures for malware detection. *ICEIS* **9**, 317–320 (2009)
 38. Saracino, A., Sgandurra, D., Dini, G., Martinelli, F.: Madam: effective and efficient behavior-based android malware detection and prevention. *IEEE Trans. Dependable Secure Comput.* **15**(1), 83–97 (2016)
 39. Saxe, J., Berlin, K.: Deep neural network based malware detection using two dimensional binary program features. In: *2015 10th International Conference on Malicious and Unwanted Software (MALWARE)*, pp. 11–20. IEEE (2015)
 40. Sen, S., Aydogan, E., Aysan, A.I.: Coevolution of mobile malware and anti-malware. *IEEE Trans. Inf. Forensics Secur.* **13**(10), 2563–2574 (2018)
 41. Sun, G., Qian, Q.: Deep learning and visualization for identifying malware families. *IEEE Trans. Dependable Secure Comput.* (2018)
 42. Surendran, R., Thomas, T., Emmanuel, S.: GSDroid: graph signal based compact feature representation for android malware detection. *Expert Syst. Appl.* **159**, 113581 (2020)
 43. Tam, K., Khan, S.J., Fattori, A., Cavallaro, L.: Copperdroid: automatic reconstruction of android malware behaviors. In: *Ndss*. (2015)
 44. Ucci, D., Leonardo, A., Roberto, B.: Survey of machine learning techniques for malware analysis. *Comput. Secur.* **81**, 123–147 (2019)
 45. Vinayakumar, R., Soman, K.P.: DeepMalNet: evaluating shallow and deep networks for static PE malware detection. *ICT Express* **4**(4), 255–258 (2018)
 46. Wang, X., Yang, Y., Zeng, Y., Tang, C., Shi, J., Xu, K.: A novel hybrid mobile malware detection system integrating anomaly detection with misuse detection. In: *Proceedings of the 6th International Workshop on Mobile Cloud Computing and Services*, pp. 15–22. (2015)
 47. Wu, W.C., Hung, S.H.: DroidDolphin: a dynamic Android malware detection framework using big data and machine learning. In: *Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems*, pp. 247–252. (2014)
 48. Xiao, X., Wang, Z., Li, Q., Xia, S., Jiang, Y.: Back-propagation neural network on Markov chains from system call sequences: a new approach for detecting Android malware with system call sequences. *IET Inf. Secur.* **11**(1), 8–15 (2017)
 49. Xu, W., Qi, Y., Evans, D.: Automatically evading classifiers. In: *Proceedings of the 2016 Network and Distributed Systems Symposium*, vol. 10. (2016)
 50. Xue, Y., Meng, G., Liu, Y., Tan, T.H., Chen, H., Sun, J., Zhang, J.: Auditing anti-malware tools by evolving android malware and dynamic loading technique. *IEEE Trans. Inf. Forensics Secur.* **12**(7), 1529–1544 (2017)
 51. Zhang, S., Xiao, X.: Cscdroid: Accurately detect android malware via contribution-level-based system call categorization. In *2017 IEEE Trustcom/BigDataSE/ICSS*, pp. 193–200. IEEE (2017)
 52. Zhou, M.: A hybrid feature selection method based on fisher score and genetic algorithm. *J. Math. Sci. Adv. Appl.* **37**(1), 51–78 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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. BT'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].

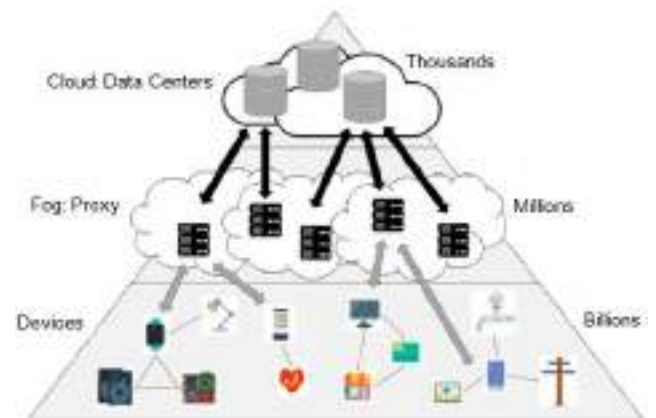


FIGURE 1. Fog computing structure [4].

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.

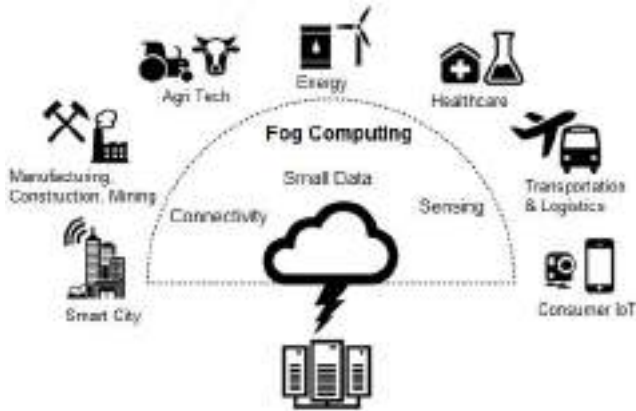


FIGURE 2. Fog computing applications.

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.

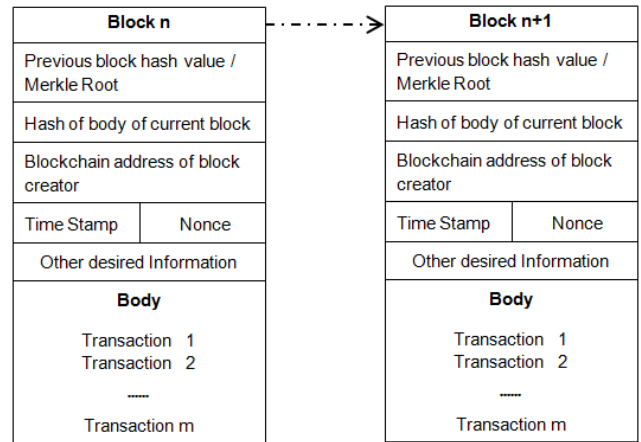


FIGURE 3. Block structure.

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.

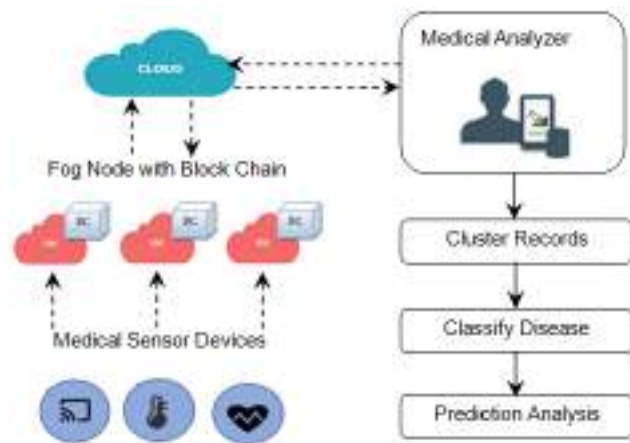


FIGURE 4. Proposed system model.

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.

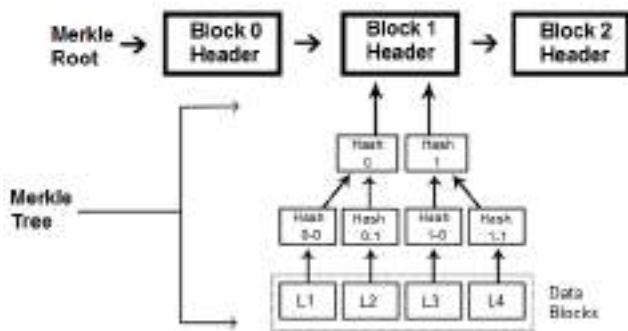


FIGURE 5. Merkle tree structure in blockchain.

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.

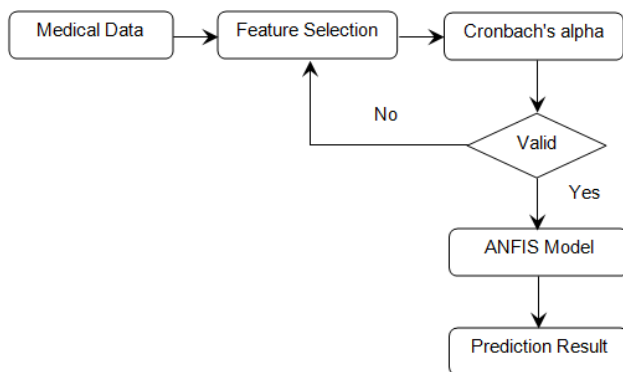


FIGURE 6. Prediction work flow.

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.

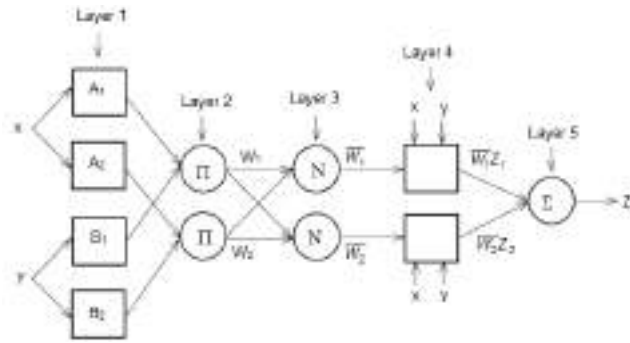


FIGURE 7. ANFIS architecture.

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

$$L_i^1 = \mu_{A_i}(x) \quad (i = 1, 2) \quad (2)$$

$$L_i^1 = \mu_{B_{(i-2)}}(y) \quad (i = 3, 4) \quad (3)$$

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

$$L_i^2 = \omega_i = \mu_{A_i}(x) \mu_{B_i}(x) \quad (i = 1, 2) \quad (4)$$

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,

$$L_i^3 = \bar{\omega}_i = \frac{\omega_i}{\sum \omega_i} = \frac{\omega_1}{\omega_1 + \omega_2}, \quad (i = 1, 2) \quad (5)$$

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

$$L_i^4 = \bar{\omega}_i \cdot f_i = \bar{\omega}_i (p_i x + q_i y + r_i), \quad (i = 1, 2) \quad (6)$$

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.

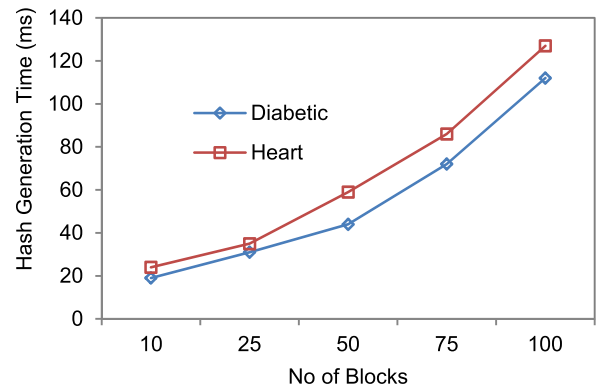


FIGURE 8. Blockchain hash generation time.

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.

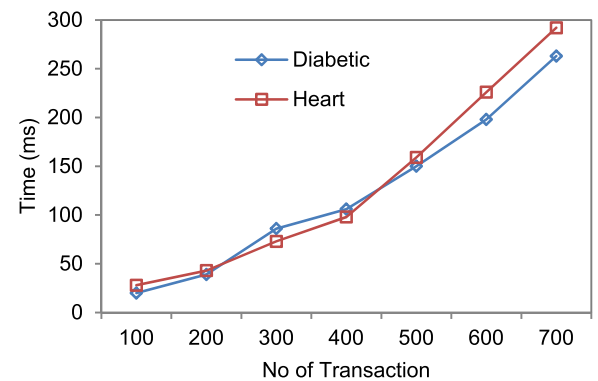


FIGURE 9. Transaction creation time.

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]

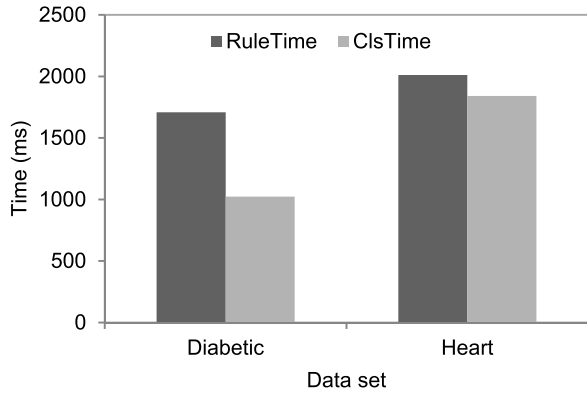


FIGURE 10. Execution time for rule and cluster formation.

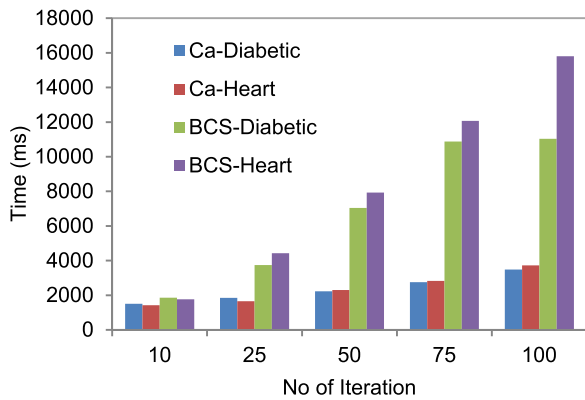


FIGURE 11. Feature selection running time.

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

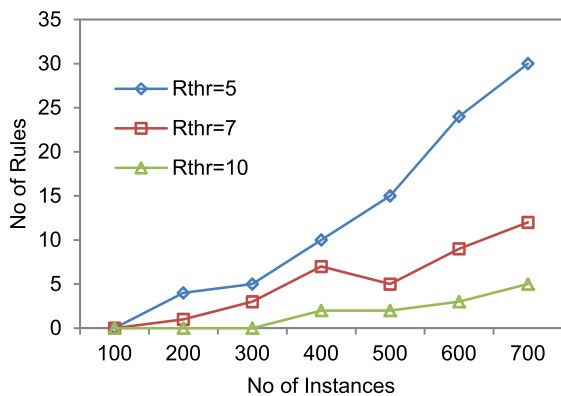


FIGURE 12. Instances vs rules for diabetic data.

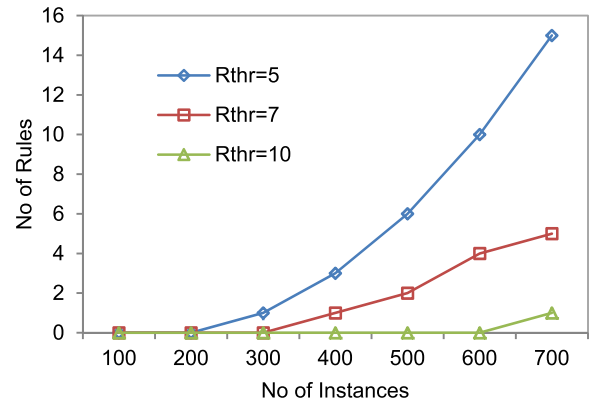


FIGURE 13. Instances vs rules for heart data.

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 $R_{thr} = 5$.

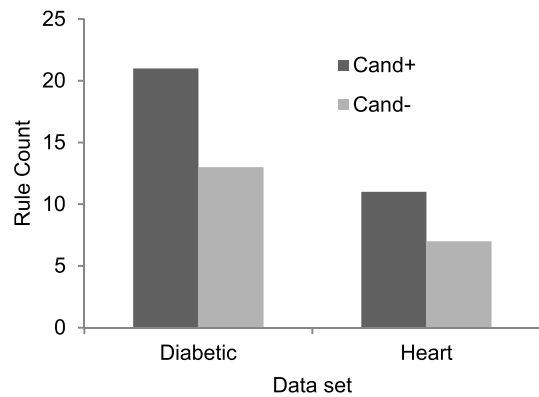


FIGURE 14. Candidate rule count $R_{thr} = 5$.

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.

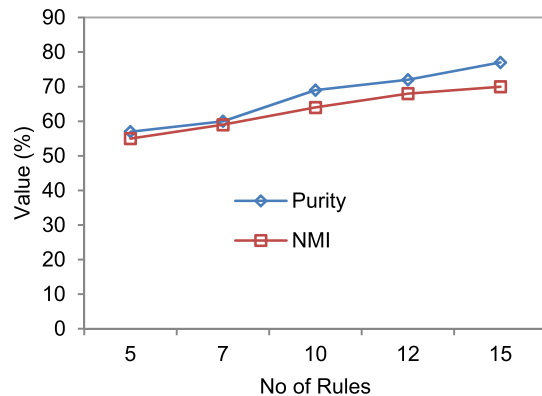


FIGURE 15. Purity and NMI for diabetic data.

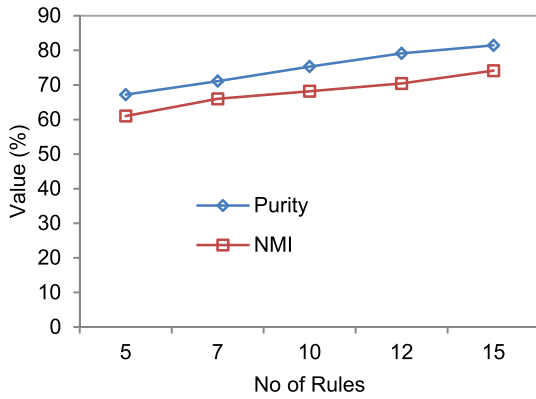


FIGURE 16. Purity and NMI for heart data.

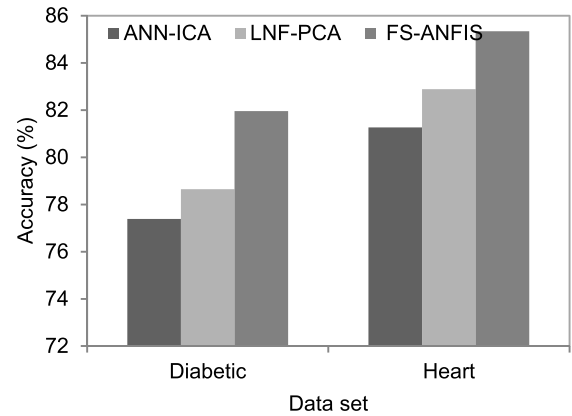


FIGURE 19. Accuracy comparison.

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.

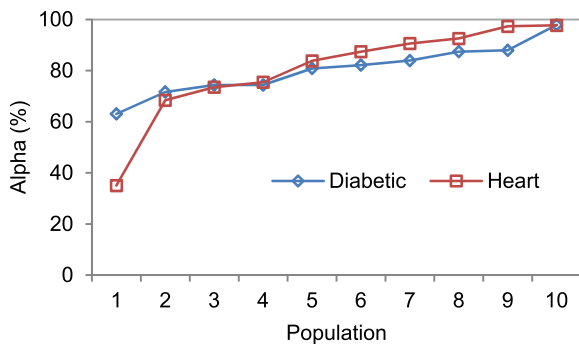


FIGURE 17. Alpha for different population.

Fig 18 shows the alpha value for 100 iterations.

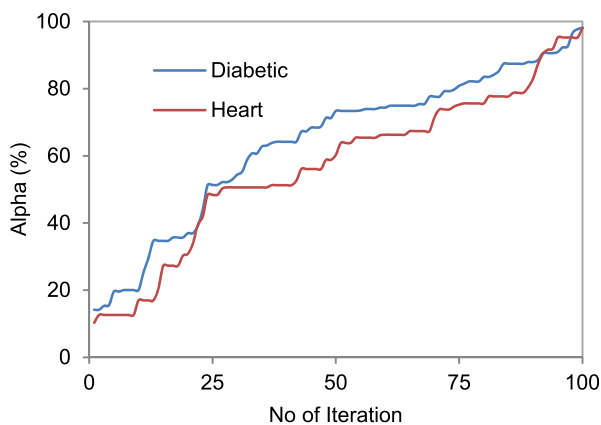


FIGURE 18. Alpha vs. no of iteration.

