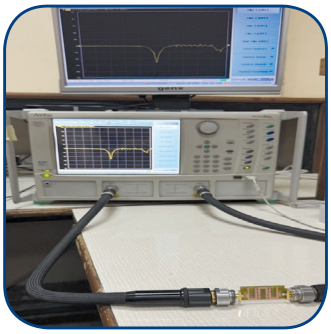
Volume 16, Number 2, 2025

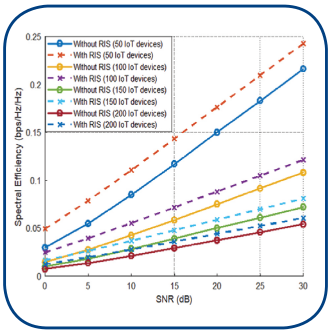
ISSN 1847-6996

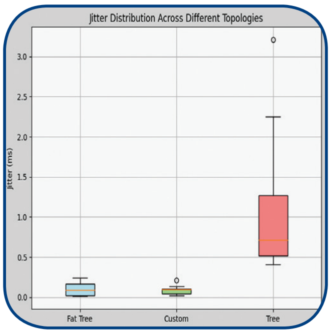
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

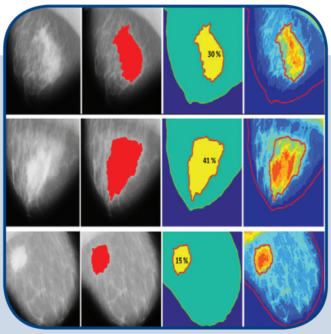
IJECES International Journal of Electrical and Computer Engineering Systems

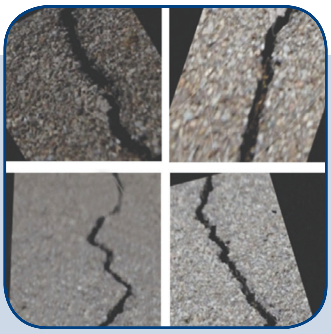
International Journal of Electrical and Computer Engineering Systems

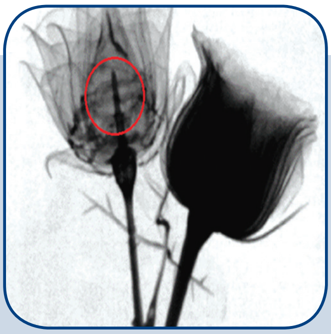



Without RIS (50 IoT devices)
With RIS (50 IoT devices)
Without RIS (100 IoT devices)
With RIS (100 IoT devices)
Without RIS (150 IoT devices)
With RIS (150 IoT devices)
Without RIS (200 IoT devices)
With RIS (200 IoT devices)
Spectral Efficiency (bps/Hz/Hz)
0.25
0.2
0.15
0.1
0.05
0
0
5
10
15
20
25
30
SNR (dB)









ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia

Osijek, Croatia | Volume 16, Number 2, 2025 | Pages 87 - 194

The International Journal of Electrical and Computer Engineering Systems is published with the financial support of the Ministry of Science and Education of the Republic of Croatia

CONTACT

International Journal of Electrical and Computer Engineering Systems (IJECES)

Faculty of Electrical Engineering, Computer
Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b, 31000 Osijek, Croatia
Phone: +38531224600, Fax: +38531224605
e-mail: ijeces@ferit.hr

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.
Giro account: 2390001 - 1100016777,
Croatian Postal Bank

Journal is referred in:

- Scopus
- Web of Science Core Collection (Emerging Sources Citation Index - ESCI)
- Google Scholar
- CiteFactor
- Genamics
- Hrčak
- Ulrichweb
- Reaxys
- Embase
- Engineering Village

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003
Published: quarterly
Circulation: 300

IJECES online

https://ijeces.ferit.hr

TABLE OF CONTENTS

Reconfigurable Intelligent Surfaces in 6G mMIMO NOMA Networks: A Comprehensive Analysis

Original Scientific Paper

Mohamed Hassan
Department of Electrical Engineering,
Omdurman Islamic University, Omdurman, Sudan,
mhbe4321@gmail.com

Khalid Hamid
Department of Electrical Engineering,
Omdurman Islamic University, Omdurman, Sudan,
khalidhamidk9@gmail.com

Rashid A. Saeed*
Department of Computer Engineering, College of
Computers and Information Technology,
Taif University, P.O. Box 11099, 21944, Taif, Saudi Arabia,
abdulhaleem@tu.edu.sa

*Corresponding author

Hesham Alhumyani
Department of Computer Engineering, College of
Computers and Information Technology,
Taif University, P.O. Box 11099, 21944, Taif, Saudi Arabia,
h.alhumyani@tu.edu.sa

Abdullah Alenizi
Department of Information Technology, College of
Computer and Information Sciences,
Majmaah University, Al-Majmaah 11952, Saudi Arabia,
aalenizi@mu.edu.sa

Abstract *– As the features and characteristics of six-generation (6G) connectivity are defined, advanced technologies such as multiple-input, multiple-output (mMIMO), non-orthogonal multiple access (NOMA), and reconfigurable intelligent surfaces (RISs) are becoming more important for many Internets of Things (IoT) uses. This study comprehensively and uniquely investigates the impact of RIS on the effectiveness of NOMA download (DL) mMIMO systems in the IoT environment within the context of the 6G network. This work aims to analyze the impact of including the RIS in the spectral efficiency (SE) and capacity performance of proposed hybrid system-enabled IoT setting device distributions, such as clustered and hotspot configurations. It highlights the ability of RIS to optimize wireless latency communication and throughput, depending on the mobility and density of IoT devices, respectively. The proposed methodologies are assessed through a simulation software application, under unstable channel conditions with varying distances and power locations while accounting for 256-quadrature amplitude modulation (256-QAM), frequency selective Rayleigh fading, and successive interference cancellation (SIC) context inside the 6G network environment. The results indicate that the four IoT groups (50, 100, 150, and 200) achieved capacity improvements of 5.84%, 5.81, 5.78, and 5.8%, and SE increases of 5.759%, 5.755%, 5.753%, and 5.84%, respectively, when utilizing RIS compared to their performance without it. The implementation of RIS yielded latency rate enhancements of 16.44%, 12.24%, 9.75%, and 8.1% across all four IoT groups, respectively, at a mobility speed of 120 Km/h. Each of the four IoT groups had throughput enhancements of 26%, 25.6%, 25.3%, and 25%, respectively, while using RIS within a coverage area of 400 square meters (sqm).*

Keywords*:* *Internet of Things, SE, massive MIMO (mMIMO), NOMA, RIS*

Received: Received: September 7, 2024; Received in revised form: October 25, 2024; Accepted: October 26, 2024

1. INTRODUCTION

The fifth generation (5G) represented a substantial advancement in network design, incorporating crucial technologies such as network slicing, centralized radio access networks (C-RAN), and improved mobile broadband. It facilitated reduced latency, increased capacity, and enhanced service flexibility with technologies including massive multiple-input multiple-output (mMIMO), beamforming, and edge computing. Nonetheless, sixth-generation (6G) aspires to surpass 5G in architectural design.

5G has yielded significant economic benefits, enabling progress in autonomous vehicles, industrial automation, and smart city development. However, 6G is expected to have a far higher impact, driving digital transformation throughout all sectors.

Numerous individuals assert that the emergence of 6G would herald a new technological epoch defined by the transmission of vast digital data among interconnected devices, humans, and automobiles. This interaction will build a self-sustaining ecosystem based on development and life. This ecosystem will utilise artificial intelligence (AI) to provide novel services.

The Internet of Things (IoT) is fundamentally transforming our way of life. It profoundly modifies human interactions with one another and their surroundings. According to the IoT concept, connections may be established across all entities and individuals [1, 2]. It is projected that by 2025, the number of IoT devices will have surpassed 100 billion, marking an unparalleled surge.

Many believe that the arrival of 6G will begin a new technological age in which interconnected devices, people, and cars will exchange vast quantities of digital data. Through this connection, an ecosystem that is self-sufficient and based on both development and life will be developed. This ecosystem will use artificial intelligence (AI) to provide novel services. The IoT is fundamentally transforming our way of life. It fundamentally reconstructs how individuals engage with one another and with their surroundings. Following the IoT paradigm, it is possible to establish connections between everything and everyone. The proliferation of IoT devices has been experiencing an unparalleled surge, with certain projections indicating that it may surpass 100 billion by the year 2025 [3-7].

Smart applications in many fields, including healthcare, smart cities, and industrial automation, are made possible by the IoT. A network is a collection of interconnected devices capable of direct communication with each other. With billions of devices relying on 6G, the IoT is anticipated to function at a large scale, necessitating low latency, and great reliability. Efficient management of spectrum resources and ensuring robust connectivity are essential for the effective service of a vast number of IoT devices [8].

To achieve the criteria of 6G, including minimal latency, optimal spectral efficiency (SE), feasible data rates, user fairness, and communication with a large number of devices [9], the non-orthogonal multiple access (NOMA) approach has been introduced, as described in [10]. NOMA is a crucial technology in wireless systems that improves spectrum efficiency. In contrast to conventional orthogonal multiple access (OMA) systems, NOMA enables several users to exploit the same frequency band by distinguishing them based on power levels [11]. This facilitates concurrent transmission to many users, enhancing system capacity and ensuring equitable treatment of users, particularly in situations characterized by a dense concentration of IoT devices [12].

The mMIMO means putting a lot of antennas at the base station (BS) so that it can serve a lot of customers at once. This improves SE and coverage. mMIMO is essential in 6G to meet the significant data throughput and low latency demands of IoT applications. It also improves the network's ability to manage the considerable connectivity required by IoT [13]. Recently, researchers have explored NOMA in mMIMO systems to improve spectrum efficiency [14]. A significant step towards better service quality has been the merging of these two technologies, especially when using huge devices in the IoT, as well as addressing two important challenges, including massive connectivity and low latency.

Reconfigurable intelligent surfaces (RISs) are surfaces that are fitted with programmable components capable of reflecting and manipulating electromagnetic waves in a desired way. In 6G, the RIS is strategically implemented to boost signal quality, expand coverage, and optimize the performance of IoT devices by dynamically adjusting the wireless environment.

Different from current solutions, RIS regulates wireless environment behaviour deterministically and programmatically. The majority of RIS implementations use 2D metasurface arrays. Tuning each element's phase shift skill fully changes signal propagation. Communication ancillary technology like amplification and forwarding relays uses more energy than RIS [14]. RIS technology offers a fresh approach to improving NOMA system performance by recreating the wireless environment; hence, we are strongly encouraged to incorporate RIS into NOMA [15]. The study examines the constraints and potential complications of traditional NOMA as user numbers rise, attributed to successive interference cancellation (SIC) at each user. Space-time block code-assisted NOMA (STBC-CNOMA) needs fewer SICs than traditional NOMA [16].

They thought about a communication setting where all users are the same, with NOMA users clustered together. In their studies, the writers of [17-19] looked at the issues with resource management in multi-cell MIMO networks. For instance, to maximize the users' total capacity, the writers in [20-22] offered a less-than-ideal plan. Their findings demonstrated that NOMA systems can still achieve considerable gains in user capacity when inferior approaches are used.

Moreover, [23] investigated the benefits of RIS in a parasitic radio (SR) system. The authors devised passive and active RIS and BS beamforming to reduce BS transmission power. These designs were based on two constraints: the rate constraint for core communication and the signal-to-interference-plus-noise ratio (SINR) for decoding backscattered signals.

The system models in references examined networks with users using a single antenna throughout the network, therefore constraining the productivity of IoT devices. For instance, in order to maximize the users' total capacity, the writers in [24- 27] offered a less-than-ideal plan. Their findings demonstrated that NOMA systems can still achieve considerable gains in user capacity when inferior approaches are used.

Moreover, [28] investigated the benefits of RIS in a parasitic radio (SR) system. The authors devised passive and active RIS and BS beamforming to reduce BS transmission power. These designs were based on two constraints: the rate constraint for core communication and the signal-to-interference-plus-noise ratio (SINR) for decoding backscattered signals.

The system models in references [29, 30] examined networks with users using a single antenna throughout the network, therefore constraining the productivity of IoT devices.

The majority of the prior work on IRS-enabled NOMA systems was based on the premise of a perfectly stable channel, which is not feasible for real-world situations. However, further investigations are still needed to improve the performance of RIS and NOMA. This work mainly highlights the following contributions:

- This study examines the effects of implementing the RIS in a mMIMO DL NOMA-enabled IoT environment. By implementing RIS, the research demonstrates substantial enhancements in both capacity and SE across different IoT user distributions, showcasing the capability of RIS to optimize wireless communication systems in 6G networks.
- The work investigates the efficiency of NOMA systems throughout various distributions of IoT devices, including clustered and hotspot setups. Incorporating user distribution patterns into the simulation enhances its realism by accurately representing the impact on system performance and exploring the system's scalability and strategies for efficiently combining mMIMO systems with RIS to enhance the overall performance of NOMA systems in IoT networks.
- The paper looks at how well IoT devices handle latency at different mobility speeds with and without RIS. It gives useful information about how well RIS works in different situations by looking at unstable channel conditions, Rayleigh fading, and SIC. We aim to enhance the system's realism, emulate the actual environment, adhere to design constraints, and boost its performance.

The following is the structured rest of the article. A concise review of the pertinent literature is presented in Section 2. The mathematical formulation procedures and the network model specifics are detailed in Section 3. The suggested system parameters, findings, and discussions are presented in Section 4. Section 5 offers final thoughts and recommendations for further research.

2. RELATED WORKS

The study in [31] provides a clear and thorough explanation of RIS technology, addressing its rationale, applications, and locations of usage, while also discussing the challenges and corresponding solutions. However, the case study was extremely inadequate in terms of the number of users. [32] provided an extensive overview of RIS systems, emphasizing their operational principles, performance assessment, development, design, and interaction with other new technologies. Nevertheless, the issues confronting RIS technology during congestion or mobility and their effects on performance were not recognized.

[33] examines NOMA utilizing RIS, concentrating on enhancing power allocation for each user and bandwidth configuration in RIS, while guaranteeing compliance with minimum rate, latency, and reliability standards. The numerical findings indicate that the suggested strategy attains an acceptable rate. The system is predicated only on optimizing power allocation, which presents a constraint as the cumulative power allocations total one and are distributed at varying rates dependent on the number of users and their respective locations. The developed NOMA system with RIS partitioning improves SE by increasing user fairness and ergodic rate [34]. The balanced sum rate, outage probability, and user fairness performance of the proposed system beat the benchmark systems. The primary concept is around the partitioning of RIS; yet, the system has not been thoroughly examined, particularly concerning significant aspects like interference.

Two successful IRS-based channel estimate algorithms for different channel parameters in a multi-user broadband communication system with orthogonal multiple access (OMA) are proposed in [35]. The findings demonstrate that the suggested channel estimation techniques and training strategies outperform comparator systems. NOMA surpasses OMA since the BS consistently determines the user's position and transmits the appropriate power, hence enhancing the implementation of these approaches and yielding better outcomes. The use of IRS to improve coverage by facilitating communication between the cell edge user device and the base station is examined in [36] for both DL and uplink (UL) NOMA and OMA networks. When compared to complete decoding and forward relay, the findings show that IRS is far better. However, the dimensions and spacing of cells for RIS to accommodate edge users remain undetermined.

For signal cancellation-based RISs in the MIMO NOMA network that supports concurrent users, a novel passive beamforming weight design is offered [37]. According to the findings of the analysis, inter-group interference may be eliminated by using a high number of components. In contrast to the anomalous reflector scenario, the diffuse dispersion scenario requires line-of-sight for the BS-RIS and RIS-user connections, which causes shortcomings in this methodology. The mMIMO BS with RIS powers IoT devices wirelessly and enables multiple data users. The BS's precoding uses dual-band transmission and an instantaneous stable channel, whereas the RIS's passive beamforming uses a progressively moving statistically stable channel. Pilot contamination and channel estimation errors were used to generate closed-form formulae for IoT device information user SE and average power [38]. Nevertheless, the mechanism for the distribution of IoT devices and the sort of network employed remains undetermined. The following Table 1 provides a very simplified and comparative analysis of our work with previous literature, with a brief explanation of our contributions.

Table 1. Comparison with previous literature and contributions to the paper

Author's Name	IOT	RIS	NOMA	mMIMO	Device Mobility	6G
C. Pan (2021)	-	✓	-	-	-	-
Y. Liu (2021)	✓	✓	✓	✓	-	✓
Z. Ding (2023)	-	✓	✓	-	-	-
E. Basar (2022)	-	✓	✓	-	-	-
B. Zheng (2020)	-	✓	-	-	-	-
Y. Cheng (2021)	-	✓	✓	-	-	-
T. Hou (2020)	✓	✓	✓	-	-	-
H. Q. Ngo (2024)	✓	✓	✓	✓	-	-
This work	✓	✓	✓	✓	✓	✓

3. SYSTEM MODELS

Assuming the IoT setup using mMIMO and downlink (DL) NOMA, we will determine the system's primary components and how they interact with each other, as shown in Figs. 1 and 2.

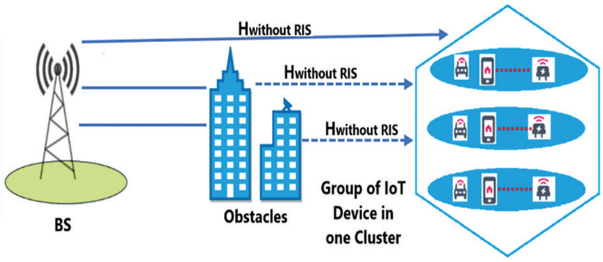


Fig. 1. IoT mMIMO DL NOMA system without RIS technology

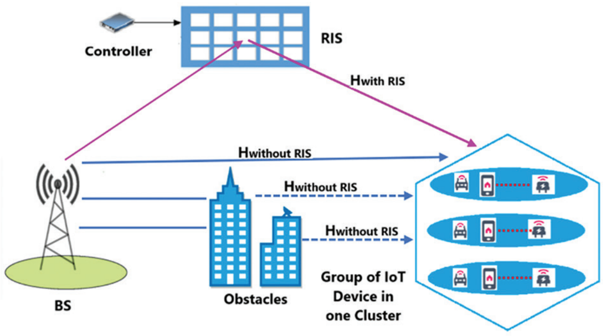


Fig. 2. IoT mMIMO DL NOMA system with RIS scheme

The BS has an array of 128 x 128 mMIMO antennas. By distinguishing between users' power levels, the BS enables numerous IoT devices to share the same frequency band through NOMA signal transmission, and each IoT device has an antenna [39]. In the given case, the clustered distribution model involves grouping devices into clusters and the Rayleigh fading channel model is utilized to characterize the wireless channel between the BS and the IoT devices. The channel model incorporates path loss, antenna gains, RIS enhancement, and mobility effects. Frequently, this assumption accurately represents the actual signal issues and challenges that devices encounter in specific regions.

Under the same prior assumptions, the proposal system engages with the RIS in the second scenario. In this configuration, the base station employs the identical mMIMO architecture as previously to engage with several IoT devices. The channel matrix H represents the mMIMO system that consists of N_{Tx} broadcast antennas and N_{Rx_r} receive antennas, $H \in C^{N_{Rx} \times N_{Tx}}$). The channel matrix H between the BS and an IoT device in the mMIMO system is denoted as:

$$H = \frac{L_p G}{N_{Rx} \times N_{Tx}} \tag{1}$$

In this equation, L_p represents large-scale path loss, whereas G represents small-scale Rayleigh fading channel coefficients. It can be expressed as:

$G \sim CN(0,1)$, where $CN(0,1)$denotes a complex Gaussian distribution characterised by a mean of zero and a variance of one.

The BS concurrently accommodates several IoT devices utilizing the NOMA concept. In NOMA, numerous users utilize the same frequency and temporal resources, differentiated by their power levels [40]. The BS assigns varying power levels to customers according to their channel conditions, guaranteeing that those with superior channels receive reduced power, while those with inferior channels receive more power. This is articulated as follows: For K users or devices, the power assigned to the *k-th* user or device is represented as P_k, with $\sum^{K}_{k=1} Pk = P_{total}$, where P_{total} signifies the total available power.

In the given case, the clustered distribution model involves grouping devices into clusters with unstable channel conditions. Frequently, this assumption accurately represents the actual signal issues and challenges that devices encounter in specific regions.

let x represent the transmit signal vector from the BS.

$$x = Ws \tag{2}$$

For each user K, the symbol vector is represented by $s \in C^{K \times 1}$. W, which is a precoding matrix applied at the BS, is defined as $W \in C^{N_t \times K}$. Channel matrices in mMIMO systems with no RIS. Adjustments can be made to the channel gain in a clustered distribution to accurately represent the increased signal intensity resulting from the proximity of users inside clusters [41].

$$G_{base} = G_{base\_factor} \times G_{cluster\_adjust} \tag{3}$$

G_{base} is total system gain, including base gain factor and clustering modifications. Without clustering effects, G_{base_factor} is the system or antenna setup's inherent gain. Such as antenna efficiency, transmission power, etc. $G_{cluster\_adjust}$ multiplicative modification to base gain. This adjustment factors in gain increases from beamforming, geographical clustering, and user clustering to optimize signal strength [42].

$$G_{base}=G_{base\_factor}\times G_{hotspot\_adjust} \quad (4)$$

$G_{hotspot\_adjust}$ is multiply adjustments for enhanced signal gain in hotspot locations. Regions with a higher user concentration or stronger signal focus may have higher gain due to localized optimizations such beamforming, power allocation, or signal upgrades [43].

The channel matrix H_{withis} improved by the RIS matrix R upon its introduction,

$$H_{with\ RIS}=H_{without\ RIS}\cdot R \quad (5)$$

where H_{with} is the channel matrix when the RIS or another transformation (beamforming, phase-shifting, etc.) is used. The channel matrix $H_{without}$ represents the direct channel between transmitter and receiver, without any modification. R is a transformation matrix used by the RIS to enhance or improve channel conditions by applying phase shifts or adjustments to incoming signals.

The diagonal matrix R, which represents RIS, contains phase shifts denoted as,

$$R=diag(e^{j\phi_1},e^{j\phi_2},\ldots,e^{j\phi_{N_{RIS}}}) \quad (6)$$

The number of RIS elements is represented by N_{RIS}. The phase shift of the *i-th* RIS is denoted by ϕ_i. The complex phase shift is referred to as the equation $e^{j\phi_i}$). For the channel gain without RIS,

$$Gain_{eff}=Gain_{base} \quad (7)$$

For the channel gain with RIS, the effective channel gain is,

$$Gain_{eff}=Gain_{base}\times RIS_{enhancement} \quad (8)$$

where $RIS_{enhancement}$ Enhancement factor provided by RIS. Examine the intermediary channel connecting the BS and the user devices. The signal y_k received by user K precisely represented as [44],

$$y_k = h_k^H W s + n_k \quad (9)$$

The user k's channel vector is represented as $h_k \in C^{N_t\times 1}$. The additive white Gaussian noise (AWGN) with variance $\sigma_n{}^2$ is denoted by $n_k \sim CN(0, \sigma_n{}^2)$. $h_k{}^H$ the Hermitian (conjugate transpose) of the h_k. In the case of RIS, the channel vector h_k is increased by the channel information signal RIS. Consider the channel $G \in C^{N_{RIS}\times N_t}$ connecting the BS to RIS, and the channel $v_k \in C^{N_{RIS}\times 1}$ connecting the RIS to user k. The effective channel h_k with RIS can be expressed as [45],

$$h_k = v_k^H \Theta G + h_{direct,k} \quad (10)$$

This is the RIS phase shift matrix: Θ=diag(θ_1, θ_2,…,$\theta_{N_{RIS}}$), where the *i-th* RIS element introduces a phase shift, denoted as θ_i. Without utilising the RIS, the direct connection from the BS to user k is represented as channel $h_{direct,\ k}$. The signal strength received by user k without RIS is [46],

$$P_{no\ RIS,k} = |h_k^H W_k|^2 P_{tx} \quad (11)$$

The received signal power changes in the RIS scenario to,

$$P_{with\ RIS,k} = \left|(v_k^H \Theta G + h_{direct,k})^H W_k\right|^2 P_{tx} \quad (12)$$

where P_{tx} is the transmit power. The user k's SNR is calculated as,

$$SNR_{no\ RIS,k} = \frac{P_{tx}\cdot\left|h_k^H W_k\right|^2}{\sigma_n^2} \quad (13)$$

$$\mathrm{SNR}_{\mathrm{with\ RIS,k}} = \frac{\mathrm{P_{tx}}\cdot\left|(\mathrm{v_k^H}\Theta\mathrm{G}+\mathrm{h_{direct,k}})^{\mathrm{H}}\mathrm{W_k}\right|^2}{\sigma_{\mathrm{n}}^2} \quad (14)$$

user's allotted transmit power is denoted as P_{tx}. By applying the Shannon-Hartley theorem, we may determine user k's capacity C_k,

$$C_k = \frac{BW}{K}\cdot\log_2(1+SNR_k) \quad (15)$$

The total bandwidth is denoted by BW. K is the number of users. The capacity normalized by the bandwidth is the SE_k for user k,

$$\mathrm{SE}_k = \frac{1}{K}\cdot\log_2(1+SNR_k) \quad (16)$$

The Capacity and SE of the System Overall with and Without RIS,

$$C_{no\ RIS,k} = \frac{BW}{K}\cdot\log_2\left(1+\frac{\left|P_{tx}\cdot h_k^H W_k\right|^2}{\sigma_n^2}\right) \quad (17)$$

$$SE_{no\ RIS,k} = \frac{1}{K}\cdot\log_2\left(1+\frac{\left|P_{tx}\cdot h_k^H W_k\right|^2}{\sigma_n^2}\right) \quad (18)$$

$$C_{with\ RIS,k} = \frac{BW}{K}\cdot\log_2\left(1+\frac{P_{tx}\cdot\left|(v_k^H \Theta G+h_{direct,k})^H W_k\right|^2}{\sigma_n^2}\right) (19)$$

$$SE_{with\ RIS,k} = \frac{1}{K}\cdot\log_2\left(1+\frac{P_{tx}\cdot\left|(v_k^H \Theta G+h_{direct,k})^H W_k\right|^2}{\sigma_n^2}\right) (20)$$

Device mobility, network load, and channel conditions all affect communication system latency. This model is used for broad analysis,

$$L = L_{base} + D_{mob} + D_{load} \quad (21)$$

The total network delay (L) comprises base latency, mobility delays, and network load delays.

L_{base} refers to the system's intrinsic latency under ideal conditions, excluding mobility and network load considerations. D_{mob}: The delay during user or device movement, determined by speed and distance. D_{load} refers to the delay caused by network traffic. As network congestion or user activity grows, this delay increases. Network load delay, this metric quantifies the influence of the present network load on the delay [47].

$$L_{without\ RIS} = L_{base} + \alpha\cdot M + \beta\cdot N \quad (22)$$

$$L_{with\ RIS} = L_{base} + \alpha\cdot M\cdot R + \beta\cdot N \quad (23)$$

where, M is network users' or devices' speed or mobility, which affects latency owing to handovers and dynamic connection quality. N is network traffic or users, with larger loads causing congestion and delay.

R scales the RIS's latency impact. α is a proportionality constant that affects latency increase with movement. The constant β measures the influence of network load on latency, by scaling congestion's contribution to overall delay.

$$D_{mob} = \gamma\cdot S\cdot d \quad (24)$$

D_{mob} is a Mobility Delay. Speed is S. Distance (D). The mobility delay coefficient is γ.

where, γ is a proportionality constant. S is the speed of the device. D is the distance the device moves in the given time. The formula typically used to determine the throughput T is,

$$T = BW \cdot \log_2\left(1 + \frac{SNR \cdot G}{N_0}\right) \tag{25}$$

where, G is Channel gain (with or without RIS), N_0 is Noise power spectral density.

4. SIMULATION RESULTS AND DISCUSSION

Table 2 displays the simulation parameters for various model networks. The graphs of IoT mMIMO DL NOMA devices illustrate the variations in capacity and SE, and SNR under different conditions. The outcomes are displayed before and following the implementation of RIS, which enhances the latency with device mobility, throughput, and coverage area of the network among the group of devices, while also tackling the difficulties posed by an abundance of IoT devices and resource-intensive 6G network applications. Fig. 3 flowchart shows the methodology process and the simulation steps required.