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.

• • •



An AI-enabled lightweight data fusion and load optimization approach for Internet of Things

Mian Ahmad Jan ^a  , Muhammad Zakarya ^a  , Muhammad Khan ^b, Spyridon Mastorakis ^c, **Varun G. Menon** ^d, Venki Balasubramanian ^e, Ateeq Ur Rehman ^a

Show more 

 Outline |  Share  Cite

<https://doi.org/10.1016/j.future.2021.03.020> 

[Get rights and content](#) 

Highlights

- A lightweight data fusion approach using stratified sampling.
- A dynamic load optimization approach using Evolutionary algorithms to maintain balanced traffic.
- A dynamic service migration technique to balance the load across several edge servers that triggers migration decisions.

Abstract

In the densely populated Internet of Things (IoT) applications, sensing range of the nodes might overlap frequently. In these applications, the nodes gather highly correlated and redundant data in their vicinity. Processing these data depletes the energy of nodes and their upstream transmission towards remote datacentres, in the fog infrastructure, may result in an unbalanced load at the network gateways and edge servers. Due to heterogeneity of edge servers, few of them might be overwhelmed while others may remain less-utilized. As a result, time-critical and delay-sensitive applications may experience excessive delays, packet loss, and degradation in their Quality of Service (QoS). To ensure QoS of IoT applications, in this paper, we eliminate correlation in the gathered data via a lightweight data fusion approach. The buffer of each node is partitioned into strata that broadcast only non-correlated data to edge servers via the network gateways. Furthermore, we propose a dynamic service migration technique to reconfigure the load across various edge servers. We assume this as an optimization problem and use two meta-heuristic algorithms, along with a migration approach, to maintain an optimal Gateway-Edge configuration in the network. These algorithms monitor the load at each server, and once it surpasses a threshold value (which is dynamically computed with a simple machine learning method), an exhaustive search is performed for an optimal and balanced periodic reconfiguration. The experimental results of our approach justify its efficiency for large-scale and densely populated IoT applications.

PDF

Help



Keywords

Internet of Things; Data fusion; Load optimization; Evolutionary algorithms; Gateway-Edge configuration; Service migration

1. Introduction

In the Internet of Things (IoT), the sensor nodes of various applications gather highly correlated data in their neighbourhoods that affect the outcome of any decision made at the cloud data centres[1], [2]. In these applications, the data are unstructured, intermittent and somewhat dynamic. The raw data gathered by the nodes need to be processed locally and analysed at the edge and cloud data centres to optimize the usage of available resources. The raw data need to be fused within the network to reduce the correlation in them. Each node, unaware of its neighbour's sensing range, gathers data in its neighbourhood. The sensing range of two or more nodes may overlap leading to the aggregation of similar data[3]. Each node needs to perform local data fusion to discard multiple copies of the same data. In-network data fusion alleviates the redundancy to trade-off the volumes of data and the available resources at the edge and cloud data centres[4]. The presence of resource-starving nodes means that a data fusion approach needs to be lightweight, robust, and scalable, based on application requirements.

Data fusion alone is not enough to optimize the usage of available network resources. The upstream fused data toward the cloud data centres need to be fairly distributed among the edge servers[5], [6]. In the existing literature[7], [8], [9], the fused data streams are offloaded to the nearest edge servers. However, this approach is not efficient as some of these servers may overload quickly in comparison to others that remain underutilized. The underlying nodes and network gateways associated with the over-utilized servers may suffer higher latency, packet drop, and bandwidth consumption. For a fair distribution of the network load, a dynamic load balancing approach needs to be adopted to assign the time-consuming tasks to underutilized servers. Based on the run-time load at the servers, a decision needs to be made for the assignment of data streams. The configuration of network entities needs to be constantly monitored for an optimized and balanced load. Artificial Intelligence (AI)-enabled algorithms can manage the complex relationship among the network entities[10], [11]. These algorithms need to be adopted for intelligent load balancing and optimization of the selected paths for reliable transmission of the fused data. They have the ability to reconfigure the devices' connectivity based on their experienced load.

Moreover, heterogeneity of the edge servers and network bandwidth may generate opportunities for application migrations (running within virtual machines) which could be beneficial in further load-balancing, avoiding stranded (wasted) resources, and performance degradation (due to overload situations). Stranded resources are those which cannot be allocated due to the unavailability of another resource e.g. CPU cores are fully utilized but memory is half utilized – half memory cannot be allocated because there are no CPU cores available to run the VM/workload/application. Here, heterogeneity refers to the speed of server to process data or network bandwidth. This is achieved through comparing the current utilization levels of the edge servers and/or the rate of transferring over a network link (channel conditions) to some pre-defined threshold values. If utilization level of an edge server or a network link surpasses a particular threshold value, migrations will happen. However, a static threshold may not be appropriate; therefore, we use a simple machine learning model to compute an adaptive threshold.

In this paper, we propose a novel data fusion and load optimization approach for the IoT-enabled applications. Our approach reduces data redundancy at the node level and fairly distributes the fused data streams among the edge servers. It is scalable and can be used by any application, provided that the threshold values for monitored data are known. It ensures the availability of high-quality data at the cloud data centres for decision-making. The main contributions of this work are as follows.

1. A lightweight data fusion approach that reduces the correlation and redundancy in the gathered data by using *MiniMax* stratified sampling. The buffer of each node is partitioned into multiple stratum, each one holding only two values, i.e., a minimum (*min*) and a maximum (*max*). A comparison with *min* and *max* decides to discard or retain any newly sensed data. After a sampling interval, the stratum of each node transmits only two data readings by discarding all other correlated readings.
2. A dynamic load optimization approach that maintains a balanced traffic in the network using a real-valued Genetic Algorithm (GA) and Discrete Particle Swarm Optimization (DPSO). A Software-Defined Networking (SDN) controller monitors the load on individual edge servers and reconfigures the current Gateway-Edge configuration if an unbalanced load is experienced. For reconfiguration, the SDN controller invokes these evolutionary algorithms to identify the transmission path for each gateway towards a prospective server.
3. The above contribution does not account for dynamic load-balancing, i.e., when on some particular resources, the data get processed quicker than others. A dynamic service migration technique is suggested to balance the load across several edge servers that triggers migration decisions, based on current resource (edge server, network channel) usage. A dynamic threshold is computed using a simple regression model in order to keep resources well-balanced.

The rest of our paper is organized as follows. In Section 2, we provide an overview of the background studies pertaining to our proposed approach. In Section 3, our proposed framework and algorithms are described in detail. This section also offers a service migration technique for load balancing across several edges. The experimental results and performance evaluation are sketched in Section 4. Finally, we provide concluding remarks and future research directions in Section 5.

2. Background

In this section, we provide the background studies pertaining to data fusion in the context of load optimization for IoT applications.

In [12], a cloud-based adaptive sensing belief propagation protocol (ASBP) was proposed. ASBP estimates the quality of links to determine the shortest routes toward the cloud for data gathered from IoT applications. The protocol exploits the spatio-temporal correlation among the data streams at cloud datacentres to reduce the energy consumption, and balance the load by keeping a subset of nodes in active states at a given time. ASBP, however, is unable to evenly distribute the load on edge servers for a large-scale IoT network. Besides, fusing massive amount of sensor data at the cloud incur a significant amount of transmission overhead. A dynamic sensor activation algorithm, SensorRank, was proposed to prioritize the deployed nodes based on their residual energy levels, their relative distance, and their links qualities [13]. SensorRank considered symmetric channels for data transmission among the neighbouring nodes. These channels may lead to an uneven load distribution among the nodes, and on the gateways and edge servers, respectively. A spatio-temporal based novel data mining approach (NDM) was proposed for the removal of redundant data, prior to upstream transmission towards the gateways [14]. NDM uses a packet classification approach to filter out redundant data to maintain the network load on the edge servers. NDM is non-scalable and its iterative nature of load distribution at the edge incurs an excessive overhead at the resource constrained sensor nodes.

In [11], an optimized mobile sink-based load balancing (OMS-LB) protocol was proposed to achieve balanced load for a large-scale IoT network. OMS-LB offloads the computationally complex tasks from data gathering devices to a Software-defined Network (SDN) controller that is interfaced with cloud datacentres. The proposed protocol uses PSO and GA to determine the optimal paths for a mobile edge server and optimal data gathering points, i.e., gateways. OMS-LB does not define any criteria for data collection from an application perspective. Besides, the presence of a single server makes this protocol non-scalable, and vulnerable to security threats. A multi-edge based architecture was proposed for seamless integration of cloud datacentres in an IoT environment [5]. The proposed architecture used a multilevel protocol for gateways selection and AI-based load balancer for the identification of an optimal load distribution. However, the proposed architecture lacks any information about the heterogeneity of

nodes, network latency and bandwidth requirements. In [15], the authors proposed a data aggregation scheme by estimating an accurate sensor matrix from the gathered raw data. A fog server is used to reconstruct the matrix that contains minimal noise and highly refined data. However, the proposed matrix does not take into account the load balancing issue and has limitations imposed on its scalability. Besides, it lacks any information on heterogeneous data fusion and interoperability of IoT devices.

All these existing approaches focused on centralized gateways and edge servers for load optimization and decision-making. The presence of centralized entities affect the scalability, fault tolerance and optimal load adjustment of a network. Besides, these approaches operate without data aggregation and fusion at the network level. As a result, they require excessive processing and storage of redundant data at the network gateways and edge servers. The transmission of redundant data ultimately deteriorate the QoS of an underlying network. In the IoT paradigms, it is inevitable to consider AI algorithms for maintaining a balanced load for various applications. There are numerous AI algorithms developed to resolve the load optimization problem. However, in this paper, we utilize the most embraced evolutionary algorithms, i.e., Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) [16], [17], [18]. GA and PSO are population-based algorithms, where the population means a group of all possible solutions, i.e., Gateway-Edge configurations (load balancing and optimization) [19].

GA is a bio-inspired search algorithm in which the population is referred to as a group of chromosomes. The genes of the fittest available chromosomes are utilized to generate new chromosomes, i.e., new optimal Gateway-Edge configurations, via mutation and crossover. On the other hand, in standard PSO, the population of all possible solutions is referred to as a swarm of particles. PSO is inspired by the social behaviour of swarms of ants, a flock of birds, a shoal of fish, etc. In all these cases, the swarm probe the search space for identifying the food with varying velocities. In the case of PSO, each particle is considered a candidate solution for the Gateway-Edge configuration problem. In the case of GA, each chromosome is considered a candidate solution. Since both these algorithms are not directly applicable to integer-based load optimization problems, we have developed a real-valued GA and a Discrete PSO (DPSO) for identifying the optimal Gateway-Edge configurations.

Load balancing is an essential part of the IoT, edge and cloud frameworks that could be achieved in two different ways: (i) dynamic service placement; and (ii) service migration. In respect of (i), two policies are suggested in [20]: cloud-only placement: place all application's modules in the server; and edge-ward placement: favour to run application's modules on various edge devices. Moreover, if allocation of an edge device is not suitable for a particular module, then either resources from other edge devices (server) could be provisioned or it could be migrated somewhere else. Empirical evaluation of both policies suggests that the edge-ward policy significantly improves the application's performance and reduces the network traffic. In respect of (ii), authors in [21], [22] suggest that if an application's performance is the worst on a particular edge device (due to more number of connected sensor devices, network congestion etc.), then its migration either to the server or to another edge device could improve its performance and reduces network traffic. Moreover, mobility management in mobile edge clouds (MECs) also involves migrations [22].

Migrations could also be triggered to balance resource utilization levels of edge nodes. For example, if the utilization of an edge node increases certain threshold value (say 80%), some of the application's module may be moved to other edges. Migration can also take place when resources are under utilized i.e. threshold of 20%. This is done to conserve and consolidate resources to save energy [23]. In the later case, energy could be saved through migrating workloads from these underutilized servers to other servers; and switching them off. However, this may cause performance issues, in particular, if demand exceeds suddenly. We, in this paper, prefer the former one as our objective is not saving energy; instead we want to balance the load across different switched on servers. Furthermore, we use a dynamic threshold-based method that estimate these threshold values periodically – using Eq. (12). Service migration could only be achieved if various sand-boxing technologies such as virtualization, containerization are being used to virtualize the server and edge device resources [24]. In practice, resources in public clouds are virtualized, which increases resource utilization levels and saves energy. If various modules of a particular application are being run in a Virtual Machine (VM) or container; then the service can be migrated either off-line or live. In live migration, the service is moved transparently while still running; however, in off-line

migration the service is stopped first, moved, and then resumed at the target edge. Using CRIU¹ technology, containers could be more quickly migrated than VMs. In case of live VM migration, where VMs data are kept on a shared storage reachable over the network, the time of migration T_{mig} depends on the total volume of memory used by the VM M_{vm} and available network bandwidth B_{total} . For virtual machine V_i the total migration time is given by:

$$T_{mig_{V_i}} = \frac{M_{V_i}}{B_{total}} \quad (1)$$

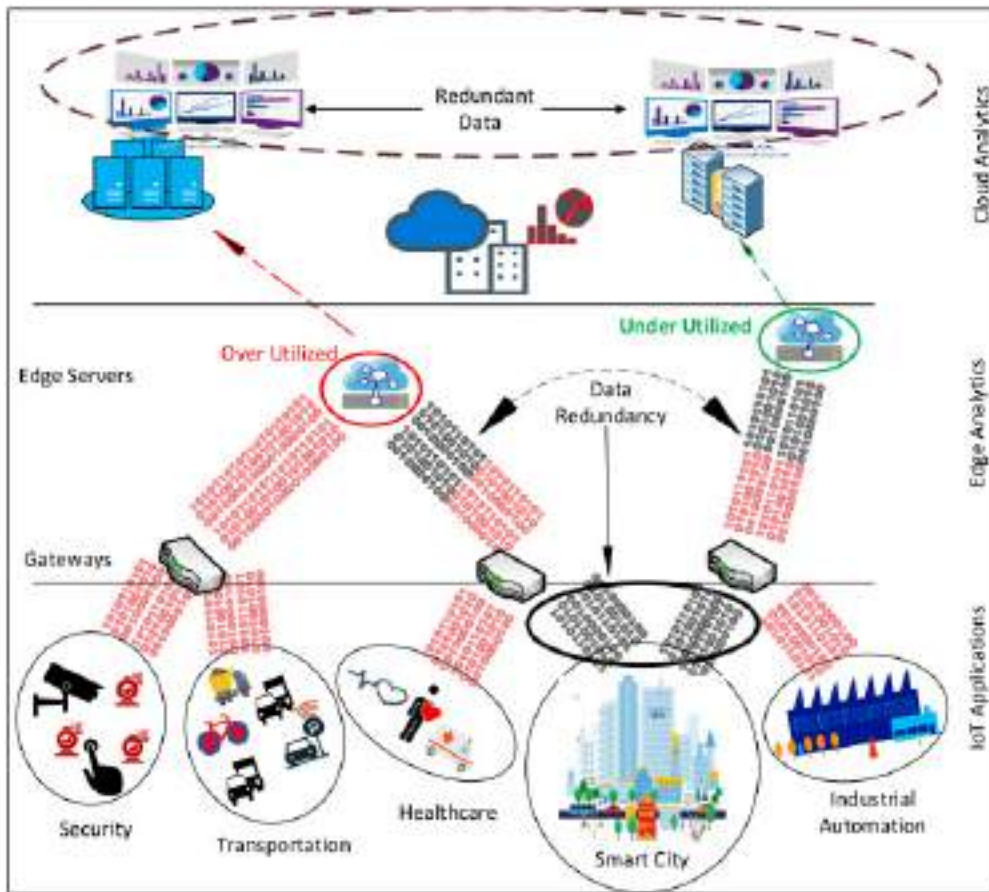
The above equation is used to compute only the migration time of a particular VM. Every VM for this T_{mig} time is considered offline, which is also called the downtime of the VM. The downtime (or performance loss) is dependent on the migration duration, as given by Eq.(1)[25]. Increased downtime results in poor performance; therefore it should be minimized for high availability of the datacenter. The performance degradation P_{deg} due to a single migration is calculated using the following formula (as given by Eq.(2)), where U_{V_i} is the CPU utilization of VM V_i , t_0 is the migration start time and 0.1 is the factor that shows the average performance degradation for web application i.e. 10% of the CPU utilization[23], [26].

$$P_{deg_{V_i}} = 0.1 \times \int_{t_0}^{t_0+T_{mig_{V_i}}} U_{V_i}(t) \cdot dt \quad (2)$$

Note that, the above performance degradation model (10% loss in workload execution time) is benchmarked in[23], [27]; and we assume that it already accounts for other time consuming activities such as: the time to initiate a VM migration; the time to transfer page files (dirty pages in case of live or online VM migration); the time to boot/spin up a new server (if there is no currently running server that can accept the VM being migrated); and the time to restart the VM (in the case of cold or offline VM migration)[28].

3. Data fusion and load optimization approach for IoT applications

In Fig.1, the sensor nodes of various applications transmit their data to cloud data centres via the network gateways and edge servers. Among these applications, the smart city nodes gather and transmit highly correlated data streams. The transmission of these streams affects the decision-making at data centres and creates bandwidth bottlenecks for time-critical and delay-sensitive applications. Moreover, these applications experience excessive latency and degradation in the network throughput if an unbalanced load is experienced at the edge servers. An uneven load distribution results in some of the servers over-utilized while others remain underutilized. The unbalanced load leads to packet loss, longer delays, and network congestion. In this section, we discuss our proposed data fusion and load optimization approach to eliminate data redundancy and maintain a balanced load at the network entities.



[Download : Download high-res image \(1MB\)](#)

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

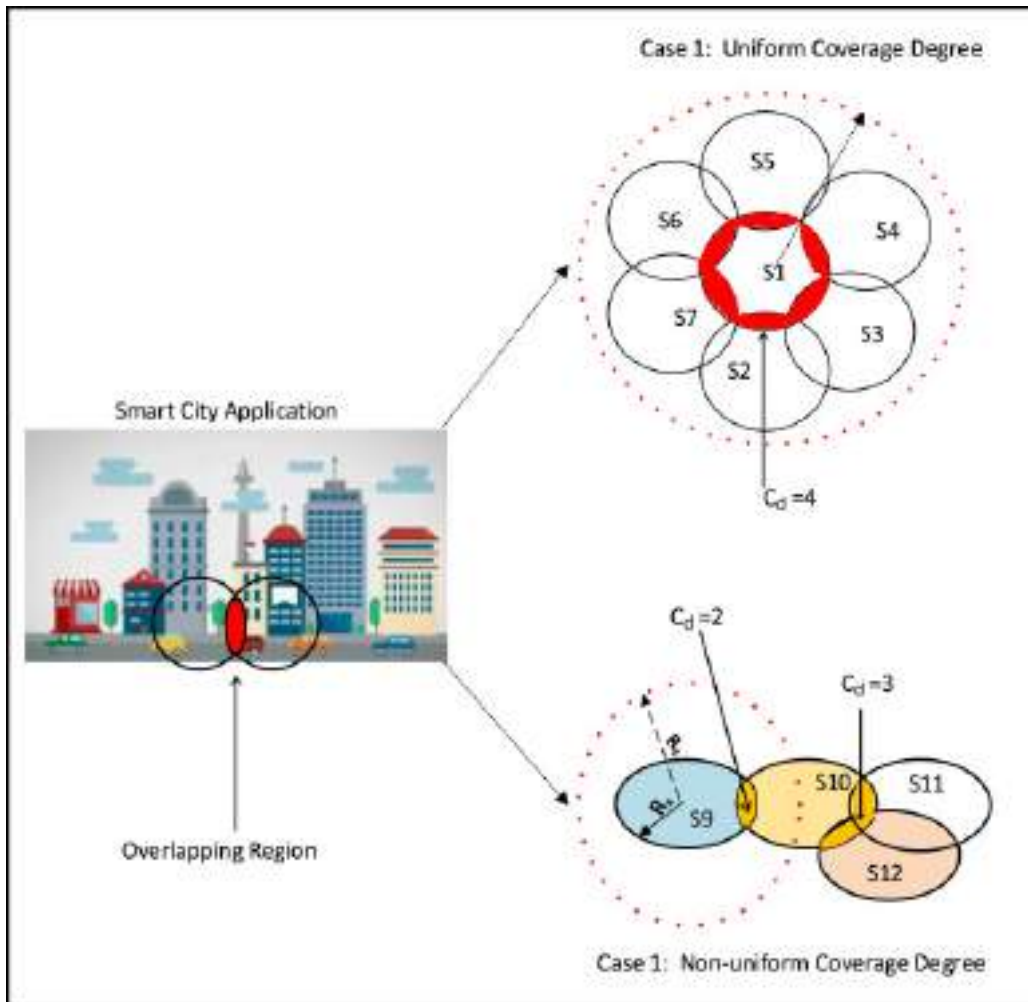
Fig. 1. Data redundancy and unbalanced load in IoT.