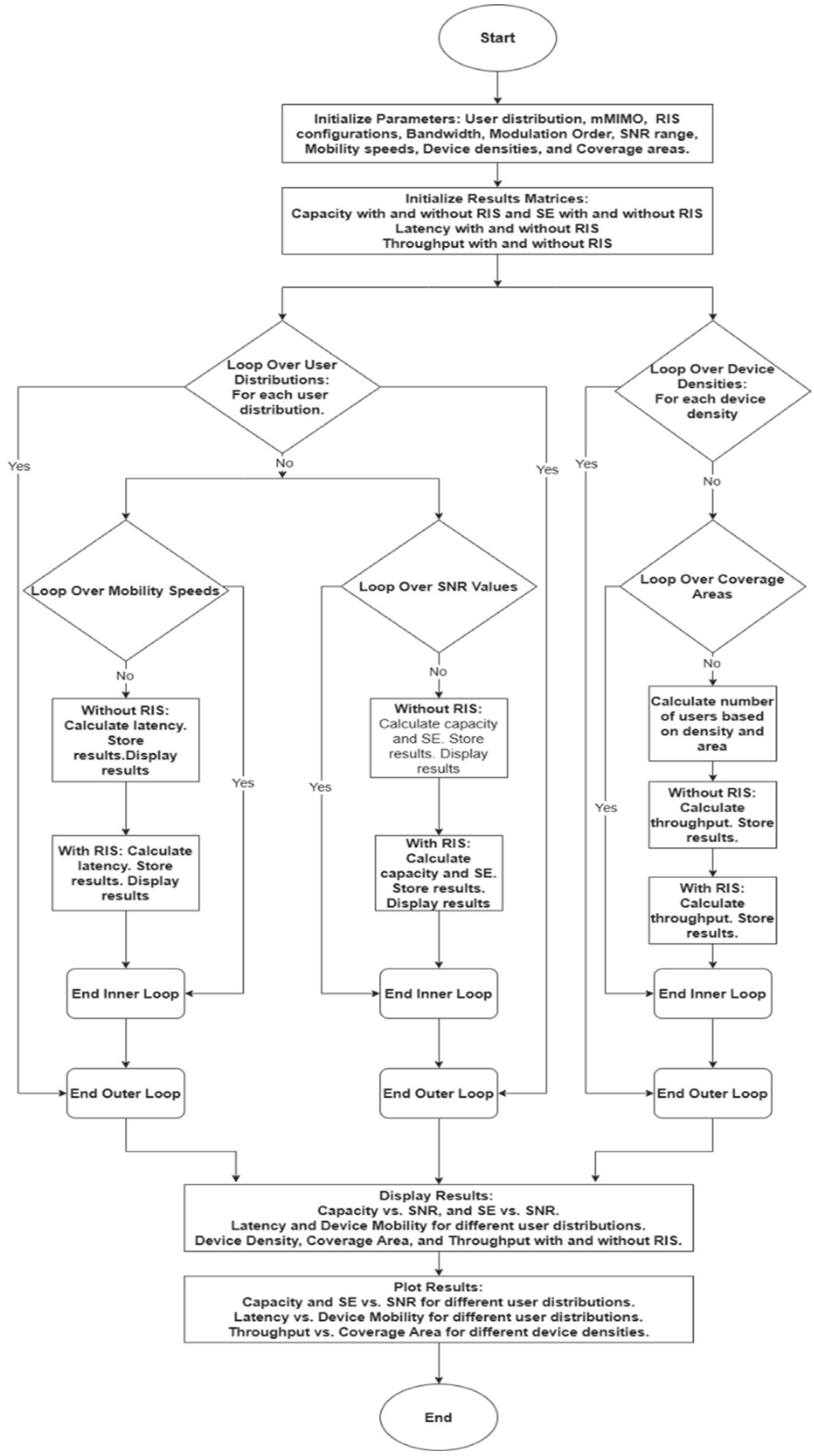

Start
Initialize Parameters: User distribution, mMIMO, RIS configurations, Bandwidth, Modulation Order, SNR range, Mobility speeds, Device densities, and Coverage areas.
Initialize Results Matrices: Capacity with and without RIS and SE with and without RIS Latency with and without RIS Throughput with and without RIS
Loop Over User Distributions: For each user distribution.
Loop Over Device Densities: For each device density
Yes
No
Yes
Yes
No
Loop Over Mobility Speeds
Loop Over SNR Values
Loop Over Coverage Areas
No
No
No
Without RIS: Calculate latency. Store results.Display results
Without RIS: Calculate capacity and SE. Store results. Display results
Calculate number of users based on density and area
Yes
Yes
Yes
With RIS: Calculate latency. Store results. Display results
With RIS: Calculate capacity and SE. Store results. Display results
Without RIS: Calculate throughput. Store results.
With RIS: Calculate throughput. Store results.
End Inner Loop
End Inner Loop
End Inner Loop
End Outer Loop
End Outer Loop
End Outer Loop
Display Results: Capacity vs. SNR, and SE vs. SNR. Latency and Device Mobility for different user distributions. Device Density, Coverage Area, and Throughput with and without RIS.
Plot Results: Capacity and SE vs. SNR for different user distributions. Latency vs. Device Mobility for different user distributions. Throughput vs. Coverage Area for different device densities.
End

Fig. 3. Three types of IoT mMIMO DL NOMA cases with and without RIS systems are shown in a flow chart

In the 6G network, Fig. 4 illustrates the interaction between the capacity rate and SNR for four mMIMO DL NOMA IoT groups, both with and without the RIS method. The IoT devices own channels that exhibit variations in both distance and power distribution. A strong association existed between the capacity rate and the SNR. The group in the cluster with fewer devices (50 devices) supports the highest achievable capacity of 1.29 Mbps/Hz. The capacity rates for the second group with (100 devices), the third group with (150 devices), and the fourth group with (200 devices) were calculated to be 0.648, 0.432, and 0.324 Mbps/Hz, respectively. All four IoT groups saw capacity improvements of 5.84%, 5.81, 5.78, and 5.8% when using RIS, compared to their performance without RIS. This was confirmed at the SNR of 30 dB.

Table 2. Presents comprehensive information on the simulators employed for simulating various models

Parameter	Value
Devices Groups	50, 100, 150, and 200
Modulations	256 QAM
Path-loss exp.	4
BW	6 G Hz
Antennas No.	128x128 mMIMO
RIS	128x128
SNR	0 to 30 dB
Mobility Speed	(0-120) km/h
Coverage Areas	(100, 200, 300, 400) m2 or sqm

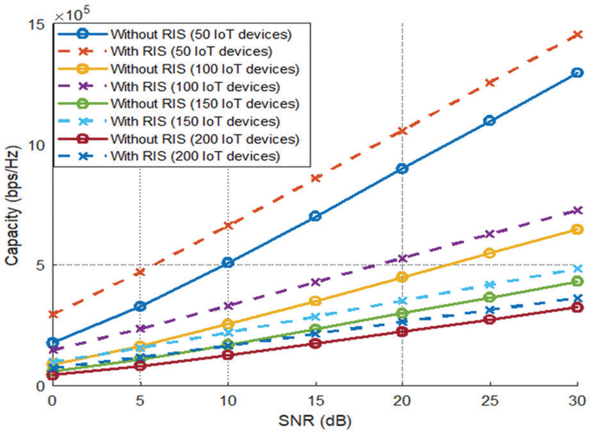


Fig. 4. Capacity rate vs. SNR for 4 groups in the IoT mMIMO DL NOMA network with and without RIS system

Within the 6G network, Fig. 5 depicts the correlation between the SE and SNR for four mMIMO DL NOMA IoT groups, both with and without the RIS architecture, and a strong correlation between SE and SNR was observed. The subgroup inside the cluster consisting of 50 devices allows for the maximum attainable SE of 0.21629 bps/Hz. The predicted SE rates for the second group consisting of 100 devices, the third group with 150 devices, and the fourth group consisting of 200 devices were determined to be 0.10815, 0.072097, and 0.054073 bps/Hz, respectively. All four IoT groups experienced SE enhancements of 5.759%, 5.755%, 5.753%, and 5.84%, respectively, when employing RIS, in comparison to their performance in the absence of RIS. This was verified at the SNR of 30 dB. The final result exceeded the results obtained according to references.

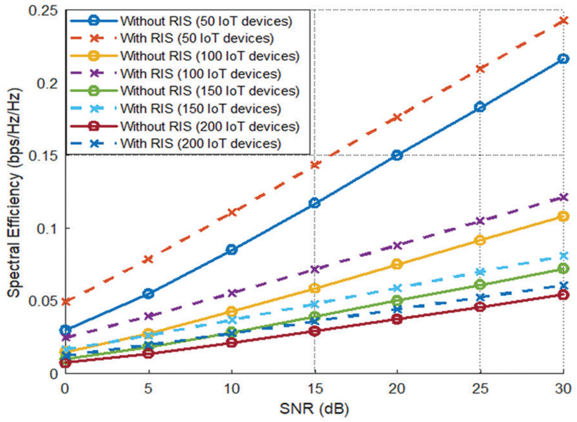


Fig. 5. SE against SNR for 4 groups in the IoT mMIMO DL NOMA network with and without RIS technique

Implementing RIS in the mMIMO system notably improves both the transmission capacity and the efficiency of spectrum utilization. The primary factor behind this enhancement is the improved effective channel gain offered by RIS, which amplifies the signal strength, hence generating more capacity and optimizing the utilization of the bandwidth. This improvement is especially noticeable at increased SNR levels and broader user dispersal, rendering RIS a valuable solution in IoT and 6G networks. The final result exceeded the results obtained according to references [48].

Fig. 6 illustrates the correlation between latency vs. device mobility with and without RIS for Different IoT Device Distributions for four mMIMO NOMA groups. The study focuses on a 6G network in different mobility scenarios: from 0 km/h to 120 km/h. During periods of slow movement or immobility, the channel conditions between the transmitter (such as BS) and the receiver (such as an IoT device) stay rather constant. The stability of the channel results in less fast fluctuations in both channel gains and signal quality, therefore enabling more consistent and dependable communication. By contrast, the channel conditions of faster-moving devices undergo rapid changes as a result of phenomena like as Doppler shift, multipath fading, and frequent handovers between various BS. Fluctuations in signal quality caused by these quick changes might result in increased delay, greater packet loss, and more frequent mistakes.

The subgroup inside the cluster, which has 50 devices, enables a minimum achievable latency of 2.44 milliseconds. The estimated latency rates for the second group, which included 100 devices, the third group, which included 150 devices, and the fourth group, which included 200 devices, were analyzed and found to be 1.72 ms, 1.48 ms, and 1.36 ms accordingly, at a velocity of 120 Km/h.

Implementing RIS resulted in latency rate improvements of 16.44%, 12.24%, 9.75%, and 8.1% for all four IoT groups, respectively.

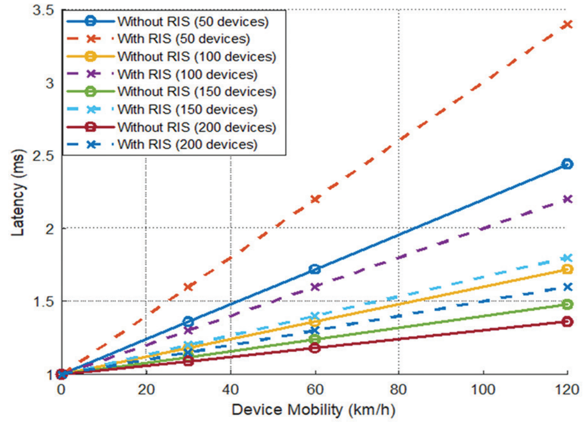
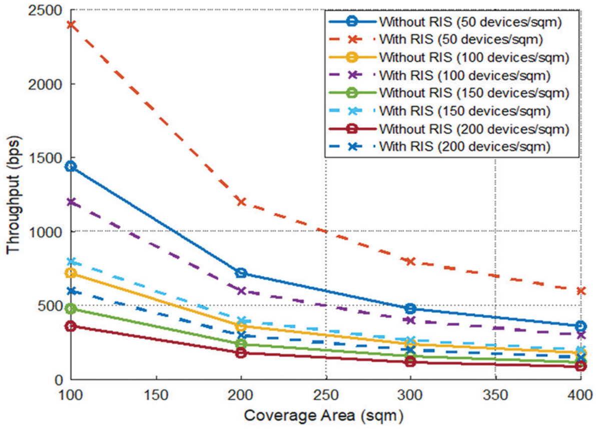


Fig. 6. Latency vs. Device Mobility with and without RIS for Different 4 groups IoT mMIMO DL NOMA Device Distributions

Fig. 7 depicts the relationship between Throughput and Coverage Area for four mMIMO NOMA groups, both with and without RIS, with varying device densities.



The study examines a 6G network under several Device Density scenarios, namely 100, 200, 300, and 400 square metres. Coefficient of inverse correlation between Throughput and coverage area. The subgroup inside the cluster, including 50 devices, allows for the maximum attainable throughput of 360 bps/Hz. The predicted throughput rates for the second group consisting of 100 devices, the third group with 150 devices, and the fourth group containing 200 devices were determined to be 180, 120, and 90 bps/Hz, respectively at the coverage area of 400 square meters (sqm). Each of the four IoT groups saw throughput improvements of 26%, 25.6%, 25.3%, and 25% correspondingly while using RIS. The augmented channel gains that RIS facilitates are primarily to blame for the increase and improvement in throughput. The RIS enhances signal propagation by dynamically regulating signal reflections, leading to improved connection quality and optimizing resource utilization, minimizing delay, and countering the effects of multipath fading. By adding NOMA, throughput is increased, making RIS-assisted NOMA a powerful method for fast communication in future 6G networks.

Fig. 7. Throughput vs. coverage area with and without RIS for different 4 groups IoT mMIMO DL NOMA device densities

5. CONCLUSIONS

This study provides a thorough evaluation of RIS technology in a 6G network environment, specifically within the mMIMO DL NOMA-enabled IoT system. The results show that the proposed approach is a promising solution for future 6G IoT networks, significantly improving capacity and SE, particularly in high-density areas. By integrating RIS with DL NOMA systems, the research highlights its ability to boost communication performance, especially in clustered user scenarios, making it essential for optimizing system performance and resource utilization in IoT networks. Additionally, RIS greatly reduces latency in NOMA systems, regardless of user distribution or mobility speeds, and improves communication efficiency. The latency model used provides key insights into the impact of RIS and mobility on system performance. Furthermore, RIS enhances data transfer rates, supporting increased device densities in various IoT environments, and the throughput model demonstrates its ability to optimize network architecture. In conclusion, RIS presents considerable benefits for improving network performance in 6G IoT systems. Future studies should explore its integration with advanced technologies like machine learning-based resource allocation and dynamic spectrum management to further enhance NOMA system performance in diverse IoT scenarios.

6. ACKNOWLEDGEMENT

The authors extended their appreciation to Taif University, Saudi Arabia, for supporting this work through project number (TU-DSPP-2024-132).

7. FUNDING

The research was funded by Taif University, Taif, Saudi Arabia (TU-DSPP-2024-132).

8. REFERENCES

[1] T. Qiu, N. Chen, K. Li, M. Atiquzzaman, W. Zhao, "How Can Heterogeneous Internet of Things Build Our Future: A Survey", IEEE Communications Surveys & Tutorials, Vol. 20, No. 3, 2018, pp. 2011-2027.

[2] B. U. Rehman, et al. "Joint power control and user grouping for uplink power domain non-orthogonal multiple access", International Journal of Distributed Sensor Networks, Vol. 17, No. 12, 2021.

[3] M. Hassan et al. "BER Improvement of Cooperative Spectrum Sharing of NOMA in 5G Network", Pro-

ceedings of the IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Benghazi, Libya, 21-23 May 2023, pp. 647-652.

[4] G. Fortino, C. Savaglio, G. Spezzano, M. Zhou, "Internet of Things as System of Systems: A Review of Methodologies, Frameworks, Platforms, and Tools", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 51, No. 1, 2021, pp. 223-236.

[5] Y. Al Mtawa, A. Haque, B. Bitar, "The Mammoth Internet: Are We Ready?", IEEE Access, Vol. 7, 2019, pp. 132894-132908.

[6] Y. Qian, D. Wu, W. Bao, P. Lorenz, "The Internet of Things for Smart Cities: Technologies and Applications", IEEE Network, Vol. 33, No. 2, 2019, pp. 4-5.

[7] N. N. Elmadina et al. "Performance of Power Allocation Under Priority User in CR-NOMA", Proceedings of the IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Benghazi, Libya, 21-23 May 2023, pp. 618-622.

[8] L. U. Khan, I. Yaqoob, M. Imran, Z. Han, C. S. Hong, "6G Wireless Systems: A Vision, Architectural Elements, and Future Directions", IEEE Access, Vol. 8, 2020, pp. 147029-147044.

[9] B. Zheng, Q. Wu, R. Zhang, "Intelligent Reflecting Surface-Assisted Multiple Access With User Pairing: NOMA or OMA?", IEEE Communications Letters, Vol. 24, No. 4, 2020, pp. 753-757.

[10] F. Solano, S. Krause, C. Wöllgens, "An Internet-of-Things Enabled Smart System for Wastewater Monitoring", IEEE Access, Vol. 10, 2022, pp. 4666-4685.

[11] Y. Zhang et al. "Performance Analysis of CF-mMIMO-Aided SWIPT IoT Networks With Nonideal RF Response and Low-Resolution ADCs/DACs", IEEE Sensors Journal, Vol. 24, No. 3, 2024, pp. 3594-3607.

[12] M. Hassan et al. "NOMA Cooperative Spectrum Sharing Average Capacity Improvement in 5G Network", Proceedings of the IEEE 3rd International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Benghazi, Libya, 21-23 May 2023, pp. 653-658.

[13] K. Senel, H. V. Cheng, E. Björnson, E. G. Larsson, "What Role can NOMA Play in Massive MIMO?", IEEE Journal of Selected Topics in Signal Processing, Vol. 13, No. 3, 2019, pp. 597-611.

[14] Y. Zhang, L. Xiao, T. Jiang, "Cloud-Based Cell-Free Massive MIMO Systems: Uplink Error Probability Analysis and Near-Optimal Detector Design", IEEE Transactions on Communications, Vol. 70, No. 2, 2022, pp. 797-809.

[15] Y. Zhang, B. Di, H. Zhang, J. Lin, Y. Li, L. Song, "Reconfigurable Intelligent Surface Aided Cell-Free MIMO Communications", IEEE Wireless Communications Letters, Vol. 10, No. 4, 2021, pp. 775-779.

[16] M. W. Akhtar, S. A. Hassan, S. Saleem, H. Jung, "STBC-Aided Cooperative NOMA With Timing Offsets, Imperfect Successive Interference Cancellation, and Imperfect Channel State Information", IEEE Transactions on Vehicular Technology, Vol. 69, No. 10, 2020, pp. 11712-11727.

[17] M. Hassan et al. "Capacity Enhancement Based on mMIMO for Cognitive Radio NOMA in Future 6G Networks", Proceedings of the IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering, Tripoli, Libya, 19-21 May 2024, pp. 387-392.

[18] J. Zuo, Y. Liu, Z. Ding, L. Song, H. V. Poor, "Joint Design for Simultaneously Transmitting and Reflecting (STAR) RIS Assisted NOMA Systems", IEEE Transactions on Wireless Communications, Vol. 22, No. 1, 2023, pp. 611-626.

[19] Z. Ding, R. Schober, H. V. Poor, "On the Impact of Phase Shifting Designs on IRS-NOMA", IEEE Wireless Communications Letters, Vol. 9, No. 10, 2020, pp. 1596-1600.

[20] R. Shafin, L. Liu, "Multi-Cell Multi-User Massive FD-MIMO: Downlink Precoding and Throughput Analysis", IEEE Transactions on Wireless Communications, Vol. 18, No. 1, 2019, pp. 487-502.

[21] Z. Chen, F. Sohrabi, W. Yu, "Multi-Cell Sparse Activity Detection for Massive Random Access: Massive MIMO Versus Cooperative MIMO", IEEE Transac-

tions on Wireless Communications, Vol. 18, No. 8, 2019, pp. 4060-4074.

[22] L. D. Nguyen, H. D. Tuan, T. Q. Duong, H. V. Poor, L. Hanzo, "Energy-Efficient Multi-Cell Massive MIMO Subject to Minimum User-Rate Constraints", IEEE Transactions on Communications, Vol. 69, No. 2, 2021, pp. 914-928.

[23] M. Amjad, L. Musavian, M. H. Rehmani, "Effective Capacity in Wireless Networks: A Comprehensive Survey", IEEE Communications Surveys & Tutorials, Vol. 21, No. 4, 2019, pp. 3007-3038.

[24] M. H. Babikir et al. "Optimization Efficiency of 5G MIMO Cooperative Spectrum Sharing NOMA Networks", Proceedings of the 9th International Conference on Mechatronics Engineering, Kuala Lumpur, Malaysia, 13-14 August 2024, pp. 183-187.

[25] S. Zhang, R. Zhang, "Intelligent Reflecting Surface Aided Multi-User Communication: Capacity Region and Deployment Strategy", IEEE Transactions on Communications, Vol. 69, No. 9, 2021, pp. 5790-5806.

[26] S. Ahmed, M. Z. Chowdhury, Y. M. Jang, "Energy-Efficient UAV-to-User Scheduling to Maximize Throughput in Wireless Networks", IEEE Access, Vol. 8, 2020, pp. 21215-21225.

[27] J. Li, X. Li, Y. Bi, J. Ma, "Energy-Efficient Joint Resource Allocation With Reconfigurable Intelligent Surfaces in Symbiotic Radio Networks", IEEE Transactions on Cognitive Communications and Networking, Vol. 8, No. 4, 2022, pp. 1816-1827.

[28] M. Hassan et al. "Modeling of NOMA-MIMO-Based Power Domain for 5G Network under Selective Rayleigh Fading Channels", Energies, Vol.15, 2022, p. 5668.

[29] P. Ramezani, Y. Zeng, A. Jamalipour, "Optimal Resource Allocation for Multiuser Internet of Things Network With Single Wireless-Powered Relay", IEEE Internet of Things Journal, Vol. 6, No. 2, 2019, pp. 3132-3142.

[30] K. Shafique, B. A. Khawaja, F. Sabir, S. Qazi, M. Mustaqim, "Internet of Things (IoT) for Next-Generation Smart Systems: A Review of Current Challenges, Future Trends and Prospects for Emerging 5G-IoT Scenarios", IEEE Access, Vol. 8, 2020, pp. 23022-23040.

[31] C. Pan et al. "Reconfigurable Intelligent Surfaces for 6G Systems: Principles, Applications, and Research Directions", IEEE Communications Magazine, Vol. 59, No. 6, 2021, pp. 14-20.

[32] Y. Liu et al. "Reconfigurable Intelligent Surfaces: Principles and Opportunities", IEEE Communications Surveys & Tutorials, Vol. 23, No. 3, 2021, pp. 1546-1577.

[33] R. Deshpande, M. V. Katwe, K. Singh, Z. Ding, "Resource Allocation Design for Spectral-Efficient URLLC Using RIS-Aided FD-NOMA System", IEEE Wireless Communications Letters, Vol. 12, No. 7, 2023, pp. 1209-1213.

[34] A. Khaleel, E. Basar, "A Novel NOMA Solution With RIS Partitioning", IEEE Journal of Selected Topics in Signal Processing, Vol. 16, No. 1, 2022, pp. 70-81.

[35] B. Zheng, C. You, R. Zhang, "Intelligent Reflecting Surface Assisted Multi-User OFDMA: Channel Estimation and Training Design", IEEE Transactions on Wireless Communications, Vol. 19, No. 12, 2020, pp. 8315-8329.

[36] Y. Cheng, K. H. Li, Y. Liu, K. C. The, H. Vincent Poor, "Downlink and Uplink Intelligent Reflecting Surface Aided Networks: NOMA and OMA", IEEE Transactions on Wireless Communications, Vol. 20, No. 6, 2021, pp. 3988-4000.

[37] T. Hou, Y. Liu, Z. Song, X. Sun, Y. Chen, "MIMO-NOMA Networks Relying on Reconfigurable Intelligent Surface: A Signal Cancellation-Based Design", IEEE Transactions on Communications, Vol. 68, No. 11, 2020, pp. 6932-6944.

[38] M. Mohammadi, H. Q. Ngo, M. Matthaiou, "Phase-Shift and Transmit Power Optimization for RIS-Aided Massive MIMO SWIPT IoT Networks", IEEE Transactions on Communications, 2024. (in press)

[39] B. Ur Rehman et al. “Uplink power control scheme for spectral efficiency maximization in NOMA systems”, Alexandria Engineering Journal, Vol. 64, 2023, pp. 667-677.

[40] N. Elmadina et al. "Downlink Power Allocation for CR-NOMA-Based Femtocell D2D Using Greedy Asynchronous Distributed Interference Avoidance Algorithm", Computers, Vol. 12, 2023, p. 158.

[41] T. Guo et al. "A Hybrid Indoor Positioning Algorithm for Cellular and Wi-Fi Networks", Arabian

Journal for Science and Engineering, Vol. 47, 2022, pp. 2909–2923.

[42] J. Fei, S. Li, "Adaptive Fractional High Order Sliding Mode Fuzzy Control of Active Power Filter", Proceedings of the Joint 10th International Conference on Soft Computing and Intelligent Systems and 19th International Symposium on Advanced Intelligent Systems, Toyama, Japan, 5-8 December 2018, pp. 576-580.

[43] H. Liu, G. Li, X. Li, Y. Liu, G. Huang, Z. Ding, "Effective Capacity Analysis of STAR-RIS-Assisted NOMA Networks", IEEE Wireless Communications Letters, Vol. 11, No. 9, 2022, pp. 1930-1934.

[44] E. Basar, I. Yildirim, "Reconfigurable Intelligent Surfaces for Future Wireless Networks: A Channel Modeling Perspective", IEEE Wireless Communications, Vol. 28, No. 3, 2021, pp. 108-114.

[45] S. Jia, X. Yuan, Y.-C. Liang, "Reconfigurable Intelligent Surfaces for Energy Efficiency in D2D Communication Network", IEEE Wireless Communications Letters, Vol. 10, No. 3, 2021, pp. 683-687.

[46] M. Hassan, M. Singh, K. Hamid, R. Saeed, M. Abdelhaq, R. Alsaqour, N. Odeh, "Enhancing NOMA's Spectrum Efficiency in a 5G Network through Cooperative Spectrum Sharing", Electronics, Vol. 12, 2023, p. 815.

[47] M. Hassan et al. "Design of Power Location Coefficient System for 6G Downlink Cooperative NOMA Network", Energies, Vol. 15, 2022, p. 6996.

[48] A. BANON, Fayez W. Zaki, M. M. Ashour, "The effect of quantized ETF, grouping, and power allocation on non-orthogonal multiple accesses for wireless communication networks", International Journal of Electrical and Computer Engineering Systems, Vol. 13, No. 8, 2022, pp. 681-693.

Design and Implementation of a Novel 5G Hairpin Bandpass Filter with Defected Ground Structure

Original Scientific Paper

Shereen Abdalkadum Shandal*
Department of Computer Techniques Engineering
Middle Technical University, Baghdad, Iraq
Shereen@mtu.edu.iq

*Corresponding author

Abstract *– In this paper, a three-pole hairpin resonator is designed, simulated, and fabricated on the top layer of the FR4 substrate. Recent trends in miniature size and improved filter performance, particularly in terms of scattering parameters and wider bandwidth, have increased demand for such filters. This filter uses two different Defect Ground Structure (DGS) techniques utilizing the top and ground layers. The first Defect Ground Structure (DGS) technique incorporates two dumbbells and rectangular slots beneath two feed lines, resulting in a unique and modified bandpass filter design. In the second DGS, a series of grooves embedded at three hairpin resonators provide a more compact size and enhanced scattering parameters with wider bandwidth, which is considered an improvement of this design over the existing works. The simulation results use High Frequency Structure Simulator (HFSS) software. Parametric optimization has been conducted; the optimized values of three significant parameters are 4mm length of tap Lt, 0,4mm space between resonators S, and (3×9)mm2 area of rectangular slot (DGS2). The presented filter resonates at 2.5 GHz center frequency with a -3dB fractional bandwidth of 22.4%. The acquired values of insertion loss (S21) and return loss (S11) at the passband are -1.6dB and -54.19dB, respectively, with a flat group delay. The design validity has been verified using Computer Simulation Technology (CST) simulation software and a fabricated prototype. The fabrication results match the simulations excellently, making the suggested filter suitable for various fifth-generation (5G) applications.*

Keywords: *DGS, insertion loss, CST and HFSS, group delay, Vector Network Analyzer (VNA)*

Received: September 8, 2024; Received in revised form: October 20, 2024; Accepted: November 25, 2024

1. INTRODUCTION

In recent days, the applications of wireless communication devices have witnessed increasing requests, which has led to further growth in the microwave device field. Radio Frequency (RF) and microwave filters are excessively utilized in radar systems, satellite communication, recent warfare devices, TV broadcasts, radio broadcasts, and mobile devices. Each wireless communication device operates at various frequencies and allocates various bands for each device [1-6]. Each receiver uses a filter, an essential component that functions as an electronic circuit, allowing a specific range of frequencies to pass while rejecting or attenuating unwanted frequencies. Moreover, filters can be classified according to their characteristics into four different types: lowpass filter (LPF), highpass filter (HPF), bandpass filter (BPF), and bandstop filter (BSF) [7-9]. Therefore, numerous methods exist for designing filters, each with pros and cons. These methods include parallel coupled lines, edge coupled lines, inter-digital lines, comb lines, hairpins, and more.

The hairpin method has the advantage of being small compared to others; it is widely used in filter design [10]. The planar hairpin filter meets the requirements of a compact, high-quality, and low-cost RF/microwave filter. A hairpin BPF passes frequency within a specific extent and rejects or attenuates frequency outside that extent [11]. Also, in hairpin BPF, the hairpin lines comprise folded parallel coupled half-wavelength resonators, which makes the area where parallel lines take up smaller. One of the most common ways to improve filter parameters like insertion loss, return loss, and harmonic suppression is to use a defect ground structure (DGS) [12, 13]. Accordingly, the microstrip technique is used in the hairpin BPF design instead of lumped elements due to its advantages, such as its small size, lightweight, affordability, ease of manufacture, and low loss [14, 15]. Consequently, microwave and RF circuits use a microstrip line as a means of power transmission. The microstrip line consists of three layers: the top and bottom layers, known as conductor strips and ground, are made from conducting material such as copper, while the middle layer is known as the dielectric substrate [16-20]. Numerous prior researchers have employed the DGS technique to design a hairpin bandpass filter, as seen in [21], where they designed a microstrip hairpin line BPF with two square-shaped DGS. Although this filter has a compact size and good insertion loss (S21), it suffers from low S11 values in the passband region and narrow bandwidth. In [22], a simple filter design for S-band radar incorporates the DGS as a square groove to minimize filter size and suppress harmonics; however, it still struggles with a large occupied area and low return loss (S11). In [23], a planar third-order hairpin BPF employs DGS slot resonators to achieve a high S11 value of -40dB. However, it suffers from a high S21 value and large size. An open stub with DGS employed in [24] is utilized to design a filter for X-band weather radar applications with 120 MHz bandwidth; the S21 value at the passband is -1,57dB. Two different shapes of DGS are utilized to design the hairpin bandpass filter presented in [25], which operates at 2.4GHz and has a moderate return loss of -26dB with sharp roll-off and wide bandwidth in the passband region. In [26], an open loop microstrip structure is used as a DGS to design a three-pole hairpin bandpass filter for a VSAT (Very Small Aperture Terminal) system, with an observed return loss S11 of -13 dB with a triple band at 10.2GHz, 12.2GHz, and 14.8GHz. In [27], a new miniature two-layer bandpass filter is designed that operates at 2.5GHz with a fractional bandwidth of 4.75%; the S21 and S11 values are -1.65dB and -45dB, respectively. From our review of previous studies in the filter design, we found that the DGS technique was used in various geometric shapes and was placed, in most studies, on the bottom layer of the substrate material. However, upon reviewing these studies, it was found that the physical size of most filters was large, the scattering parameter values needed to be higher, and a lack of parametric optimization and traditional hairpin structures were used. These gaps can be addressed by reducing the filter size without affecting its performance, improving the filter's response in terms of scattering parameter values and bandwidth, switching the dielectric material to a low-loss material, conducting a parametric study of several parameters that influence the filter's performance, and modifying the shape and position of the DGS.