3.1. The Proposed Fog-IoT Framework

The proposed model consists of three layers i.e. the cloud layer, the edge layer and the local layer, as shown in Fig. 1. The local layer is responsible to gather important data (related to traffic, healthcare, and crowd, etc.) through various IoT devices and sensors. Once the data is collected, it is processed and/or stored at the edge layer through edge clouds [29]. The edge level processing may also include aggregation that could be achieved through removing redundant data. The filtered data can, then, be sent to the remote cloud layer for further processing such as storage, monitoring and resource management. Transferring the gathered data at local layer directly to the cloud layer, or through edge, may introduce significant delays in the cloud network, which is optimized through data fusion and load balancing methods as discussed in Sections 3.2 Data fusion, 3.3 Load optimization.

The edge infrastructure is of great use when reading the stored data for processing through machine learning approaches. For example, real-time prediction of the traffic flow might happens at the edge layer, however, prediction for monitoring services (load, service migration) can be performed at the cloud layer [30]. Moreover, if real-time prediction, for example, shortest or safe route estimation, is carried out on the remote cloud, then it will incur significant delays depending on the network quality and capacity. In that scenario, the nearest edge cloud can predict the road conditions, congestion and distance; if the required data is stored locally. However, due to the least storage and processing capabilities of the edge clouds [31], the data may not be available or processed locally. In that case, there are three various options: (i) move the required data from the cloud to the edge, process, take decisions, and discard; (ii) perform the prediction at the remote cloud (in whole); and (iii) train the prediction model at the remote cloud and predict at edge layer (distributed fashion computation). Similarly, the huge amount of collected data may consist of duplicate values that could create network congestion and, therefore, affect the prediction process. The edge cloud can use fusion and aggregation technique to send only appropriate data to the remote cloud.

PDF
Help



[Download : Download high-res image \(591KB\)](#)

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

Fig. 2. Data Correlation of varying coverage degrees.

3.2. Data fusion

In a densely deployed smart city, energy hole problem is a common issue faced by the one-hop neighbours of the base station [32]. These one-hop neighbours not only transmit their own data but also relay the data of downstream nodes to the base station. As a result, their energy is depleted rapidly as compared to other nodes. To fill the void left by energy-depleted nodes and to maintain seamless network connectivity, one or more nodes may either sense multiple regions or move around the field to fill this gap. These nodes continuously sense and aggregate data in their neighbourhood, as shown in Fig.2. Each node S maintains a coverage area based on its sensing range (R_s), and a radio coverage based on its communication range (R_c), respectively. The R_s enables efficient data monitoring, whereas the R_c ensures the upstream data transmission. These nodes can have a uniform or non-uniform coverage degree. The coverage degree represents the number of nodes actively monitoring a particular region, i.e., an overlapped region. For uniform coverage, the value of correlation degree (C_d) remains constant for all the nodes. On the other hand, the value of C_d varies for some or all the nodes in a non-uniform coverage. A larger value of C_d represents highly correlated and redundant data as multiple nodes monitor a particular event in the overlapped region.

To eliminate the correlated and redundant data, we use a lightweight data fusion approach at the node level. The proposed approach uses a *minimax* function for the identification and removal of redundant data. In a smart city, each node is equipped with multiple sensors based on an application requirement. In this paper, we restrict our discussion to the temperature sensors only. However, the flexibility of our approach enables it to be extended for any application provided that the threshold values of monitored data are known. We classify the sensed

temperature readings based on the correlation and similarity index among them. Each class, also known as stratum, contains a particular range of temperature readings. For each node, we define ten different stratum that are dynamic and depend on application requirements. They are designed using the concept of stratified sampling, a probability-based sampling technique [33], [34].

The strata² are defined within the buffer of each node and can store temperature readings ranging from 20°C to 39.99°C. Based on these readings, each stratum holds a different range of varying values of up to 2°C. For instance, the range of first stratum S_{t_1} is $\{20.00, \dots, 21.99\}$, and the last stratum $S_{t_{10}}$ is $\{38.00, \dots, 39.99\}$, respectively. The outcome of each stratum can either be a *min* or *max* value. Each stratum has a mean value m that defines its *min* and *max*, respectively. In the beginning, when a new temperature reading T_i is sensed by a node, it is checked against strata of the given node to identify a destination stratum. Once a match is found, T_i is compared against m of the stratum. If T_i is less than m , it becomes the *min*, otherwise, it becomes the *max*, as shown in Eq.(3). The next time a new reading T_{i+1} is sensed within the range of the same stratum, it is compared against m . If T_{i+1} is less than m , a comparison is made with the new *min*. If T_{i+1} is less than *min*, the former turns out to be the new *min*, otherwise, it is discarded. A similar comparison is made with *max*. If T_{i+1} is greater than *max*, it turns out to be the new *max*, otherwise, it is discarded. Irrespective of T_i , T_{i+1} or any other subsequent readings, an exact match with the values of m , *min* or *max* means that these readings will be discarded.

$$f(S_{t_1}, \dots, S_{t_{10}}) = \begin{cases} \min, & \text{if } T_i < m, \\ 0, & \text{if } T_i == m, \\ \max, & \text{if } T_i > m. \end{cases} \quad (3)$$

The *max* and *min* of a given stratum can be plotted as a stationary point on a curve. A point $P(x_0, f(x_0))$ is considered a stationary point of a function $f(x)$ if $(\frac{df}{dx})$ is 0 at $x=x_0$. Suppose a function $y=f(x)$ is a stationary point with $x=x_0$. Then

- if $\left[\frac{d^2 f}{dx^2}\right]_{x=x_0} < 0$, then $x=x_0$ is the *max* of a stratum.
- if $\left[\frac{d^2 f}{dx^2}\right]_{x=x_0} > 0$, then $x=x_0$ is the *min* of a stratum.
- if $\left[\frac{d^2 f}{dx^2}\right]_{x=x_0} == 0$, then
 - if $\left[\frac{df}{dx}\right]_{x=x_0} < 0$ for $x > x_0$, and $\left[\frac{df}{dx}\right]_{x=x_0} > 0$ for $x < x_0$, then $x=x_0$ is the *max* of the given strata.
 - if $\left[\frac{df}{dx}\right]_{x=x_0} > 0$ for $x > x_0$, and $\left[\frac{df}{dx}\right]_{x=x_0} < 0$ for $x < x_0$, then $x=x_0$ is the *min* of the given strata.

In our data fusion approach, the sampling rate of each node is S_r packets per second, where $S_r \geq 1$. The stratum of each node transmits only two packets, i.e., a *min* and a *max*, after every one minute. If a node constantly maintains its sampling rate at $S_r = 1$ for one minute, our approach achieves a maximum of 3 times reduction in the number of transmitted packets to the gateways. For $S_r > 1$, the number of transmitted packets is reduced even further minimum of 3 times. In our approach, the S_r of a node and the number of transmitted packets from its strata to gateways are inversely proportional to each other. These strata significantly reduce data redundancy, network latency, packet collision probability, and ultimately the network congestion. In our proposed approach, each node maintains a similarity index (Ω) for the data gathered over the S_r interval. The value of Ω ranges from 0.03 to 0.1. If Ω is equal to 0.03, it means that among two temperature packets within the range of 0.03°C, only one will be retained. For example, in case of two packets with values 20°C and 20.03°C, only one will be retained in the stratum. Hence, larger the value of Ω , higher will be the rate at which the data are fused.

3.3. Load optimization

Upon data fusion, each node transmits the refined data to cloud data centres via the network gateways and edge servers. The gateways are relay nodes that need to be monitored for maintaining a balanced load at the edge servers. For this purpose, an optimal Gateway-Edge configuration is required. We use various Key Performance

Indicators (KPIs) for an in-depth analysis of the network traffic to identify the optimal configuration. An SDN controller is used for identifying the transmission route for each gateway. It monitors the load on each server, and once it surpasses a threshold value, an alarm is raised to re-configure the current Gateway-Edge connection. If the Gateway-Edge configuration is known at a particular time t , then finding the optimal balanced Gateway-Edge configuration at time $t + 1$ is a primary challenge. If N is the number of gateways, and M is the number of edge servers, then the Gateway-Edge configuration at a particular time t can be represented by a vector $G^t = \{G_1^t, G_2^t, \dots, G_n^t\}$, where $G_n^t \in \{1, 2, \dots, N\}$. As an example, $G_n^t = m$ means that the n th gateway is transmitting to an m th server at time t . Finding the optimal Gateway-Edge configuration vector at time $t + 1$, i.e., $G^{t+1} = \{G_1^{t+1}, G_2^{t+1}, \dots, G_n^{t+1}\}$, is a prime objective. To solve the Gateway-Edge configuration problem, we consider two KPIs, i.e., Average Residual Energy (KPI_{ARE}) of the network and Load Fairness Index (KPI_{LFI}) of the servers.

The LFI is monitored based on Jain's Fairness Index[35]. The normalized weighted sum of these two KPIs is taken into account to maximize the network performance (NP) at time t as shown in Eq.(4).

$$Max(NP) = \alpha KPI_{ARE} + \beta KPI_{LFI}. \quad (4)$$

Here, NP is a primary objective function for optimization problems, and α and β are the weights assigned to each KPI. These weights represent the priority level of each KPI in the objective function, as shown in Eq.(5).

$$NP = \alpha \left(\frac{1}{N_n} \sum_{i=1}^{N_n} \frac{R_i(t)}{\hat{E}} \right) + \beta \left(\frac{1}{M} \frac{\left(\sum_{i=1}^M \sum_{n=1}^N I_{n,i} \varphi_n(t) \right)^2}{\sum_{i=1}^M \left(\sum_{n=1}^N I_{n,i} \varphi_n(t) \right)^2} \right). \quad (5)$$

where, $\frac{R_i(t)}{\hat{E}}$ is the residual energy of a sensor node i at time t and is defined as the remaining energy ($R_i(t)$) of node i to the initial energy (\hat{E}) of each node at time t . For all the nodes in the network, \hat{E} is similar at the time of deployment. For the second KPI, we consider the load fairness at the edge servers. $I_{n,i}$ is a binary indicator, i.e., $I_{n,i}$ is 1, if an n th gateway transmits φ_n packets to i th server at time t , otherwise, $I_{n,i}$ is 0.

To find an optimal Gateway-Edge configuration, an SDN controller needs to perform an exhaustive search for all possible gateway to edge combinations. Literally, it means that the size of the search space is equivalent to M^N , where M represents the number of edge servers and N represents the number of active gateways at a particular time t . The number of possible configurations increases exponentially with an increase in the number of M and N , respectively. To resolve the Gateway-Edge configuration as an optimization problem, we use the evolutionary algorithms, i.e., GA and DPSO. The following steps are executed for these algorithms to achieve an optimal configuration and a balanced load.

1. Generate a random population R^0 of size Δ . The best possible position for each particle, i.e., Gateway, is initiated such that $Pbest_i^0 = r_i^0, \forall 1 \leq i \leq \Delta$.
2. Discover the fitness value of each particle for DPSO and each chromosome for GA in R^0 (using Eq.(5)) and identify its global best position $Gbest^0$, using Eq.(6).

$$Gbest^0 = argmax_{1 \leq i \leq \Delta} F(Pbest_i^0).$$

3. For GA, if the best candidate solution for Gateway-Edge configuration is attained or the maximum number of generations has reached, then the search ends, otherwise, Step 4 is executed. For DPSO, if the best candidate solution is achieved, then the velocities of particles in the current population need to be updated using Eq.(7).

$$v_i^I = j_w v_i^{I-1} + a_1 r_1 (Pbest_i^I - x_i^I) + a_2 r_2 (Gbest_i^I - x_i^I). \quad (7)$$

Here, x_i^I represents the current position of particle i at I^{th} iteration, r_1 and r_2 are random variables within the (0, 1) range, a_1 and a_2 are acceleration constants used for pulling the particles toward the best position, and j_w reflects the inertia effect of preceding particle's velocity over the updated particle's velocity.

4. Next, a set of the best available γ chromosomes are extracted from the current population for GA. The current population is R^I and the selection probability is P_g . For DPSO, the iteration number is simply updated, i.e., $I=I+1$.
5. In case of GA, crossover and mutation are performed on γ . All infeasible solutions, i.e., $R^I - \gamma$ are replaced with μ . Here, μ represents the newly generated chromosomes. In the case of DPSO, if the best candidate solution is attained for Gateway-Edge configuration, then the search ends; otherwise, Step 6 is executed.
6. For GA, all the steps from Step 2 are repeated. For DPSO, the personal best position for each particle is updated using Eq.(8).

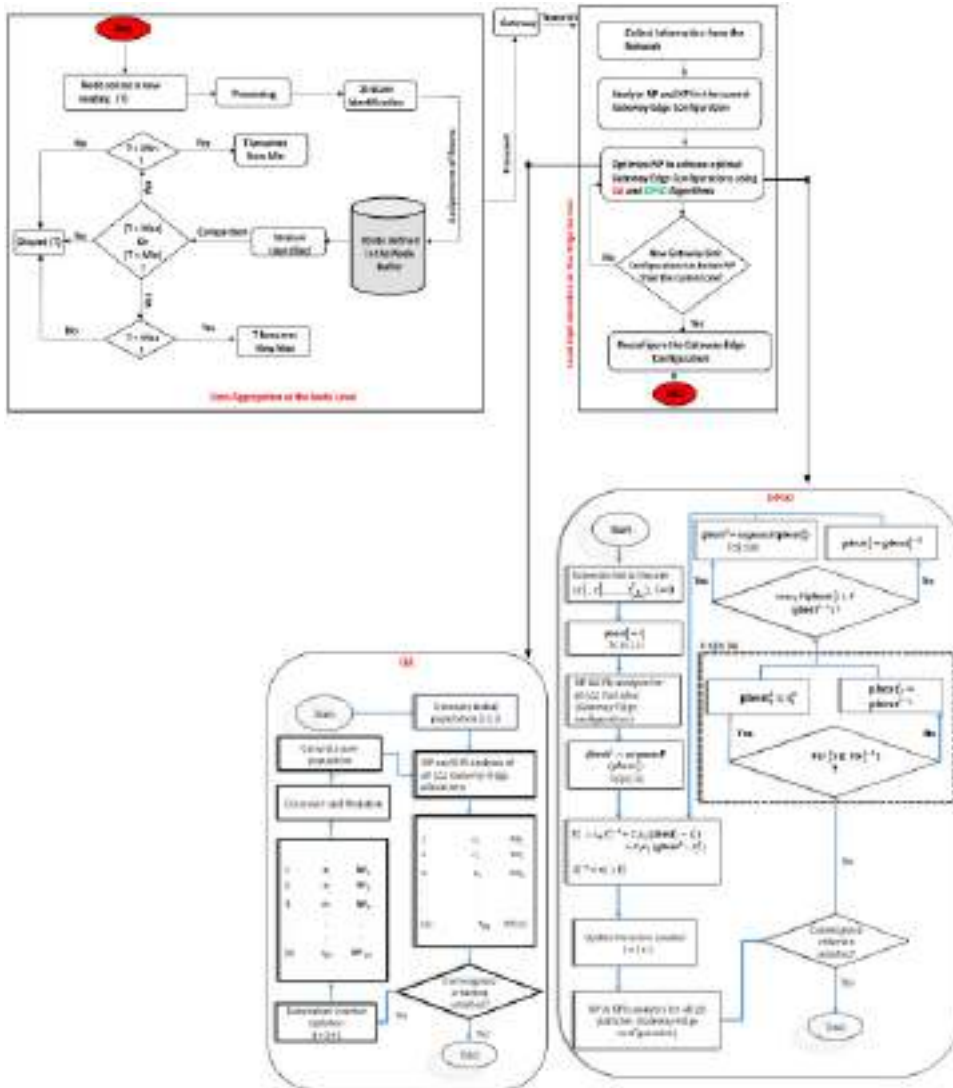
$$Pbest_i^I = \begin{cases} Pbest_i^{I-1}, & \text{if } F(r_i^I) \leq F(Pbest_i^{I-1}) \\ r_i^I, & \text{otherwise.} \end{cases} \quad (8)$$

7. For DPSO, the global best position is updated using Eq.(9).

$$Gbest_i^I = \begin{cases} \operatorname{argmax}_{1 \leq i \leq \Delta} F(Pbest_i^I), & \text{if } F(Pbest_i^I) > F(Pbest_i^{I-1}) \\ Gbest_i^{I-1}, & \text{otherwise.} \end{cases} \quad (9)$$

8. For DPSO, repeat all the steps from Step 1.

The flowchart of our proposed approach is shown in Fig.3. The SDN controller constantly performs the Gateway-Edge configuration based on the NP and KPI values. Since the controller needs to collect various information from the network to analyse the NP for the current Gateway-Edge configuration, we have highlighted the data fusion at the node level as well. The above solution can be used for homogeneous edge servers or that have capabilities to execute at approximate equal times. Furthermore, it does not account for dynamic scenarios where some data get processed earlier than the other edges. Therefore, to further balance the load, a service migration algorithm is suggested – which is feasible as, largely, edge servers run Linux-based operating systems.



Download : [Download high-res image \(693KB\)](#)
 Download : [Download full-size image](#)

Fig. 3. Flowchart of the proposed approach.

3.3.1. Service migration technique

In our framework, load balancing can bring, at least, two benefits: performance improvement; and infrastructure energy efficiency. We trigger service migration either if: the utilization level of a particular edge node; and/or the data transfer rate on a particular link (channel condition), exceed certain pre-defined threshold values (steps 1 to 5)[36]. Once a module is being moved to another edge, it will immediately start receiving packets on another, perhaps, less utilized route. This could be achieved, after copying memory contents of the VM or container through sending a complimentary ARP (address resolution protocol) reply packet to inform the routing devices, within the network, to send data packets to its new location. As a result, both goals, i.e. balanced workload on various edges and reduced network traffic, could be achieved. Once a migration decision is triggered from a particular edge, next is to select a module of a suitable application to move. We move the application's module which can utilize the destination edge more i.e. priority is given to the module which is receiving more packets than other modules (step 6). Lastly, the module is migrated to the least utilized, neighbouring, edge platform; in order to diminish the migration performance impacts over the application's module and migration time (step 7). Finally, the selected module of the application m and the destination server are added to the migration map (step 8–10). The migration map is, then, passed to the load optimization module, as shown in Fig. 3, in order to reconfigure the Gateway-Edge



configuration, periodically. The migration steps are described in Algorithm 1:

Algorithm 1: Service migration technique

Input: dynamic CPU and memory utilization threshold values i.e. U_{tc} and U_{tm} , respectively - computed periodically using Eq. (12); channel condition C_t

Output: migration list $map \rightarrow$ input for load optimization module in Fig. 3

- 1 compute CPU utilization level of the edge node (EN_{uc});
- 2 compute memory utilization level of the edge node (EN_{um});
- 3 compute channel condition (C_t);
- 4 **for** each node \in edge **do**
- 5 **if** $EN_{uc} \geq U_{tc}$ or $EN_{um} \geq U_{tm}$ or $C_t \geq C_r$ **then**
- 6 select application m from edge node;
- 7 choose edge node n as destination node;
- 8 $map \leftarrow m, n$;
- 9 **end if**
- 10 **end for**
- 11 **return** map

[Download : Download high-res image \(186KB\)](#)

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

where EN_{uc} is the total provisioned CPU resources (cores) and EN_{um} is the total provisioned memory resources (RAM) with respect to their total capacities. Note that, U_{tc} and U_{tm} refer to the utilization level of a particular resource i.e. CPU, memory, respectively. Network resources such as bandwidth can also be considered in this formulation. Moreover, the channel condition B_{ij} is estimated using the transmission rate T_r , as given by:

$$T_r = B_{ij} \cdot \log_2 \left(1 + \frac{P_{ij} \cdot h_{ij}}{N} \right) \quad (10)$$

where B_{ij} represents the bandwidth between edge server i and gateway j , h_{ij} denotes the channel gain for gateway j at edge server i and P_{ij} is the transmission power of gateway j . Furthermore, N is the background noise [37]. Note that, Alg. 1 will approximately take $\mathcal{O}(mn \log(n))$ - where m denotes the total number of edges, n denotes the number of edge nodes and $\log(n)$ is the time needed to compute configuration states such as resource utilization levels and channel conditions. The best case occurs at $\mathcal{O}(\log(n))$ plus the time needed to complete all possible migrations. However, complexity would increase up to $\mathcal{O}(mn)^2$ for large number of edges, hosts and application requests – if unluckily an application cannot be placed or, in case, enough resources are not available. Note that, from security point of view, service migration in the IoT and VM or container migration in infrastructure clouds are completely different [36]. Usually, in infrastructure clouds, the migration data is transferred over dedicated networks; however, in IoT the data is transferred over the internet. This makes it essential to encrypt the migrated data and to authenticate the service migration messages that are exchanged among various edge devices.

Using static values for thresholds may not be feasible to trigger effective migrations in platforms with dynamic, heterogeneous and unpredictable workloads. This is due to the fact that resources that falls within the range of the least and most utilized (lower and upper thresholds) resources could not be reconfigured i.e. all hosts are equally loaded. In such scenario, threshold values can either be decreased or increased to balance the load amongst the edge nodes. Therefore, threshold values are needed to be adaptive and dynamically estimated using some sort of statistical techniques on historical data [23]. For example, we can adjust the threshold values based on the strength of the deviation of the edge or link utilization levels because higher deviations increase the likelihood of rising utilization levels. In other words, the higher the deviation, the lower the value of the threshold. Various methods such as local regression (LR), median absolute deviations (MAD), and entropy can be used to measure the statistical dispersion. For implementational simplification purposes, we prefer to use the MAD that describes the median of absolute values of deviations (residuals) from the data's median. For a particular dataset

$\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \dots, \mathcal{D}_n\}$, the MAD can be computed as:

$$MAD = \text{median}_i (\text{abs} [\mathcal{D}_i - \text{median}_j (\mathcal{D}_j)]) \quad (11)$$

The adaptive threshold value \mathcal{T}_v is given by:

$$\mathcal{T}_v = 1 - \lambda \cdot MAD \quad (12)$$

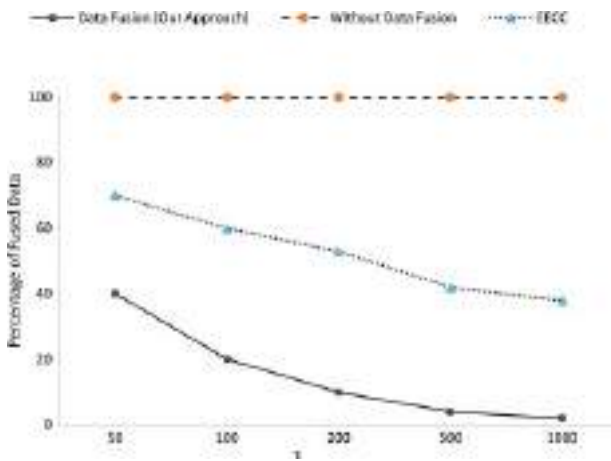
where λ is a parameter that describes how strongly the system tolerates edges over utilizations – lower λ results in higher tolerance to variations in utilization level. Once \mathcal{F}_v is computed, the utilization levels of the node and link are compared to it in order to trigger appropriate migration decisions.

4. Experimental results

In this section, we evaluate the efficiency of our proposed data fusion and load optimization approach in terms of various experimental metrics. For data fusion, we developed a Java-based simulator that utilizes the data collected from sensors, a setup similar to the one adopted at Intel Berkeley Research Lab[38]. Upon fusion, the simulator feeds the refined data to the gateways. For optimal Gateway-Edge configuration, we use Matlab 2018a interfaced with Java. Moreover, we added several Java class files to mimic the notion of containers that simulate a containerized fog infrastructure. The classes were taken from the well-known fog simulator iFogSim[20]. The service migration technique uses either: (i) a static threshold value of 80%; or (ii) dynamic thresholds computed using Eq.(12) in order to trigger migrations of application modules across edges. We further assume that overload (i.e. upper threshold) will not happen due to service placement constraint. To carry out this, the iFogSim default policies for selecting over-loaded servers, containers and target servers were used.

In Fig.4, the percentage of fused packets transmitted to the gateways is shown for different values of τ . Here, τ represents the number of readings sensed by each node over its sampling interval (S_r). The percentage of transmitted packets is calculated as $\frac{S_p}{\tau} \times 100$, where S_p denotes the number of fused packets sent from the strata of each node. The efficiency of our data fusion approach enhances with an increase in the value of τ . The percentage of transmitted fused packets from the strata of each node drops to 1% for 1000 packets, sensed during S_r . Our approach conserves the energy of resource-starving nodes and at the same time, reduces the burden on the network gateways. In comparison to our approach, the existing schemes deliver higher percentage of redundant data to the gateways. For example, EECC[34] transmits multiple copies of the same data from the strata of each node after S_r interval. As a result, the percentage of fused data delivered at the gateways is proportionally high. Moreover, without data fusion, all the sensed packets need to be transmitted to the gateways that will adversely affect the decision-making at the data centres.

During data fusion, each node examines the similarity index (Ω) in the data gathered over the S_r interval. This index further reduces the redundancy and at the same time, lowers the processing burden on the nodes and the network gateways. In Fig.5, the percentage of fused data for varying values of Ω is shown. In this figure, the values of τ varies from 200 to 1000 and Ω from 0.03 to 0.1, respectively. If Ω is 0.03, it means that among multiple readings having a similarity lower than or equal to 0.03, only one reading will be retained and the rest will be discarded. As a result, a higher percentage of readings will be discarded with an increase in the value of Ω . Moreover, our approach achieves a higher percentage of fusion when the value of τ increases. This figure shows that with higher values of Ω and τ , the processing and transmission burdens on the edge nodes and gateways decreases, significantly. In the absence of data fusion technique, a higher percentage of data is delivered to the gateways that in turn increases the processing and transmission burden on the nodes.

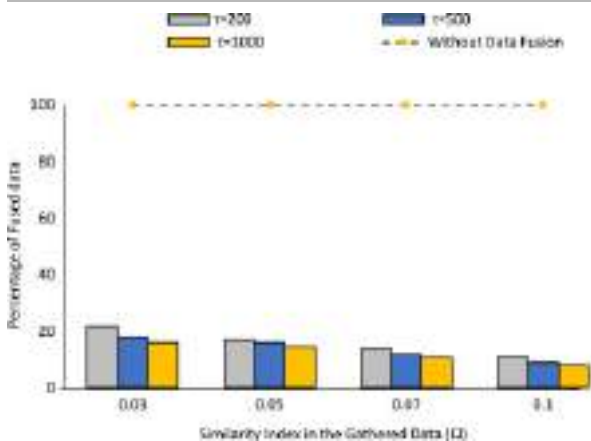


[Download : Download high-res image \(113KB\)](#)

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

Fig. 4. Percentage of fused data.

The optimal Gateway-Edge configurations achieved by GA and DPSO are shown in Table 1. We considered three benchmark problems P_1 , P_2 , and P_3 with two, three, and four edge servers, respectively. P_1 , P_2 , and P_3 contain 1200 sensor nodes including 200 gateways, distributed over the sensing field. In P_1 , GA converges to an optimal Gateway-Edge configuration after the 14th generation. DPSO, on the other hand, achieves the optimal configuration after the 11th iteration. Please note that iteration and generation are similar terms. The former is used in PSO and the latter in GA, as discussed in Section 3.3. GA achieves 185 optimal Gateway-Edge configurations over 200 generations with a convergence rate of 0.925, whereas, in DPSO, there are 188 optimal configurations with a convergence rate of 0.94. In P_2 and P_3 , the convergence rate of GA and DPSO decreases and larger values of iterations and generations are required to achieve an optimal Gateway-Edge configuration. It is mainly due to an increasing number of edge servers in these benchmark problems. These results show that DPSO reaches an optimal solution in fewer iterations as compared to GA.



[Download : Download high-res image \(102KB\)](#)

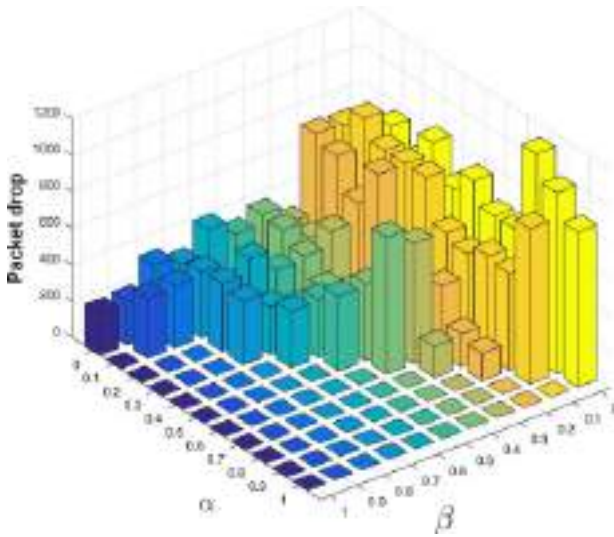
[Download : Download full-size image](#)

Fig. 5. Data fusion with varying values of similarity index.

We assessed the network performance (NP) for all optimal solutions in term of packet drop by modifying the KPIs such that α and β are set in the ordered form of 0,0.1, ...,1 with a constraint $\alpha + \beta \leq 1$. To properly tune α and β , an exhaustive search is performed on P_1 using these parameters to achieve an optimal Gateway-Edge configuration. When α and β are set to 0.8 and 0.2, respectively; a minimum packet drop is observed, as shown in Fig.6. The selection of proper weights for the optimization function, i.e., NP, is challenging and essential for achieving optimal results in the context of evolutionary algorithms.

PDF

Help



[Download : Download high-res image \(467KB\)](#)

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

Fig. 6. Packet loss with varying values of α and β .

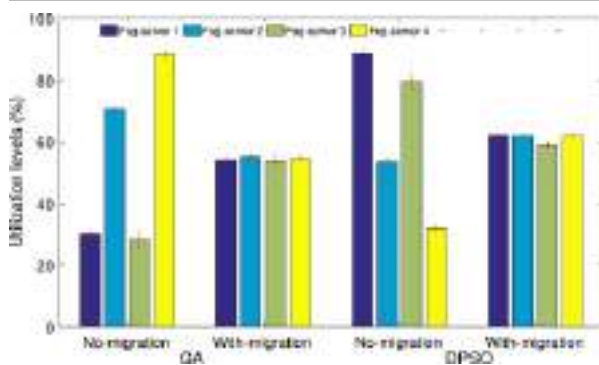
Table 1. Optimal gateway-edge configurations.

Number of edge servers	Convergence rate		Number of iterations	
	GA	DPSO	GA	DPSO
2	0.925	0.94	14	11
3	0.86	0.895	28	21
4	0.805	0.85	43	31

The service migration technique, as suggested in Section 3.3.1, was implemented to balance the load across the edge servers. Increasing the total number of edge servers decreases the utilization levels and vice versa, as shown in Table 2. Moreover, the number of migrations happened is proportional to the amount of fog servers. The standard deviation actually represents how the current load on each server differs from other servers – the higher this value, the more less balanced is the workload and vice versa. We observed significant reduction in utilization levels, that, essentially translate to greater energy savings and improved levels of performance. We observed, as shown in Table 2, that using static threshold values reduces the migration opportunities. Moreover, increased levels of variations were observed in server's utilization levels. This means that either resources were utilized more or the least due to less migration opportunities. Using dynamic threshold values, the variations in utilization levels may decrease significantly – as more migrations will occur subsequently. Varying the threshold values will essentially result in variations of outcomes and differences in Gateway-Edge reconfiguration. Fig. 7 shows average utilization levels (along with error bars at five minute intervals) when four fog servers are taken into account. Moreover, number of migrations would have some impacts on network traffic and processing performance. For four fog servers when migrations are not taken into account, we achieved 48.75 ± 23.67 and 41.28 ± 28.56 average utilization levels for GA and DPSO, respectively. The high variations (standard deviations) show the imbalance load across various servers which might happened due to dynamics in workloads execution or processing patterns. With migrations, we were able to significantly reduce these variations as shown in Fig. 7.

Table 2. Average utilization levels (%) of the edge servers and number of migrations (using static and dynamic threshold values for load balancing and resource re-configuration).

Number of servers	GA		DPSO	
	Utilization	Migrations	Utilization	Migrations
Static threshold values				
2	62.9±5.99	101	61.4±6.45	89
3	60.4±6.31	105	58.7±8.01	103
4	59.6±6.44	147	59.1±7.32	122
Dynamic threshold values				
2	71.2±1.27	134	75.3±2.02	155
3	62.8±1.71	178	68.4±3.89	189
4	54.3±1.09	201	59.7±2.99	217



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

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

Fig. 7. Load balancing for four fog servers [error bars denote standard deviations from the means].

5. Conclusions and future work

In this paper, we proposed a lightweight data fusion and AI-enabled load optimization approach for reconfigurable IoT applications. The buffer of each node is partitioned into strata that hold and transmit only non-correlated fused data towards the network gateways and edge servers. We used GA and DPSO to optimize the usage of available resources by identifying the optimal routes for upstream transmission of refined data from the gateways to edge servers. These algorithms monitored the load at the servers, and if an unbalanced load is experienced, the current Gateway-Edge configuration is reconfigured. For load monitoring at the edge, various Key Performance Indicators (KPIs) were used. Our experimental results significantly reduced the processing and transmission burden at the nodes for large-sized data streams. Our approach achieved optimal gateway-edge configurations for varying number of edge servers in a densely populated network setup. Moreover, a migration approach was used to balance the load across different edge servers. Our evaluation of the proposed migration approach demonstrated that all edge servers are relatively utilized uniformly while having lower standard deviations in their utilization levels. Subsequently, this ensures that data is processed at edge which increases performance. In the future, we aim to analyse the network performance by maintaining a balanced load at the network gateways. It will enable the gateways to automate the downstream transmission links towards the nodes. Moreover, we are keen to see the impact of migrations in dynamic scenarios, particularly, on network traffic and transmission delays.

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

This work was partially supported by a pilot award from the Center for Research in Human Movement Variability, USA and the NIH, USA ([P20GM109090](#)) and the National Science Foundation, USA under award [CNS-2016714](#).

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

References

- [1] Shen Y., Zhang T., Wang Y., Wang H., Jiang X.
Microthings: A generic IoT Architecture for flexible data aggregation and scalable service cooperation
IEEE Commun. Mag., 55 (9) (2017), pp. 86-93
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [2] Khan F., Jan M.A., Rehman A.U., Mastorakis S., Alazab M., Watters P.
A secured and intelligent communication scheme for IIoT-enabled pervasive edge computing
IEEE Trans. Ind. Inf. (2020)
[Google Scholar](#) ↗
- [3] Ding W., Jing X., Yan Z., Yang L.T.
A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion
Inf. Fusion, 51 (2019), pp. 129-144
 [View PDF](#) [View article](#) [View in Scopus](#) ↗ [Google Scholar](#) ↗
- [4] Alam F., Mehmood R., Katib I., Albogami N.N., Albeshri A.
Data fusion and IoT for smart ubiquitous environments: a survey
IEEE Access, 5 (2017), pp. 9533-9554
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [5] Twayej W., Khan M., Al-Raweshidy H.S.
Network performance evaluation of M2M with self organizing cluster head to sink mapping
IEEE Sens. J., 17 (15) (2017), pp. 4962-4974
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [6] Jan M.A., Khan F., Khan R., Mastorakis S., Menon V.G., Watters P., Alazab M.
A lightweight mutual authentication and privacy-preservation scheme for intelligent wearables devices in industrial-CPS
IEEE Trans. Ind. Inf. (2020)
[Google Scholar](#) ↗
- [7] Mazza D., Tarchi D., Corazza G.E.
A unified urban mobile cloud computing offloading mechanism for smart cities
IEEE Commun. Mag., 55 (3) (2017), pp. 30-37
[View in Scopus](#) ↗ [Google Scholar](#) ↗
- [8] El-Sayed H., Sankar S., Prasad M., Puthal D., Gupta A., Mohanty M., Lin C.-T.
Edge of things: The big picture on the integration of edge, IoT and the cloud in a distributed computing environment

IEEE Access, 6 (2017), pp. 1706-1717

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

- [9] Sharma S.K., Wang X.
Live data analytics with collaborative edge and cloud processing in wireless IoT networks
IEEE Access, 5 (2017), pp. 4621-4635

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

- [10] Yao H., Li M., Du J., Zhang P., Jiang C., Han Z.
Artificial intelligence for information-centric networks
IEEE Commun. Mag., 57 (6) (2019), pp. 47-53

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

- [11] Al-Janabi T.A., Al-Raweshidy H.S.
A centralized routing protocol with a scheduled mobile sink-based AI for large scale I-Iot
IEEE Sens. J., 18 (24) (2018), pp. 10248-10261

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

- [12] Bijarbooneh F.H., Du W., Ngai E.C.-H., Fu X., Liu J.
Cloud-assisted data fusion and sensor selection for internet of things
IEEE Internet Things J., 3 (3) (2016), pp. 257-268

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

- [13] Nesa N., Banerjee I.
Sensorrank: An energy efficient sensor activation algorithm for sensor data fusion in wireless networks
IEEE Internet Things J. (2018)

[Google Scholar](#) ↗

- [14] Kumar S., Chaurasiya V.K.
A strategy for elimination of data redundancy in internet of things (IoT) based wireless sensor network (WSN)
IEEE Syst. J. (2018)

[Google Scholar](#) ↗

- [15] Sanyal S., Zhang P.
Improving quality of data: IoT data aggregation using device to device communications
IEEE Access, 6 (2018), pp. 67830-67840

[CrossRef](#) ↗ [Google Scholar](#) ↗

- [16] Vose M.D.
The Simple Genetic Algorithm: Foundations and Theory, Vol. 12
MIT press (1999)

[Google Scholar](#) ↗

- [17] Trelea I.C.
The particle swarm optimization algorithm: convergence analysis and parameter selection
Inf. Process. Lett., 85 (6) (2003), pp. 317-325

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

- [18] 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 (1) (2020), p. 6

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

[19] Kaur K., Garg S., Kaddoum G., Ahmed S.H., Atiquzzaman M.

KEIDS: Kubernetes-based energy and interference driven scheduler for industrial IoT in edge-cloud ecosystem

IEEE Internet Things J., 7 (5) (2020), pp. 4228-4237

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

[20] Gupta H., Vahid Dastjerdi A., Ghosh S.K., Buyya R.

iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments

Softw. - Pract. Exp., 47 (9) (2017), pp. 1275-1296

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

[21] Bittencourt L.F., Diaz-Montes J., Buyya R., Rana O.F., Parashar M.

Mobility-aware application scheduling in fog computing

IEEE Cloud Comput., 4 (2) (2017), pp. 26-35

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

[22] Machen A., Wang S., Leung K.K., Ko B.J., Salonidis T.