The research problem involves improving the scattering parameter values (S11 and S21), obtaining a wider bandwidth in the passband region while maintaining a small size and low cost for the latest compact devices. Filter designers encounter challenges when reducing size, including increased losses, narrower bandwidth, fabrication tolerance sensitivity, and cost. A tradeoff between size reduction with low cost and high performance should be reached. Therefore, in this research, I will study the design of a small-sized filter with unique hairpin-shaped resonators, improving the scattering parameter values, including a parametric optimization of three important parameters that significantly affect the filter's performance (tap length, space between resonators, and rectangular slot area DGS2). This study improves a novel DGS added at the substrate's top and bottom layers, which differs from previous studies and results in enhanced filter performance regarding in-band and out-of-band response, better-scattering parameters, and smaller overall size. The suggested filter has been fabricated and tested to verify the results, and the measured results agree well with the simulated ones. FR4-epoxy is the substrate used, with 4.4 permittivity, 0.02 loss tangent, and 1.6 mm thickness. Sections 2 and 3 arrange the remaining parts. Section 2 provides a brief overview of the design steps of the proposed filter, while Section 3 conducts a parametric study. Then, the simulation results and discussion are presented in Section 4. Section 5 includes the experimental results, followed by Section 6, which compares the proposed filter with other references.

2. PROPOSED FILTER DESIGN

This section will demonstrate constructing a bandpass filter using a three-hairpin resonator with the DGS technique at the top and bottom layers. Fig. 1 displays the equivalent circuit schematic for three resonators at the top layer, which consists of three inductors and capacitors; C12 and C23 represent coupling capacitors between adjacent resonators.

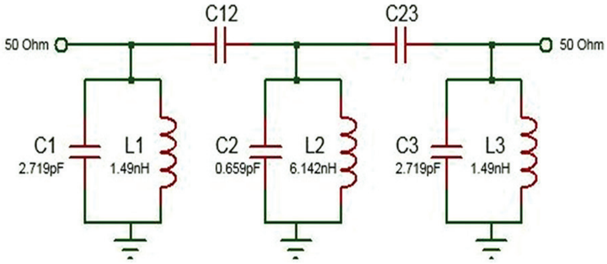


Fig. 1. Lumped elements of the hairpin resonators

The design of a three-pole hairpin bandpass filter aims to strike a balance between performance, size, and selectivity. The shape of the hairpin was chosen mainly for its compactness, easy coupling, and good frequency response. Three hairpin resonators are often chosen instead of five to balance size, complexity, cost, and performance. It is an appropriate choice when moderate bandwidth and selectivity are satisfactory and reducing loss, cost, and size is crucial. A hairpin resonator comprises a U-shaped structure, a substrate, a ground plane, and coupling sections. These parts work together to provide excellent filter performance regarding selectivity, size, and bandwidth. It must first determine the filter and substrate specifications, as shown in Table 1. The FR4 substrate is chosen because it is widely available and cost-effective. The Chebyshev table will yield the following low pass prototype values for the proposed filter (order three, ripple 0.1, and Chebyshev filter response): $g0 = g4 = 1, g1 = g3 = 1.03$, and $g2 = 1.14$. The equations (1-10) were used to design the proposed bandpass filter [8]. The following steps must be followed:

- Step (1), calculate the external quality factor at input and output ports denoted, respectively, by using the following equations (1-2)

$$Q_{e1} = \frac{g_0 g_1}{FBW} \;, Q_{en} = \frac{g_n g_{n+1}}{FBW} \tag{1}$$

$$FBW = \frac{f_h - f_l}{f_o} \tag{2}$$

Where f_h, f_l represents the higher and lower cutoff frequencies, f_o is the center frequency, and FBW is the fractional bandwidth of the filter.

- Step (2), the width of each resonator is calculated by the following equation (3) and (4)

$$B = \frac{60\,\pi^2}{z_c\,\sqrt{\in r}} \tag{3}$$

$$\begin{gathered} w/h = \frac{2}{\pi}\Big\{\big((B-1)\big) - \ln(2B-1) + \\ + \frac{\in r - 1}{2 \in r}\Big[\ln(B-1) + 0.39 - \frac{0.61}{\in r}\Big]\Big\} \end{gathered} \tag{4}$$

Where h is the thickness of the FR4 substrate 1.6mm and z_c represents a characteristic impedance equal to 50 ohm, the calculated width of each resonator is 2mm. Furthermore, the resonator length is equal to 16.5mm which is calculated by equations (5-7)

$$k_0 = \frac{2\pi\, f_0}{c} \tag{5}$$

$$\in_e = \frac{\in r + 1}{2} + \frac{\in r - 1}{2}\sqrt{\frac{1}{1 + 12(h/w)}} \tag{6}$$

$$L = \frac{\frac{\pi}{180^0}}{\sqrt{\in_e} k_0} \tag{7}$$

Where c represents the speed of light in free space 3×10^8 m/s and $\in e$ is an effective dielectric constant.

- Step (3), calculate the mutual coupling between resonators using equation (8)

$$M_{i,i+1} = \frac{FBW}{\sqrt{g_i g_{i+1}}} \tag{8}$$

The calculated values of two mutual couplings are $M_{12} = M_{23} = 2.067$, the spacing between two adjacent resonators is assumed to be an initial value, and then the filter is designed by using the software simulator HFSS to observe the insertion loss S21 curve and take the two peaks of that curve to calculate coupling coefficients k using equation (9)

$$k = \frac{{f_h}^2 - {f_l}^2}{{f_h}^2 + {f_l}^2} \tag{9}$$

These two peaks on the S21 curve are represented by f_h and f_l, when the value of k is close to the value of M, then the assumed separation distance between adjacent resonators will be considered, equal to 0.4mm in our proposed design.

- Step (4), calculate the tapping point t that represents the position of the feed line at both ports, using equation(10)

$$t = \frac{2l}{\pi}\sin^{-1}\left(\sqrt{\frac{\pi}{2}\left(\frac{z_0}{Q_e z_r}\right)}\right) \tag{10}$$

Where z_0 represents the impedance needed for terminals, z_r impedance for hairpin line, and external quality factor Q_e, the calculated tapping point value is 6mm. Fig. 2 depicts the top layer of the suggested filter, while Table 2 displays its dimensions.

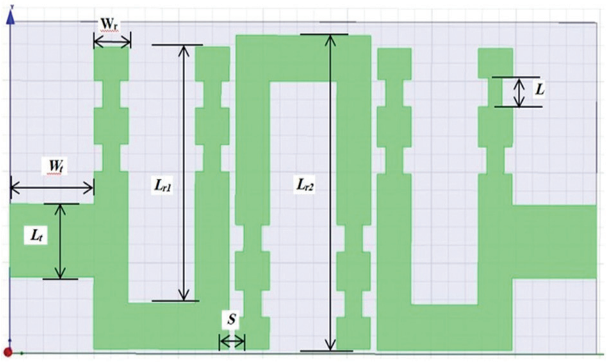


Fig. 2. The top layer of the proposed filter

The -3dB fractional bandwidth measures the filter bandwidth relative to its center frequency (2.5GHz). Where filter bandwidth is 560MHz at -3dB points at the insertion loss (S21) curve. The dielectric constant, thickness, loss tangent, and thermal stability of the FR4 substrate all play a big part in the design of the hairpin resonator. These properties affect how well the filter works. A high dielectric constant leads to more size compactness but reduces the quality factor.

The first DGS is added at the top layer as consecutive grooves with dimensions (0.5×1.5) mm^2 to enhance filter response regarding selectivity with size reduction. On the other hand, the second DGS is employed at the bottom layer of the suggested filter to improve the scattering parameters, which consist of two geometric shapes, as shown in Fig. 3.

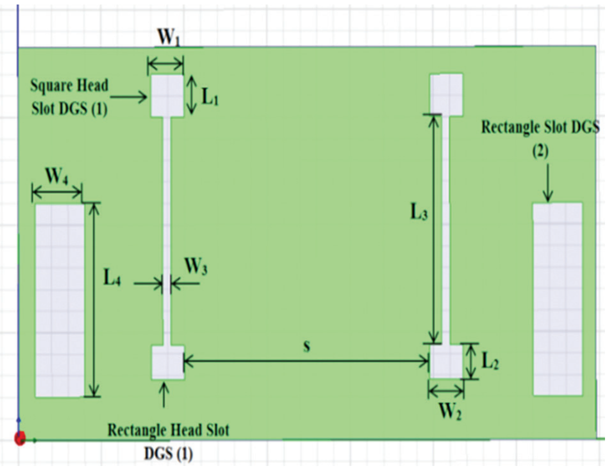


Fig. 3. The bottom layer of the proposed filter

The first one (DGS1) consists of two dumbbells separated by a distance of 14.8 mm. In contrast, the second one (DGS2) consists of a pair of rectangular slots located below the feed lines with dimensions (3×9) mm2, as shown in Fig. 4. These two different DGS techniques differ in their application to the filter design by etching one DGS in the top layer while the other is etched at the bottom layer. Both of them have a simple structure that can be easily implemented. Their impact on the filter includes enhanced scattering parameters, sharper roll-off, and improved passband performance without an increase in the filter size. Table 3 also illustrates all the dimensions of the bottom layer. Fig. 4 displays the ultimate 3D view of the suggested filter.

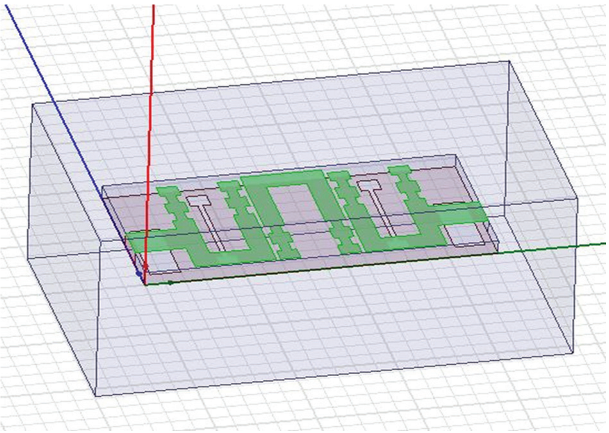


Fig. 4. 3D view of the proposed filter

Table 1. Filter and substrate specifications

Filter parameter	Value
Lower cutoff frequency f_l	2.2GHz
Upper cutoff frequency f_h	2.8GHz
Center frequency f_0	2.5GHz
Order	3
Filter response	Chebyshev
Return loss S11	≤-10
Insertion loss S21	>-2
Substrate type	FR4$_{epoxy}$
Substrate thickness h	1.6mm
Permitivity	4.4
Loss tangent	0.02

Table 2. Top layer dimensions

Filter dimensions (top layer)	Value (mm)
Substrate length	18.25
Substrate width	34.8
Resonator 1 length $Lr1$	16.59
Resonator 2 length $Lr2$	17.25
Resonator 3 length $Lr3$	16.59
Resonator width Wr	2
Tap line length	4
Tap line width	5
Space among three resonators S	0.4

Table 3. Bottom layer dimensions

Filter dimensions (bottom layer)	Value (mm)
Length of square head slot $L1$	2
Width of square head slot $W1$	2
Length of rectangle head slot $L2$	1.6
Width of rectangle head slot $W2$	2
Length of path $L3$	10.6
Width of path $W3$	0.5
Length of rectangle shape slot $L4$	9
Width of rectangle shape slot $W4$	3
Space between two rectangular shape heads (s)	14.8

3. PARAMETRIC STUDY

The suggested filter has been simulated using the finite element method based on HFSS software. This software will provide accurate simulation results, parametric optimization, extraction of the scattering parameters (S11 and S21), and quality factor calculation. The values of insertion loss (S21) and return loss (S11) affect how well the filter works. When insertion loss (S21) is close to zero in the passband region, the signal is attenuated as little as possible. Low return loss (S11 ≤ -10) indicates good impedance matching, and the reflected signal is reduced, which increases the filter's efficiency. The designed filter will experience a series of variations in the tap length, spacing between resonators, and various areas of DGS2. The optimum goal is to get the scattering parameter values (S11 and S21) in an adequate range. The tap length is altered, as shown in Fig. 5, which depicts that the minor increase in tap length changes S11 and S21 values and fractional bandwidth. As shown in the figure, when the tap length increases, the scattering parameter values are enhanced due to sufficient coupling and excellent impedance matching between the resonator and external circuitry. The best return loss (S11) is obtained when Lt is 4 mm, equal to -54.19 dB, as illustrated in Table 4. Optimizing the space between resonators enhances the filter response. The calculated space value S of 0.2mm concerning coupling coefficient k is altered, as indicated in Fig. 6. The figure depicts that the center frequency will not be affected. In contrast, the scattering parameters and fractional bandwidth are noticeably affected, as illustrated in Table 5. The interpretation of this behaviour is that when the resonators are placed close, the coupling increases, increasing the fractional bandwidth and reducing losses due to strong signal transfers between resonators. Finally, the area of the second shape of DGS denoted as DGS2, can be included in the parametric study. The area of that slot is changed from (2×8) mm^2 to (4×10) mm^2 to enhance filter response, as indicated in Fig. 7. The figure indicates a slight effect on the insertion loss (S21), fractional bandwidth, and center frequency. However, Table 6 illustrates a significant impact on the return loss value (S11) when the rectangular slot area equals (3×9) mm2 due to increasing the capacitive and inductive effects when increasing the area of DGS2 and reducing the center frequency.

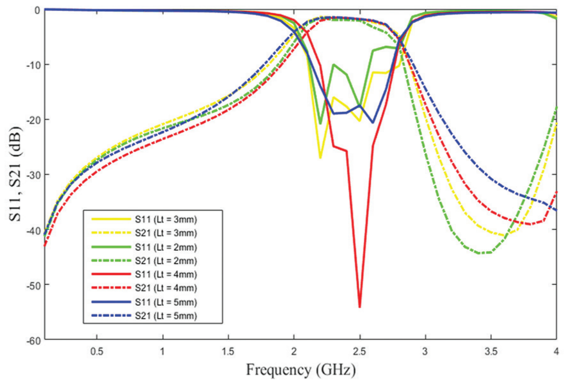


Fig. 5. Simulated S-parameter response when changing the tap length

Table 4. Different tap lengths L_t for the suggested filter

Length of tap *Lt* (mm)	S21 (dB)	S11 (dB)	Center frequency (GHz)	FBW
2	-1.4, -2	-21, -18	2.2, 2.5	6.8%, 12%
3	-1.4, -1.7	-27, -20	2.35	27%
4 (proposed)	-1.6	-54.19	2.5	22%
5	-1.6	-19, -21	2.5	22.8%

Table 5. The proposed filter comparison based on the spacing between resonators

Space between resonators (mm)	S21 (dB)	S11 (dB)	Center frequency (GHz)	FBW
0.2	-1.4, -3	-30, -23, -15	2.23, 2.8	26%, 4.2%
0.4 (proposed)	-1.6	-54.19	2.5	22.4%
0.6	-1.8	-20, -21	2.5	19.6%
0.8	-2	-23.5	2.5	14%

Table 6. The proposed filter comparison based on the variation of rectangular slot area

Area of rectangular slot (DGS2) (mm)2	S21 (dB)	S11 (dB)	Center frequency (GHz)	FBW
2×8	-1.6	-19, -16	2.5	25.2%
3×9 (proposed)	-1.6	-54.19	2.5	22.4%
4×10	-1.7	-17, -22	2.38	27.3%

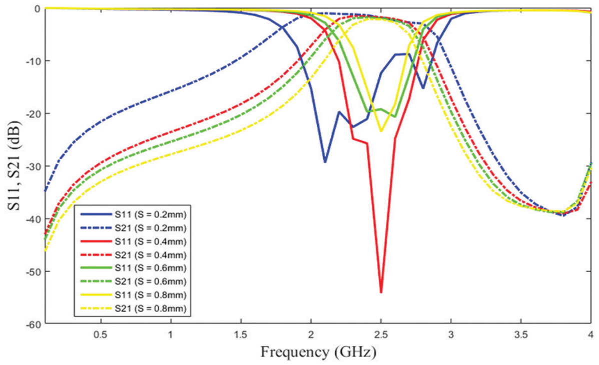


Fig. 6. Simulated S-parameter response when changing space between the resonator

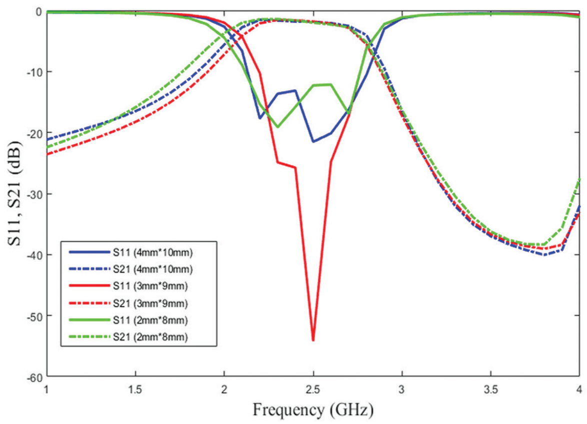


Fig. 7. Simulated S-parameter response when changing the rectangular slot area (DGS2)

4. SIMULATION RESULTS AND DISCUSSION

The proposed filter was initially designed by HFSS software; The results of the simulation for different aspects of the filter design have been extracted, as detailed below:

4.1. VARIOUS MODEL CONSTRUCTIONS

Three models have been designed to demonstrate the effect of DGS on the performance of the filter response. Model 1 incorporates both proposed DGS shapes at both layers. Model 2 encompasses a full ground plane with only defects at the top layer, whereas Model 3 includes top-layer defects and only the first shape of DGS (DGS1) at the bottom layer. Table 7 presents a comparison of the three proposed models. The simulation results for these three models are depicted in Fig. 8. It is clear that the first model has the best scattering parameters (S21, S11) and fractional bandwidth compared to the others, which are -1.6 dB, -54.19 dB, and 22.4%, respectively. Where the lower S11 values are below -10dB, and the closer the S21 values are to zero, the better, as this improves the filter's efficiency. The resonator performance is significantly enhanced by using the second DGS, which leads to improved impedance matching observed when return loss S11 becomes less than -50dB. Furthermore, the second and third models have moderate values for S21 and S11. The center frequency for all three proposed models does not change significantly because the three hairpin resonators at the top layer of the suggested filter have not changed; the only change occurs at the ground layer of the suggested filter.

4.2. CST SIMULATION RESULTS

The CST simulator verifies the results previously obtained from the HFSS. Fig. 9 compares scattering parameters between two software simulators, providing a means to assess the accuracy of the simulated results. The CST simulated results indicate excellent agreement, especially near lower and higher cut-off frequencies, where the lower cut-off frequency is 2.15 GHz and the higher cut-off frequency is 2.7 GHz, with a center frequency of 2.47 GHz. These variations between the two software are due to several factors, such as the numerical method, boundary conditions and excitations, meshing, and solver technique es.

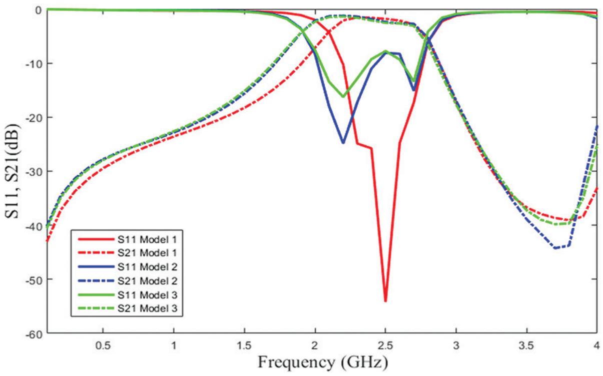


Fig. 8. Simulated S11 and S21 responses of Three different models to demonstrate the effect of DGS

4.3. CURRENT DISTRIBUTION AND GROUP DELAY

The surface current distribution of the proposed filter at the center frequency of 2.5 GHz is indicated in Fig. 10. The filter's surface demonstrates a current flow that appears at the three hairpin resonators in the bandpass region. Another important parameter for the filter design is a group delay, ensuring the signal passes through the filter without distortion and preserving its integrity. This design manages a group delay by optimizing DGS and conducting simulation. The more stable group delay at the passband region, the better. A flat group delay is shown in Fig. 11, which indicates a minimal signal distortion at the passband range.

Table 7. Comparison among three proposed models of the filter design

Model	S21 (dB)	S11 (dB)	Center frequency (GHz)	FBW
model 1 (proposed)	-1.6	-54.19	2.5	22.4%
model 2	-1.3, -3	-25, -15	2.25, 2.7	22.2%, 5.5%
model 3	-1.4, -3	-16, -12	2.25, 2.7	22.2%, 7.4%

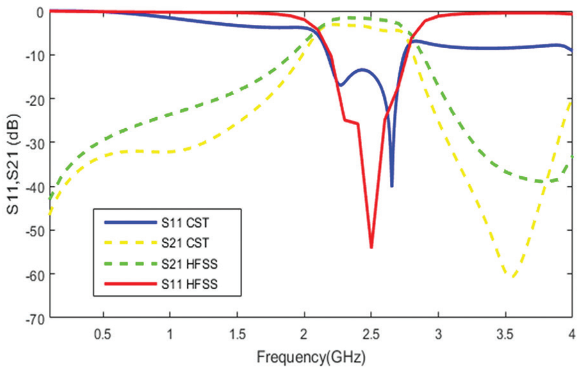


Fig. 9. Simulated S11 and S21 responses of the CST and HFSS simulators for the proposed filter

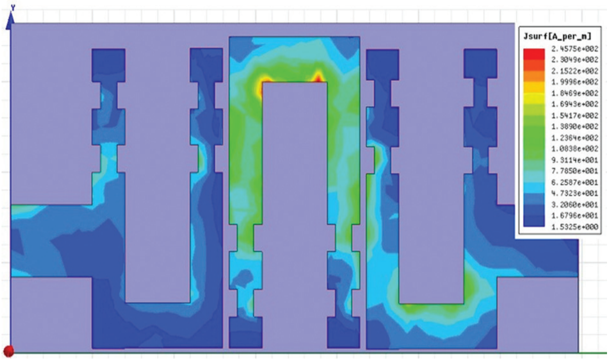


Fig. 10. Surface current distribution of the proposed filter at 2.5GHz

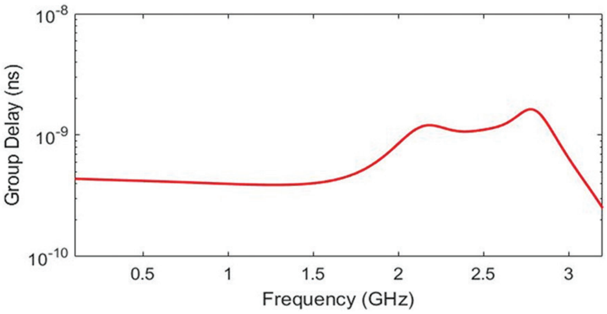
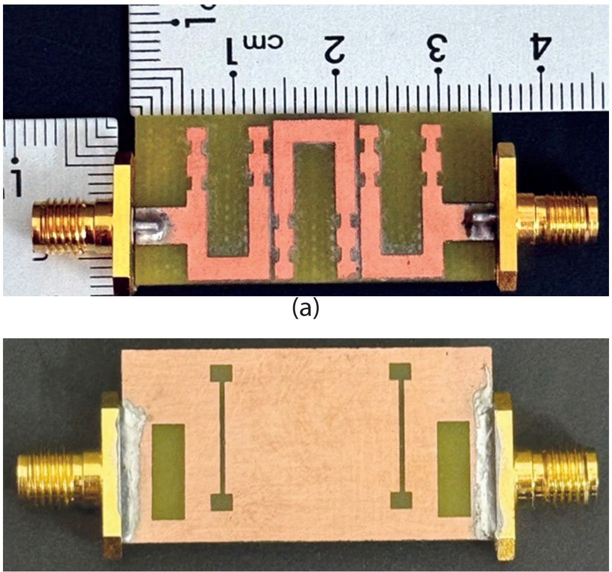


Fig. 11. Group Delay of the proposed filter

5. EXPERIMENTAL RESULTS

The proposed hairpin bandpass filter has been fabricated and tested to verify previously obtained simulation results from CST and HFSS software simulators. The main purpose of verifying the design with CST and the fabricated prototype is to ensure that the theoretical performance of the filter aligns with real-world behaviours. This dual verification process is essential for identifying any discrepancies, design optimization, and impacts of the manufacturing process and ensuring that the final product meets design specifications and industry standards. Fig. 12 illustrates a photograph of the fabricated prototype. Two SMA connectors were utilized to measure the fabricated results. The Vector Network Analyzer (VNA) MS4642A has been used to obtain the measurements, as shown in Fig. 13. A comparison of the simulated and measured results is depicted in Fig. 14. From the figure, it is obvious that there is an outstanding agreement between the simulation and measurement results. Still, slight variations occur due to fabrication tolerance, SMA connector mismatch, substrate material properties, and environmental conditions. Temperature and electromagnetic interference are the two main environmental factors influencing the measurement results. The simulation results achieved a bandwidth of (2.2-2.8) GHz with a center frequency. In measurement results, the bandwidth of (2.4-2.7) GHz with a center frequency of 2.6GHz . The simulated and measured return losses (S11) are -54dB and -30dB for the bandpass region, respectively, while the simulated and measured insertion losses (S21) are -1,17dB and -1,20dB. The challenges encountered during the fabrication process of this filter are pattern precision and variability of substrate material; these were addressed by using engravings with dimensions that are compatible with the cutting machine and simple geometric structures of DGS to avoid design complexity. The proposed filter is appropriate for many modern wireless communication systems that require filters with high selectivity and sharp cut-offs to separate the intended signal band from noise and interference. 5G applications such as enhanced mobile broadband, Internet of Things (IoT), and autonomous vehicles can utilize the suggested filter.



(a)

(b)

Fig. 12. Photograph of the fabricated filter. (**a**) Top view, (**b**) Bottom view

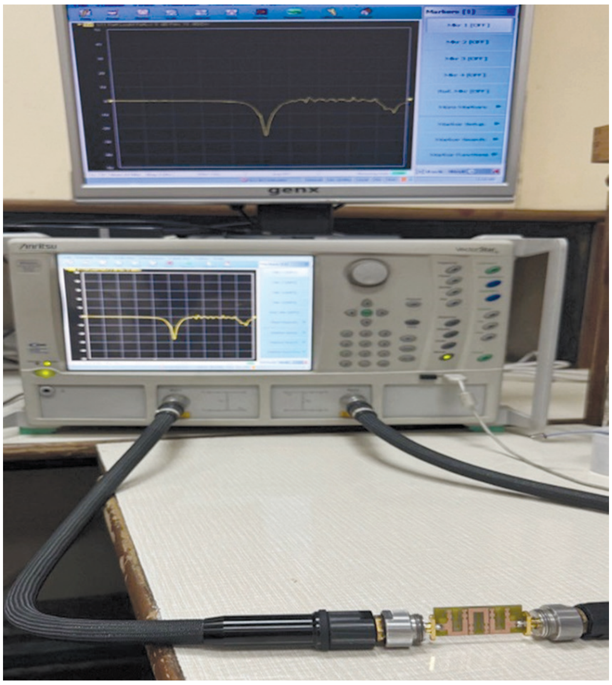


Fig. 13. VNA for measuring the fabricated filter

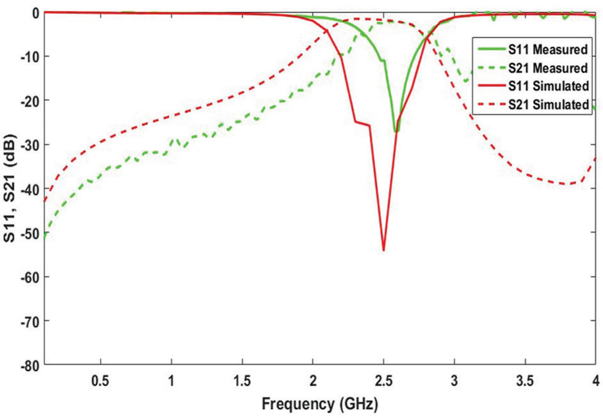


Fig. 14. Comparison between simulated and measured results

6. COMPARISON WITH OTHER REFERENCES

The proposed filter is compared with the other references [21-27] at various ranges of frequency, focusing on insertion loss (S21), return loss (S11), bandwidth, and center frequency highlighted in Table 8. As illustrated in the table, although the suggested filter has a simple structure, it provides the optimum S21, S11, and wider fractional bandwidth equal to -1.6 dB, -54.19 dB, and 22.4%, respectively.

Table 8. Comparison between the proposed filter and other references

Ref No.	Year of pub.	S21 (dB)	S11 (dB)	Bandwidth (BW) (MHz)	Center frequency (GHz)	Filter size (mm)^2
[9]	2017	-2.1	-35	389	2.45	40×38
[17]	2018	-1.7	-34	620	2.7	280×140
[19]	2018	-0.37	-34.03	107.4	2.45	20.2×13.3
[21]	2017	-0.2946	-46.64	460	2.22	18.2×34.8
[22]	2018	-0.76	-29	200	3	53.7×17.6
[23]	2019	-3	-41	250	2.35	70×45
[24]	2018	-1.57	-29.9	100	9.5	
[25]	2018	-0.1	-24	600	2.2	49×25
[26]	2021	-1.9, -0.9, -1.2	-13.25, -12.8, -14.72	400, 700, 300	10.28, 12, 14.62	
[27]	2022	-1	-31	130	2.55	21.8×21.6
This Work	2024	-1.6	-54.19	560	2.5	18.25×34.8

7. CONCLUSION

In this paper, a novel hairpin bandpass filter with defective ground structures is analyzed, simulated, and fabricated. 5G applications such as enhanced mobile broadband, Internet of Things (IoT), and autonomous vehicles can utilize the suggested filter. It can also be utilized in wi-fi networks and radar communications. This study utilizes two distinct defect ground structures at the top and bottom layers to enhance the scattering parameters, widen the bandwidth, and maintain compact size. The suggested filter was first created using HFSS software, and a parametric study was highlighted on the best filter response by changing the space between resonators, the length of the tap, and the area of the rectangular slot (DGS2). The optimized values of three significant parameters are 4mm length of tap Lt, 0,4mm space between resonators S, and (3×9)mm^2 area of rectangular slot (DGS2). The CST simulation results and the fabricated prototype confirm the previous HFSS results. The fabricated prototype's measurement results differed slightly from the simulated ones due to fabrication tolerance and SMA connector mismatch. This filter provides a wider fractional bandwidth of 22.4% in the passband region, excellent return loss (S11) of -54 dB, and insertion loss (S21) of -1.6dB with a short group delay. Future work can apply various ways to the proposed filter to achieve further enhancements such as more miniaturization, conducting parametric optimization with another parameter, and changing the type of substrate material.