Live service migration in mobile edge clouds

IEEE Wirel. Commun., 25 (1) (2018), pp. 140-147

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

[23] Beloglazov A., Buyya R.

Managing overloaded hosts for dynamic consolidation of virtual machines in cloud data centers under quality of service constraints

IEEE Trans. Parallel Distrib. Syst., 24 (7) (2013), pp. 1366-1379, [10.1109/TPDS.2012.240](#)

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

[24] Khan A.A., Zakarya M., Khan R.

H² – a hybrid heterogeneity aware resource orchestrator for cloud platforms

IEEE Syst. J., 13 (4) (2019), pp. 3873-3876

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

[25] Zakarya M., Gillam L.

Managing energy, performance and cost in large scale heterogeneous datacenters using migrations

Future Gener. Comput. Syst., 93 (2019), pp. 529-547

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

[26] Zakarya M., Gillam L., Ali H., Rahman I., Salah K., Khan R., Rana O., Buyya R.

EpcAware: A game-based, energy, performance and cost efficient resource management technique for multi-access edge computing

IEEE Trans. Serv. Comput. (2020), pp. 1-14, [10.1109/TSC.2020.3005347](#)

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

[27] Khan A.A., Zakarya M., Buyya R., Khan R., Khan M., Rana O.

An energy and performance aware consolidation technique for containerized datacenters

IEEE Trans. Cloud Comput. (2019), [10.1109/TCC.2019.2920914](#)

[Google Scholar](#)

PDF

Help

- [28] Zakarya M., Gillam L., Khan A.A., Rahman I.U.
Perficientcloudsim: a tool to simulate large-scale computation in heterogeneous clouds
J. Supercomput. (2020), pp. 1-55, [10.1007/s11227-020-03425-5](https://doi.org/10.1007/s11227-020-03425-5) [↗](#)
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [29] Mastorakis S., Mtibaa A., Lee J., Misra S.
Icedge: When edge computing meets information-centric networking
IEEE Internet Things J., 7 (5) (2020), pp. 4203-4217
[CrossRef](#) [↗](#) [View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [30] Nour B., Mastorakis S., Mtibaa A.
Compute-less networking: Perspectives, challenges, and opportunities
IEEE Netw., 34 (6) (2020), pp. 259-265
[CrossRef](#) [↗](#) [View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [31] Mastorakis S., Mtibaa A.
Towards service discovery and invocation in data-centric edge networks
2019 IEEE 27th International Conference on Network Protocols (ICNP), IEEE (2019), pp. 1-6
[CrossRef](#) [↗](#) [Google Scholar](#) [↗](#)
- [32] Wang Q., Lin D., Yang P., Zhang Z.
An energy-efficient compressive sensing-based clustering routing protocol for WSNs
IEEE Sens. J., 19 (10) (2019), pp. 3950-3960
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [33] Li T., Bolic M., Djuric P.M.
Resampling methods for particle filtering: classification, implementation, and strategies
IEEE Signal Process. Mag., 32 (3) (2015), pp. 70-86
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [34] Jan S.R.U., Jan M.A., Khan R., Ullah H., Alam M., Usman M.
An energy-efficient and congestion control data-driven approach for cluster-based sensor network
Mob. Netw. Appl., 24 (4) (2019), pp. 1295-1305
[CrossRef](#) [↗](#) [View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [35] Jain R., Durresi A., Babic G.
Throughput fairness index: An explanation
ATM Forum Contribution, Vol. 99 (1999)
[Google Scholar](#) [↗](#)
- [36] Mach P., Becvar Z.
Mobile edge computing: A survey on architecture and computation offloading
IEEE Commun. Surv. Tutor. (2017)
[Google Scholar](#) [↗](#)
- [37] Zhang H., Guo F., Ji H., Zhu C.
Combinational auction-based service provider selection in mobile edge computing networks
IEEE Access, 5 (2017), pp. 13455-13464
[View in Scopus](#) [↗](#) [Google Scholar](#) [↗](#)
- [38] Madden S.

Intel berkeley research lab data

(2003)

[Google Scholar](#) ↗

Cited by (24)

AI augmented Edge and Fog computing: Trends and challenges

2023, Journal of Network and Computer Applications

[Show abstract](#) ✓

Energy-efficient secure data fusion scheme for IoT based healthcare system

2023, Future Generation Computer Systems

[Show abstract](#) ✓

An IoT-based resource utilization framework using data fusion for smart environments

2023, Internet of Things (Netherlands)

Citation Excerpt :

...This includes consuming more resources and processing time for the massive and redundant data, downgrading resource utilization accuracy as well as availability and reliability due to the faulty, real-time, and heterogenous IoT data. In such cases, data fusion appears to be a solution for handling IoT data features that would implicitly resolve the resource utilization concerns in IoT-based systems [18]. Data fusion is the process of handling data from multiple data sources to produce more consistent, accurate, and useful information than those provided by any individual data source [19]...

[Show abstract](#) ✓

Smart-3DM: Data-driven decision making using smart edge computing in hetero-crowdsensing environment

2022, Future Generation Computer Systems

Citation Excerpt :

...Similarly, authors in [14] develop a Virtualized Network Function (VNFs) edge cloud architecture that considers the dynamicity of the network changes and the latency-sensitivity of the applications while automatically allocate the resources in the edge and the cloud. Also, the authors in [15] consider the application with low time tolerance and high delay-sensitivity where they tried to improve the application QoS through proposing a lightweight data fusion optimization. The proposed mechanism relies on eliminating the correlation and redundancy of the sensed data and adopting a dynamic migration technique that allows reducing the overhead while maintaining a periodic re-configuration of the ES...

[Show abstract](#) ✓

Multi-tier delay-aware load balancing strategy for 5G HC-RAN architecture

2022, Computer Communications

[Show abstract](#) ✓

The central role of data repositories and data models in Data Science and Advanced Analytics

2022, Future Generation Computer Systems

[Show abstract](#) ✓

PDF

Help

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



Mian Ahmad Jan is an Assistant Professor at the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. He received his Ph.D. in Computer Systems from the Faculty of Engineering and Information Technology (FEIT), University of Technology Sydney (UTS), Australia. His research interests include energy-efficient and secured communication in Wireless Sensor Networks and Internet of Things. He has been guest editor of numerous special issues in various prestigious journals e.g. Future Generation Computer Systems, IEEE Transactions on Industrial Informatics, Springer Neural Networks and Applications, IEEE Sensor, etc. are few to mention. He is an IEEE senior member.



Muhammad Zakarya received the Ph.D. degree in Computer Science from the University of Surrey, Guildford, U.K. He is currently a Lecturer with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. His research interests include cloud computing, mobile edge clouds, Internet of Things (IoT), performance, energy efficiency, algorithms, and resource management. He has deep understanding of the theoretical computer science and data analysis. Furthermore, he also owns deep understanding of various statistical techniques which are, largely, used in applied research. He is an Associate Editor for the IEEE Access Journal and has served as TPC member in various international conferences and workshops.



Muhammad Khan received his Ph.D. degree in wireless communications from Brunel University London, United Kingdom. He is associated with the Wireless Network and Communication Centre (WNCC) at Brunel University London, and the Computer Networks (ComNets) Lab at New York University (NYU). He is currently working on Projects related to 5G mmWave Congestion Control in collaboration with the University of Contabria, Spain and University of Limerick, Ireland. His main research interest is next generation wireless communications, Cloud Radio Access networks, Artificial Intelligence, Machine learning and Congestion Control.



Spyridon Mastorakis (smastorakis@unomaha.edu) is an Assistant Professor in Computer Science at the University of Nebraska Omaha. He received his Ph.D. in Computer Science from the University of California, Los Angeles (UCLA) in 2019. He also received an M.S. in Computer Science from UCLA in 2017 and a 5-year diploma (equivalent to M.Eng.) in Electrical and Computer Engineering from the National Technical University of Athens (NTUA) in 2014. His research interests include network systems and protocols, Internet architectures (such as Information-Centric Networking and Named-Data Networking), edge computing and IoT, and security.



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.

PDF

Help



Venki Balasubramanian received the Ph.D. degree in body area wireless sensor network (BAWSN) for remote healthcare monitoring applications. He is the Pioneer in building (pilot) remote healthcare monitoring application (rHMA) for pregnant women for the New South Wales Healthcare Department. His research establishes a dependability measure to evaluate rHMA that uses BAWSN. His research opens up a new research area in measuring time-critical applications. He contributed immensely to eResearch software research and development that uses cloud-based infrastructure and a core member for the project sponsored by Nectar Australian research cloud provider. He contributed heavily in the field of healthcare informatics, sensor networks, and cloud computing. He also founded Anidra Tech Ventures Pty Ltd a smart remote patient monitoring company.



Ateeq Ur Rehman is currently working as a Lecturer at the department of Computer Science, Abdul Wali Khan University Mardan, KPK, Pakistan. He received his Ph.D. degree from the University of Southampton in 2017. His main research interests are next-generation wireless communications and cognitive radio networks, Internet of Things, Internet of Vehicles, blockchain technology and differential privacy.

1 https://criu.org/Main_Page ↗.

2 Plural of stratum.

[View Abstract](#)

© 2021 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.




PDF

Help

[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)

Instant access to the full article PDF.

[Rent this article via DeepDyve.](#)

[Learn more about Institutional subscriptions](#)

References

1. Ahmed H, Younis EM, Hendawi A, Ali AA (2020) Heart disease identification from patients' social posts, machine learning solution on spark. *Fut Gener Comput Syst* 111:714–722
2. Kumar PM, Gandhi UD (2018) A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Comput Elect Eng* 65:222–235
3. Kwon J-M, Kim KH, Jeon K-H, Park J (2019) Deep learning for predicting in-hospital mortality among heart disease patients based on echocardiography. *Echocardiography* 36(2):213–218
4. Hao Y, Usama M, Yang J, Hossain MS, Ghoneim A (2019) Recurrent convolutional

neural network-based multimodal disease risk prediction. *Fut Gener Comput Syst* 92:76–83

5. Jonnagaddala J, Liaw S-T, Ray P, Kumar M, Chang N-W, Dai H-J (2015) Coronary artery disease risk assessment from unstructured electronic health records using text mining. *J Biomed Inform* 58:S203–S210

6. Melin P, Miramontes I, Prado-Arechiga G (2018) A hybrid model based on modular neural networks and fuzzy systems for classification of blood pressure and hypertension risk diagnosis. *Exp Syst Appl* 107:146–164

7. Al-Makhadmeh Z, Tolba A (2019) Utilizing IoT wearable medical device for heart disease prediction using higher-order Boltzmann model: a classification approach”. *Measurement* 147:1–12

8. Ali F, Kwak D, Khan P, El-Sappagh S, Ali A, Ullah S, Kim KH, Kwak K-S (2019) Transportation sentiment analysis using word embedding and ontology-based topic modeling. *Knowl Based Syst* 174:27–42

9. Yadav A, Singh A, Dutta MK, Travieso CM (2019) Machine learning-based classification of

cardiac diseases from PCG recorded heart sounds. *Neural Comput Appl* 32:1–14

10. Ali F, El-Sappagh S, Islam SR, Kwak D, Ali A, Imran M, Kwak K-S (2020) A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Inform Fusion* 63:208–222

11. Kishore A, Jayanthi V (2018) Neuro-fuzzy based medical decision support system for coronary artery disease diagnosis and risk level prediction. *J Comput Theor Nanosci* 15(3):1027–1037

12. Ali F, Islam SR, Kwak D, Khan P, Ullah N, Yoo SJ, Kwak KS (2018) Type-2 fuzzy ontology-aided recommendation systems for IoT-based healthcare. *Comput Commun* 119:138–155

13. Garate-Escamila AK, El-Hassani AH, Andres E (2020) Classification models for heart disease prediction using feature selection and PCA. *Inform Med Unlocked* 19:1–15

14. Diwakar M, Tripathi A, Joshi K, Memoria M, Singh P et al (2020) Latest trends on heart disease prediction using machine learning and

image fusion. Mater Today Proceed 37:3213–3218

15. Haq AU, Li JP, Memon MH, Nazir S, Sun R (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mob Inf Syst 2018:1–21

16. Mohan S, Thirumalai C, Srivastava G (2019) Effective heart disease prediction using hybrid machine learning techniques. IEEE Access 7:81542–81554

17. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A (2020) Heart disease identification method using machine learning classification in e-healthcare. IEEE Access 8:107562–107582

18. Khan MA, Algarni F (2020) A healthcare monitoring system for the diagnosis of heart disease in the IOMT cloud environment using msso-anfis. IEEE Access 8:122259–122269

19. Fitriyani NL, Syafrudin M, Alfian G, Rhee J (2020) Hdpm: an effective heart disease prediction model for a clinical decision

support system. *IEEE Access* 8:133034–133050

20. Abdeldjouad FZ, Brahami M, Matta N (2020) A hybrid approach for heart disease diagnosis and prediction using machine learning techniques. *International conference on smart homes and health telematics*. Springer, New York, pp 299–306

21. Ansari MF, AlankarKaur B, Kaur H (2020) A prediction of heart disease using machine learning algorithms. *International conference on image processing and capsule networks*. Springer, New York, pp 497–504

22. Shahid AH, Singh MP, Roy B, Aadarsh A (2020) Coronary artery disease diagnosis using feature selection-based hybrid extreme learning machine. In: *Proceedings of the 2020 3rd international conference on information and computer technologies (ICICT)*, IEEE, pp 341–346.

23. Javeed A, Rizvi SS, Zhou S, Riaz R, Khan SU, Kwon SJ (2020) Heart risk failure prediction using a novel feature selection method for feature refinement and neural network for classification. *Mob Inf Syst 2020*

24. Costa W, Figueiredo L, Alves E (2019) Application of an artificial neural network for heart disease diagnosis. In: Proceedings of the XXVI Brazilian congress on biomedical engineering, pp 753–758. Springer, New York
-
25. Vivekanandan T, Iyengar NCSN (2017) Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease. *Comput Biol Med* 90:125–136
-
26. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220
-
27. Rish I, et al (2001) An empirical study of the naive Bayes classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, vol 3, pp 41–46.
-
28. Meyer D, Leisch F, Hornik K (2003) The support vector machine under test. *Neurocomputing* 55(1–2):169–186
-

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	Accepted	Published
21 February 2021	11 May 2021	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

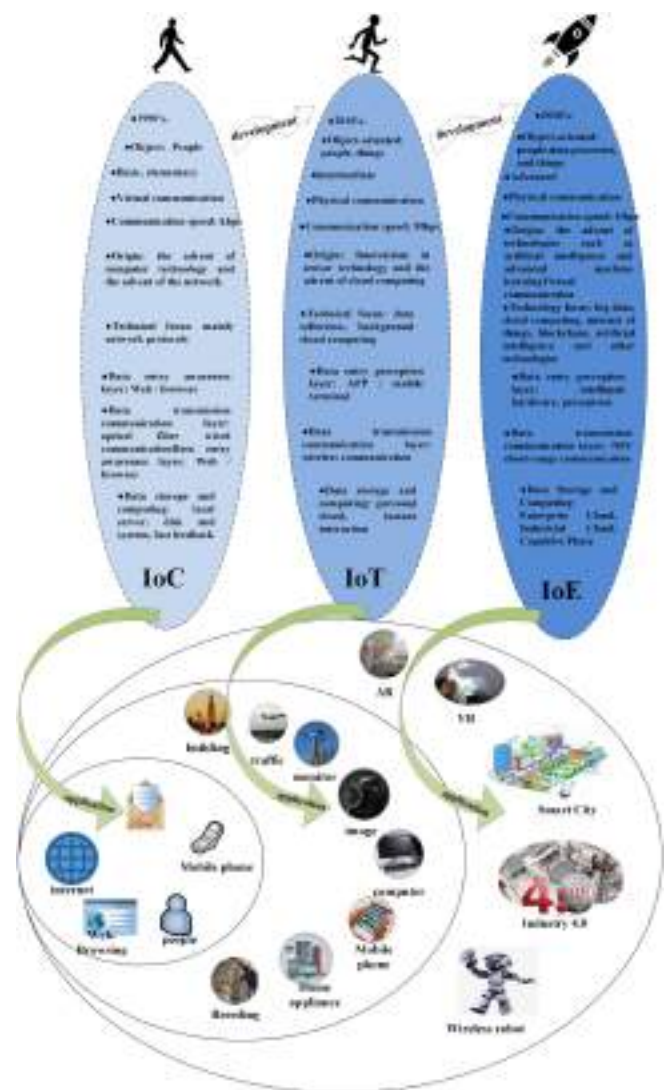


Fig. 1. Development comparison of IoC, IoT, and IoE.

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

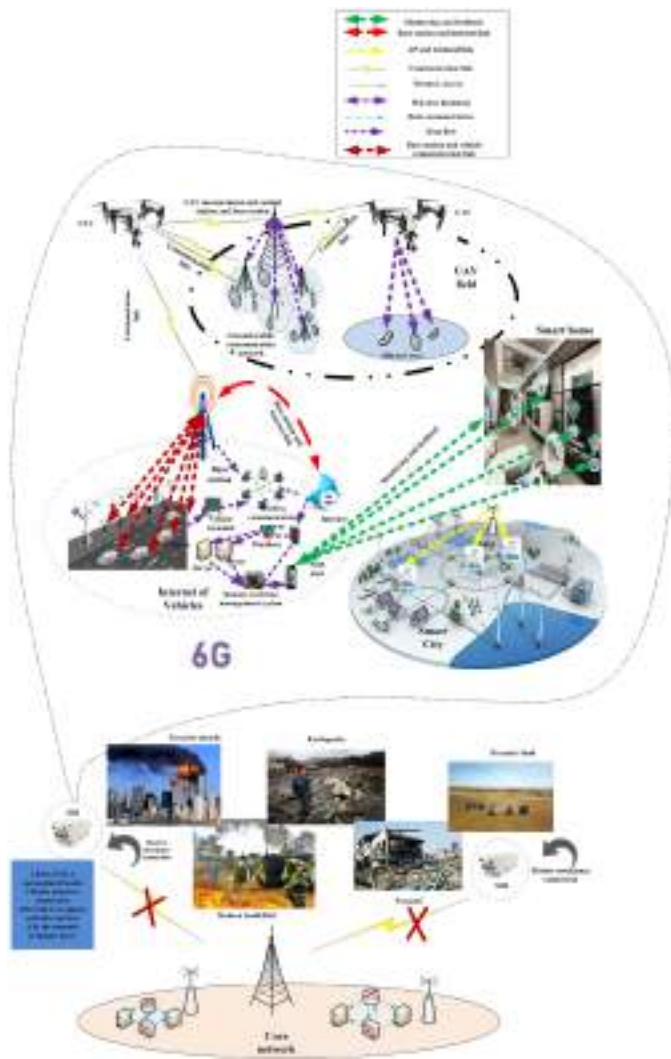


Fig. 2. Key technologies of 6G enabled CI as a part of NIB based on IoE.

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.

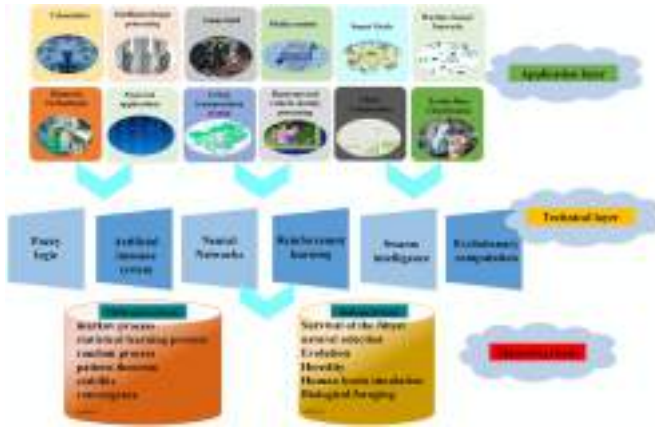


Fig. 3. CI theory, technology and application.

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 technologies used	Advantages	Issues
[14] [27]	Basic coverage	6G network deployment and network coverage	Wide network coverage in rural areas, basic services coverage and network coverage and network coverage	Basic network deployment and network coverage, network coverage and network coverage
[14] [27]	Emergency	Emergency network and network coverage	Emergency network coverage and network coverage	Emergency network coverage and network coverage
[14]	Disaster relief	Disaster relief network and network coverage	Disaster relief network coverage and network coverage	Disaster relief network coverage and network coverage
[27]	Disaster relief	Disaster relief network and network coverage	Disaster relief network coverage and network coverage	Disaster relief network coverage and network coverage
[14]	Intelligent manufacturing	Intelligent manufacturing network and network coverage	Intelligent manufacturing network coverage and network coverage	Intelligent manufacturing network coverage and network coverage
[14]	Wearable devices	Wearable devices network and network coverage	Wearable devices network coverage and network coverage	Wearable devices network coverage and network coverage
[14]	Smart grid	Smart grid network and network coverage	Smart grid network coverage and network coverage	Smart grid network coverage and network coverage
[14] [27]	Healthcare services	Healthcare services network and network coverage	Healthcare services network coverage and network coverage	Healthcare services network coverage and network coverage
[14]	Smart manufacturing	Smart manufacturing network and network coverage	Smart manufacturing network coverage and network coverage	Smart manufacturing network coverage and network coverage
[14]	Disaster relief	Disaster relief network and network coverage	Disaster relief network coverage and network coverage	Disaster relief network coverage and network coverage

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
Location	Mobile devices (MD), Small Cell (SC) and edge servers	Terminal devices, Small Cell and edge servers	Edge
Implementation	Applications are the terminal and edge servers. It is a distributed system.	Terminal devices, Small Cell and edge servers. It is a distributed system.	Edge servers are the terminal and edge servers. It is a distributed system.
Architecture	Multi-tier architecture. Applications are the terminal and edge servers. It is a distributed system.	Multi-tier architecture. Applications are the terminal and edge servers. It is a distributed system.	Single-tier architecture. Applications are the terminal and edge servers. It is a distributed system.
Network	Network is the terminal and edge servers. It is a distributed system.	Network is the terminal and edge servers. It is a distributed system.	Network is the terminal and edge servers. It is a distributed system.
Security	Security is the terminal and edge servers. It is a distributed system.	Security is the terminal and edge servers. It is a distributed system.	Security is the terminal and edge servers. It is a distributed system.
Performance	Performance is the terminal and edge servers. It is a distributed system.	Performance is the terminal and edge servers. It is a distributed system.	Performance is the terminal and edge servers. It is a distributed system.
Scalability	Scalability is the terminal and edge servers. It is a distributed system.	Scalability is the terminal and edge servers. It is a distributed system.	Scalability is the terminal and edge servers. It is a distributed system.
Flexibility	Flexibility is the terminal and edge servers. It is a distributed system.	Flexibility is the terminal and edge servers. It is a distributed system.	Flexibility is the terminal and edge servers. It is a distributed system.
Reliability	Reliability is the terminal and edge servers. It is a distributed system.	Reliability is the terminal and edge servers. It is a distributed system.	Reliability is the terminal and edge servers. It is a distributed system.
Availability	Availability is the terminal and edge servers. It is a distributed system.	Availability is the terminal and edge servers. It is a distributed system.	Availability is the terminal and edge servers. It is a distributed system.
Interoperability	Interoperability is the terminal and edge servers. It is a distributed system.	Interoperability is the terminal and edge servers. It is a distributed system.	Interoperability is the terminal and edge servers. It is a distributed system.
Portability	Portability is the terminal and edge servers. It is a distributed system.	Portability is the terminal and edge servers. It is a distributed system.	Portability is the terminal and edge servers. It is a distributed system.
Extensibility	Extensibility is the terminal and edge servers. It is a distributed system.	Extensibility is the terminal and edge servers. It is a distributed system.	Extensibility is the terminal and edge servers. It is a distributed system.
Modifiability	Modifiability is the terminal and edge servers. It is a distributed system.	Modifiability is the terminal and edge servers. It is a distributed system.	Modifiability is the terminal and edge servers. It is a distributed system.
Reusability	Reusability is the terminal and edge servers. It is a distributed system.	Reusability is the terminal and edge servers. It is a distributed system.	Reusability is the terminal and edge servers. It is a distributed system.
Compliance	Compliance is the terminal and edge servers. It is a distributed system.	Compliance is the terminal and edge servers. It is a distributed system.	Compliance is the terminal and edge servers. It is a distributed system.
Interoperability	Interoperability is the terminal and edge servers. It is a distributed system.	Interoperability is the terminal and edge servers. It is a distributed system.	Interoperability is the terminal and edge servers. It is a distributed system.

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].

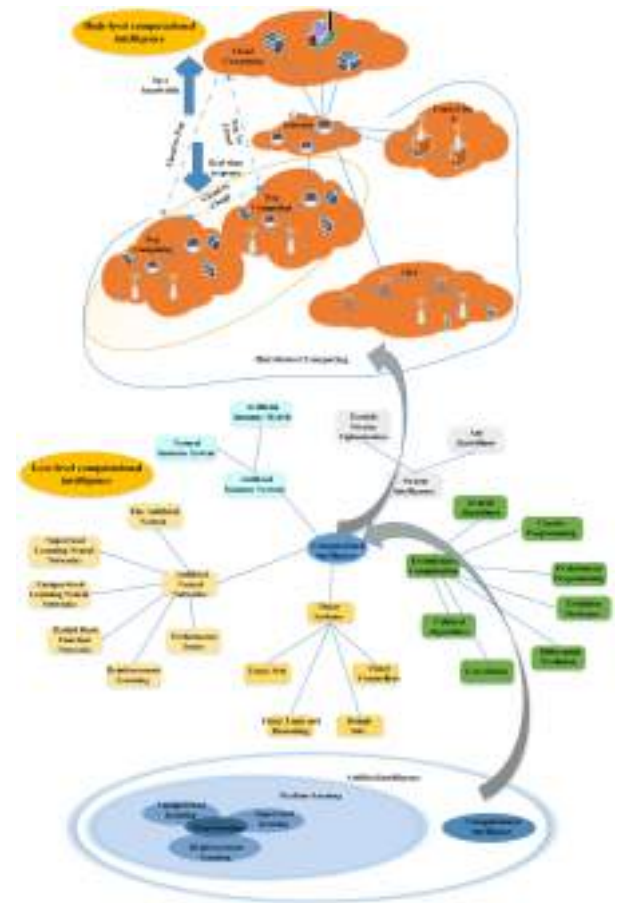


Fig. 6. Relationship for various technologies.

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 ROADMAP OF NEXT-GENERATION WIRELESS TECHNOLOGY FOR 6G-ENABLED VEHICULAR NETWORKS

Mainak Adhikari, Abhishek Hazra, Varun G Menon, Brijesh K Chaurasia, and Shahid Mumtaz

ABSTRACT

Ultra-reliable connectivity and delay-sensitive data transmission is the ever increasing demand for the next generation transportation system. However, substantial network congestion and long-distance data travel always create significant delays even with conventional 4G or 5G cellular technology. Situations become critical even in autonomous transportation systems in various domain-specific applications, where smart vehicles continuously generate an enormous amount of data to make optimal decisions, for example, self-driving cars, autonomous robots, and industrial transportation. Thus, the importance of beyond 5G or 6G technology in the transportation system is prominent and requires a different level of investigation. With this motivation, in this article, we briefly analyze next-generation 6G-supported technologies and their advantages on adopting 6G into the Internet of Vehicles (IoV). We also shed light on several 6G-based frameworks in IoV networks. Finally, we conclude our discussion by addressing current open research challenges and future directions of solving those challenges with 6G technology.

INTRODUCTION

With its strength to produce a single platform empowering quality of assistance, such as improved mobile broadband, augmented reality, the Internet of Things (IoT), and smart connectivity, 5G signifies a discovery in cellular networks. However, looking into the progressive development of social, industrial, and individual service demands regarding new services and offers, it is possible to envision the need for beyond 5G communication and the design of new innovative technologies for digital transportation systems [1, 2]. The key drivers of the 6G initiative are not only the limited performance of 5G technology, but also a continuous technology-driven paradigm shift in wireless networks. For example, high-tech and autonomous industry vehicles continuously demand ubiquitous mobile ultra band (uMUB), ultra-high density data (uHDD), and ultra-high-speed low-latency-communication (uHSLLC) services from cellular operators. However, the existing 5G infrastructure fails to provide such requirements. This article aims to motivate researchers toward 6G communication, starting from the requirement of 6G communication, associated 6G-enabled technologies such as terahertz communication, artificial intelligence (AI), visible light communication (VLC), and ultra-wide communication (UC) in rural areas, and several advantages of using 6G in transportation systems.

RESEARCH INITIATIVES OF 6G

Even though 5G communication technology started being deploying in 2019, research initiatives on 6G are under different organizations and countries. The International Telecommunication Union (ITU) has initiated 6G research to develop new networking technologies for the future of 2030 and beyond. This initiative includes several technology associations of 6G, for example, Internet protocols, new standards, network architecture, and several new services.

Mainak Adhikari and Brijesh K Chaurasia are with IIIT Lucknow, India.

Abhishek Hazra is with the Indian Institute of Technology (Indian School of Mines), India.

Varun G. Menon is with the SCMS School of Engineering and Technology, India.

Shahid Mumtaz is with the Institute of Telecommunications, Portugal.

Digital Object Identifier: 10.1109/IOTM.001.2100075

The University of Oulu, Finland, has a novel 6G research program, which mainly focuses on enabling 6G-based wireless technology, ultrafast data transmission, and deployment of new 6G-supported technologies in the existing wireless networks. The Federal Communication Commission in the United States opened new spectrums from 95 GHz to 3 THz to allocate new licenses to network operators. IEEE Future Networks launched a 6G initiative to define new technologies for 6G communication. Several other 6G research initiatives in the European Union and South Korea, such as EU Terranova and Networking Research beyond 5G, will also start deploying 6G in the near future. On the other hand, several other industrial initiatives from LG Electronics, Samsung, Nokia, and Ericsson began to collaborate with organizations and academia on 6G network research and deployment. The key technologies for 6G-enabled communication are depicted in Fig. 1.

LIMITATION OF TRADITIONAL COMMUNICATION: WHY 6G?

Even though 5G is not yet fully deployed, it started out not offering quality of service (QoS) applications to the consumer, which has encouraged network researchers to think about 6G requirements [3]. The history of cellular networks and the industrial revolution shows that the enlargement of necessary requirements usually begin a new technological revolution. For example, the revolution from Apple raised the provision of mobile Internet and 4G technology.

On the other hand, industrial transportation and vehicular communication require a ubiquitous network, where smart vehicles and intelligent robots expect ultra-broadband cellular networks and ultra-high uplink and downlink data rates with negligible latency and extreme reliability, encouraging new types of 6G as a service. Furthermore, intelligent devices, intelligent vehicles, drones, and other vehicles used in military operations require high data rate, extended coverage, and all-weather on-demand service, where a large amount of data are in the form of 4K/8K video streaming data, augmented reality/virtual reality, gaming data, and 3D holographic video. Hence, the need for not only beyond enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC) also require uMUB, uHDD, and uHSLLC services. These types of services cover

6G REVOLUTION AND KEY TECHNOLOGIES FOR VEHICULAR NETWORKS

Each new generation of wireless communication contributes more enduring speed and appends additional functionality to mobile subscribers. 1G first introduced wireless technology, where radio signals are encoded into analog signals. 2G adds three more benefits: digital communication, mobile data service, and text messaging. 3G introduced the first mobile broadband for intelligent devices (e.g., CDMA and EV-DO). 4G added LTE, LTE-Advanced, and WiMax technology with high data rate. 5G extends the functionality of 4G by adding New Radio, where mobile devices get higher data rate and stable wireless connectivity [4]. As expected, 6G will incorporate full artificial intelligence (AI)-supported technology, wireless rural area coverage, and secure communication. The revolution of 6G technology in cellular communication is depicted in Fig. 2.

MAIN FEATURES OF 6G COMMUNICATION

As of now, we have discussed various requirements for 6G development. But *what is the nature of 6G, and what are the main features and technologies 6G should contain?* The 3rd Generation Partnership Project (3GPP) is mainly responsible for defining standards and technologies for 6G wireless communication worldwide. It is expected that 6G could include several new technologies such as sub-terahertz communication, energy-efficient communication, an AI-enabled fully automatic system, edge computing, blockchain-based communication, and several other combinations of hybrid technologies [5]. By adding these new technologies, 6G will be capable of extending base station density up to 1000 km², user experience > 1 Gb/s, and < 0.001–0.1 ms end-to-end delay [6]. Moreover, this technology can provide 10 ns processing density, more than 5× the reliable connectivity than 5G, and provide more than 5× the user experience. By adopting this technology, transportation systems get more benefit from tactile Internet, satellite communication, MTC, and Industry 5.0 technology.

REQUIREMENTS OF 6G IN TRANSPORTATION SYSTEMS

The open nature of intelligent transportation systems as wireless communication leads to stable communication, computation, security, and privacy-related issues. Although 5G communication technology was designed to meet such challenges, recent 5G architectures and developments fail to address high mobility and transportation-related challenges. For example, 5G supports 20 Gb/s uploading data rate and 10 Gb/s downloading data rate, and 10–50 ms delay [7]. However, AI-enabled automotive transportation systems require more reliable connectivity while vehicles move fast. Hence, the need to incorporate 6G into the transportation system where both uploading and downloading data rates increase up to 1 Tb/s and delay is minimized to 1 ms.

Consider a 6G-enabled intelligent transportation system where smart vehicles connect with personal devices, other vehicles, and existing infrastructure, which will change our transportation system into an interoperable and safe transportation system. This new vision for the 6G-enabled transportation system brings new opportunities, business scope, and countless benefits.

- **Scenario 1:** Advanced vehicles are capable of 360° vision within the range of 300 m and keep personal information private. By incorporating 6G into the intelligent transportation system, drivers can get quick alerts regarding congestion. Sometimes they may get it through visual display, tone, or vibration, which helps drivers stop potential crashes. On the other hand, vehicles can communicate with 6G-enabled roadside infrastructure to quickly alert drivers regarding vehicle crossing, which also helps alert drivers during bad weather conditions or in icy conditions, or when roads are slippery. Moreover, faster sharing of information from multiple vehicles about natural hazards and warns drivers before they

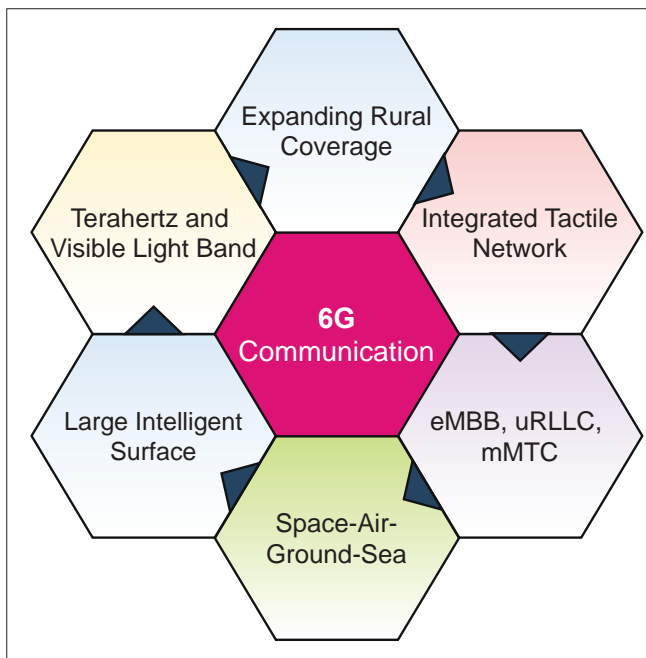


FIGURE 1. Key technologies for 6G-enabled communication.

Parameter	5G	6G
Data rate	20 Gb/s	1 Tb/s
Security	Low	High
Frequency	30–300 GHz	.3–3 THz
Energy-efficient communication	Partial	Full
Traffic	10 Mb/s/m ²	1–10 Gb/s/m ³
Uploading: data rate	20 Gb/s	1 Tb/s
Downloading: data rate	10 Gb/s	1 Tb/s
Rural coverage	Medium	High
Delay	10–50 ms	.1 ms
Congestion	Minimal	Negligible
Mobility support	Up to 500 km/h	Up to 1000 km/h
Maximum spectral efficiency	30 b/s/Hz	100 b/s/Hz

TABLE 1. Comparison between 5G and 6G technologies.

satellite, plane, bullet trains, and underwater communication with negligible delay. Zong *et al.* defined these requirements in a set of six-F trends: full spectrum, full coverage, full dimension, full coverage, full photonic, and full intelligence.

The design of 5G was meant to encourage and cover several domains, including telecommunication, robotics, smart transportation, extended reality, and other disciplines. Further, 5G covers the five most promising technologies such as millimeter-wave (mmWave), small cell tower, massive multiple-input multiple-output (MIMO), beamforming, and full duplex. However, in reality these technologies suffer from their own limitations. For example, mmWAVE cannot travel through obstacles and is absorbed by trees and houses; small cell towers are not suitable for rural area coverage; and massive MIMO leads to serious network interference. All these challenges motivate researchers to think about new 6G-based cellular technology, which is expected to solve all such challenges and provide better services to users. A comparative analysis between 5G and 6G technologies in vehicular network is depicted in Table 1.

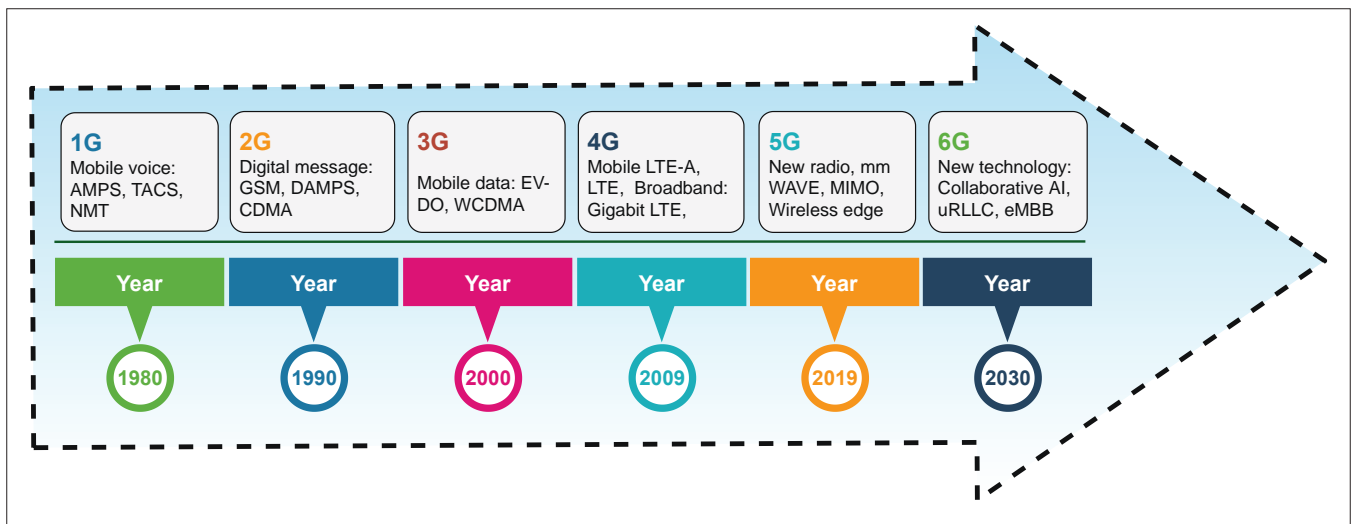


FIGURE 2. Revolution of 6G technology in cellular communication.

experience them, which can be achieved by 6G technology in the vehicular network.