8. REFERENCES

[1] E. G. Ouf, E. A. Abdallah, A. S. Mohra, H. El-Hennawy, “Electronically switchable ultra-wide band/dual-band bandpass filter using defected ground structures”, Progress In Electromagnetics Research C, Vol. 91, 2019, pp. 83-96.

[2] M.-M. Ma, Z.-X. Tang, X. Cao, T. Qian, “Tri-band cross-coupling bandpass filter with rectangular defected ground structure array”, Journal of Electromagnetic Waves and Applications, Vol. 32, No. 11, 2018, pp. 1409-1415.

[3] J. B. Jadhav, P. J. Deore, “A compact planar ultra-wideband bandpass filter with multiple resonant and defected ground structure”, AEU - International Journal of Electronics and Communications, Vol. 81, 2017, pp. 31-36.

[4] A. Djaiz, “A new compact microstrip two-layer bandpass filter using aperture-coupled SIR-hairpin resonators with transmission zeros”, IEEE Transactions on Microwave Theory and Techniques, Vol. 54, No. 5, 2006, pp. 1929-1936.

[5] L.-H. Hsieh, K. Chang, “Compact, low insertion-loss, sharp-rejection, and wide-band microstrip bandpass filters”, IEEE Transactions on Microwave Theory and Techniques, Vol. 51, No. 4, 2003, pp. 1241-1246.

[6] R. P. Verma, B. Sahu, A. Gupta, “A miniaturized UWB bandpass filter with tunable cut-off frequencies employing rectangular open-loop defected ground structure”, International Journal of Microwave and Wireless Technologies, Vol. 15, No. 10, 2023, pp. 1689-1697.

[7] E. Sghir, A. Errkik, J. Zbitou, O. Oulhaj, A. Lakhssassi, M. Latrach, “Miniaturized ultra-wideband coplanarwaveguide lowpass filter with extended stop band”, Indonesian Journal of Electrical Engineering and Computer Science, Vol. 19, No. 3, 2020, p. 1415.

[8] D. M. Pozar, "Microwave engineering", John Wiley & Sons, 2011.

[9] S. Mora, Y. Alonso, N. Vargas, J. Vera, J. Avendano, “Design of a bandpass filter using microstrip Hairpin resonators”, Proceedings of the Chilean Conference on Electrical, Electronics Engineering, Information and Communication Technologies, Pucon, Chile, 18-20 October 2017, pp. 1-5.

[10] A. Ramdedovic, “Tight-coupled microstrip hairpin bandpass filter”, Journal of Engineering Research, Vol. 11, No. 1A, 2023.

[11] A. B. Abdel-Rahman, A. K. Verma, A. Boutejdar, A. S. Omar, “Control of bandstop response of Hi-Lo microstrip low-pass filter using slot in ground plane”, IEEE Transactions on Microwave Theory and Techniques, Vol. 52, No. 3, 2004, pp. 1008-1013.

[12] N. Abd Wahab, W. N. W. Muhamad, M. M. A. M. Hamzah, S. S. Sarnin, N. F. Naim, “Design a microstrip hairpin band-pass filter for 5GHZ unlicensed WiMAX”, Proceedings of the International Conference on Networking and Information Technology, Manila, Philippines, 11-12 June 2010, pp. 183-186.

[13] A. Boutejdar, A. Elsherbini, A. Balalem, J. Machac, A. Omar, “Design of new DGS hairpin microstrip bandpass filter using coupling matrix method”, Progress in Electromagnetic Research Symposium, 2007, pp. 261-265.

[14] S.-W. Ting, K.-W. Tam, R. P. Martins, “Miniaturized microstrip lowpass filter with wide stopband using double equilateral U-shaped defected ground structure”, IEEE Microwave and Wireless Components Letters, Vol. 16, No. 5, 2006, pp. 240-242.

[15] S. Sun, L. Zhu, “Wideband microstrip ring resonator bandpass filters under multiple resonances”, IEEE Transactions on Microwave Theory and Techniques, Vol. 55, No. 10, 2007, pp. 2176-2182.

[16] K. Srisathit, J. Tangjit, W. Surakampontorn, “Miniaturized microwave bandpass filter based on modified hairpin topology”, Proceedings of the IEEE International Conference of Electron Devices and Solid-State Circuits, Hong Kong, China, 15-17 December 2010, pp. 4-7.

[17] K. Kavitha, M. Jayakumar, “Design and performance analysis of hairpin bandpass filter for satellite applications”, Procedia Computer Science, Vol. 143, 2018, pp. 886-891.

[18] S. A. Shandal, Y. S. Mezaal, M. A. Kadim, M. F. Mosleh, “New compact wideband microstrip antenna for wireless applications”, Advanced Electromagnetics, Vol. 7, No. 4, 2018.

[19] S. M. K. Azam, M. I. Ibrahimy, S. M. A. Motakabber, A. K. M. Z. Hossain, “A compact bandpass filter using microstrip hairpin resonator for WLAN applications”, Proceedings of the 7th International Conference on Computer and Communication Engineering, Kuala Lumpur, Malaysia, 19-20 September 2018, pp. 313-316.

[20] M. A. Kadhim, M. F. Mosleh, S. A. Shandal, “Wideband Square Sierpinski Fractal Microstrip Patch Antenna for Various Wireless Applications”, IOP Conference Series: Materials Science and Engineering, Vol. 518, No. 4, 2019, p. 42001.

[21] V. S. Kershaw, S. S. Bhadauria, G. S. Tomar, “Design of Microstrip Hairpin-Line Bandpass Filter with Square Shape Defected Ground Structure”, Asia-Pacific Journal of Electronic and Electrical Engineering, Vol. 1, No. 1, 2017, pp. 21-30.

[22] N. Ismail, T. S. Gunawan, T. Praludi, E. A. Hamidi, “Design of microstrip hairpin bandpass filter for 2.9 GHz-3.1 GHz s-band radar with defected ground structure”, Malaysian Journal of Fundamental and Applied Sciences, Vol. 14, No. 4, 2018, pp. 448-455.

[23] M. Nedelchev, A. Kolev, “Synthesis of Planar Filters Using Defected Ground Structure Miniaturized Hairpin Resonators”, Engineering, Technology & Applied Science Research, Vol. 9, No. 1, 2019, pp. 3734-3738.

[24] T. Hariyadi, S. Mulyasari, Mukhidin, “Design and Simulation of Microstrip Hairpin Bandpass Filter with Open Stub and Defected Ground Structure (DGS) at X-Band Frequency”, IOP Conference Series: Materials Science and Engineering, Vol. 306, No. 1, 2018.

[25] H. Sajjad, A. Altaf, S. Khan, L. Jan, “A compact hairpin filter with stepped hairpin defected ground structure”, Proceedings of the IEEE 21st International Multi-Topic Conference, Karachi, Pakistan, 1-2 November 2018, pp. 1-5.

[26] N. Ambati, G. Immadi, M. V. Narayana, K. R. Bareddy, M. S. Prapurna, J. Yanapu, “Parametric Analysis of the Defected Ground Structure-Based Hairpin Band Pass Filter for VSAT System on Chip Applications”, Engineering, Technology & Applied Science Research, Vol. 11, No. 6, 2021, pp. 7892-7896.

[27] N. Chami, D. Saigaa, A. Djaiz, “A New Miniature Micro-Strip Two-Layer Band-Pass Filter Using Aperture-Coupled Hairpin Resonators”, Engineering, Technology & Applied Science Research, Vol. 12, No. 4, 2022, pp. 9038-9041.

Performance Enhancement in OFDM System Using Preamble-Based Time Domain SNR Estimation

Original Scientific Paper

Shahid Manzoor
UCSI University,
Faculty of Engineering, Technology and Built Environment, Department of Electrical and Electronics Engineering,
Cheras, 56000, Kuala Lumpur, Malaysia.
shahid@ucsiuniversity.edu.my

Noor Shamsiah Othman*
Universiti Tenaga Nasional,
Department of Electrical and Electronics Engineering, College of Engineering & Institute of Power Engineering,
Jalan IKRAM-UNITEN, Kajang, 43000, Selangor, Malaysia.
shamsiah@uniten.edu.my

*Corresponding author

Abstract *– This work proposes a time domain signal-to-noise ratio (SNR) estimator for a single input-single output (SISO) orthogonal frequency division multiplexing (OFDM) system using a pre-fast Fourier transform (pre-FFT) SNR estimator. The pre-FFT SNR estimator requires no additional overhead since it reuses the preamble for synchronization in the OFDM system. In this work, a preamble structure proposed by Morelli and Mengali to overcome carrier frequency offset (CFO) due to Doppler effects is utilized. The proposed pre-FFT SNR estimator has been employed to estimate SNR for the SISO-OFDM system, and its performance has been evaluated against the corresponding frequency domain SNR estimator, also known as a post-FFT SNR estimator. The normalized mean square error (NMSE) of the pre-FFT SNR estimator has also been evaluated against the normalized Cramer-Rao bound (NCRB). The simulation results show that for the additive white Gaussian noise (AWGN) and Stanford University Interim-5 (SUI-5) channels, the pre-FFT SNR estimator exhibits 0.41 dB and 0.66 dB difference between the estimated SNR and the actual SNR, respectively. The NMSE of the pre-FFT SNR estimator outperforms the benchmarker post-FFT SNR estimator, which is close to the NCRB. The proposed pre-FFT SNR estimator achieved bit error rate (BER) improvements of about 1 dB and 2 dB for AWGN and SUI-5 channels, respectively, over the post-FFT SNR estimator at BER= 10^{-4}. Moreover, there is an approximately 50% reduction in complexity between the proposed pre-FFT SNR estimator and the benchmarker post-FFT SNR estimator.*

Keywords: *Adaptive modulation, Efficiency, OFDM, Preamble-based SNR estimator, Constant Amplitude Zero Autocorrelation preamble, Channel capacity*

Received: Received: August 13, 2024; Received in revised form: October 4, 2024; Accepted: November 2, 2024

1. INTRODUCTION

The most widely used multicarrier modulation technology underpinning the fifth-generation (5G) mobile communications networks is orthogonal frequency division multiplexing (OFDM). It offers strong performance in frequency-selective channels and facilitates the effective use of the available channel capacity [1]. Adaptive transmission can significantly enhance an OFDM system's performance in the presence of a frequency-selective channel. Due to this, the signal-to-noise ratio (SNR) is a critical component of adaptive transmission. The SNR value denotes the channel quality, and an adaptive modulation and coding (AMC) scheme adapts parameters like modulation and coding schemes by the channel condition [2]. In the AMC scheme, SNR is computed at the receiver to assess channel quality, and its value is sent back to the transmitter for parameter adjustment [3, 4]. This process imposes feedback overhead.

Unmanned aerial vehicle (UAV) communication is an application in which where the AMC technique is utilized to alleviate problems in dynamically changing communication environments. Recent studies on AMC for UAV communication have utilized machine learning methods to assess channel quality for the modulation and coding scheme parameters selection [5-7]. The authors in [7] studied the AMC scheme in a UAV-to-ground free space optical communication system, which employed a machine learning-based channel estimator considering turbulence effects. In [8], deep reinforcement learning combined with a neural network was used to predict channel conditions for underwater acoustic OFDM communication systems, resulting in an improved bit-error rate (BER) and spectral efficiency. However, the proposed machine learning-based AMC schemes depend on the quality of the training data, such as the estimated SNR and other relevant channel atmospheric parameters. Improving SNR estimation accuracy would therefore prove advantageous.

SNR estimation is beneficial in a high-mobility environment, in which the channel condition is rapidly changing, and the Doppler effect is generated. Such fluctuations in the time domain require a well-estimated SNR for an adaptive transmission to achieve a significant throughput gain, resulting in an improved spectrum efficiency [3]. Therefore, a highly accurate SNR estimation method should ensure the intended level of communication performance by invoking the adaptive adjusted communication rate, modulation, and coding schemes [9]. However, a complicated SNR estimate method could result in feedback delays and worsen throughput performance.

There are two categories of SNR estimation, namely data-assisted (DA) and non-data-assisted (NDA) SNR estimation. The NDA-SNR estimator overcomes feedback overhead issues at the expense of lower accuracy because the transmitted signal's past information is not used to estimate SNR. This accuracy shortfall can be eliminated by employing a DA-SNR estimator, albeit at the expense of a throughput penalty, which increases the system's overhead. Nonetheless, research on DA-SNR that does not result in a throughput penalty has been done, such as the preamble-based SNR estimator of [10,11]. Thus, a preamble-based SNR estimator is considered in this study.

A preamble-based SNR estimator uses the OFDM system's synchronization preamble to estimate SNR. The OFDM system is sensitive to timing errors and frequency offsets, and various preamble structures have been proposed to address these limitations [12-15]. The SNR estimators proposed in [16-18] consider the carrier interference generated by the frequency offset, while most of the proposed SNR estimators assume perfect frequency synchronization.

Two main issues to be considered in developing SNR estimator are: (i) complexity and (ii) throughput penalty. There are two factors that can contribute to the computational complexity of a preamble-based SNR estimator, namely the type of SNR estimation domain and algorithm. In OFDM systems, SNR estimation is performed either in the frequency domain, also known as post-fast Fourier transform (FFT) estimation, or in the time domain, known as pre-FFT estimation. In the post-FFT estimation, the SNR is estimated after the FFT block of the OFDM system. In the pre-FFT estimation, SNR is estimated at the front-end of the receiver before demodulation of the received data. Thus, a pre-FFT estimation has lower complexity than a post-FFT estimation. In addition, pre-FFT estimation is less prone to carrier offset errors, hence avoiding losses in subcarrier orthogonality [19, 20]. On the other hand, some estimation algorithms use probabilistic approaches, which have higher computational complexity, in contrast to autocorrelation-based SNR estimation algorithms [11, 21].

Motivated by the advantages of the preamble-based SNR estimation algorithm in [11, 21], this study aims to develop an SNR estimator that has low computational complexity and low training symbol overhead. More specifically, this study developed a pre-FFT SNR estimation algorithm based on autocorrelation of the received signal at the receiver front-end and utilizing one preamble. However, the SNR estimation algorithm developed in [11, 21] used synchronization preamble structure in [14], which does not consider carrier frequency offset (CFO) in the algorithm. Therefore, this paper investigates an SNR estimation algorithm with a frequency offset. The proposed pre-FFT SNR estimator exploits the preamble structure of [12], where SNR is estimated using the autocorrelation function, and its algorithm utilizes one synchronization preamble. The performance of the proposed pre-FFT SNR estimator is contrasted with the benchmarker post-FFT SNR estimator of [16], which utilizes the second-order moment criterion for SNR estimation. The benchmark SNR estimator is referred to as the Millan post-FFT SNR estimator. The SNR estimator performance has also been evaluated against the normalized Cramer-Rao bound (NCRB) to assess how well the developed pre-FFT SNR estimator could approach the theoretically best achievable performance, thus ensuring system efficiency is not compromised.

The contributions of this paper are as follows:

1. A pre-FFT SNR estimator that utilizes the preamble structure for synchronization in the OFDM system of [12] is contrived. Hence, there is no throughput penalty associated with the proposed SNR estimate. Moreover, this SNR estimator utilizes one preamble that reduces the training symbol overhead.
2. The proposed SNR estimation algorithm has low computational complexity for two reasons: (i) the SNR estimation is done at the front end of the receiver, prior to the demodulation; and (ii) it is derived from the autocorrelation function.

The remainder of the paper is structured as follows: The review of the related work is presented in Section 2. Section 3 discusses the description of the single-input single-output (SISO)-OFDM system that incorporates the proposed pre-FFT SNR estimator. Section 4 elaborates on both the proposed pre-FFT SNR estimator and Milan's post-FFT SNR estimator benchmark. Section 5 compares the performance of the proposed pre-FFT SNR estimator with that of Milan post-FFT SNR estimator and the NCRB to assess how well the proposed SNR estimator can come close to the theoretical best performance. This section also presents the computational complexity analysis of various preamble-based SNR estimators. Finally, Section 6 offers conclusions.

2. RELATED WORK

Many preamble-based SNR estimation methods have been developed over the years. The application of machine learning to the SNR estimate algorithm has garnered more attention lately [22, 23]. In [22], a deep-learning-based SNR estimator was reported that provided accurate estimation and improved the system performance at the expense of computational complexity both during training and inference. Supervised learning requires a sufficient set of training data, including SNR values [23], for a reliable model.

Table 1 provides an overview of previous studies on preamble-based SNR estimators in OFDM systems. These SNR estimators took advantage of various synchronization preamble structures.

SNR is an important parameter that reflects channel quality. Accurate SNR estimation plays a crucial role in ensuring a desired communication performance in a rapidly changing environment, such as in UAV communication systems. As discussed in Section 1, the benefits of a preamble-based SNR estimator are two-fold: (i) It is a DA-SNR estimator that has higher estimation accuracy, and (ii) it utilizes synchronization preamble, which eliminates throughput penalty. However, it is also favorable to ensure that the SNR estimation algorithm has low computational complexity. Therefore, per Table 1, the pre-FFT SNR estimation algorithms in [11, 19, 21, 24, 25, 28] are less complex since SNR is estimated at the receiver's front-end prior to demodulation of received data.

Table 1 also shows that the computational complexity of a preamble-based SNR estimator is highly dependent on the estimation algorithm, which uses maximum likelihood, second-order moment criteria, correlation, circular correlation, and autocorrelation function. For example, compared to the SNR estimator based on autocorrelation, second-order moment, which uses probabilistic approaches, has higher computational costs since it includes more multiplication and addition operations.

Table 1. Summary of preamble-based SNR estimations in the literature

Year	Author(s)	SNR Estimation Domain	SNR estimation algorithm	Contribution	Challenges
2009	Zivkovic, M. & Mathar, R. [16]	Post-FFT	Second-order moment	• Proposed SNR estimator that exploited preamble structure in [12]. • It used only one preamble to minimize the transmission overhead	• Showed poor normalized mean square error (NMSE) performance in the low region of channel SNR
2010	Zivkovic, M. & Mathar, R. [17]	Post-FFT	Second-order moment	• Extension of SNR estimator in [16]. • Improved SNR estimation for all SNRs • Utilized the method for adaptive selection of significant channel impulse response	• Post-FFT estimator performance degraded in the presence of inter-carrier interference • High complexity
2014	Ijaz, A. et all. [24]	Pre-FFT	Correlation	• Low complexity time domain SNR estimation is proposed for the OFDM system • It used one synchronization preamble proposed by Schmidl and Cox [13]	• Poor performance at low SNR
2014	Zivkovic, M. & Mathar, M. [25]	Pre-FFT	Second-order moment	• Improved Zadoff-Chu Preamble-based SNR estimation in the time domain is proposed for the OFDM system. Improved SNR estimation compared to [16, 17] using one preamble with Q > 2 equal parts • Results are robust if Q > 8	• High complexity
2016	Ishtiaq, N. et al [26]	Post-FFT	Maximum likelihood	• Data-aided SNR estimation is done in the frequency domain using maximum likelihood • Use one preamble of [13] with known pilot value insertion • The accuracy of the estimates shows improvement in the lower region	• Higher bandwidth utilization and higher complexity
2018	Aloui, A. et al. [27]	Post-FFT	Expectation statistical method	• SNR estimation is proposed for IEEE 802.15.4g OFDM • Use two preambles, one for synchronization and one for SNR estimation, as proposed by Schmidl and Cox [13]	• In lower SNR values, the estimates show a higher bias than the actual SNR • High complexity.
2018	Abid, M.K. et al. [28]	Pre-FFT	Circular correlation	• Pilot data-aided time domain SNR estimation is proposed for the OFDM system • Known pilot values inserted in the preamble signal	• Complexity is higher to achieve accurate SNR estimates
2018	Manzoor, S. & Othman, N, [21]	Pre-FFT	Autocorrelation	• The time synchronization preamble of [14] is used for time domain SNR estimation in cooperative systems. • Enhanced performance at low SNR region	• Did not consider CFO
2020	Rao, B.N. et al. [19]	Pre-FFT	Second order moment	• Preamble-based noise power estimation for the OFDM system is proposed • Time domain SNR estimation is less prone to frequency offset • Use one preamble having a preamble structure of [12]	• Did not consider CFO • Poor NMSE performance at low SNR region
2022	Manzoor, S. & Othman, N, [11]	Pre-FFT	Autocorrelation	• Utilized a modified synchronization preamble from [14] • Enhanced perfromance at low region as compared with the SNR estimator in [21] • Improved system performance with SNR estimation-based adaptive modulation scheme	• Did not consider CFO

In [12], Morelli and Mengali proposed a preamble structure for an algorithm to estimate frequency offset in an OFDM system with a reduction of training symbol overhead by employing a single preamble. Thus, the authors in [16] exploited the beneficial feature of this preamble structure in the proposed post-FFT SNR estimator. The post-FFT SNR estimator exploited the periodic nature of the preamble structure and used only one preamble, therefore minimizing the transmission overhead. The SNR estimation algorithm was derived using the second-order moment algorithm. However, the SNR estimator in [16] showed poor normalized mean squared error (NMSE) performance in the low region of channel SNR. The post-FFT SNR estimation algorithm was further improved by utilizing the method for adaptive selection of significant channel impulse response, which resulted in an improved SNR estimation for all SNRs, as presented in [17]. Similarly, a low complexity pre-FFT SNR estimator proposed in [24] also struggled with accuracy in the low region of channel SNR.

As a further enhancement, the authors in [26] proposed a post-FFT SNR estimation using a maximum likelihood approach, which improved channel SNR in the low region. However, the performance of the post-FFT SNR estimator degraded in an imperfectly synchronized system due to inter-carrier interference (ICI) caused by carrier frequency offset [25].

The preamble structure from [14] was utilized by the authors in [21] to construct an SNR estimator invoked in a cooperative SISO-OFDM system. In [11], the proposed adaptive modulation with SNR estimator utilized the modified OFDM synchronization preamble structure developed in [14]. Both SNR estimators are categorized as pre-FFT SNR estimators, in which the SNR is estimated in the time domain, and the SNR estimation algorithm utilizes autocorrelation. Thus, the SNR estimators developed in [11, 21] are attractive due to these two criteria, which result in low computational complexity.

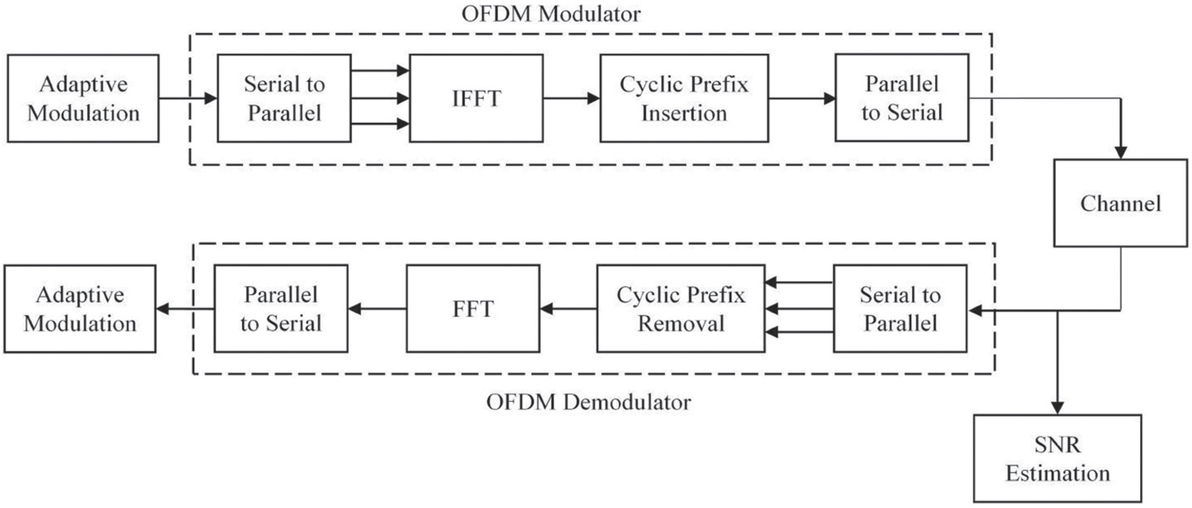


Fig. 1. SISO-OFDM system block diagram

3. SYSTEM DESCRIPTION

This paper considers a SISO-OFDM system that invokes the pre-FFT SNR estimator, as shown in Fig. 1. At the receiver, SNR estimation is performed before FFT processing.

The input data is mapped into symbols using quadrature phase shift keying (QPSK) which are then converted from serial-to-parallel stream. Next, the symbols are transformed into time domain symbols using the inverse fast Fourier transform (IFFT). More specifically, at the transmitter, the time domain OFDM signal after applying IFFT is given as:

$$x(n) = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} X(k) e^{j2\pi nk/N}, n = 0, \dots, (N-1) \quad (1)$$

where N is the size of IFFT, $X(k)$ represents the QPSK constellation point modulated data on the kth subcarrier, while in the time domain, $x(n)$ denotes the nth data sample.

A cyclic prefix, C_p is inserted into each QPSK symbol as the guard interval to avoid inter-symbol interference to form an OFDM symbol. Then, the OFDM signal is transmitted to the receiver via a wireless channel.

The OFDM signal at the input of the receiver in the presence of the CFO can be described as follows:

$$r(n) = x(n). e^{j2\pi\varepsilon n/N} + w(n) \quad (2)$$

where $w(n)$ is the noise signal at the receiver antenna, and ε is the CFO normalized to the subcarrier spacing. The SNR estimation is performed on the CFO compensated received signal before FFT processing.

4. PROPOSED PRE-FFT SNR ESTIMATION

As discussed in Section 1, the proposed pre-FFT SNR estimator performs SNR estimation in the time domain. It utilizes the preamble structure that was proposed in [12], which comprised of one OFDM symbol with Q equal sections that are all having N/Q length, where N is the IFFT size and Q>2. Fig. 2 shows the preamble, which has a repetitive structure is utilized, where P_N represents the pseudo-noise sequence, Q=4 and N=256 bits. The same preamble is also used for synchronization in the OFDM system. Thus, its use in SNR estimation does not penalize the system throughput. For benchmarking, the proposed pre-FFT SNR estimator's performance is evaluated against the Milan post-FFT SNR estimator [16]. Both estimators use the same preamble structure as shown in Fig. 3, with Q=4, N=256 bits and C_p=64 bits for the SNR estimation algorithm.

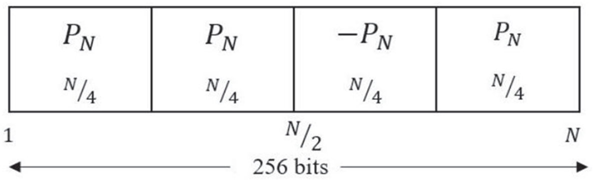


Fig. 2. Morrelli synchronization preamble structure [12]

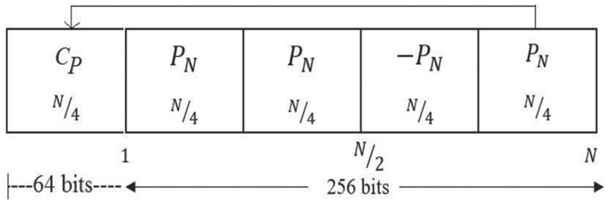


Fig. 3. Morrelli synchronization preamble structure with cyclic prefix

The Morelli and Mengali preamble structure in [12] has a periodic structure in the time domain, which corresponds to a comb-type structure in the frequency domain. Thus, for the SNR estimation algorithm in [16], the total number of N subcarriers are divided into Q parts. Each part consists of N_p=N/Q subcarriers. In each part, starting from the zeroth, every Q-th subcarrier was modulated with a QPSK signal, $X_p(n)$ for n=0,…,(N_p-1), while the remainder of the subcarriers are not used (null). The same Morelli preamble structure and the modulation technique are utilized in the proposed pre-FFT SNR estimator.

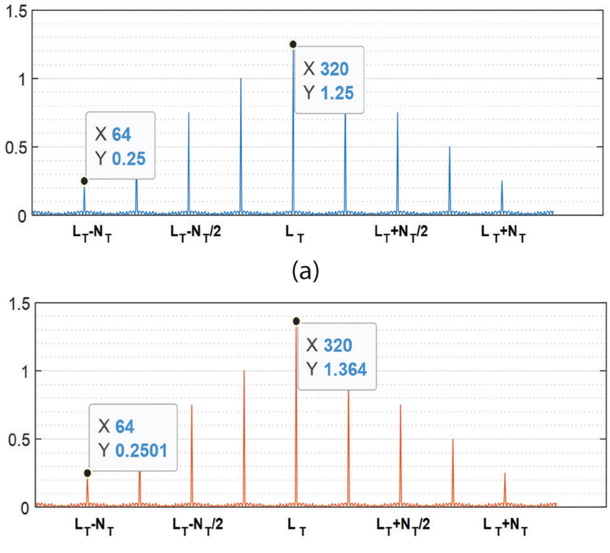
The SNR estimation algorithm of the proposed pre-FFT SNR estimator utilizes the autocorrelation function of the received signal of Eq. 2 to estimate the signal and noise power. The autocorrelation of the received signal with additional noise from the channel, $r_{rx}(t)$, can be written as:

$$r_{rx}(t) = r_{tx}(t) + r_{nw}(t) \tag{3}$$

where $r_{tx}(t)$ denotes the autocorrelation of the transmitted OFDM signal. For the noise signal's autocorrelation, $r_{nw}(t)$ for the AWGN channel with the noise variance of σ^2, can be expressed as follows:

$$r_{nw}(t) = \sigma^2 \delta(t) \tag{4}$$

where $\delta(t)$ is Dirac delta function. Similarly, the transmitted OFDM signal's autocorrelation can be written as $r_{tx}(t)$=$P_{tx}\ \delta(t)$, where P_{tx} is the signal power. Hence, at zeroth lag, the received OFDM signal's autocorrelation consists of both the signal and noise power. On the other hand, the transmitted OFDM signal's autocorrelation consists of signal power only. Thus, the difference between the received OFDM signal's autocorrelation value at zeroth lag and the estimated signal power can be used to estimate noise power.



(a)

(b)

Fig. 4. Autocorrelation illustration at SNR=10 dB (**a**) The transmitted OFDM signal (Transmitted Signal Autocorrelation at SNR = 10 dB). (**b**) The corresponding received OFDM signal (Received Signal Autocorrelation at SNR = 10 dB).

Fig. 4 shows the autocorrelation plot of the transmitted and the received OFDM signals at 10 dB channel SNR. In Fig. 4, the X-axis represents the lag between the signal and its shifted version, while the Y-axis represents the autocorrelation values at each lag. There is one prominent peak at L_T, and there are four side peaks on its right and left sides. The four side peaks on the left side appeared at specific lags of (L_T-N_T), (L_T- (3/4) N_T), (L_T- (1/2) N_T) and (L_T- (1/4) N_T).

Fig. 5 illustrates the autocorrelation stages resulting in the plot in Fig. 4. As a result, the estimation of signal power can be written as:

$$P^*_{tx} = 3\, r_{rx}\left(L_T - \frac{3}{4}N_T\right) - r_{rx}(L_T - N_T) \tag{5}$$

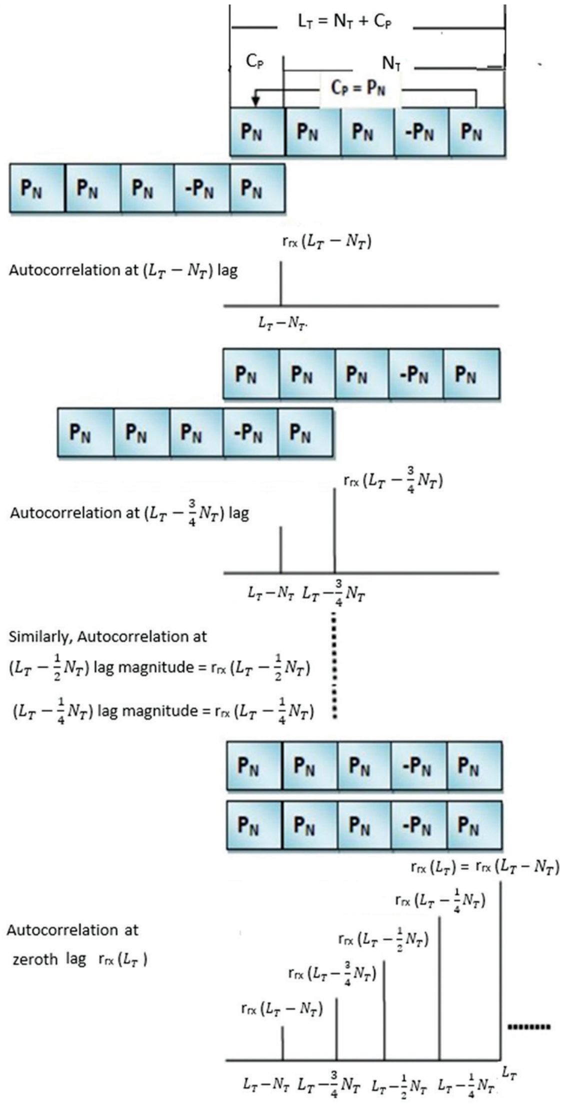
where N_T is the OFDM signal length and L_T= N_T + C_p is the total length, which includes C_p = L_T-N_T.

Having the estimated signal power defined by Eq. 5, the noise power can be estimated as:

$$\sigma^2_{est} = r_{rx}(L_T) - P^*_{tx} \tag{6}$$

where $r_{rx}(L_T)$ is the maximum peak indicating the received OFDM signal's autocorrelation value at zeroth-lag. Therefore, the estimated SNR can be calculated by utilizing Eq. 5 and Eq. 6, which can be written as:

$$SNR_{Est} = \frac{P^*_{tx}}{\sigma^2_{est}} \tag{7}$$



$L_T = N_T + C_P$
C_P
N_T
$C_P = P_N$
P_N P_N P_N $-P_N$ P_N
P_N P_N P_N $-P_N$ P_N
$r_{rx}(L_T - N_T)$
Autocorrelation at $(L_T - N_T)$ lag
$L_T - N_T$
$r_{rx}(L_T - \frac{3}{4}N_T)$
Autocorrelation at $(L_T - \frac{3}{4}N_T)$ lag
$L_T - N_T$ $L_T - \frac{3}{4}N_T$
Similarly, Autocorrelation at
$(L_T - \frac{1}{2}N_T)$ lag magnitude = $r_{rx}(L_T - \frac{1}{2}N_T)$
$(L_T - \frac{1}{4}N_T)$ lag magnitude = $r_{rx}(L_T - \frac{1}{4}N_T)$
$r_{rx}(L_T) = r_{rx}(L_T - N_T)$
$r_{rx}(L_T - \frac{1}{4}N_T)$
$r_{rx}(L_T - \frac{1}{2}N_T)$
$r_{rx}(L_T - \frac{3}{4}N_T)$
$r_{rx}(L_T - N_T)$
Autocorrelation at zeroth lag $r_{rx}(L_T)$
$L_T - N_T$ $L_T - \frac{3}{4}N_T$ $L_T - \frac{1}{2}N_T$ $L_T - \frac{1}{4}N_T$ L_T

Fig. 5. The transmitted OFDM signal's autocorrelation function, which includes the cyclic prefix

4.1. POST-FFT SNR ESTIMATION BENCHMARKER

Milan's post-FFT SNR estimator benchmarker of [16] is used for comparison with the proposed pre-FFT SNR estimator. More specifically, the SNR estimation in the benchmarker SNR estimator is performed in the frequency domain after FFT processing. For a fair comparison, both the proposed pre-FFT SNR estimator and the benchmarker SNR estimator utilize the Morelli and Mengali preamble structure of [12], as shown in Fig. 2, which consists of Q parts, each containing $N_p = N/Q$ samples.

Similarly, the same OFDM modulation method is used for both the proposed pre-FFT SNR estimator and the benchmarker SNR estimator, where the QPSK signal, $X_p(m)$ for $m=0,\dots,(N_p-1)$ is loaded in every Q-th subcarrier. The remainder (N-N_p) of the subcarriers are not used (nulled). Therefore, the transmitted signal on the kth subcarrier can be expressed as [16]:

$$X(k) = X(mQ+q) = \begin{cases} X_P(m), & q = 0 \\ 0, & q = 1, \cdots, Q-1 \end{cases} \quad (8)$$

where $k=mQ+q$, with $m=0,\dots,(N_p-1)$ and $q=0,\cdots,(Q-1)$. Hence, the index of the loaded subcarriers is $k=mQ$, with $m=0,\dots,(N_p-1)$ and $q=0$.

Milan's post-FFT SNR estimation algorithm was developed based on the second-order moment of the demodulated OFDM signal to estimate the SNR at the receiver. After the CFO compensation, the received signal on the loaded subcarrier can be expressed as [16]:

$$Y(k) = Y(mQ) = \sqrt{SQ}X_P(m)H_P(m) + \sqrt{W}\sigma(m) \quad (9)$$

where SQ is the total transmit power and H_p (m) is the channel response on the loaded subcarriers. W is the noise power on each subcarrier, and $\sigma(m)$ is the corresponding sampled zero-mean AWGN with unit variance.

The received signal on the nulled subcarriers consists of only noise signal and is given as [16]:

$$Y(k) = Y(mQ+q) = \sqrt{W}\sigma(mQ+q) \quad (10)$$

where $q=1,\dots,(Q-1)$.

The second-order moment is applied to the received signal, $Y(mQ)$ on the loaded subcarriers as shown in Eq. 10, using expressions of [16]:

$$P_{RS} = \frac{1}{N_p}\sum_{m=0}^{N_p-1} |Y(mQ)|^2 \quad (11)$$

Similarly, the received noise power from the nulled subcarrier is given as [16]:

$$P_{RN} = \frac{1}{N_p(Q-1)}\sum_{m=0}^{N_p-1}\sum_{q=1}^{Q-1} |Y(mQ+q)|^2 \quad (12)$$

Thus, the SNR estimation can be determined using the following equation:

$$SNR_{Est} = \frac{1}{Q}\left(\frac{P_{RS} - P_{RN}}{P_{RN}}\right) \quad (13)$$

5. RESULTS AND DISCUSSION

In this section, the performance of the proposed pre-FFT SNR estimator is characterized when it is invoked in the SISO-OFDM system, as described in Section 3. The comparison performance of the proposed pre-FFT SNR estimator is also investigated against the post-FFT SNR estimator benchmarker using estimated SNR, BER, NMSE, and computational complexity. Table 2 displays the simulation settings for the SISO-OFDM system, which were selected from the IEEE802.16d standard [29].

Table. 2. IEEE802.16d Standard Parameters for OFDM [29]

Parameters	Value
$IFFT_{Length}$, N_{ifft}	256
$Sampling_{Frequency}$, F_s	20 MHz
$SubCar_{Spacing}$, $\Delta f = F_s/_N$	1×10^5
$Symbol_{Time}$, $T_{sy} = 1/\Delta f$	1×10^{-5}
$Guard_{Interval}$, $T_{gi} = G_i \times T_{sy}$	2.5×10^{-6}
$OFDM_{Symb\text{-}time}$, $T_s = T_{sy} + T_{gi}$	1.25×10^{-5}
$Channel_{Used}$	AWGN, SUI-5

As discussed in Section 4, the preamble structure with Q=4, N=256 bits and C_P length of N/4=64 bits are utilized in this work. Thus, the total frame length is L_T=N+C_P=320 bits. The simulation results consider the advocated scheme when communicating over the AWGN and SUI-5 channels. The SUI-5 channel parameters used in this study are shown in Table 3, indicating the multipath delay and power profile. The SUI channels are designed to model three outdoor-terrain categories, as shown in Table 3, and have been adopted by IEEE802.16d standard [30]. Table 4 also shows that the SUI-5 channel models type A terrain, which deals with huge path loss and it is most suited for hilly terrain with high densities of foliage. The estimated SNR is obtained for both the SNR estimators, and estimates of SNR are obtained by averaging over M_t=2000 iterations.

Table. 3. Channel Description of SUI-5 Wireless Channel [30]

SUI5 channel	Path 1	Path 2	Path 3	Unit
$Path_{Delay}$	0	4	10	*μsec*
$Path_{Power}$	0	-11	-22	dB
K_{Factor}	2	0	0	–

Table. 4. Types of Terrain Corresponding to SUI Channels [30]

TerrainTypes	SUIChannels
C	SUI-1, SUI-2
B	SUI-3, SUI-4
A	SUI-5, SUI-6

Fig. 6 shows the estimated SNR performance of the proposed pre-FFT SNR estimator and the corresponding benchmarker post-FFT SNR estimator for transmission over the AWGN channel. The proposed pre-FFT estimator is further compared against the preamble-based pre-FFT estimators proposed in [11]. More specifically, in [11], the preamble pre-FFT estimators exploit the synchronization preamble structure proposed in [14], referred to as the CAZAC pre-FFT SNR estimator.

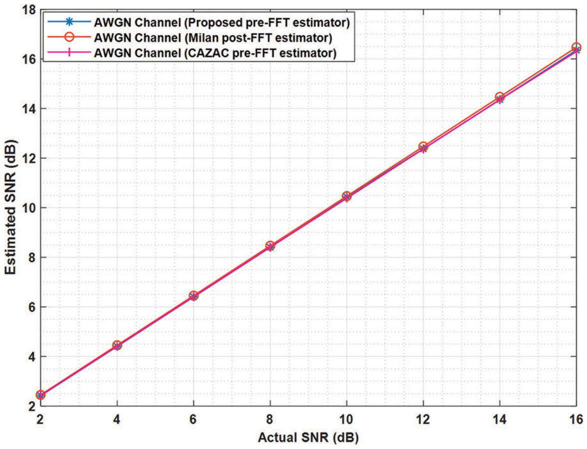


Fig. 6. Performance of the pre-FFT SNR estimator invoked in QPSK-SISO-OFDM system for transmission over AWGN channel in terms of estimated SNR

The close-up of Fig. 6 is shown in Fig. 7, where it can be observed that the proposed pre-FFT SNR estimator exhibited 0.41 dB difference from the actual SNR, which is referred to as bias. The benchmarker Milan post-FFT SNR estimator exhibited bias of 0.454 dB. On the other hand, the CAZAC pre-FFT SNR estimator exhibited bias of approximately 0.419 dB. The SNR estimation performance was estimated with the presence of CFO.

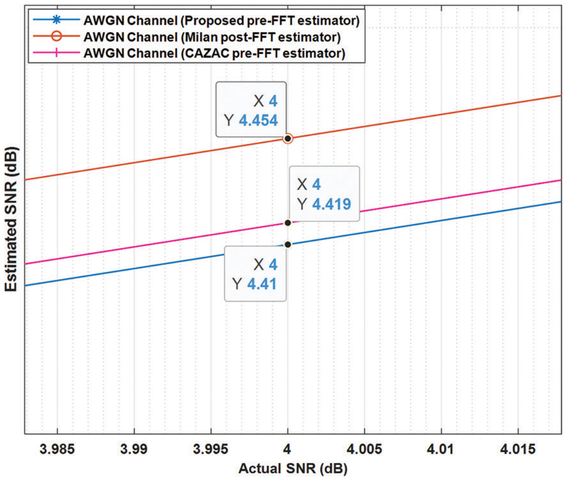


Fig. 7. Close-up of Fig. 6

Fig. 8 shows the estimated SNR performance of the proposed pre-FFT SNR estimator when communicating over the SUI-5 channel. The close-up of Fig. 8 is shown in Fig. 9, and it can be observed that the proposed pre-FFT SNR estimator exhibited a 0.66 dB difference between the estimated SNR and the actual SNR. The benchmarker Milan post-FFT SNR estimator exhibited 0.72 dB bias. Hence, the proposed pre-FFT SNR estimator outperformed its corresponding benchmarker. The CAZAC pre-FFT SNR estimator of [11] was outperformed by exhibiting a 0.663 dB difference between the estimated and actual SNR values.

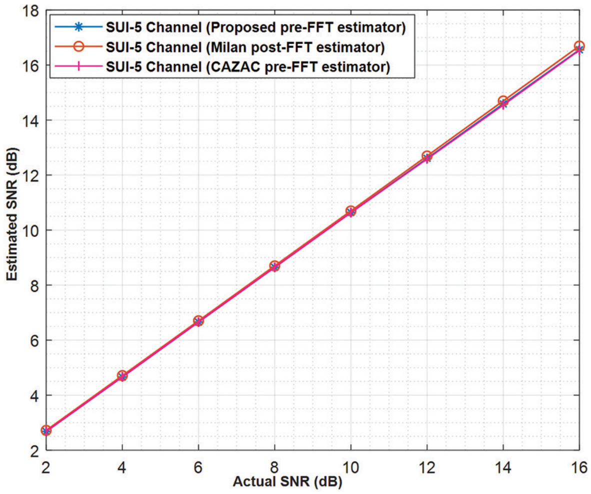


Fig. 8. Performance of the proposed pre-FFT SNR estimator invoked in the QPSK-SISO-OFDM system for transmission over the SUI-5 channel in terms of estimated SNR

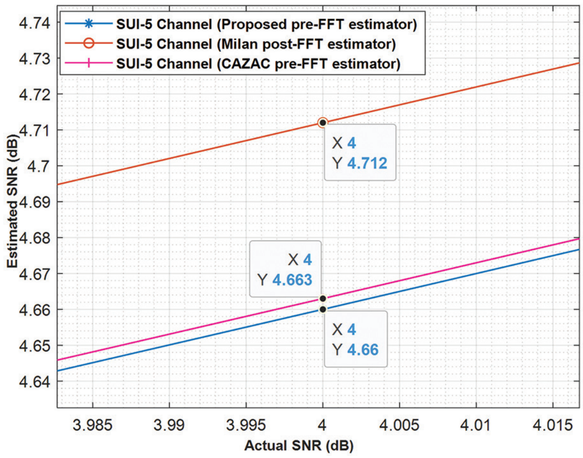


Fig. 9. Close-up of Fig. 8

In both cases, the preamble-based pre-FFT SNR estimators demonstrated better SNR estimation than the Milan post-FFT SNR estimator. Thus, the pre-FFT SNR estimators exhibit beneficial performance for dynamic environments in the presence of the CFO due to Doppler effects, such as in UAV or vehicular communication systems [18]. Thus, employing such an SNR estimator is beneficial for applying AMC to maximize throughput performance in the dynamic fading of wireless channels. Moreover, the proposed pre-FFT SNR estimator performed better in terms of SNR estimation under the CFO scenario.

Fig. 10 shows the average difference between the estimated SNR and the actual SNR, which is referred to as bias. It is observed that the average bias of the pre-FFT SNR estimator exhibits better performance in the region of low values of the actual SNR for transmission over the SUI-5 channel than that of its corresponding benchmarker post-FFT SNR estimator.

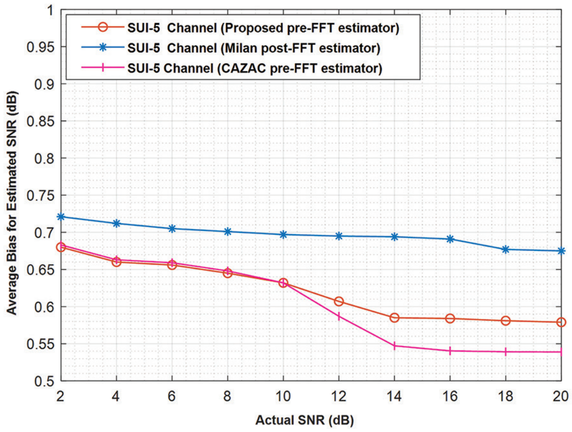


Fig. 10. Estimated SNR bias versus actual SNR performance

The pre-FFT SNR estimator performance has also been evaluated in terms of NMSE. Hence, the NMSE performance is quantified using Eq. 14, where SNR_{act} denotes the average value of the actual SNR, while SNR_{Est} for the proposed pre-FFT SNR and the benchmarker post-FFT SNR estimators can be calculated using Eq. 7 and Eq. 13, respectively:

$$NMSE = \frac{1}{M_t}\sum_{n=1}^{M_t}\left[\frac{SNR_{Est} - SNR_{act}}{SNR_{act}}\right]^2 \tag{14}$$

The performance of the proposed SNR estimator is evaluated against the NCRB for frequency selective channel to assess how effectively the proposed SNR estimator performance approaches the theoretical optimum. The CRB was derived in [31] as follows:

$$CRB_{(SNR_{Est})} = \left[\frac{1 + Q(SNR_{act})}{\sqrt{N(Q-1)}}\right]^2 \tag{15}$$

where N=256 bits is the preamble length, Q=4 is the number of preamble parts, as discussed in Section 4.

The variance of CRB can be found by taking the inverse of the Fisher information matrix (FIM) [31]. Hence, the NCRB can be obtained by dividing Eq. 15 by $(SNR_{act})^2$ and written as follows:

$$NCRB(SNR_{Est}) = \frac{CRB_{(SNR_{Est})}}{(SNR_{act})^2} \tag{16}$$

where SNR_{Est} for the proposed pre-FFT SNR and the benchmarker post-FFT SNR estimators can be calculated using Eq. 7 and Eq. 13, respectively.

Figs. 11 and 12 illustrate the NMSE comparison performance of the proposed pre-FFT SNR estimator and the benchmark post-FFT SNR estimator for AWGN and SUI-5 channels, respectively. The pre-FFT SNR estimator outperforms its benchmarker post-FFT SNR estimator, as seen in Fig. 11. In the region of high values of an actual SNR of more than 6 dB, the NMSE performance of the pre-FFT SNR estimator improves with an increase in actual SNR for the AWGN channel. On the other hand, the NMSE performance of the benchmarker post-FFT SNR estimator shows no further improvement in the region of actual SNR of more than 12 dB. It can also be seen that the NMSE performance of the proposed pre-FFT SNR approaches the NCRB and outperforms the considered post-FFT SNR estimator.

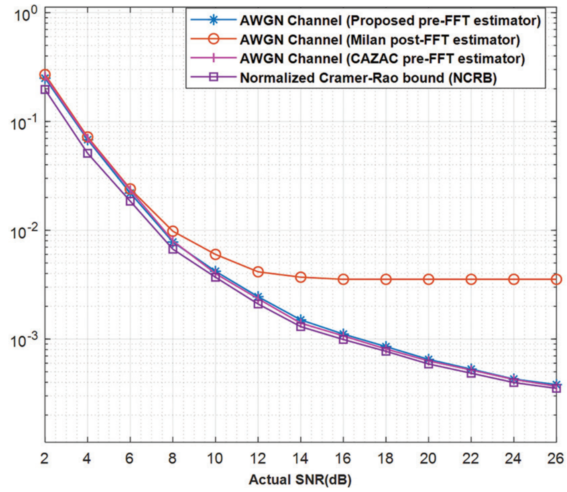


Fig. 11. Comparison of NMSE performance of the pre-FFT SNR estimator and that of its corresponding benchmarker for transmission over AWGN channel

Similarly, for transmission over SUI-5 channel, there is no more NMSE improvement in the region of high values of actual SNR of more than 16 dB for post-FFT SNR estimator, as shown in Fig. 12. The NMSE performance of the proposed pre-FFT SNR outperforms the post-FFT SNR estimator at all regions of SNR. It is observed that the pre-FFT SNR NMSE performance approaches the NCRB. In both cases, the preamble-based pre-FFT SNR estimators demonstrated similar NMSE performance with the CAZAC pre-FFT SNR estimator. This observation aligns with the SNR estimated performances shown in Fig. 7 and Fig. 9.

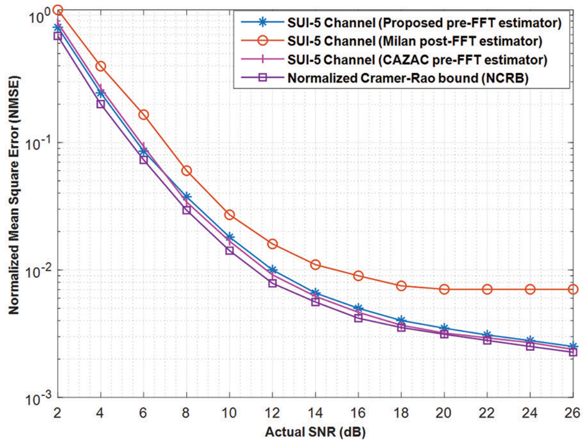


Fig. 12. Comparison of NMSE performance of the pre-FFT SNR estimator and that of its corresponding benchmarker for transmission over the SUI-5 channel

Fig. 13 and Fig. 14 compare the BER performance of the proposed pre-FFT SNR estimator and that of its corresponding post-FFT SNR estimator benchmark for the AWGN and SUI-5 channels, respectively. From Fig. 13, the QPSK-SISO-OFDM scheme that invokes the pre-FFT SNR estimator outperforms the benchmark scheme with post-FFT SNR estimator by about 1.0 dB at $BER=10^{-4}$. Similarly, it can be seen in Fig. 14 that the pre-FFT SNR estimator outperforms the post-FFT SNR estimator benchmark scheme by about 2.0 dB at $BER=10^{-4}$.

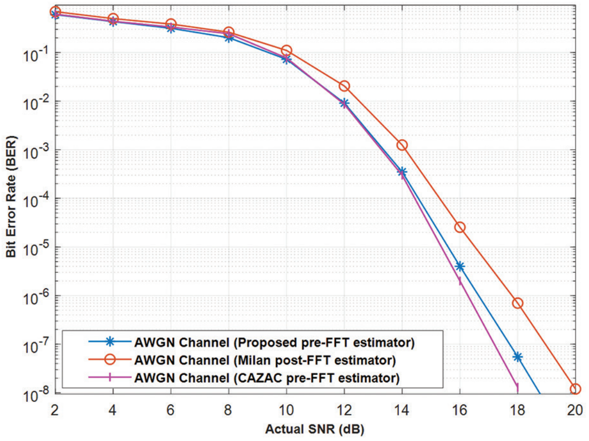


Fig. 13. Performance of the QPSK-SISO-OFDM system for transmission over AWGN channel in terms of BER

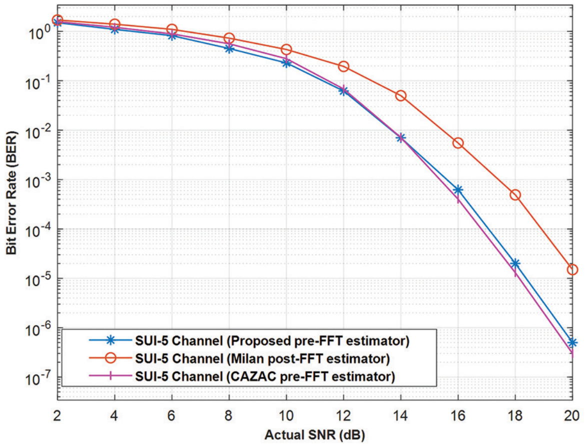


Fig. 14. Performance of the QPSK-SISO-OFDM system for transmission over the SUI-5 channel in terms of BER

The simulation results show that the SNR estimation in time domain is less prone to CFO error than the SNR estimation in frequency domain.

5.1. COMPLEXITY ANALYSIS

The floating point operations per second (FLOP) complexity metric is used to assess the complexity of the pre-FFT SNR estimate. Table 5 depicts the complexity comparison. Generally, FLOPs refer to the number of computations needed for a single SNR estimate.

Table. 5. Complexity Analysis

SNR Estimator	SNR estimation algorithm	SNR Estimation Domain	FLOPs
PTD [28]	Circular correlation	Pre-FFT	$2N_p(Q_p)+3N_p+L$
TLSE [24]	Second order moment	Pre-FFT	$5N$
TPSE [20]	Second order moment	Pre-FFT	$7.5N+4$
TDZCE [25]	Second order moment	Pre-FFT	$4N+2$
Milan [16]	Second order moment	Post-FFT	$4N+2$
Proposed	Autocorrelation	Pre-FFT	$2N$-1

The pre-FFT SNR estimator proposed in this study is based on the autocorrelation of the received signal in the time domain. As explained in Section 4, the signal power and noise power are estimated using Eq. 5 and Eq. 6, respectively. Generally, the autocorrelation function calculates the product of the received signal and its lagged version at each time step and then sums these products for all time steps within the overlapping range. Moreover, the autocorrelation function is computed at zeroth lag; the computational complexity is only based on the multiplications of N bits and $(N-1)$ additions. Hence, the pre-FFT SNR estimator required $(N+N\text{-}1= 2N–1)$ FLOPs for one SNR estimation.

The complexity of the proposed pre-FFT SNR estimator is compared with SNR estimators developed in previous studies as [16], [20] [24, 25], [28]. According to [28], the pilot-based time domain SNR estimator (PTD) requires ($2N_p(Q)+3N_p+L_p$) FLOPs, and the complexity relies on the number of pilot subcarriers $N_p = N/Q$., and L_P channel taps, in which Q represents the number of preamble parts, as discussed in Section 4.

The time domain low complexity SNR estimator (TLSE) was investigated in [24]. This requires $5N$ FLOPs and depends on the number of periodic parts Q. The authors in [20] introduced the time domain preamble-based SNR estimator (TPSE), which requires ($7.5N + 4$) FLOPs to perform estimation of one SNR estimate. Time domain Zadoff-Chu preamble-based SNR estimator (TDZCE) presented in [25], needs ($4N + 2$) FLOPs. The Milan SNR estimator [16], presented in Section 4.1, involves ($4N + 2$) FLOPs to compute one SNR estimate.

Therefore, Table 5 shows that the proposed SNR estimator has the lowest complexity for estimating SNR. It can also be observed that the benchmarker post-FFT Milan SNR estimator requires ($4N + 2$) FLOPs, in comparison with the proposed pre-FFT SNR estimator, which allows a 50% reduction in FLOPs.

6. CONCLUSION

This work presents a pre-FFT SNR estimator that utilizes preamble structure for synchronization in OFDM system in [12]. The performance comparison between the proposed pre-FFT SNR estimator and Milan post-FFT SNR estimator is presented, in which both SNR estimators utilize the same preamble structure. On the other hand, the auto-correlation function is the basis for the pre-FFT SNR estimation algorithm. The second-order moment is utilized in the post-FFT SNR estimator algorithm, which incurs higher computational complexity. The estimated SNR using the pre-FFT SNR estimator exhibited 0.41 dB and 0.66 dB bias when communicating over AWGN and SUI-5 channels, respectively. The benchmark post-FFT Milan SNR estimator exhibited 0.454 dB bias and 0.72 bias over AWGN and SUI-5 channels, respectively. Similarly, the pre-FFT SNR estimator outperformed the benchmarker Milan post-FFT SNR estimator in terms of NMSE. More specifically, the NMSE performance of Milan SNR estimator showed no further improvements in the region of actual SNR of more than 12 dB and 16 dB for AWGN and SUI-5 channels, respectively. It was also demonstrated that the NMSE performance of the proposed pre-FFT SNR estimator approached the theoretical limit set by the NCRB. Moreover, the NMSE performance of the benchmarker post-FFT in comparison to the post-FFT Milan SNR estimator, the proposed pre-FFT SNR estimator achieved BER improvements of about 1 dB and 2 dB, respectively, at BER=10^{-4} for transmission over AWGN and SUI-5 channels. There is about a 50% reduction in complexity between the proposed pre-FFT SNR estimator and the benchmarker post-FFT SNR estimator. Further studies should consider the development of a preamble-based SNR estimator for UAV communication, which considers the Doppler effect due to the flight speed. OFDM technology can successfully be leveraged into UAV communication. However, knowledge of exact UAV communication channels is required.

7. REFERENCES

[1] T. Liu, “A Review on the 5G Enhanced OFDM Modulation Technique”, Proceedings of the 3rd Asia-Pacific Conference on Communications Technology and Computer Science, Shenyang, China, 25-27 February 2023, pp. 677-683.

[2] S. Kojima, K. Watanabe, K. Maruta, C. J. Ahn, “Joint Adaptive Modulation and Transmit Power Control on FSS-OFDM Mobile Relay System”, Journal of Signal Processing, Vol. 23, No. 3, 2019, pp. 83-93.

[3] T. E. Bogale, “Adaptive Beamforming and Modulation Design for 5G V2I Networks”, Proceedings of the 10th Annual Computing and Communication Workshop and Conference, Las Vegas, NV, USA, 6-8 January 2020, pp. 90-96.

[4] S. Kojima, K. Maruta, C. J. Ahn, “Adaptive Modulation and Coding Using Neural Network Based SNR Estimation”, IEEE Access, Vol. 7, 2019, pp. 183545-183553.

[5] L. S. Chen, C.-H. Ho, C.-C. Chen, S.-Y. Kuo, "Learning Scheme for Adaptive Modulation and Coding in 5G New Radio", Proceedings of the 6th International Conference on System Reliability and Safety, Venice, Italy, 23-25 November 2022, pp. 430-434.