- **Scenario 2:** Road weather data from vehicles can be sent to the traffic management system quickly with the cooperation of 6G-enabled roadside units, providing real-time information to help monitor or manage transportation system performance. Centers can take action by adjusting traffic and speed lights. Drivers can monitor real-time road conditions and weather conditions before leaving home. Awareness can be enhanced about police activity and medical emergencies. An accident zone alerts car drivers regarding the danger zone, and suggests slowing down and changing routes within a fraction of times using 6G communication. Moreover, the 6G-enabled intelligent transportation system connects public transportation and travelers more efficiently.

6G-ENABLED TECHNOLOGIES FOR IOV

The 6G-enabled intelligent transportation system is aimed at solving transportation-related problems and improving the overall performance of the transportation network. This transportation system falls under the class of active mobility within the framework of smart cities, which have been gaining strength in the past few decades. 6G-enabled communication, with advanced computing technologies such as edge/fog computing, plays a major role by supporting massive interconnectivity between the IoV and remote computing devices with highly diverse service requirements and meets the exceptional demands of ubiquitous connectivity in future vehicular networks. The main stakeholders of the 6G-enabled vehicular network are the Internet of Vehicles (IoV), roadside units, and remote computing devices. The framework of 6G-enabled vehicular networks is depicted in Fig. 3. Several advancements and technological developments are going on in this particular domain, and a vast scope of improvement is occurring. However, this field still suffers from crucial technical issues related to intelligent and autonomous communication and ultra-fast data transmission. 6G-enabled technologies can be a suitable solution for improving efficiency and smart traffic management in transportation networks. A short summary of 6G-based technologies is explained below:

- **AI-Enabled Wireless Network Model:** AI is one of the essential requirements for any smart system. This system includes a wide range of applications and domains [8]. Similarly, in the transportation system, AI helps drivers to make quicker decisions, optimizes traveling costs by measuring proper distance, and provides an alert system regarding crashes and road conditions. This sort of intelligence also allows the driver to run a car with zero human intervention. Nowadays,

deep learning and machine learning models are widely used for intelligent decision making, predictive maintenance, and monitoring devices' health conditions, hence reducing extra overhead from the driver [9]. Recent studies also show that 6G enabled AI technology is suitable for vehicle-to-everything (V2X) communication, where vehicles can intelligently communicate with other devices with optimized networking technologies [7]. Besides real-time tracking, data analysis and finding drivers' driving patterns are also measured and monitored through AI-enabled techniques.

- **Ultra-Reliable Low-Latency Communication:** Low latency and reliability of communication are challenging for delay-critical applications, such as healthcare, industrial applications, and intelligent transportation systems, where data needs to be sent to the targeted device within the desired delay. This requires modifying existing standards and adopting new technologies such as THz communication and VLC. 6G technology brings new opportunities by providing seamless connectivity and high data rate for real-time applications. Several 6G-based communication technologies such as massive MIMO, beamforming, and full-duplex communication reduce information loss and propagation delay among the communicating devices.
- **Faster Mobility and High Dynamicity:** Mobility management is one of the key functions of the transportation system, aiming to track and monitor the vehicle efficiently. It can also be seen that smart vehicles are highly mobile and dynamic in nature, creating difficulties in communicating with other vehicle and roadside base stations. Moreover, high dynamicity in the manufacturing industry restricts vehicles from interoperably communicating with other vehicles, thus limiting the sharing of information and experience. 6G communication technology can be an alternative solution for the mobility issue in which vehicles can take advantage of mobility management techniques and tools directly from the network operator.
- **Secure and Privacy Solutions for IoV:** In IoV, vehicles including smart sensors and IoT devices help to obtain environment information and transfer the sensed data to remote computing devices for further analysis with advanced communication technology such as 6G. This information can be helpful in safe navigation, detecting hurdles, optimizing routes, and traffic management. A key concern when relying on vehicle data is vulnerability to data. Therefore, the need is to develop various efficient strategies and protocols that provide security, trust, and privacy to both communicating entities and secure vehicle data from malicious entities. With numerous connected devices on the vehicles getting direct access to sensitive information, ensuring the security and privacy of transmitted data is challenging. GPS jamming, cargo sys-

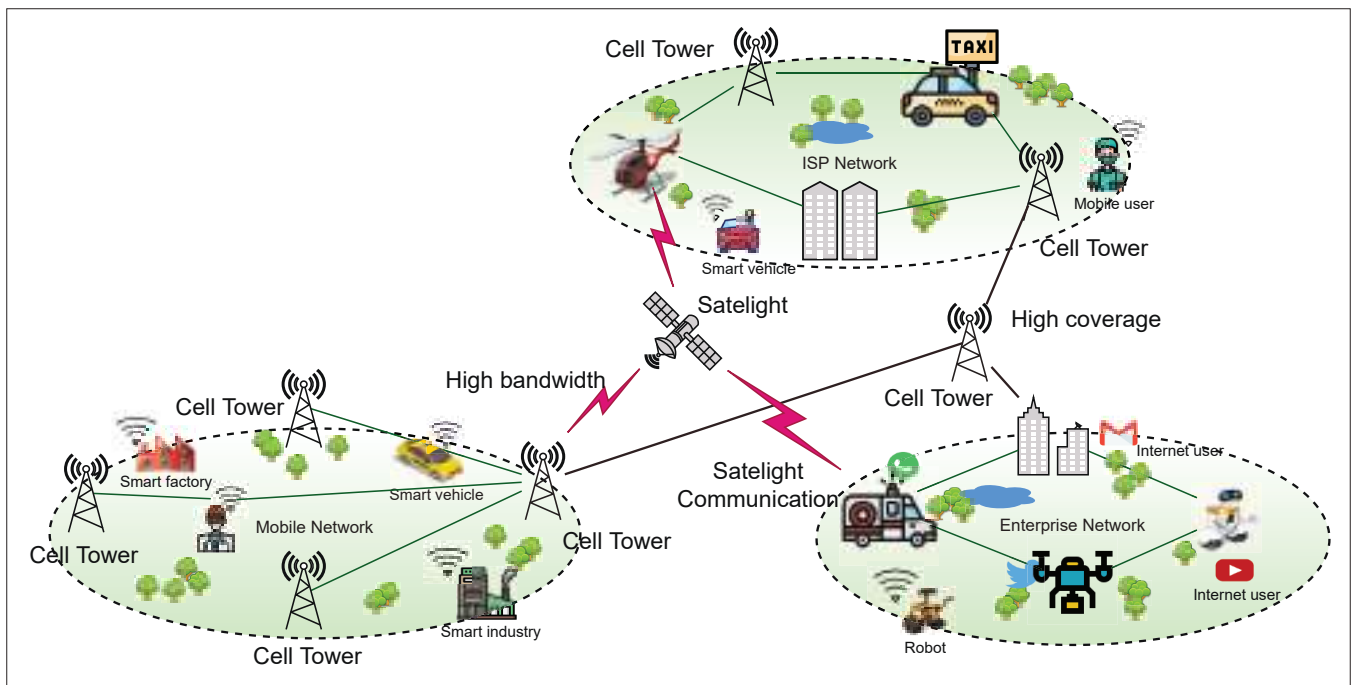


FIGURE 3. Framework for 6G-enabled vehicular networks.

tem manipulation, and ransomware attacks are a few recent cybersecurity threats faced by the industry [10]. Also, due to the lack of infrastructure, their open nature, the high mobility of vehicles, and data privacy regulatory constraints, security is one of the most considerable topics in vehicular networks. Thus, providing security of vehicular network traffic with advanced communication technology and securing the stored data at the remote computing servers for further decision making or data analysis are important requirements for developing a reliable transportation system.

- Dynamic Transportation Networks:** The recent emergence of new technologies in vehicular networks, including automated and connected vehicles, intelligent incentive and routing platforms, shared mobility services, and so on, has a significant impact on traffic flow in road networks. The rapid development of vehicular networks with the autonomous transportation system and advanced vehicular sensors with new capabilities in data collection and communication brings great opportunities and new challenges for managing and controlling the transportation/vehicular network efficiently. As a result, it is important to integrate the emerging computation and communication technologies such as edge/fog computing and 6G technology into dynamic transportation network analysis and large-scale computation with dynamic traffic assignment models. Further, dynamic transportation initiates new research challenges on multi-modal dynamic transportation networks with emerging technologies such as 6G communication as essential components. Also, extending the scope of the dynamic traffic assignment model to emerging mobility trends of vehicular networks is a research challenge of utmost importance for the transportation system.

USE CASE SCENARIOS

The invention of 6G technology and the increasing maturity of disruptive technological paradigms such as urban computing, IoT, sensor networks, connected vehicles, and edge/fog computing play a major role in revolutionizing next-generation IoV applications including autonomous driving, precise fleet management, real-time video analytics, and so on. This sharp increase of computing and communication technologies is being enabled by designing intelligent IoV frameworks for mobility and transportation, as service, processes, and applications. To

demonstrate the effectiveness of 6G-enabled IoV networks, we highlight the requirements and characteristics of two important use cases in IoV networks: an edge-centric smart transportation system and intelligent unmanned aerial vehicles (UAVs) for military operation. The potentials and challenges of these use cases over 6G-enabled technologies are depicted in Table 2.

EDGE-CENTRIC SMART TRANSPORTATION SYSTEM

The intelligent transportation system (ITS) is significantly changing the form of the traditional transport system by incorporating advanced computation and communication technologies such as edge computing and 6G technology. The transportation system is typically represented as a network with a set of roadside equipment (RSE) and the size of IoV, where vehicles can communicate with each other in a multihop manner or access the connected services with or without the support of the infrastructure installed in the RSE. However, this framework faces a lot of real-time challenges due to varying locations, long road networks, and the flow of a large volume of data traffic. Therefore, efficient computation and communication cooperation on the road including vehicle-to-vehicle and vehicle-to-RSE is a significant challenge in the transport system. Furthermore, several advanced machine learning and deep learning algorithms have been widely used to improve network efficiency and decision making for the connected transportation system and autonomous vehicles. However, such models have failed to instantly analyze, evaluate, and predict network efficiency.

An edge-centric and intelligent framework is emerging as a promising solution to provide real-time services to connected and autonomous vehicles. One of the possible edge-centric smart transportation frameworks is depicted in Fig. 4. In this framework, wearable devices or sensors retrieve the environmental parameters and send the information to the distributed edge devices through a reliable 6G-enabled communication channel. These sensory data are analyzed on the local edge devices with minimum processing and communication delay, and decide to take specific actions using the set of actuators. Due to the limited storage capacity, the analyzed data are further stored at centralized cloud servers for future analysis. Thus, the integrated edge computing and 6G technology help to analyze the sensory data locally with minimum latency and energy consumption while making faster decisions for transportation systems.

Use case scenario	Technologies	Potentials	Challenges
Intelligent transportation system	AI-enabled wireless network model	<ul style="list-style-type: none"> • Intelligent decision making • Optimize traffic congestion 	<ul style="list-style-type: none"> • Stable connectivity, Prediction, analytics
	Ultra-reliable low-latency communications	<ul style="list-style-type: none"> • High transmission rate • High throughput over the network 	<ul style="list-style-type: none"> • Cost optimization, Bandwidth utilization • Reduce propagation delay
	Faster mobility and high dynamicity	<ul style="list-style-type: none"> • Efficient Hand-off mechanism • Forward pointer based routing 	<ul style="list-style-type: none"> • High dynamicity among the vehicles • efficient mobility management techniques
	Security and privacy solutions for IoV	<ul style="list-style-type: none"> • Blockchain based spectrum sharing • Quantum based secure communication 	<ul style="list-style-type: none"> • Privacy, authentication, encryption techniques • Data sharing, public key generation
	Dynamic transportation networks	<ul style="list-style-type: none"> • New communication technology • Seamless data transmission 	<ul style="list-style-type: none"> • Rural area coverage, On-device processing • Optimized dynamic network parameters
Intelligent UAV system	AI-enabled wireless network model	<ul style="list-style-type: none"> • Measure height and distance • Intelligent decision making 	<ul style="list-style-type: none"> • Intelligent decision making • Energy efficient communication
	Ultra-reliable low-latency communications	<ul style="list-style-type: none"> • Light weight communication protocol • High distance coverage 	<ul style="list-style-type: none"> • Traffic management • Path planning, and scheduling
	Faster mobility and high dynamicity	<ul style="list-style-type: none"> • Efficient communication protocol • Optimized energy management technique 	<ul style="list-style-type: none"> • Spectrum allocation, and networking • UAV communication
	Security and privacy solutions for IoV	<ul style="list-style-type: none"> • Proper authentication technique • blockchain based authentication system 	<ul style="list-style-type: none"> • Efficient encryption algorithm • UAV certification and verification
	Dynamic transportation networks	<ul style="list-style-type: none"> • Distributed communication technology • Contingency management techniques 	<ul style="list-style-type: none"> • Multi-UAV coordination • Distributed UAV control

TABLE 2. Use case scenarios and relevance of 6G-enabled technologies.

Driven by the massive number of connected and autonomous vehicles, and the stringent requirements of the data-intensive applications, it is difficult to analyze the data at the centralized cloud servers with high reliability and efficiency. Thus, to meet the requirements of IoV including low latency and energy consumption, and handle the large-scale sensory data, the potential solution is to analyze the data at the edge of the network with advanced communication technology such as 6G technology. In addition, advanced AI technology with 6G-enabled edge networks helps to make the transportation system intelligent enough by supporting intelligent traffic control, adaptive resource allocation, intrusion, and misuse detection, and so on. Therefore, the main objective of designing a smart transportation framework is to incorporate 6G communication technology for THz communication with high transmission speed and AI-enabled edge-centric computation for processing the sensory data at the edge of the network with minimum delay and high reliability. The major research challenges of the 6G-enabled smart transportation system are:

- Design a 6G-enabled energy-efficient edge-centric IoV platform for intelligent decision making with minimum latency and communication energy consumption.
- Incorporate AI-enabled technologies for real-time traffic management and control in 6G-enabled intelligent autonomous transport systems in edge networks.
- Develop different types of blockchain, security, and vulnerability scanning technologies in 6G-enabled edge networks for analyzing various types of IoV applications with high reliability.
- Design different types of resource optimization, service provisioning, and learning fusion techniques for IoV applications at 6G-enabled networks for higher prediction analysis and minimum error rate.
- Integrate knowledge discovery and traffic predictivity at 6G-enabled edge networks from the aspect of edge intelligent IoV applications for accurate decision making.

INTELLIGENT UAV FOR MILITARY OPERATION

An unmanned aerial vehicle (UAV) plays a key role in a wide range of military applications including surveillance, attack roles, battle assessment damage, spying enemy territory, and

so on. A UAV is an aircraft with no pilot and limited sensing, control, and communication capabilities due to UAV payload constraints. It is autonomously controlled by a ground control station with pre-programmed functional plans and complex dynamic automation systems. Due to the inefficient functional operations including real-time data gathering, sensing, communication, and computing, achieving efficient and reliable communication with smart coordination, positioning, and trajectory design is challenging in UAV networks. Furthermore, due to the forthcoming dense operations of UAVs, particularly over urban regions, ensuring airspace safety is becoming an urgent issue.

A fully autonomous UAV has to predict other real-time observations such as collisions, fault diagnosis, tracking, and failure control. In particular, this vehicle must integrate with other UAVs through various communication techniques to achieve different mission-critical military applications. One of the possible smart UAV frameworks for military operation is depicted in Fig. 5. In this framework, the UAVs retrieve the environmental information through a set of sensors and cameras, and analyze the sensory data or images locally using built-in embedded devices or transmit the data to the local edge servers for further analysis through a 6G-enabled communication channel. This can reduce overall latency and processing delay. Also, the cloud servers are integrated with this framework for analyzing computation-intensive applications using advanced AI technologies and store the data securely for future analysis.

With the recent advancement of communication including 6G technology and advanced AI technologies, UAVs become fully autonomous, more maneuverable, and smarter without human intervention. A smart UAV incorporates a centralized or decentralized traffic management policy, and addresses the scalability issue considering the large volumes of UAVs of variable, uncertain, and heterogeneous mobility features. Furthermore, the advanced AI-enabled technologies help to deal with environmental uncertainties in UAV traffic management decision making, and air and ground infrastructures together support UAV traffic management. Therefore, the main objective of developing an intelligent UAV is to incorporate 6G communication technology for reliable and fast data transmission to the remote devices for further analysis and incorporate AI-en-

abled technology for efficient traffic management and decision making with minimum error. The major research challenges of intelligent UAVs are:

- Develop different types of resource allocation and mobility models of UAVs for energy management at the 6G-enabled network for military operation.
- Incorporate different optimization, learning, and AI techniques to manage application deployment and decision making for UAV-assisted military operation.
- Design different cost-effective cooperative computing and scheduling strategies for UAV-assisted military applications in 6G networks with minimum latency and energy consumption.
- Introduce 6G-enabled software-defined networking and energy harvesting techniques for better traffic monitoring, surveillance, weather monitoring, firefighting, and so on during UAV-assisted military operation.
- Design ultra-reliable low-latency protocols for minimizing latency and providing high reliability for mission-critical military applications in the 6G-enabled UAV framework.

FUTURE PROSPECTS

6G has several potential advantages over standard cellular technology, and a large scope of research and engineering is required to make a proper standardized 6G network. Among other future directions, in this article, we discuss the open development scope in the context of the 6G-enabled transportation system.

- **Integrated AI in Transportation System:** AI is one of the vital tools to add intelligence into a non-intelligent system. AI in the field of smart transportation and communication systems can assemble transaction data to decrease congestion and increase the scheduling of public transportation networks [11]. Specifically, public transportation with 6G technology is influenced by traffic movement, and AI can provide smooth traffic models, more intelligent traffic information, and real-time tracking to manage more precious and decreased travel patterns. The effectiveness of AI techniques in a 6G-enabled transportation system relies on the ability of these models to effectively handle the dynamics of vehicular networks, providing the algorithmic means to learn the patterns within the information flows exchanged from/to vehicles, infrastructure, and pedestrians. Important research work related to AI techniques and 6G-enabled vehicular networks are transfer learning, online learning, and federated learning for intelligent decision making, routing, security, and privacy.
- **High Mobility in the Transportation System:** Future mobility incorporates all forms of breathing transport and will build an ITS with 6G-enabled vehicular networks. A growing amount of smart vehicles such as cars, vans, and trucks have better self-sufficiency and are running toward the intention of them driving us instead of us driving them [12]. Electrical self-driving vehicles carrying people and goods throughout cities are essential to monitoring downtown pollution and transportation congestion levels. The culture to innovate in this field has completely begun in 6G-enabled vehicular networks. Nowadays, drones with a 6G communication mode, presently used in assessment and monitoring tasks, can transform industrial sectors such as the building and retail sectors by transporting heavy cargoes to inaccessible areas.
- **Interoperable Connectivity in the Transportation System:** Interoperability is vital to guarantee appropriate device system connectivity. Interoperability focuses on allowing the effective communication of intelligent transportation system components with other components in automobile systems, devices, transport systems, and devices as necessary, regardless of when and how they are built and used. The application of 6G-enabled vehicular communication systems and the emergence of automatic vehicle transport networks will make interoperability essential, as device interconnections boost variety and complexity into the 6G-enabled networks.

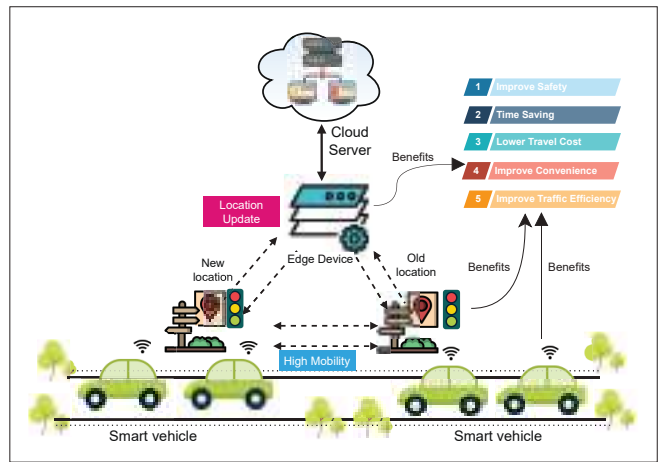


FIGURE 4. Use case scenario 1: edge-centric smart transportation system.