[6] Z. Wang, Y. Tang, S. Song, H. Chen, X. Lu, F. Liu, "SI-AMC: Integrating DL-Based Scenario Identification into Adaptive Modulation and Coding in Vehicular Communications", Proceedings of the IEEE Wireless Communications and Networking Conference, Dubai, United Arab Emirates, 21-24 April 2024, pp. 1-6.

[7] Q. Zhang, B. Liu, G. Chen, S. Zhan, Z. Li, J. Zhang, N. Jiang, B. Cao, Z. Li, "An Improved Adaptive Coding and Modulation Scheme with Hybrid Switching Standard for UAV-to-Ground Free Space Optical Communication", IEEE Photonics Journal, Vol. 16, No. 1, 2024, pp. 1-8.

[8] X. Cui, P. Yan, J. Li, S. Li, J. Liu, “Deep reinforcement learning-based adaptive modulation for OFDM underwater acoustic communication system”, EURASIP Journal of Advance Signal Processing, Vol. 1, 2023, pp. 1-23.

[9] S. Vappangi, V. M. Vakamulla, “Channel Estimation in ACO-OFDM Employing Different Transforms for VLC”, AEU - International Journal of Electronics and Communications, Vol. 84, 2018, pp. 111-122.

[10] D. R. Pauluzzi, N. C. Beaulieu, “A Comparison of SNR Estimation Techniques for the AWGN Channel”, IEEE Transactions on Communications, Vol. 48, 2000, pp. 1681-1691.

[11] S. Manzoor, N. S. Othman, “Adaptive Modulation with CAZAC Preamble-Based Signal-to-Noise-Ratio Estimator in OFDM Cooperative Communication System”, IEEE Access, Vol. 10, 2022, pp.126550-126560.

[12] M. Morelli, U. Mengali, "An improved frequency offset estimator for OFDM applications," IEEE Communications Letters, Vol. 3, No. 3, 1999, pp. 75-77.

[13] T. M. Schmidl, D. C. Cox, “Robust Frequency and Timing Synchronization for OFDM”, IEEE Transactions on Communications, Vol. 45, No. 12, 1997, pp.1613-1621.

[14] S. A. Suparna, S. Sekhar, S. P. Sakuntala, “An Efficient Preamble Design for Timing Synchronization in MIMO-OFDM Systems”, Proceedings of the International Conference on Control, Instrumentation, Communication and Computational Technologies, Kumaracoil, India, 18-19 December 2015, pp. 84-88.

[15] K. Yağlı, S. A. Çolak, “Preamble-Based Symbol Timing Algorithms in OFDM Systems”, The European Journal of Research and Development, Vol. 2, No. 2, 2022, pp. 445-458.

[16] M. Zivkovic, R. Mathar, “Preamble-Based SNR Estimation in Frequency Channels for Wireless OFDM Systems”, Proceedings of the IEEE 69th Vehicular Technology Conference, Barcelona, Spain, 26-29 April 2009, pp. 1-5.

[17] M. Zivkovic, R. Mathar, “An Improved Preamble-Based SNR Estimation Algorithm for OFDM Systems”, Proceedings of the 21st Annual IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Istanbul, Turkey, 26-30 September 2010, pp. 172-176.

[18] J. Li, M. Liu, N. Tang, B. Shang, "Non Data-Aided SNR Estimation for UAV OFDM Systems", Algorithms, Vol. 13, No. 1, 2020, pp. 1-11.

[19] B. N. Rao, M. V. Raghunadh, R. Sudheer, “Noise Power Estimation for OFDM System”, Proceedings of the 11th International Conference on Computing, Communication and Networking Technologies, Kharagpur, India, 1-3 July 2020, pp. 1-6.

[20] F. Yang, X. Zhang, Z. P. Zhang, “Time-Domain Preamble-Based SNR Estimation for OFDM Systems in Doubly Selective Channels”, Proceedings of the IEEE Military Communications Conference, Orlando, FL, USA, 29 October - 1 November 2012, pp. 1-5.

[21] S. Manzoor, N. S. Othman, “Signal to Noise Ratio Estimation Using CAZAC Time Synchronization Preamble in Cooperative Communication System”, Proceedings of the IEEE 4th International Symposium on Telecommunication Technologies, 26-28 November 2018, pp. 1-6.

[22] S. Zheng, S. Chen, T. Chen, Z. Yang, Z. Zhao, X. Yang, "Deep Learning-Based SNR Estimation", IEEE Open Journal of the Communications Society, Vol. 5, 2024, pp. 4778-4796.

[23] S. Kojima, K. Maruta, C. Ahn, "Adaptive Modulation and Coding Using Neural Network Based SNR Estimation", IEEE Access, Vol. 7, 2019, pp. 183545-183553.

[24] A. Ijaz, B. Awoseyila, B. G. Evans, “Low-Complexity Time-Domain SNR Estimation for OFDM Systems”, Electronics Letters, Vol. 47, No. 20, 2011, pp. 1154-1156.

[25] M. Zivkovic, R. Mathar, “Zadoff-Chu Sequence Based Time-Domain SNR Estimation for OFDM Systems”, Proceedings of the IEEE 15th International Workshop on Signal Processing Advances in Wireless Communications, Toronto, ON, Canada, 22-25 June 2014, pp. 110-114.

[26] N. Ishtiaq, S. A. Sheikh, “Maximum Likelihood SNR Estimation for QAM Signal Over Slow Flat Fading Rayleigh Channel”, KSII Transactions on Internet and Information Systems, Vol. 10, No. 11, 2016, pp. 5365-5380.

[27] A. Aloui, O. B. Rhouma, C. Rebai, “Preamble based SNR estimation for IEEE 802.15.4g MR-OFDM”, Proceedings of the 25th IEEE International Conference on Electronics, Circuits and Systems, Bordeaux, France, 9-12 December 2018, pp. 325-328.

[28] M. K. Abid, J. Varun, M. Z. Ur Rahman, T. J. Muhammad, O. Chugtai, M. H. Rehmani, “Pilot-Based Time Domain SNR Estimation for Broadcasting OFDM Systems”, Journal of Computer Networks and Communications, 2018, pp. 1-8.

[29] M. Weiss, “WiMAX General Information about the Standard 802.16”, Rohde & Schwartz Application Note, Germany, 2006.

[30] K. Hari, D. Baum, A. Rustako, R. Roman, D. Trinkwon, “Channel Models for Fixed Wireless Applications”, IEEE 802.16 Broadband Wireless Access Working Group, 2003.

[31] M. Morelli, M. Moretti, “Joint Maximum Likelihood Estimation of CFO, Noise Power, and SNR in OFDM Systems”, IEEE Wireless Communications Letters, Vol. 2, No. 1, 2013, pp. 42-45.

Multipath Routing Algorithm to find Optimal Path in SDN with POX Controller

Original Scientific Paper

Deepthi Goteti*
Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation,
Green Fields, Vaddeswaram, A.P, India
2102031088@kluniversity.in

Imran Rasheed
Department of Computer Science & Engineering,
Koneru Lakshmaiah Education Foundation,
Green Fields, Vaddeswaram, A.P, India
imran.rasheedamu@kluniversity.in

*Corresponding author

Abstract *– The past decade witnessed a tremendous increase in network usage, and traditional network architecture is needed to sustain modern requirements with high throughput and minute delay. This leads to the introduction of software-defined networks. Congestion is a critical problem that needs attention, so identifying the optimal path is required to eliminate the congestion. Researchers introduce rigorous studies to identify optimal paths, some resulting in less data loss and delay. Identifying multiple paths between nodes may eliminate congestion. When the first best path is congested, selecting the second best path between nodes can solve the congestion problem. With this ideology, the multipath routing algorithm is developed and tested on Fat Tree, Custom, and Tree topologies, and performance is measured using quality of service factors. Considering Throuhput, Fat Tree produced 27.15% better throughput than the tree topology and 17.57% better than the custom topology, Whereas in the case of jitter, fat tree topology reduces jitter by about 90.36%compared to the tree topology, but custom topology reduces jitter by about 12.24% compared to the fat-tree topology. In the packet delivery ratio, fat tree topology reduces packet loss by about 77.87% compared to the tree topology. Fat tree topology reduces packet loss by about 55.62% compared to the custom topology. Fat tree performs best overall, with the highest throughput, lowest packet loss, and significantly reduced jitter compared to the tree and custom topologies.MiniNet is used to perform simulations. TCP and UDP flows are calculated with the iperf tool and tested on the POX Controller.*

***Keywords**:* *SDN, Congestion, POX, Multipath, MiniNet*

Received: September 6, 2024; Received in revised form: October 28, 2024; Accepted: November 26, 2024

1. INTRODUCTION

The network's infrastructure is based on hardware, and a multitude of devices, like switches and routers, work for data forwarding and operate with rules and requirements. This is resisting convolution networks' upgrades to services [1]. Network usage has immensely increased over the past decade, and traditional networks are unable to withstand due to the unadoptable nature of their architectures [2].

To meet modern requirements, layered architecture is unsuitable, so the architecture is designed into planes in software-defined networks. Every plane is dedicated to a particular responsibility and can add programmability to SDN, which enhances the chances of opting for the SDN. The data plane is tightly packed with various data forwarding devices, and the control plane works as intelligence to network [3].

Data planes are designed to work with switches and to create topologies. POX controller is used for routing decisions, traffic monitoring, and identifying the optimal path between two nodes and computing the shortest route; multipath algorithm is implemented with the help of POX controller on Fat Tree and Custom topologies.

The main aim of separating planes is to use resources efficiently and control and secure them. Due to that, numerous controllers like Floodlight, RYU, POX, OpenDayLight, and POX have been developed. SDN has a wide range of Controllers, and selecting the optimum controller depends on the application. Gupta et al. assess and contradict various SDN controllers, and simulations are conducted with Mininet [4].

Control planes are responsible for routing decisions, using open flow protocol as an interface. Using the open-flow protocol, the controller will identify paths across switches for data packets [5]. Centralized and distributed path computations are well-known approaches with prominent results due to their dynamic resource handling. The author analyzed these techniques and performed simulations to test critical factors such as latency, throughput, and fault tolerance in various traffic patterns with varying loads. The research provided valuable insights into optimized path computation with an extensive network and low latency. Results reveal that the centralized approach is superior but needs more scalability. Whereas distributed approaches face challenges with higher latency, the author concludes that hybrid models perform best in scalability and fault tolerance, but security issues must be addressed [6].

Caria et al. proposed a model that combines distributed routing with a centralized control plane, which allows centralized decision-making. Results are noted on parameters like control overhead and scalability. The author concluded that this approach could manage large-scale features [7].

Denar et al. mentioned that handling massive data by switches will downgrade the performance, so the author suggested threading and multiprocessing, which are parallel programming methods that improve controller performance on CPU time consumption, memory usage, and execution time and it concludes that Ryu produced superior results over POX [8]. The application plane uses logical programs handled by APIs and deploys software, routing, and policies. Planes are communicated with two sets of interfaces: southern and northern [9]. SDN architecture, which includes various planes, is shown in Fig. 1.

If data transmission needs to occur between two dedicated nodes over a network, then we need to identify the optimal path between nodes. Then transmission will happen. The main drawback of selecting the same optimal path every time is that it leads to congestion, which leads to network performance degradation. We address this problem by identifying multiple paths between nodes, regardless of the cost and length of the paths. We measured network performance with Throughput, jitter, and packet loss. The controller used in the control plane influences these metrics and the overall network design. The multipath algorithm meticulously tests three distinct topologies using the POX controller within the innovative Software-Defined Networking (SDN) framework. We carry out simulations using a MiniNet simulator.

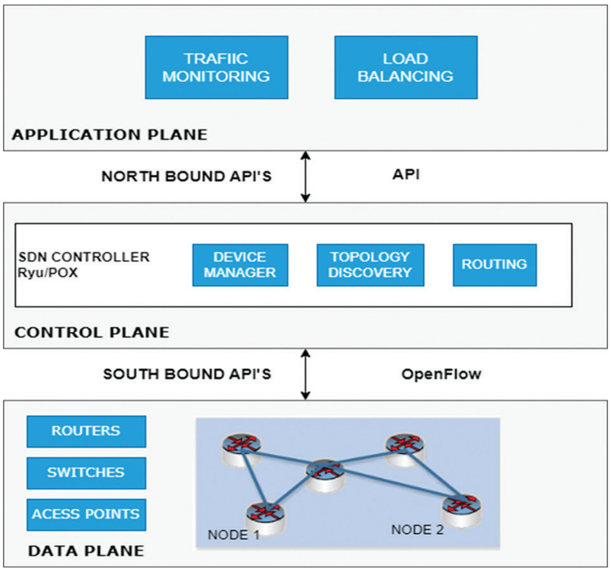


Fig. 1. SDN architecture

The Fat Tree topology performed well over custom and Tree topologies in Throughput and packet loss. On the other hand, a custom topology also demonstrated slightly improved performance, with less jitter than the Fat Tree topology.

The remaining sections explain the methodology followed by the simulations taken over topologies with the POX controller. The result analysis showcases the QOS factor over topologies through extensive benchmarking using the iPerf tool.

2. LITERATURE REVIEW

2.1. SDN CONTROLLERS

This research mainly concentrates on increasing the throughput of network with optimal path algorithm, Optimal path will allow packet loss routing from source to given destination without congestion. static routers are not suitable if there is a drastic increase of load in network also fail to maintain if nodes are added routing tables are fail to control this may be result link failure and congestion in SDN. SDN controller computes path and maintains network information. So, it is called a central part of the SDN. The network is more efficiently managed with the help of OpenFlow protocol. SDN controller forwards the packets from source to destination to establish the path we are using. Multi-path routing is tested with a POX controller, and performance is measured with QoS factors.

2.2. EXPLORING THE FUNCTIONALITY OF SDN CONTROLLER OVER DIFFERENT APPROACHES

POX is inherited from NOX Controller, written in python code and implemented in OpenFlow protocol. It is mainly used for academic research due to its ease of use and flexibility to develop network system and control applications. MiniNet is used to build POX controller code available from GitHub [10, 11]. This can run multiple programs like switch, load balance and hub.

Mohammadi et al. compared conventional and SDN based on throughput, packet loss, and delay on three typologies with Wire shark and concluded that linear topology performed better than tree topology in delay and throughput [12]. According to Salman et.al., with OpenFlow protocol, POX can directly access forwarding devices which is easy and perfect for experiments and demonstrations [13]. In this paper, we are going to test the shortest path computing algorithm on custom and Fat Tree typologies with QoS as measuring factors.

Many researchers analyzed the performance of Ryu, POX, ONOS, OpenDaylight, and NOX on various typologies and with common factors like throughput, jitter, and packet loss.

Table 1. shows a literature review

Author	Year	Topology/Approach	Controller	QoS Parameters	Simulators
Patel et al. [14]	2023	Comparative evaluation of SDN controllers	POX, Ryu, OpenDaylight	Response time, resource management	Mininet-WiFi
Smith et al. [15]	2023	Enhancements in POX SDN topologies	POX	Latency, throughput, reliability	Mininet
Koulouras et.al. [16]	2022	Abilene Network, GEANT Network	ONOS, Ryu OpenDaylight	Throughput, RTT latency, packet loss, and jitter	Mininet
Wilson et al. [17]	2022	Ryu SDN controller for scalable topologies	Ryu	Scalability, resource efficiency	Mininet
Liehuang Zhu et al. [18]	2020	IoT and VANETS	POX, Ryu,Nox,ONOS, Floodlight, and OD	Latency and throughput	CBench, PktBlaster, and OFNet
Lee et al. [19]	2020	Ryu SDN framework	Ryu	Latency, throughput, resource management	Mininet
Numan et al. [20]	2019	Single	POX	RTT, Jitter, Delay	Ping, iperf, MiniNet
Sajid et al. [21]	2018	Round Robin & Random Algorithm	POX	response time and transaction per second	MiniNet
Duque et al. [22]	2018	Custom	POX Floodlight	QoS Metrics	Mininet
Farrugia et al. [23]	2018	Geant, Butterfly,Optimized peer- Multipath	OpenDaylight	throughput, jitter	NS-3.26
Abdul-hafiz et al. [24]	2017	Abilene network/Improved Dijkstra's	Ryu	throughput and latency	MinNet,Iperf
Bholebawa et al.[25]	2016	OpenFlow-enabled network topologies	POX	Latency, throughput	Mininet
Yahya et al. [26]	2015	Abilene network/Extended Dijkstra's	Ryu	end-to-end latency, and throughput.	MinNet,Iperf
Zhang et al. [27]	2015	Tree-based topology design	POX	Throughput, delay	Mininet
Stancu et al. [28]	2015	Comparison of SDN controllers	Various	Performance metrics (response time, resource utilization)	Mininet

Table 1 shows a literature review of different articles. A few authors have worked on Dijkstra's performance evaluation on multiple controllers, as noted in the table below. Numerous authors worked on finding optimal paths and analyzing the performance of various topologies and controllers observed from the last decade; however, identifying multiple paths and testing them is presented in this work with the help of the POX controller on topologies like Custom and Fat Tree. The multipath algorithm will be tested on various controllers with further available topologies.

Ram et al. analyze SDN performance in wireless and wired networks over single, linear, and tree topologies. They measure POX and RYU controllers' metrics, such as jitter, Bitrate, and packet loss, with D-ITG, Mininet, and Mininet-WiFi simulators. The research found that wireless networks showed performance inconsistency through SDN-optimized resource management. However, this research has limited application in real-world applications, and more dynamic topologies need to be tested.

N. Ullah et al. worked on evaluating the performance of Dijkstra's algorithm on POX and Ryu controllers. They measured Jitter, throughput, packet loss, and packet delivery ratio with iperf, Wireshark, and the MiniNet simulation tool. Where RYU outpaced POX in terms of witnessing low Jitter and higher throughput, which is well suited for dynamic networks. In the case of packet loss, both have near results. Research can be done on more dynamic and hybrid topologies to know the adoptability of this approach [29].

Several distributed but logically centralized controllers are available, including POX, Ryu, Floodlight, and OpenDayLight; the author compared these with Ryu. Ryu performed well due to its scalability and modularity; when testing the MniNet emulator, the author exhibited the controller's scalability and reliability and studied network performance under heavy loads over linear topology. The study briefly touches on operational styles but suggests further research is needed to fully understand how different modes impact controller efficiency in various environments [30]. Therefore, we must test any algorithm that needs more in-depth testing with various topologies.

Koulouras et al. worked on the evaluation of various SDN controllers customized for wireless networks. To assess controllers, the authors specially used an analytic hierarchy process, and they found Ryu and ONOS to be the best among other controllers. They concluded that selecting the controller plays a crucial role in evaluating the performance of any approach. Still, studies must fully explore wireless protocols like 5G and their adaptability to massive networks and dynamic conditions. The type of topology also makes a difference in various quality of service factors [31].

This is proved by exploring optimized tree-based network topology in SDN and applying various optimization strategies like routing to address inefficiencies in traditional topologies and manage traffic in a Dynamic way. Data centers use tree topology because it performs better under higher loads. Author prospered teaching produced a 10% increase in throughput and a 15 %reduction in latency. However, the author stuck to only tree topology and did not do rigorous testing on dynamic networks [32]. Identifying the optimal path is achieved by Dijkstra's algorithm, and with the help of the POX controller, path computation and network performance are enhanced. Real-time network conditions drive dynamic decisions on routing. Algorithm efficacy is measure with QOS factors. It also addressed critical issues in traditional routing and achieved noticeable throughput and less jitter. However, it still lacks energy efficiency and is stuck to general topologies like mesh, star, and ring topologies where the structure is straightforward. Calculating the optimal path is simple [33].

Cabarkapa et al. conducted a performance analysis of the Ryu-POX controller in various tree-based Software-Defined Networking (SDN) topologies. The methodology involves simulations in a controlled environment where different tree-based topologies, such as binary and balanced trees, are evaluated [34]. The quality-of-service factors measures the effectiveness of the Ryu and POX controllers in handling the load over these topologies and performance. The quality-of-service factors measures the effectiveness of the Ryu and POX controllers in handling the load over these topologies and performance. This work proves that selecting topology depends on the traffic type and network demand. Most of the researchers' work concluded that selecting a proper controller and testing with multiple topologies will help them evaluate any approach's performance. Also, for finding an optimal path, Dijkstra is one role model, though congestion is a significant issue. So, in this paper, we have used multipath routing tested on multiple topologies on a pox controller.

3. PROPOSED MECHANISM

Congestion will occur when two dedicated nodes select the same optimal path for every transmission and must identify alternative paths to avoid network traffic. Calculating multiple paths between all available nodes in the network is necessary. The algorithm will choose an alternative path if the load floods the selected path. This proposed mechanism discusses selecting multiple paths based on bandwidth, link cost, and hop number between dedicated nodes. Multipath routing algorithm start with finding the routes with the help of the Depth First Search algorithm, whose working functionality, is to go deep from other adjacent nodes of the last visited node of a graph [35]. DFS algorithm is showed in table. With the DFS algorithm, every node is visited once to compute paths from each node to every other node. Pseudocode shows the Multipath routing algorithm. It starts with executing topology at one end; on another end, The system initializes the POX controller and executes the Multipath algorithm. After initializing the get path function, we will use DFS to acquire paths between dedicated nodes, as we expect paths between the source and destination. The routing table stores all paths, and Get-link-cost will be used to calculate path cost.

Algorithm:

```
RecurDFS($G_i$, root):
Traversed <- set all nodes false initially
DFS(root)
function DFS($u$):
if Traversed [$u_i$] = true:
return
print($u_i$)
Traversed [$u_i$] <- true
for each vi in $G[u_i]$. neighbors ():
DFS($v_i$)
Input: $G$ is graph in Adjacency list where root
        is starting node
Output: DFS order nodes in that graph are printed
```

At the start, bandwidth is initialized. B1 is the minimum bandwidth requirement between si and sj, ewi is the bandwidth capacity of the edge between si and sj, and reference bw/bl refers to the baseline value. It ensures B1, the minimum bandwidth requirement met by the bandwidth available between si and sj, compared to the reference BW value.

We identify all possible paths and calculate the cost. The shortest Path will return a less costly option. Get Path will display several available paths and sort them in order. Upon selecting the Path,add_ports_to paths functions add ports for communication between hosts, which involves various link and switch handling functions. In the data plane, topology is created with a set of nodes and switches from the control plane. The Multipath algorithm computes paths between source to destination, and performance is measured with iPerf and wire shark with MiniNet as the simulation environment. The following sections present the simulations, network setup, and results.

Pseudocode

```
1.  Define function $get\_paths(self, s_r, dj)$:
      return $paths$
      DFS algorithm to find $paths$
```

```
2.  Define function $get\_link\_cost(self, s_i, s_j)$:
     set $e_r1$ to $self.adjacency[s_i][s_j]$
     set $e_r2$ to $self.adjacency[s_i][s_j]$
     set $bl$ to $min(self.bandwidths[s_i][s_j],$
     $self.bandwidths[s_i][s_j])$
     set $ew_i$ to $reference\_bw/bl$
     return $ew_i$
3.  Define function $get\_path\_cost(self, path)$:
     return $cost$
4.  Define function $get\_optimal\_paths(self, s_r, d_j)$:
     # $s_r$ is switch, $d_j$ is switch
     set paths to $self.get\_paths$(sr, dj)
     set $paths\_count$ to $length(paths)$
     if $length(paths) < maxi\_paths$
          else $maxi\_paths$
     return $sorted(all\_set\_of\_paths)$
5.  Define function $add\_ports\_to\_paths(self, paths,$
    $first\_port, last\_port)$:
          # assigning $ports$ to $path$
6.  return $paths\_p$
```

Fig. 2 represents the flow of work. This includes Topologies applied and simulations taken on multipath routing algorithm with implementation on POX controller.

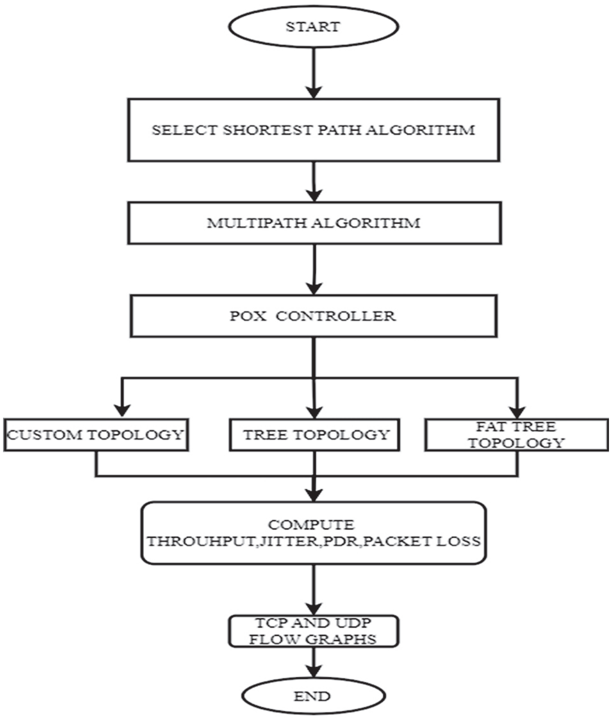


Fig. 2. Methodology

4. SIMULATION SETUP

We conduct simulations using MiniNet [36]. Stanford University develops it. It includes routers, switches, end-hosts, and SDN controllers, which are of OpenFlow and allow users to create networks with the MiniEdit tool via a graphical interface. Users can test different topologies, and it has a feature of CLI support to create and test switches and routers. Mininet supports wide-ranging controllers for flexible simulations. It gears up its usage in various sectors like education, research, and development by allowing Operating system virtualization with hundreds of nodes [37, 38].

We worked with Mininet on the 2.3.0 version of the default operating system with ubuntu-20.04.4. controller installed with POX-2.0, switch is over 2.5.4, SBI is OpenFlow 1.3 with iperf 2.0.13 also used Wireshark with 4.0.6 to analyze the network.

4.1. TOPOLOGIES

Topologies can be created either by using commands or through MiniEdit. MiniNet also supports creating topology with Python code. Here, we have used all three ways to create topologies, and this work presents a performance evaluation of the Multipath routing algorithm with Fat tree, Tree, and Custom Topologies. Python code creates a Fat Tree topology, Tree Topology with MiniEdit, and Custom topology through commands, but we have drawn all topologies in MiniEdit for better visual clarity. Fig. 3 shows the Fat Tree topology. It has three levels, where the core level is to create redundant paths, connect to aggregate switches, and ensure multiple paths between pairs of edge switches. The middle level is the aggregation level. It distributes traffic to the core level received from edge-level switches. The edge level has end host devices that directly connect to the network. In fat tree topology, each end node connects to the top of the rack switch. Fat Tree was chosen because of its identical bandwidth for bisections; each layer has the same aggregated bandwidth. Also, each port has the same speed at the end host.

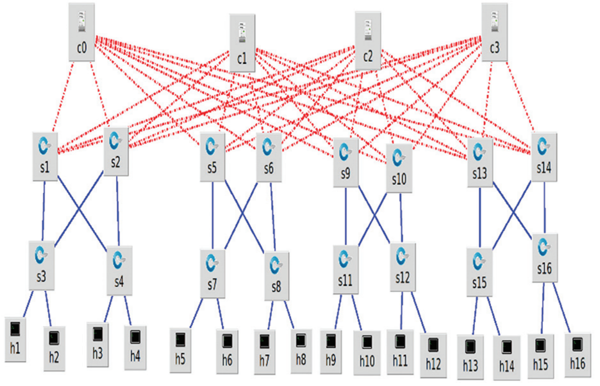


Fig. 3. Tree Topology

Observations are also taken on custom topology with superuser privileges with a remote controller located at 127.0.0.1. This means the controller can run on the same machine as MiniNet. Open vSwitch is a switch type, and the system sets the OpenFlow protocol version to 1.3. A custom tree topology creator designs it with a fan-out of 3, which means each node has three child nodes, and depth means several levels. Fig. 4 shows the custom topology. The controller creates the tree topology with nine switches, connecting them to 10 nodes in various ways.

The system generates these three topologies at the data plane while testing occurs at the control plane using a POX controller. A multipath routing algorithm calculates a path between dedicated nodes, as shown in Pseudocode.

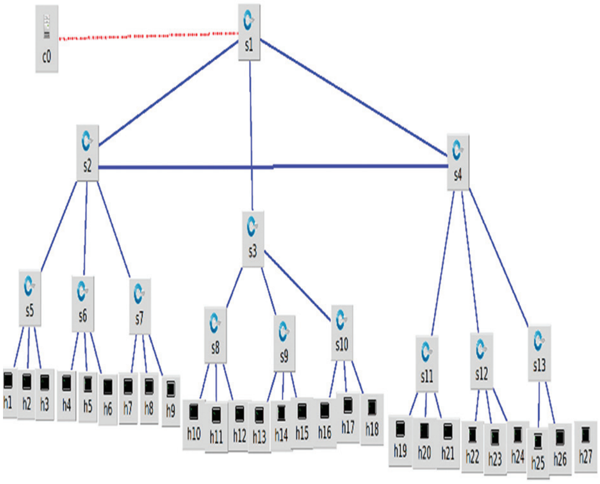


Fig. 4. Custom Topology

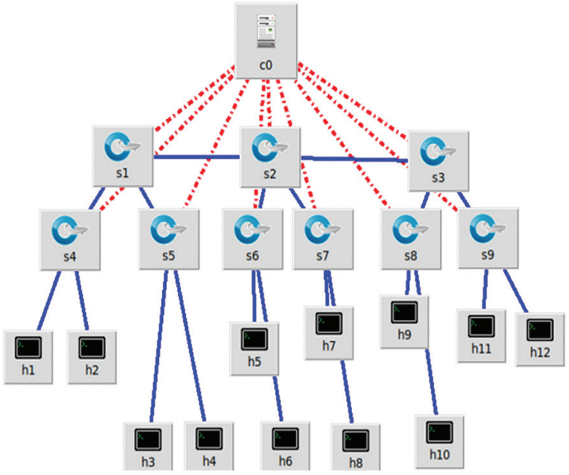


Fig. 5. Tree Topology

5. SIMULATION SETUP

Simulations are carried out with a POX controller on three topologies under three test cases. In every test case, the common factor is finding the multiple paths between node one and node two and calculating throughput, jitter, and packet loss ratio with the help of a network analyzing tool. Measured TCP and UDP flows. Test case 1 includes Testing with Fat Tree topology, Test case 2 is with custom topology, and Test case 3 is with Tree topology.

5.1. TEST CASE 1: FAT-TREE

POX implements the multipath algorithm on fat-tree topologies, and We use the iPerf benchmark utility to obtain TCP and UDP flows and observe bandwidth performance. When we execute "Fat Tree.py" in the terminal, it creates topologies according to the SDN architecture in what can be called a data plane. In the POX environment, the system opens another terminal to execute the Multipath algorithm. Links start observed between dedicated nodes after executing h1 ping h16. or we use Xtrem h1, h16, and ping from h1 to h16.with set of perf commands and note TCP and UDP flows.

Table 2. TCP and UDP Flow-Fat Tree with Multipath Routing Algorithm

C.T	TCP FLOW	UDP FLOW		
Interval (sec)	Bandwidth (Gbits/sec)	Bandwidth (Mbits)	Jitter (ms)	Lost/Total Datagram
0.0- 1.0	17.2	10.7	0.014	56/ 984
1.0- 2.0	18.7	10.6	0.084	8/ 894
2.0- 3.0	19.7	10.5	0.099	5/ 983
3.0- 4.0	14.5	10.4	0.104	9/ 898
4.0- 5.0	18.6	10.6	0.184	1/ 882
5.0- 6.0	17.3	10.5	0.211	4/ 888
6.0- 7.0	17.8	10.7	0.245	5/ 862
7.0- 8.0	17.6	10.2	0.017	31/ 898
8.0- 9.0	18.6	10.5	0.011	26/ 899
9.0-10.0	17.18	10.6	0.039	3/898
0.0-10.0	17.71	10.53	0.098	148/9085

We measure throughput from TCP and UDP flow and observe jitter and packet loss from the UDP flow. Table 2 shows the Fat Tree TCP flow and UDP flow.

5.2. TEST CASE 2: CUSTOM TOPOLOGY

Switch connected to 3 more switches. Each switch at the third level connects to three hosts (Hosts "h1" to "h8"). In the sudo command, we specified the controller as remote and IP 127.0.0.1; this specifies the controller for the MiniNet network. Fig. 4 shows the custom topology when Ping is executed between the H1 and H27 respective switches, making the path from source to destination. The multipath algorithm is applied, and TCP and UDP flow are noted in Table 3. Congestion can be avoided by carefully designing the custom topology.

Table 3. TCP and UDP Flow-Custom topology with Multipath Routing Algorithm

C.T	TCP FLOW	UDP FLOW		
Interval (sec)	Bandwidth (Gbits/sec)	Bandwidth (Mbits)	Jitter (ms)	Lost/Total Datagram
0.0- 1.0	18.98	10.7	0.106	123/ 983
1.0- 2.0	16.3	10.2	0.084	26/ 867
2.0- 3.0	17.5	10.3	0.029	44/ 895
3.0- 4.0	16.5	10.2	0.208	66/ 899
4.0- 5.0	16.6	10.2	0.134	52/ 865
5.0- 6.0	18.6	10.2	0.101	2/ 892
6.0- 7.0	18.5	10.1	0.035	2/ 899
7.0- 8.0	18.3	10.2	0.019	5/ 883
8.0- 9.0	19.4	10.2	0.099	6/ 898
9.0-10.0	19.5	10.1	0.102	2/ 899
0.0-10.0	18.01	10.24	0.086	328/8980

Table 4. TCP and UDP Flow-Tree topology with Multipath Routing Algorithm

C.T	TCP FLOW	UDP FLOW		
Interval (sec)	Bandwidth (Gbits/sec)	Bandwidth (Mbits)	Jitter (ms)	Lost/Total Datagram
0.0- 1.0	18.98	10.7	0.965	449/889
1.0- 2.0	16.3	10.2	3.208	21/ 867
2.0- 3.0	17.5	10.3	2.252	39/ 895
3.0- 4.0	16.5	10.2	0.803	75/ 899
4.0- 5.0	16.6	10.2	0.542	49/ 865
5.0- 6.0	18.6	10.2	0.412	22/ 892
6.0- 7.0	18.5	10.1	0.619	2/ 899
7.0- 8.0	18.3	10.2	1.369	12/ 883
8.0- 9.0	19.4	10.2	0.446	82/ 898
9.0-10.0	19.5	10.1	0.505	86/ 899
0.0-10.0	18.01	10.24	1.019	651/8886

5.3. TEST CASE 3: TREE TOPOLOGY

Controller, The system connects controller c0 to switches s1 through s9, facilitating node connections. The nodes communicate with each other via the switches and controllers. Fig. 5 displays the structure of the tree topology. Table 4 lists TCP and UDP flows after executing the tree topology in the data plane and running the multipath algorithm at the control plane with the POX controller. When the system executes the ping command from node 1 to node eight, it computes multiple paths, including one optimal path selected for transmission. Tree topology experiences congestion due to its architecture, and most packets queue at the root node. Tree topology experiences congestion due to its architecture, and most packets queue at the root node.

6. RESULT ANALYSIS

The simulation section notes the results of executing the multipath algorithm on Fat Tree, Custom, and Tree topologies. In this section, the quality-of-service parameters, which are throughput, jitter, and packet loss, will be calculated from the results.

6.1. THROUGHPUT

Throughput is measured in the context of the data rate at which data is successfully transmitted between source and destination or how many packets are delivered per second. This section discusses the Throughput of Fat Tree, Custom, and Tree topologies and their performances upon applying the multipath routing algorithm. The choice of topology plays a crucial role in identifying network performance.

6.1.1 Fat Tree Topology

This solution suits multipath connections between nodes best and organizations mainly use it in data centers requiring the highest throughput and crucial load balancing. Fat Tree is designed to produce high bandwidth by allowing traffic to distribute across multiple paths between dedicated nodes, reducing congestion and improving the service's overall quality. POX serves as a controller that effectively leverages multipath routing to maximize throughput. Table 4 shows the transfer rate of Fat Tree topology with ten intervals produced an average of 2.81GBytes of throughput with TCP flow. In the case of UDP, a throughput rate of 1.256 Mbytes of data is noted.

6.1.2 Custom Topology

Custom topology performance depends on the layout, how it supports multipath, and how it balances the load in various paths. Due to its adequate redundancy with path multiplicity, it achieves a high throughput, though less than Fat Tree's. If multiple paths are available between dedicated nodes, optimal performance is achieved by carefully designing configurations to eliminate bottlenecks and enhance throughput. Table 5 shows TCP flow and UDP flows noted over custom topology. TCP flow noted an average of 2.39 GBytes of throughput, whereas, in the case of UDP flow, it is 1.39 Mbytes

6.1.3 Tree Topology

Tree topology is more prone to bottlenecks and limited redundancy, so throughput is low compared to well-designed custom and fat-tree topologies

Near the root, congestion may occur. However, multipath routing will distribute traffic across available paths. However, there needs to be more path diversity, and it cannot utilize the benefits of multipath routing. There is a higher chance of congestion at the root, which limits the throughput. Table 6 displays the transfer rates of TCP and UDP flows, observing rates of 2.21 GBytes and 1.21 MBytes. Ultimately, we can conclude that the POX controller achieves better throughput with Fat Tree than with custom and tree topologies when applying multipath routing. The design of the Fat Tree itself supports multiple paths that are sometimes redundant and have high bandwidth. Suppose we design a custom topology with redundancy and multiple paths. In that case, a tree topology with minimal bottlenecks and careful path handling can also enable these topologies to produce high throughput. The TCP flow of the three topologies is compared in Fig. 6, while Fig. 7 displays the UDP flow.

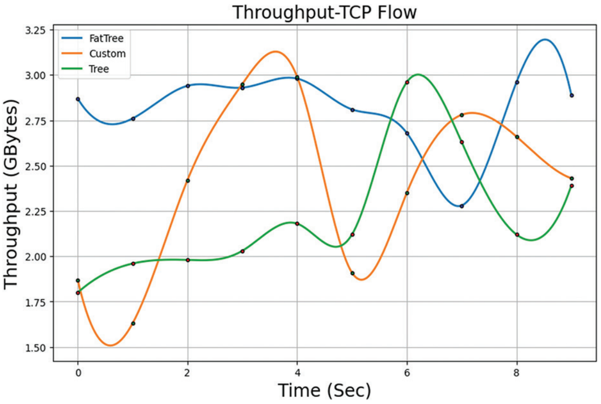


Fig 6. Throughput -TCP flow over various topologies

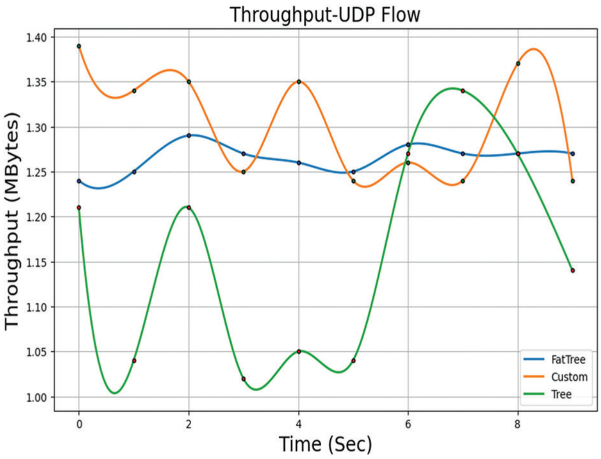


Fig 7. Throughput -UDP flow over various topologies

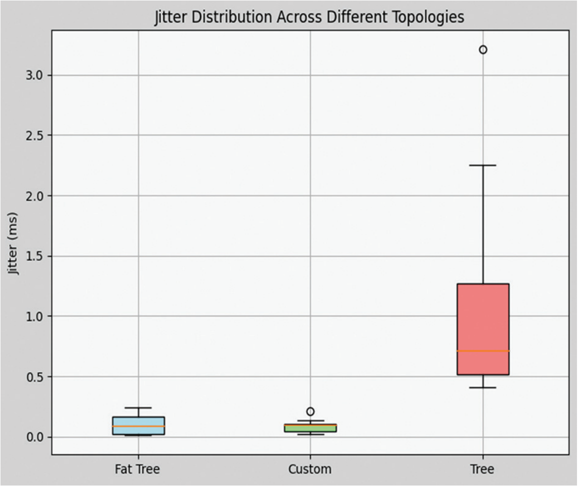


Fig 8. Jitter over various topologies

6.2. JITTER

Inconsistency of arrival time at the destination is known as jitter. Here, packets have irregular interval times. This variability we will denote as jitter. The main reasons for jitter may include congestion in the route, which may be due to load packets that may take different routes to reach the destination or a controller that introduces delays to process packets when it encounters enormous traffic. The POX controller is programmed and configured to manage the jitter. Path selection, dynamic path adjustment, and monitoring traffic influence the process. POX includes a real-time module that monitors traffic by maintaining a threshold for jitter. If it encounters massive traffic, the module reroutes the traffic to a less congested route to avoid congestion.

6.2.1 Fat-Tree Topology

This topology balances traffic by allowing it to flow on multiple paths in a balanced way, which reduces congestion on a single path. So, it produces less jitter than tree topology and custom topology. It also maintains consistent latency across paths, which minimizes jitter. The main reason is that packets are routed to the destination in a similar path length to synchronize packet delivery. POX controller avoids paths with high inconsistency. The fat tree structure maintains consistency in fluctuating traffic conditions, ensuring less jitter. Up on applying multipath routing on fat-tree, noted an average of 0.09ms of jitter.

6.2.3 Custom Topology

The design of a custom topology affects jitter. Careful topology design may yield less jitter. Path lengths and capacities produce varying jitters. Paths are balanced in terms of capacity and latency, which achieves low jitter. The POX controller performs traffic management and path selection to control traffic across paths to minimize jitter. We note an average of 0.86ms of jitter over ten intervals.

6.2.3 Tree Topology

As Tree topology suffers from limited path redundancy, the route node gets congested which leads to delays. Limited number of paths makes the queues for packets. Potential length differences can introduce inconsistency in packet delivery times, resulting in higher jitter. Managing jitter is more challenging due to the limited path range and latent congestion. The POX controller may struggle to maintain consistent path performance, leading to inconstant latency and increased jitter. Table V shows an average of 1.09 ms of jitter. Jitter is managed if multiple paths between source and destination are the same length. Among all topologies, the custom topology achieves less jitter, though the fat tree is designed well. Figure 8 shows a comparison of fat tree, custom, and tree topology.

6.3. PACKET DELIVERY

The packet delivery ratio is calculated from the number of packets lost to send. To find the number of packets delivered, subtract the lost packets from the sent packets. Measuring the network performance will be helped by a smaller number of lost packets. Traffic congestion and link failures are a few reasons for packet loss. Buffering also may lead to packet loss. To achieve less packet loss with the POX controller, the critical factor is to choose a topology with redundancy, and the controller should able to manage traffic over multiple paths

6.3.1 Fat-Tree Topology

As it has high redundancy and availability of multiple paths between any pair of nodes, it helps to distribute traffic more evenly across paths, which reduces congestion. In case of link failure, redundancy will help to minimize packet loss. Multipath routing prevents load on a single path and balances by enrooting to another path, which reduces packet loss. Path diversity in fat-tree topologies helps mitigate packet loss even under varying traffic conditions. Table 6 shows a packet delivery ratio of 10 intervals with 148 packets lost over 9085 packets, resulting in a 1.62 % packet loss ratio.

6.3.2 Custom Topology

Packet loss in custom topology depends on design with adequate redundancy, and a balanced path can achieve less packet loss. However, its limitations with redundancy suggest that multipath routing could be more effective in foreseeing packet loss. The configuration of the POX controller effectively utilizes multipath routing, although performance depends on traffic patterns and the network design. Still applying multipath routing algorithm on custom topology acquired a loss ratio of 3.65%, where 328 packets were lost during the transmission of 8980 packets over ten intervals from source to destination.

6.3.3 Tree Topology

Tree topology mostly has a problem of enormous packet loss due to a hierarchical structure where traffic is congested at the root, which increases packet loss when the network hits heavy traffic. Multipath routing helps distribute the traffic over alternative paths, significantly reducing packet loss, though lack of path diversity and redundancy limits its effectiveness. Multiple paths may exist, but they differ in factors like Capacity and not equal size.

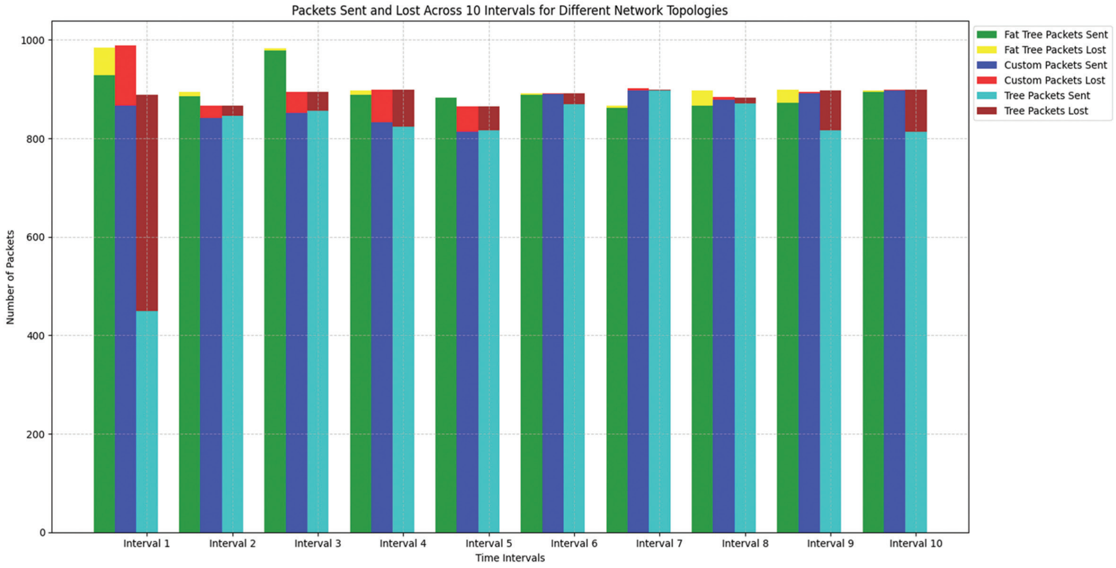


Fig 9. Packet delivery over Fat-Tree, Custom, Tree

The POX controller may need to help distribute traffic efficiently enough to avoid these losses. So, it witnessed more packet loss than the other topologies. During ten intervals, the transmission lost 651 out of 8886 packets, producing a 7.32% packet loss ratio. Figure 9 shows comparisons of fat-tree, custom, and tree topologies regarding packet loss and packets sent.Fat-tree topology showed better performance, as shown in Figure 9. The bar chart over interval 1 to interval ten shows packet loss, where each bar represents the packets delivered at the bottom and the top.at interval one, the tree topology had a packet loss of 440 packets and sent 449 packets. Later, they reduced the packet loss percentage in the tree topology. All three topologies sent packets without a considerable loss in a few intervals. In the end, fat-tree performed well when multipath routing was applied with the POX controller over topology with quality-of-service parameters like throughput, jitter, and packet loss ratio.

7. CONCLUSION

Software Defined Network (SDN) enable developers to build SDN applications due to architecture compatibility, which separates the control plane and data plane and allows centralized network management with programmability. POX controller is an open-source and Python-based controller that works with MiniNet as it supports Python-based coding to create a network. Congestion is the most significant problem; We need immediate solutions to avoid data loss. Selecting the optimal path repeatedly between dedicated nodes may lead to overload and congestion. Multipath routing has been introduced in SDN to address this problem. It identifies multiple paths between all nodes with Depth First Search and selects one path among available paths. If the selected path becomes congested, the system will reroute packets to the following path. This action can reduce congestion, and tested this algorithm on Fat-Tree, custom, and Tree topologies using the POX controller.

Throughput, jitter, and packet loss ratio are the parameters used to measure the performance of the multipath routing algorithm. Fat-Tree showed improved performance over remaining topologies due to redundancy and path diversity, which are limited in remaining topologies. If a custom topology is designed well, then this can also produce better throughput. In the case of a tree topology, the root itself is getting congested, so packet loss is more with tree topology. In the future, the Multipath routing algorithm is going to be tested with the Ryu controller.

This work helps the researcher who wants to work with optimal path identification over various topologies and load balancing over the POX controller, as well as with POX and various topologies. Fat-tree topology can be selected when the choice of performance metric is throughput and less packet loss. They can even test custom topology with proper design. Tree topology clearly shows the occurrence of congestion at the root itself.

8. REFERENCES

[1] A. Ram, S. K. Chakraborty, "Analysis of Software-Defined Networking (SDN) Performance in Wired and Wireless Networks Across Various Topologies, Including Single, Linear, and Tree Structures", Indian Journal of Information Sources and Services, Vol. 14, No. 1, 2024, pp. 39-50.

[2] Y. Zhang, M. Chen, "Performance evaluation of Software-Defined Network (SDN) controllers using Dijkstra's algorithm", Wireless Networks, Vol. 28, 2022, pp. 3787-3800.

[3] J. Ma, R. Jin, L. Dong, G. Zhu, X. Jiang, "Implementation of SDN traffic monitoring based on Ryu controller", Proceedings of the International Symposium on Computer Applications and Information Systems, 19 May 2022.

[4] N. Gupta, M. S. Maashi, S. Tanwar, S. Badotra, M. Aljebreen, S. Bharany, "A Comparative Study of Software Defined Networking Controllers Using Mininet", Electronics, Vol. 11, No. 17, 2022, p. 2715.

[5] T. H. Obaida, H. A. Salman, "A novel method to find the best path in SDN using firefly algorithm", Journal of Intelligent Systems, Vol. 31, No. 1, 2022, pp. 902-914.

[6] M. C. Saxena, M. Sabharwal, P. Bajaj, "Exploring path computation techniques in Software-Defined Networking: A review and performance evaluation of centralized, distributed, and hybrid approaches", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 11, No. 9s, 2023, pp. 553-567.

[7] M. Caria, A. Jukan, M. Hoffmann, "SDN partitioning: A centralized control plane for distributed routing protocols", IEEE Transactions on Network and Service Management, Vol. 13, No. 3, 2016, pp. 381-393.

[8] D. R. Akbi, W. Suharso, "A comparison of Ryu and Pox controllers: A parallel implementation", Journal of Intelligent Systems, Vol. 9, No. 1, 2024, pp. 1-9.

[9] E. Adedokun, A. O. Adesina, O. O. Olabiyisi, "Improved extended Dijkstra's algorithm for software defined networks", Proceedings of the International Conference on Computing, Networking and Informatics, Lagos, Nigeria, 2017, pp. 1-6.

[10] "POX controller manual current documentation", https://noxrepo.github.io/pox-doc/html/ (accessed: 2022)

[11] B. Lantz, N. Handigol, B. Heller, V. Jeyakumar, "Introduction to Mininet", Mininet Project, https://github.com/mininet/mininet/wiki/Introduction-to-Mininet (accessed: 2022)

[12] R. Mohammadi, A. Nazari, M. Nassiri, M. Conti, "An SDN-based framework for QoS routing in Internet of Underwater Things", Telecommunication Systems, Vol. 78, No. 2, 2021, pp. 253-266.

[13] M. I. Salman, "A hybrid SDN-multipath transmission for a reliable video surveillance system", Association of Arab Universities Journal of Engineering Sciences, Vol. 29, No. 2, 2022, pp. 46-54.

[14] R. Patel, N. Gupta, "Comparative evaluation of SDN controllers: POX, Ryu, and OpenDaylight", IEEE Access, Vol. 11, 2023, pp. 29015-29028.

[15] J. Smith, L. Zhang, "Enhancements in POX SDN topologies for improved network management", Journal of Network and Computer Applications, Vol. 75, No. 1, 2023, pp. 45-56.

[16] I. Koulouras, S. V. Margariti, I. Bobotsaris, E. Stergiou, C. Stylios, "Assessment of SDN controllers in wireless environments using a multi-criteria technique", Information, Vol. 14, No. 9, 2023.

[17] A. Wilson, C. Lee, "Advancements in Ryu SDN controller for scalable network topologies", IEEE Transactions on Network and Service Management, Vol. 19, No. 4, 2022, pp. 789-800.

[18] L. Zhu, M. M. Karim, K. Sharif, C. Xu, F. Li, X. Du, M. Guizani, "SDN controllers: A comprehensive analysis and performance evaluation study", ACM Computing Surveys, Vol. 53, No. 6, 2020.

[19] S. Lee, K. Park, "Ryu SDN framework: Design and performance", IEEE Transactions on Network and Service Management, Vol.17, No. 2, 2020, pp. 123-135.

[20] P. E. Numan, K. M. Yusof, M. N. B. Marsono, S. K. S. Yusof, M. H. B. M. Fauzi, S. Nathaniel, M. A. B. Baharudin, "On the latency and jitter evaluation of software defined networks", Bulletin of Electrical Engineering and Informatics, Vol. 8, No. 4, 2019, pp. 1507-1516.

[21] A. S. Sajid, S. F. N. Niloy, K. Hossain, T. Rahman, "Comprehensive evaluation of shortest path algorithms and highest bottleneck bandwidth algorithm in software-defined networks", Report, Department of Computer Science and Engineering, BRAC University, Bangladesh, 2018.

[22] J. P. Duque, D. D. Beltrán, G. P. Leguizamón, "OpenDaylight vs. Floodlight: Comparative analysis of a load balancing algorithm for software defined networking", International Journal of Communication Networks and Information Security, Vol. 10, 2018, pp. 348-357.

[23] N. Farrugia, V. Buttigieg, J. Briffa, "A globally optimized multipath routing algorithm using SDN", in Proceedings of the IEEE International Conference on Innovation Networking and Services, Paris, France, 19-22 February 2018, pp. 1-8.

[24] A. Abdul-hafiz, E. A. Adedokun, S. Man-Yahya, "Improved extended Dijkstra's algorithm for software defined networks", International Journal of Applied Information Systems, Vol. 12, No. 8, 2017, pp. 22-26.

[25] I. Z. Bholebawa, U. D. Dalal, "Design and performance analysis of OpenFlow-enabled network topologies using Mininet", International Journal of Computer Communication Engineering, Vol. 5, No. 6, 2016, pp. 419-429.

[26] W. Yahya, A. Basuki, J. R. Jiang, "The extended Dijkstra's-based load balancing for OpenFlow network", International Journal of Electrical and Computer Engineering, Vol. 5, No. 2, 2015, pp. 289-296.

[27] X. Zhang, Y. Lu, Q. Wu, "Tree-based topology design in software-defined networks", Proceedings of the IEEE Global Communications Conference, 2015, pp.1-6.

[28] A. L. Stancu, S. Halunga, A. Vulpe, G. Suciu, O. Fratu, E. C. Popovici, "A comparison between several software defined networking controllers", Proceedings of the 12th International Conference on Telecommunication in Modern Satellite, Cable and Broadcasting Services, Nis, Serbia, 14-17 October 2015, pp. 223-226.

[29] N. Naimullah, M. Imad, M. Hassan, M. Afzal, S. Khan, A. Khan, "POX and RYU controller performance analysis on Software Defined Network", EAI Endorsed Transactions on Internet of Things, Vol. 9, 2023.

[30] M. N. A. Sheikh, I.-S. Hwang, M. S. Raza, M. S. Ab-Rahman, "A qualitative and comparative performance assessment of logically centralized SDN controllers via Mininet emulator", Computers, Vol. 13, No. 4, 2024, p. 85.

[31] I. Koulouras, I. Bobotsaris, S. V. Margariti, E. Stergiou, C. Stylios, "Assessment of SDN Controllers in Wireless Environment Using a Multi-Criteria Technique", Information, Vol. 14, 2023, p. 476.

[32] Y. Wang, L. Li, Y. Zhao, "Optimizing tree-based topologies in SDN: Techniques and applications", Computer Communications, Vol. 217, 2023, pp. 10-22.

[33] Y. Chen, J. Lee, "Integrating Dijkstra's algorithm with the POX SDN controller for network optimization and path computation", Computer Networks, Vol. 211, 2023, p. 108160.

[34] D. Cabarkapa, D. Rancic, "Performance analysis of Ryu-POX controller in different tree-based SDN topologies", Advances in Electrical and Computer Engineering, Vol. 21, No. 3, 2021, pp. 47-56.

[35] S. Sryheni, "Introduction to depth first search algorithm (DFS)", Baeldung on Computer Science, www.baeldung.com (accessed: 2023)

[36] "Mininet commands", available online: http://mininet.org/ (accessed: 2022)

[37] K. K. Sharma, M. Sood, "Mininet as a container-based emulator for software defined networks", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, 2014, pp. 681-685.

[38] K. K. Sharma, M. Sood, "Mininet as a container-based emulator for software defined networks", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, 2014, pp. 681-685.

Enhanced Patch-wise Maximal Intensity Prior for Deblurring Neutron Radiographic Images

Original Scientific Paper

K. Yazid
Malaysian Nuclear Agency, Technical Support Division,
43000 Bangi, Selangor, Malaysia
khairiahyazid@gmail.com

H. Ibrahim
Universiti Sains Malaysia, School of Electrical and Electronic Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
haidi@usm.my

Mohd Zaid Abdullah*
Universiti Sains Malaysia, School of Electrical and Electronic Engineering Campus,
14300 Nibong Tebal, Penang, Malaysia
mza@usm.my

*Corresponding author

Abstract *– In neutron radiographic imaging, generally, the collimation ratio is assumed to be sufficiently large to ensure a valid approximation for parallel beam geometry. However, this assumption is difficult to apply in small nuclear reactors due to the low-intensity neutron flux. For this reason, these reactors produced inherently blurry neutron images. In this paper a blind deconvolution technique is investigated for the enhanced visual quality of neutron images through the reduction of blurring artefacts. Technically, this approach is extremely challenging because it requires an unknown point spread function. To solve this problem, scholars employ the gradient minimization strategy under the framework of a maximum a posterior, which leads to the development of an improved deblurring method, referred to in this paper as the enhanced patch-wise intensity prior. Experimental results demonstrate that the high competitiveness of the proposed method in terms of blind or no-reference evaluation measure, with an average of 46.1 for six neutron images used in this study. This value is considerably lower compared with those of existing deblurring techniques, which implies a more accurate restoration. Additionally, the proposed method resulted in the highest, and hence, the best entropy and contrast values, averaging at 7.09 and 1.05 respectively. The proposed method is also the second fastest technique witd mean time of 180 s.*

***Keywords**: Image Restoration, Point Spread Function, Blind Deconvolution, Neutron Radiography, Collimation Ratio*

Received: July 17, 2024; Received in revised form: October 7, 2024; Accepted: October 4, 2024

1. INTRODUCTION

Neutron radiography (NR) uses neutron radiation to probe the internal structures of objects, and it shows similarity to X-ray radiography. However, different from X-ray, neutrons are easily attenuated by light elements, such as hydrogen and boron, but can easily penetrate numerous heavy metals. These unique properties render NR a highly useful technique for nondestructive testing and quality control inspection. Despite being powerful and unique, the NR images produced by low-power nuclear reactors exhibit inherent degradation due to blurring. Examination of the collimation system, which is one of the core elements in NR image capture, can be used to explain the main source of blurring (Fig. 1). Referring to this figure the collimator system, which comprises an aperture and detector, directs a neutron beam to an object. Similar to a pinhole camera, the aperture prevents neutrons from entering the beam except through the hole, which concentrates neutrons within a small area, and hence reduces image distortion and chromatic aberration. Meanwhile, the detector converts neutrons into a two-dimensional (2D) image that depicts the internal structure of an object. This geometry determines the collimation ratio, which is an important characteristic of NR. In this case, the collimation ratio refers to the ratio of collimation length L to the effective diameter of the aperture D or L/D.