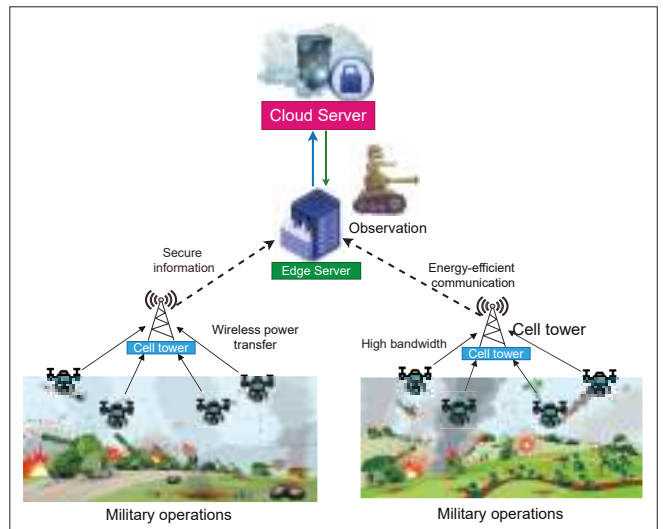


FIGURE 5. Use case scenario 2: intelligent UAVs for military operation.

Standards and architectures must endure development to speculate technological improvements and prepare the needed agreement and interoperability.

- **Cognitive Networking for Intelligent Transportation Systems:** Cognitive networking is expected to analyze and utilize a variety of information for improving the intelligence of transportation systems. To address the advanced demands of a 6G-enabled intelligent vehicular network, which cannot be met by the traditional technologies such as high scalability, low latency, high mobility, and throughput, innovative cognitive networking technologies have been applied to the vehicular network for raising the user experience through providing high-performance communications in vehicular networks by optimizing latency, energy consumption, network security, and coverage, and so on. Furthermore, cognitive computing can support location-based services, navigation, entertainment, and so on, while meeting users' expectation ratio, and even effectively guarantee traffic safety and avoid traffic accidents by monitoring the driver's physiological and psychological state.
- **Quantum Communication and Computing in the Transportation System:** 6G requires highly secure connectivity and processing capabilities to enable real-time applications [13]. Quantum computing is a promising technology that can boost the communication and computation capabilities in 6G by incorporating the quantum superposition theorem. More-

over, quantum communication is capable of providing high data security, which mainly follows the principles of uncertainty and no-cloning. Therefore, quantum computing can remarkably expedite and improve AI techniques that demand big data and extensive training. As a result, incorporating quantum computing with transportation systems is an important research aspect in the 6G-enabled vehicular network.

- **Blockchain-Based Information Sharing in the Transportation System:** Unlicensed spectrum sharing is a promising technology to share more numbers of user data through a single spectrum. It also helps to reduce the requirement of large spectrum for information sharing. Recently, blockchain technology has gained a lot of information due to its distributed and secure data sharing mechanism over the network [5]. Specifically, blockchain technology stores its data across the local computing devices of the network instead of storing the data in the centralized server. Whenever a new block is added, every computer updates its blockchain to reflect the changes. Mainly, the blockchain technology applies the hash encryption (SHA-256) technique to secure its information. As a result, blockchain technology provides data security and protects data from intruders, which can help to develop a reliable 6G-enabled transportation system in vehicular networks.

CONCLUSION

This article introduces the unique concept of integrating 6G communication into the next generation of transportation systems. Along with the development and significance of 6G technology, we have also covered several 6G-enabled networking technologies, such as predictive maintenance, distributed learning, blockchain-enabled security, and wireless power allocation, and their advantages in the transportation system [14]. Two practical use-cases are also investigated to confer the importance of 6G communications to transportation networks. These

technologies are not market-ready, which intimates broad research opportunities on 6G-enabled vehicular networks for the forthcoming digital civilization of 2030 and beyond. Finally, the influence of diverse research challenges and future directions for 6G-enabled transportation systems are also explained.

REFERENCES

- [1] B. Ji *et al.*, "A Survey of Computational Intelligence for 6G: Key Technologies, Applications and Trends," *IEEE Trans. Industrial Informatics*, 2021.
- [2] J. Navarro-Ortiz *et al.*, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Commun. Surveys & Tutorials*, vol. 22, no. 2, 2020, pp. 905–29.
- [3] K. David and H. Berndt, "6G Vision and Requirements: Is There Any Need for Beyond 5G?," *IEEE Vehic. Tech. Mag.*, vol. 13, no. 3, 2018, pp. 72–80.
- [4] M. H. C. Garcia *et al.*, "A Tutorial on 5G NR V2X Communications," *IEEE Commun. Surveys & Tutorials*, 2021.
- [5] Z. Zhang *et al.*, "6G Wireless Networks: Vision, Requirements, Architecture, and Key Technologies," *IEEE Vehic. Tech. Mag.*, vol. 14, no. 3, 2019, pp. 28–41.
- [6] A. Hazra *et al.*, "Stackelberg Game for Service Deployment of IoT-Enabled Applications in 6G-aware Fog Networks," *IEEE IoT J.*, 2020.
- [7] F. Tang *et al.*, "Future Intelligent and Secure Vehicular Network Toward 6G: Machine-Learning Approaches," *Proc. IEEE*, vol. 108, no. 2, 2020, pp. 292–307.
- [8] E. Calvanese Strinati *et al.*, "6G: The Next Frontier: From Holographic Messaging to Artificial Intelligence Using Subterahertz and Visible Light Communication," *IEEE Vehic. Tech. Mag.*, vol. 14, no. 3, 2019, pp. 42–50.
- [9] J. Du *et al.*, "Machine Learning for 6G Wireless Networks: Carrying Forward Enhanced Bandwidth, Massive Access, and Ultrareliable/Low-Latency Service," *IEEE Vehic. Tech. Mag.*, vol. 15, no. 4, 2020, pp. 122–34.
- [10] G. Aceto, V. Persico, and A. Pescapé, "A Survey on Information and Communication Technologies for Industry 4.0: State-of-the-Art, Taxonomies, Perspectives, and Challenges," *IEEE Commun. Surveys & Tutorials*, vol. 21, no. 4, 2019, pp. 3467–3501.
- [11] A. Hazra *et al.*, "Collaborative AI-Enabled Intelligent Partial Service Provisioning in Green Industrial Fog Networks," *IEEE IoT J.*, 2021.
- [12] F. Zhu *et al.*, "Parallel Transportation Systems: Toward IoT-Enabled Smart Urban Traffic Control and Management," *IEEE Trans. Intelligent Transportation Systems*, vol. 21, no. 10, 2020, pp. 4063–71.
- [13] S. J. Nawaz *et al.*, "Quantum Machine Learning for 6G Communication Networks: State-of-the-Art and Vision for the Future," *IEEE Access*, vol. 7, 2019, pp. 46,317–50.
- [14] M. Adhikari *et al.*, "Security and Privacy in Edge-centric Intelligent Internet of Vehicles: Issues and Remedies," *IEEE Consumer Electronics Mag.*, 2021.



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

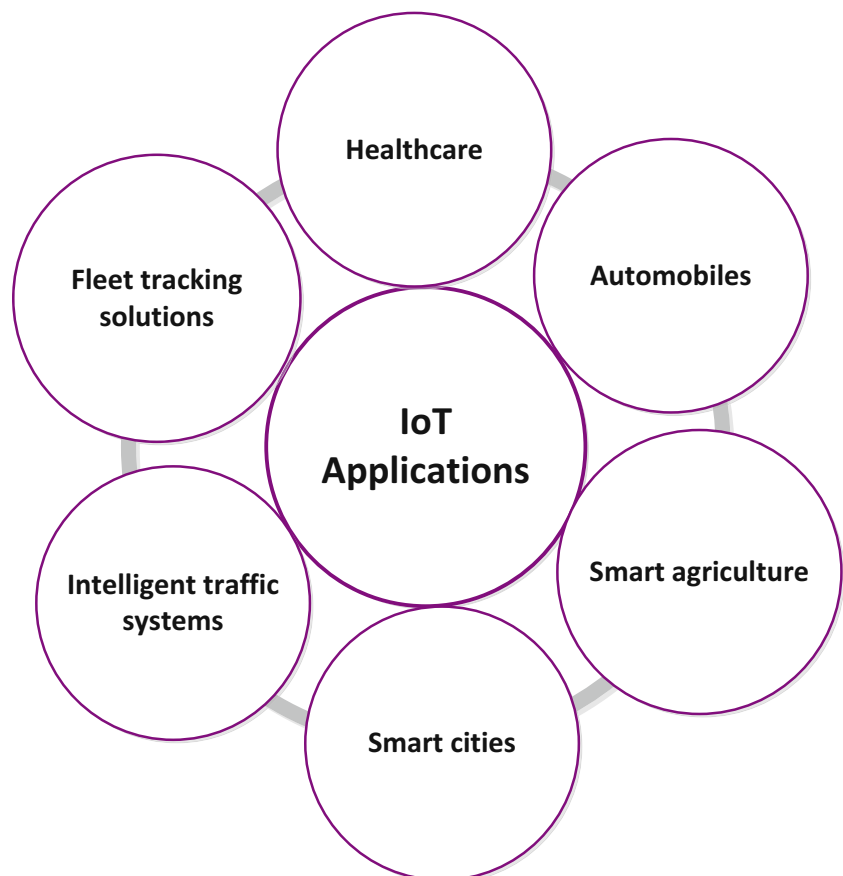
[3, 4]. These applications include smart healthcare [5], smart cities [6], smart agriculture [7], crowd sensing [8, 9], and crowd sourcing [10] etc., as shown in Fig. 1.

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

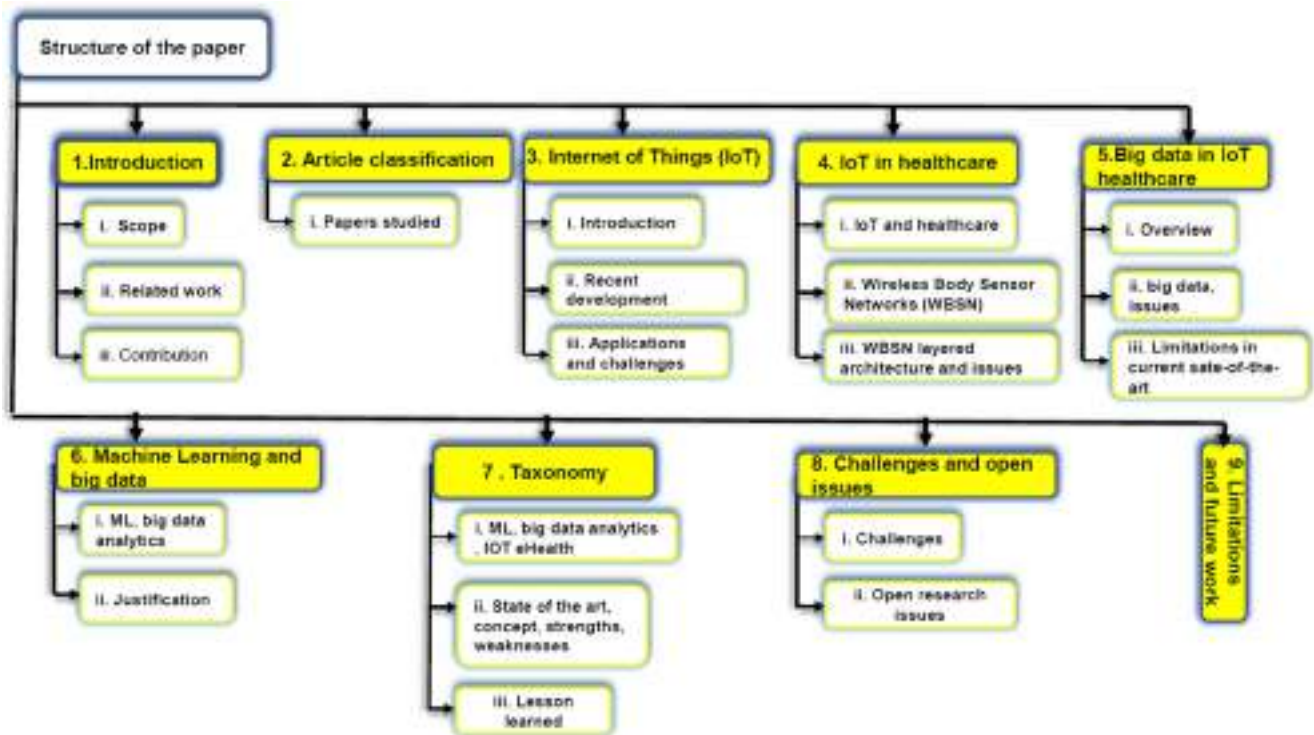


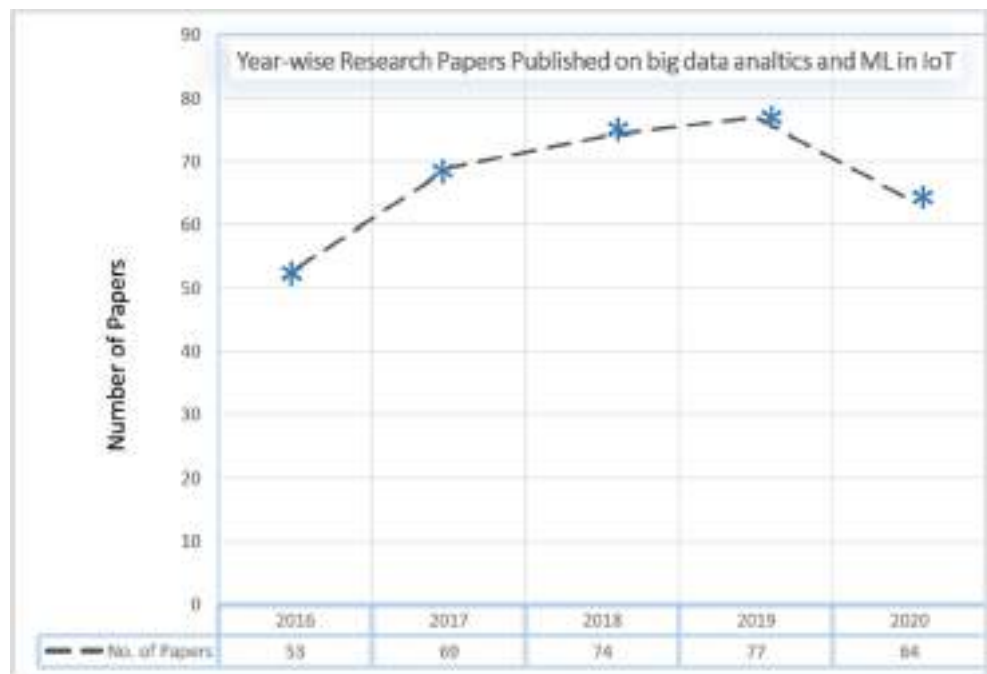
Fig. 2 Structure of the paper

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

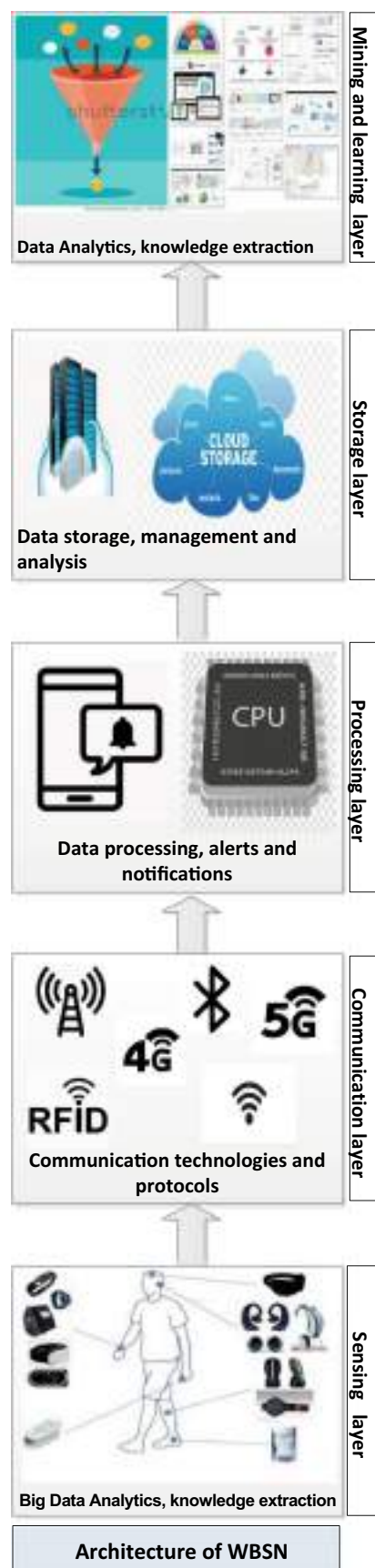


Fig. 4 Layered Architecture of Wireless Body Sensor Network

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

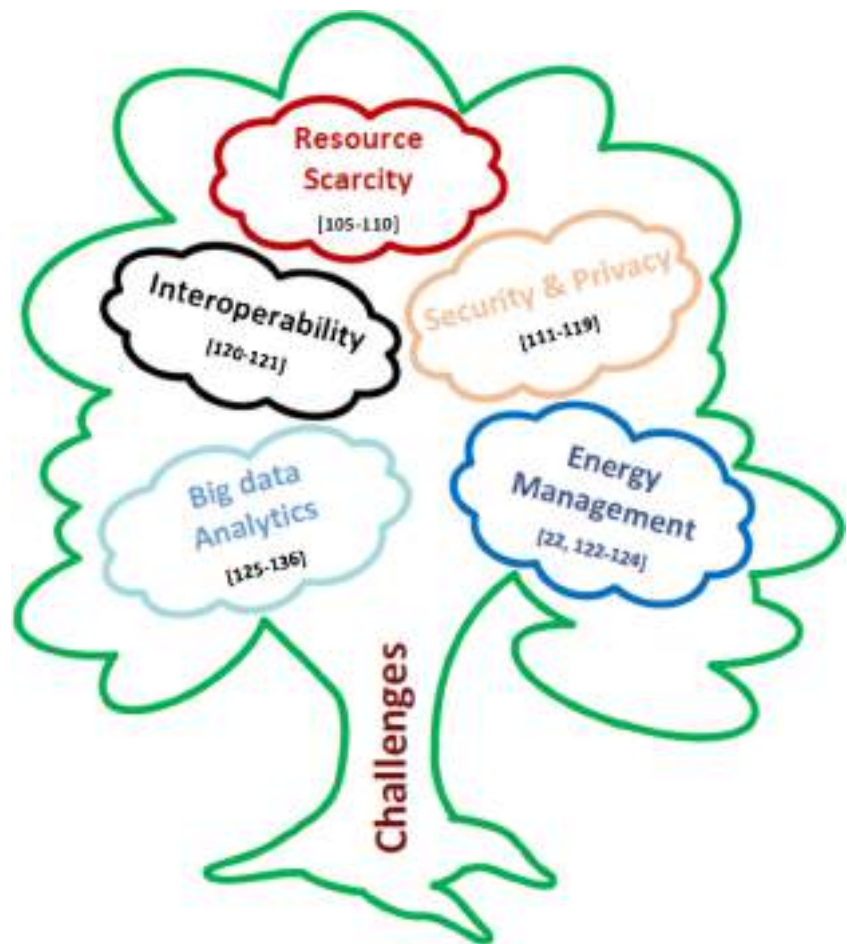
In IoT, most devices such as sensors, smart phones, microcontrollers actuators, RFIDs, and gateways have limited energy with lower computational and processing power [105–107]. Moreover, data generated from these densely

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 VJHS (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.