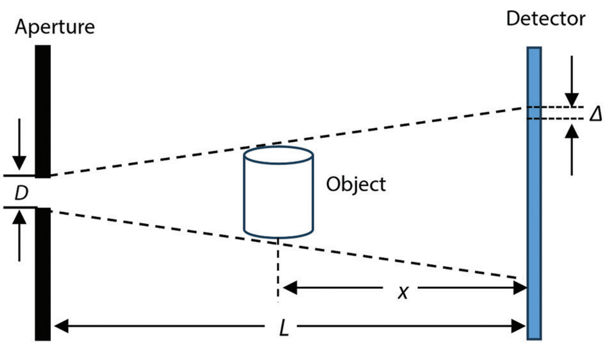


Fig. 1. Typical collimation system used for NR image capture

Following the pinhole camera analogy, geometric blurring can be defined as $\Delta= x / ((L/D))$, where Δ directly determines the resolution or sharpness of NR images. An optimal resolution is obtained when $\Delta \rightarrow 0$. This requirement can be attained either through placement of the object close to the detector ($x \rightarrow 0$) or with the use of a larger collimation ratio ($L/D \rightarrow \infty$). However, the contrast exhibits a rapid drop with the decrease in x due to the reduced intensity of neutron flux through an inverse-square relationship. For this reason, x is maintained at a reasonable distance, whereas L/D is maintained as high as possible to produce NR images with adequate contrast and acceptable resolution. Typically, the L/D of high-power neutron reactors ranges between 125 to 500 [1]. By contrast, the L/D of small reactors are generally considerably lower than this range. At Malaysia Nuclear Agency (MNA), the L/D of NR facilities approximates 105, which is moderately low compared with those of high-power or large-scale reactors. Using an arctan inverse relationship (i.e. ($\tan^{-1}(1/(L/D))^{-1}$) this ratio is equivalent to a beam divergence of approximately1.82^0. The low L / D or high beam divergence serves as the primary cause of the blurring for images captured at this facility. Hence, image deblurring or restoration constitutes one of the important tasks in post image processing activities in this reactor. For linear, shift-invariant systems, image restoration can be modeled as a convolution operation. Mathematically:

$$B=k*L+n \tag{1}$$

where B represents the blurred image, L refers to an unknown latent or sharp image, k denotes the PSF, "*" represents the 2D convolution operator and n corresponds to additive noise. In most deblurring applications, n can be ignored because it is small and uncorrelated. Given that k is generally unknown, the image restoration methods transforms into a blind-deconvolution problem. Among all available solutions to this problem, the maximum a posterior (MAP) is the most popular and widely used technique. An earlier work in this field is a paper published elsewhere [2]. Their algorithm is effective when dealing with small-sized images and hence less complex PSFs. Multilayer iterative estimation techniques are usually deployed for large images with relatively complex PSFs. Importantly, this algorithm exhibits sensitivity to local minima, which led to inaccurate estimation of PSF and in turns affected the deblurring results. Hence, regularization is performed to increase the probability of finding good local minima. In general, this step is introduced into an optimization problem to prevent overfitting and reduce complexity [3, 4]. They developed an L_0-based image smoothing algorithm by retaining large structures and removing minute details. They used L_0 and L_2 norms for image gradient prior and kernel prior, respectively. However, such an algorithm is time consuming because the solutions require solving a complicated joint optimization problem. Moreover, it requires sophisticated priors and thus considerably more complex optimizers. Therefore, the superior performance of this method is compromised by a high computational cost. These problems were addressed, which resulted in development of an improved technique [5]. These authors assumed that not all edges in the latent image are significant and useful. This assumption allowed them to enforce L_0 regularization to constrain the sparsity of image prior and use L_2 to regularize the kernel prior. Such regularization strategies not only improve the quality of image deblurring but also reduce the runtime. A new channel prior called the enhanced local maximum intensity has also been investiogated [6]. Though effective, however, this algorithm has only been tested in the restoration of text documents with a uniform background. Compared with text images, NR images exhibit a more complex intensity distributions because they contain many brightly illuminated pixels due to the strong penetration of neutrons. Therefore, the direct application of intensity priors is less satisfactory for NR images. In another work. In another research a sparse prior based on a collection of local minimal pixels in non-overlapping patches has been proposed [7]. Referred to as patchwise minimal pixels (PMPs), this method involves the calculation of the low intensity of dark pixels in non-overlapping patches. Despite the remarkable performance of this method, especially when an image contains many dark pixels, its effectivity decreases when dealing with large-sized images or complex PSFs. Solving this method led to the development of a new sparse channel prior that considers the relationship between dark and bright channel priors [8]. Even though the authors reported improved performance, however, the method is very time consuming as it requires more than 115 s to process a small image with a size of 256×256 pixels. Recently, the deep learning approach for restoring neutron image has also been reported [9,10]. However, due to the unavailability of standard neutron image dataset, the authors have resorted to using X-ray images as substitutions for training and testing. Though the results are quite promising, however, it's difficult to evaluate the actual performance of the algorithm because of different type of images used in the investigation.

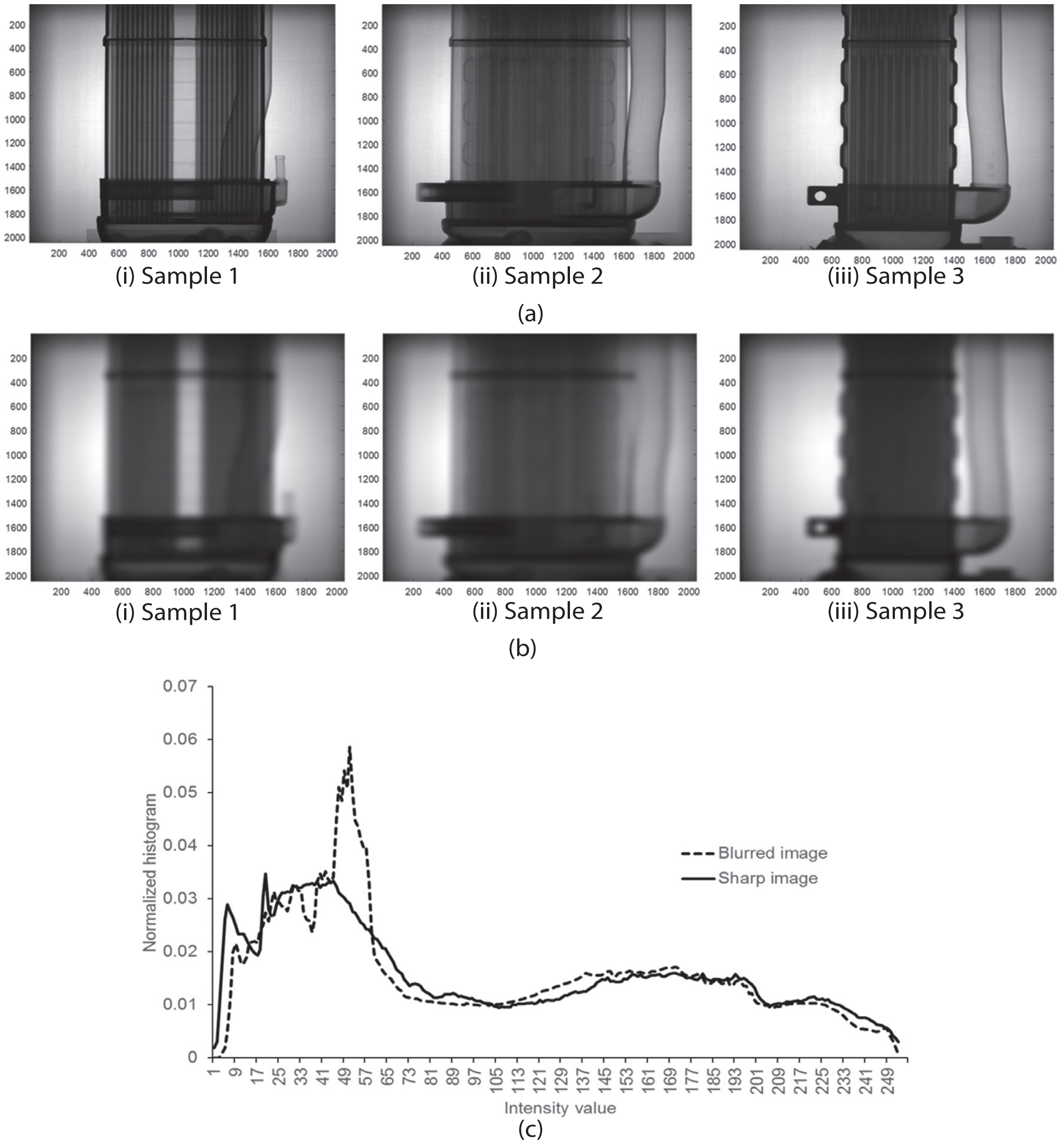
Following the above discussion, this paper proposes an alternative strategy based on high-intensity bright pixels. It also addresses large and complex PSFs as evident from the ensuing discussion. The basic idea behind this idea is first published at a conference meeting held recently [11]. Following this publication, this paper presents detailed information of the proposed method including the use of a much more efficient and accurate PSF estimator. Compared to the trial-and-error method as in our previous publication, here, a more systematic approach in determining important parameters of the algorithm is devised and discussed. Also the performance of the proposed solution is evaluated critically by comparing results with the state-of-the-art methods. Our contributions are as follows:

1. A new effective and simple image prior that uses bright pixels in nonoverlapping patches. This new image prior is referred to as the enhanced patchwise intensity (EPI).
2. A new cost-effective penalty function to enforce sparsity of the prior. During induced sparsity, the regularization term facilitates recovery of sharp images, which are in turn used in PSF estimation. The method helps in accelerating the runtime because only the nonzero elements are used in the computation.

2. MATERIALS AND METHODS

2.1. INTENSITY DISTRIBUTIONS

In consideration of the above discussion, two important assumptions are accounted for in the method proposed in this paper. First, most NR images contain a high number of brightly illuminated pixels, which is principally due to intense radiation, with neutrons easily penetrating most objects except hydrogenated materials. Second, brightly illuminated pixels show a drop in intensity due to blurring. Fig. 2 illustrates these assumptions using three different NR images.


(i) Sample 1 (ii) Sample 2 (iii) Sample 3

(a)

(i) Sample 1 (ii) Sample 2 (iii) Sample 3

(b)

(c)

Fig. 2. The effect of blurring on image brightness and intensity distributions.
(a) Sharp images, **(b)** Blured images, **(c)** Histogram

The images in Fig. 2(a) appear sharp and clear because they are captured using a high-power nuclear reactor. In this case, Figs. 2(a)(i-iii) include the original sharp images captured from three different projections, and Figs. 2(b)(i-iii) display their corresponding blurry counterparts. Blurring is simulated using a simple low-pass filter. Comparison of Figs. 2(a) and 2(b) shows the reduction in brightness after blurring, which implies the considerable drop in intensity. To further prove this observation, we examined the characteristics of sharp and blurred images on their histograms. In so doing the probability density function (PDF) for sharp and blurry images are first calculated, second normalized, and then averaged. Fig. 2(c) displays the results. As shown in the figure, the histogram of blurred image shift to the left, which results in more pixels occupying low gray-scale values and implies the reduced image brightness due to blurring.

2.2. ENHANCED PATCH-WISE INTENSITY PRIOR

On the basis of the above assumptions, the proposed prior employs high-intensity bright channel pixels in non-overlapping patch. The proposed method can be explained by referring to a sharp image L of size $m \times n$ and partitioned into d non-overlapping patches; with each size $r \times r$, the patch size r can be varied by ratio formula: $r=SF \times((m + n)/2)$ where SF is a scaling factor. Meanwhile $d=\lceil m / r\rceil \times \lceil n / r\rceil$, and $\lceil\cdot\rceil$ denotes the ceil operator. For gray scale image the EPI prior can be defined as follows:

$$EPI(L)(i) = \max_{(x,y)\in\Omega_i} L(x,y) \tag{2}$$

where (x, y) denotes the pixel coordinates, and Ω_i denotes the i-th non-overlapping patch with i=1,2,…,d. Therefore, EPI(L)(i) represents the collection of high-intensity or bright pixels of i-th non-overlapping patch. Similarly, the EPI prior of blurred image B can be expressed as follows:

$$EPI(B)(i) = \max_{(x,y)\in\Omega_i} B(x,y) \tag{3}$$

As discussed previously, the brightness of an image drops as a result of blurring. Hence, the EPI prior of the blurred image is much less than that of a sharp latent image. Mathematically, the following inequality holds:

$$EPI\ (B) \leq EPI\ (L) \tag{4}$$

Substituting Equation (1) into Equation (4) gives

$$\max_{(x,y)\in\Omega_i} B(x,y) = \max_{(x,y)\in\Omega_i} L(x,y) * k \tag{5}$$

where k>0 and $\sum k$=1. Following Equations (4) and (5), the maximum intensity value of a blurred image is also significantly less than that of a sharp image. With the assumption that the patch size for B and L is the same, the following inequality is also valid:

$$\max_{(x,y)\in\Omega_i} B(x,y) \leq \max_{(x,y)\in\Omega_i} L(x,y) \tag{6}$$

As explained, this paper exploits high-intensity pixels to distinguish sharp from blurry images. With this assumption, the deblurring model is developed and presented in the following subsections.

2.3. DEBLURRING MODEL

With the use of Equation (1), the deblurringit's model based on the standard MAP estimation framework is developed as follows:

$$\min_{L,k} \|L * k - B\|_2^2 + \gamma P(L) + \mu P(k) \tag{7}$$

The first term $\|L * k\text{-}B\|^2_2$ is a data fidelity term that constrains the convolution of L and k so that the result is consistent with B. The regularized terms $P(L)$ and $P(k)$ are priors related to latent image and PSF kernel, respectively. μ and γ are positive regularizing parameters that balance the weight relation between the fidelity and priors. The deblurring problem is non-convex; therefore, regularization helps constrain the priors to increase the probability of producing a good local solution. As explained previously, the gradient of a natural image is sparse. Consequently, $P(L)$ is regularized such that

$$P(L)= \|L\|_0 \tag{8}$$

where $\|.\|_0$ indicates the zero norm. Meanwhile, the kernel prior is formulated as

$$P(k) = \|k\|_2^2 \tag{9}$$

where $\|.\|_2$ denotes the second norm. The deblurring model can be formulated by combining Equations (8) and (9). Mathematically,

$$\min_{L,k} \|L * k - B\|_2^2 + \mu\|\nabla L\|_0 + \gamma\|k\|_2^2 \tag{10}$$

where $\nabla=(\nabla_h, \nabla_v)$ denotes the image gradients calculated in the horizontal and vertical directions. In this case the $\|.\|_2$ is also used to constraint the data term because this norm is known to be optimal for Gaussian noise. Moreover it enables the solution to be calculated using a standard fast Fourier transform (FFT) algorithm. Introducing the EPI, Equation (10) can be rewritten as:

$$\min_{L,k} \|L * k - B\|_2^2 + \gamma\|k\|_2^2 + \mu\|\nabla L\|_0 + \alpha\|EPI(L)\|_0 \tag{11}$$

where α, μ, and γ are positive weight parameters. Traditionally, an iterative-based Half Quadratic Splitting (HQS) algorithm is used in solving Equation (10), such as in [5] and [10]. However, this algorithm is complex and time consuming. Thus, an alternative strategy is employed in this study. Exploiting the sparsity of the EPI in non-overlapping patch, Equation (9) is solved directly via soft thresholding. The following condition is introduced to constrain the solutions in such a way that

$$\min_{L,k} \|L * k - B\|_2^2 + \mu\|\nabla L\|_0 + \gamma\|k\|_2^2 \tag{12}$$

subject to $EPI(L)(i) \sim p(x)$, for $i \in \{1,\dots,d\}$

where $p(x)$ is a PDF. The thresholding of the minimum and maximum pixels of non-overlapping patches with a constraint value of 0.9 produces distributions whose shape is approximately hyper-Laplacian. This type of output together with the sparsity of high-intensity pixels enhances the distinguishability between sharp from blurred images. The next step is applying the alternating optimization rule to Equation (12), which splits the cost function into two subproblems. The first subproblem characterizes L using the following cost function:

$$L = \min_{L} \|L * k - B\|_2^2 + \mu \|\nabla L\|_0 \tag{13}$$

subject to $EPI(L)(i) \sim p(x)$, for $i \in \{1,\ldots,d\}$

The second subproblem describes the unknown k as follows:

$$k = \min_{k} \|\nabla L * k - \nabla B\|_2^2 + \gamma \|k\|_2^2 \tag{14}$$

Equations (13) and (14) are principally non-convex. Hence, the ideal solution may not exist. Therefore, these equations are solved through the minimization technique to produce approximate solutions. In so doing, similar approaches published by [5], [7], [12] and [13] are adopted. Interested readers are referred to these publications for further details. A summary of the minimization procedures is presented for the sake of completeness and thoroughness of discussion. The first step in the minimization process is to estimate k. The Gaussian function is used as a first estimate of k because the blurring is essentially low-pass filtering. Then, Equations (13) and (14) are solved alternately using methods and procedures discussed in the following subsections.

2.4. *L* SUB-PROBLEM

In solving Equation (13), a constraint is imposed to induce the sparsity on EPI(L), indirectly speeding up the minimization process. Given a previous estimation of k^i, the latent image is updated via iterative thresholding. Mathematically,

$$L^{i+1} = \min_{L} \left\|L * k^i - B\right\|_2^2 + \mu \|\nabla L\|_0 \tag{15}$$

subject to $EPI(L)(i) \sim p(x)$, for $i \in \{1,\ldots,d\}$

Equation (14) comprises two important regularizers: L_2 and L_0. The data fidelity term is smooth and convex, whereas the gradient term is non-convex. With the use of an auxiliary variable G with respect to the image gradient ∇L, Equation (15) is reformulated to yield

$$\min_{L,G} \left\|L * k^i - B\right\|_2^2 + \beta \|\nabla L - G\|_2^2 + \mu \|G\|_0 \tag{16}$$

subject to $EPI\,(L)(i) \sim p(x)$, for $i \in \{1,\ldots,d\}$

where β is a positive and sufficiently large penalty parameter to enforce $\|\nabla L\text{-}G\|^2 \approx 0$,and $\nabla L \approx G$. As a result of additional constraints, L and G cannot be solved directly using popular algorithm such as the block coordinate descent. Similar to [7], an alternative soft thresholding technique is applied to solve Equation (16) iteratively. The EPI subset of $L^{t,j}$ is denoted as $L_s^{t,j} := EPI(j)$ for t^{th} latent image at j^{th} iterative step, then the subsequent latent image is calculated by direct thresholding as follows:

$$\tilde{L}_s^{t+1,j}(i) = \begin{cases} 0, \left|L_s^{t+1,j}(i)\right| < \lambda \\ L_s^{t+1,j}(i), else \end{cases} \tag{17}$$

for $i \in \{1,\ldots d\}$

where λ is the thresholding value, which is greater than zero. With $\Omega^{t+1,j}$ denoted as the index set of EPI, the binary mask corresponding to EPI subset is calculated as follows:

$$M^{t+1,j}(i,j) = \begin{cases} 1 \;\; if\,(i,j) \in \Omega^{t+1,j} \\ \;\; 0 \;\;, else \end{cases} \tag{18}$$

where $M \in R^{m \times n}$ is the binary mask corresponding to the EPI subset of L. Here EPI(L): $R^{m \times n} \rightarrow R^d$; thus, the inverse of EPI(L) is equivalent to EPI^T (L): $R^d \rightarrow R^{m \times n}$ for any $z \in R^d$. Consequently, L can be presented by:

$$L_s := EPI^T\big(EPI(L)\big) = L \circ M \tag{19}$$

where $\circ$ is the dot product. In this case $M(i, j)=1$ is the maximal pixel in the non-overlapping patch. For other pixels, $M(i, j)=0$. With the use of the results of Equation (19), the intermediate latent image at j^{th} iterative step is updated as follows:

$$\tilde{L}^{t+1,j} \;=\; L^{t+1,j} \circ (1 - M^{t+1,j}) + EPI^T\big(\tilde{L}_s^{t+1,j}\big) \tag{20}$$

Substituting Equation (20) into Equation (16), the gradient subproblem of G is reformulated as follows:

$$G^{t+1,j+1} = \min_{G} \beta \left\|\nabla \tilde{L}^{t+1,j} - G\right\|_2^2 + \mu \|G\|_0 \tag{21}$$

where $\nabla=(\nabla_h, \nabla_v)$ and $G=(G_h, G_v)$. G_h and G_v are the image gradient in the horizontal and vertical directions, respectively. Following [5], Equation (21) is solved using proximal minimization. Mathematically,

$$G^{t+1,j+1} = \begin{cases} 0, (\nabla \tilde{L}^{t+1,j}(i,j))^2 < \mu/\beta \\ \nabla \tilde{L}^{t+1,j}(i,j), else \end{cases} \tag{22}$$

Using results calculated from Equation (22), the final update formula for L is defined as follows:

$$L^{t+1,j+1} = \min_{L} \left\|k^i * L - B\right\|_2^2 + \left\|\nabla L - G^{t+1,j+1}\right\|_2^2 \tag{23}$$

The closed-form solution of Equation (23) can be obtained using the FFT algorithm. Mathematically,

$$L^{t+1,j+1} = \mathcal{F}^{-1}\left(\frac{\overline{\mathcal{F}(k^i)}\mathcal{F}(B) + \beta\big(\overline{\mathcal{F}(\nabla_h)}\mathcal{F}\big(G_h^{t+1,j+1}\big) + \overline{\mathcal{F}(\nabla_v)}\mathcal{F}\big(G_v^{t+1,j+1}\big)}{\overline{\mathcal{F}(k^i)}\mathcal{F}(k^i) + \beta\left(\overline{\mathcal{F}(\nabla_h)}\mathcal{F}(\nabla_h) + \overline{\mathcal{F}(\nabla_v)}\mathcal{F}(\nabla_v)\right)}\right) \tag{24}$$

where $F(.)$ and $F^{-1}(.)$ denote FFT and inverse FFT, respectively. $\overline{F(\cdot)}$ is the complex conjugate operator. ∇_v, ∇_h

represent vertical and horizontal differential operators, respectively.

2.5. K SUB-PROBLEM

The k subproblem is solved in the gradient space, similar to the approach of [5], [10] and [11]. The update formula for k is given by

$$k^{i+1} = \min_{k} \left\|\nabla L^{i} * k - \nabla B\right\|_{2}^{2} + \gamma\|k\|_{2}^{2} \tag{25}$$

As before, the solution to Equation (24) is obtained through the FFT algorithm. The result is as follows:

$$k^{i+1} = \mathcal{F}^{-1}\left(\frac{\overline{\mathcal{F}(\nabla_h L^i)}\mathcal{F}(\nabla_h B) + \overline{\mathcal{F}(\nabla_v L^i)}\mathcal{F}(\nabla_v B)}{\overline{\mathcal{F}(\nabla_h L^i)}\mathcal{F}(\nabla_h L^i) + \overline{\mathcal{F}(\nabla_v L^i)}\mathcal{F}(\nabla_v L^i) + \gamma}\right) \tag{26}$$

The above optimization procedures are implemented using coarse-to-fine multi-scale structure. In this way, k is always non-negative, thus fulfilling the constraint requirement. The main steps involved in solving L and k subproblems are summarized in Appendix I.

Respectively the algorithm is named as Algorithm 1 and Algorithm 2. Referring to Algorithm 1, the method performs two-layer loop iterative calculations. Theoretically, β must be large for the algorithm to work. However, a large β means that regularization is a time-consuming process. One solution to this problem is to first use smaller β and then iteratively update this figure until it reaches a stable target value. For this method to work, a >1. In this research, the following parameters are chosen for Algorithm 1: a=2 , β_0=2μ, β_{max}= 10^5, μ=4 x 10^{-3}, J=3. The threshold value λ is initially set to 0.1. This value is reduced gradually until it equals a mean value of EPI. Here β_0 determines the starting strength of the regularization. Meanwhile, β_{max} is set to an upper limit to ensure a controlled regularization.

Together, they prevent excessive deblurring while preserving important details and avoiding over-smoothing.

2.6. IMPLEMENTATION

The alternate minimization described in Equations (24) and (26) iterates between latent image and kernel estimation. The blurred image and Gaussian function are used as first-guess solutions for L and k, respectively. Large kernels are estimated using the multilayer pyramid scheme combined with iterative minimization strategy. The scheme prevents optimization from being isolated in a local minimum. The strategy works gradually from the coarsest layer to the finest layer. Fig. 3 summarizes the working principle of the proposed method.

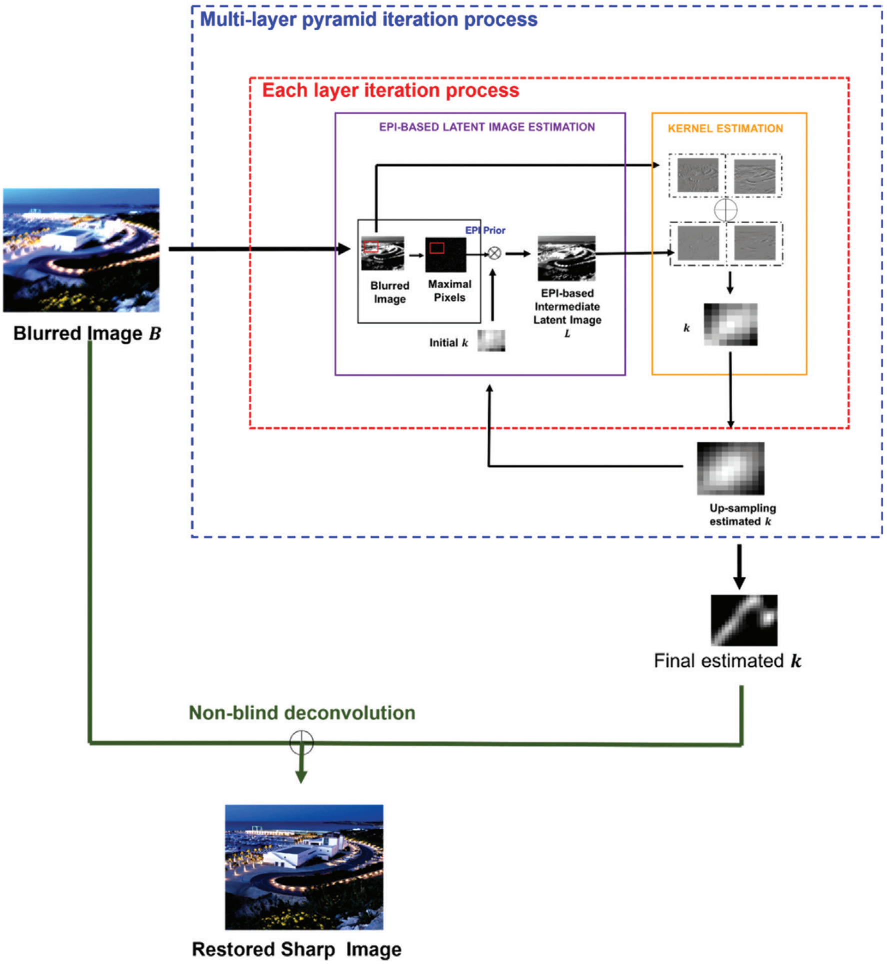


Fig. 3. Overview of the proposed EPI method

The figure shows the multilayer PSF processing unit represented by several elements inside the blue dotted square. The red-dotted rectangle refers to a single-layer pyramid PSF processing unit. In the unit, the square block on the left is the latent image calculation engine, and that on the right is the PSF estimator unit. Once k and L have been estimated in the low-resolution layer, k is upscaled for the next layer. L and k are refined iteratively in each layer. In this study, the maximum number of iterations is rigidly fixed to five. This value is selected heuristically because it produces the best solutions for all images used in the present work. During operation, the input image is first transformed into a grayscale one and then downscaled a few times. Here, the extent of downscaling depends on the size of the input PSF, which also determines the number of pyramid layers. The PSF size is used in the retrieval of a part of the kernel alone, which prevents noise accumulation in subsequent estimation processes.

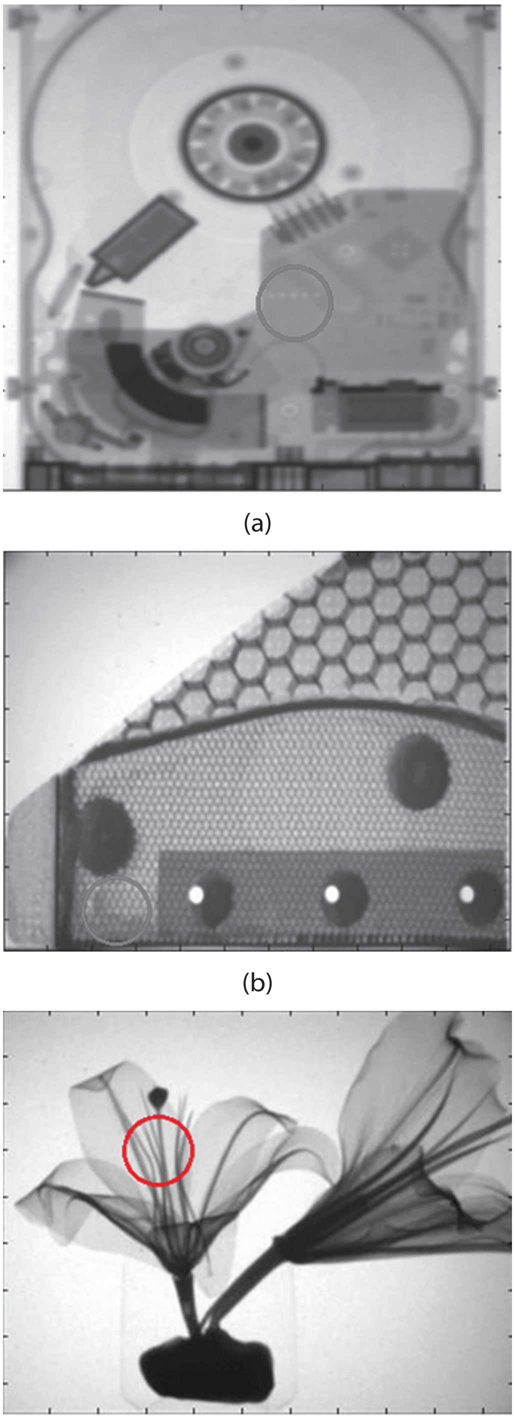
Referring again to Fig. 3, in the minimization process, the blurred image is deconvoluted using the k estimated from previous iteration. The size of k in the coarsest layer is rigidly fixed to 7×7. At the start of minimization, weight μ is set to a high value to ensure restoration of strong edges and removal of details. During each iteration cycle, a coarse L is computed with an EPI prior for each nonoverlapping patch. Theoretically, the EPI prior shows increased sparsity. Hence, L exhibits more details with the increase in the number of layers.

The orange rectangular block in Fig. 3 contains the kernel estimation algorithm. The weight parameter γ in Equation (26) is constantly set to a positive value to penalize large Fourier coefficients, which ensures a smooth PSF distribution. Then, the estimated k is upscaled by a factor of 2, and the result serves as the initial prediction for the next estimation layer. The procedure is repeated until the intended size of k is reached. Equation (26) is solved, and all negative-value pixels in k are set to zero, centered, and finally normalized to 1.

The k estimated from a previous step is used in image restoration, with the blurred image serving as input. The restoration is non-blinded; therefore, such a problem can be solved using various image deconvolution techniques. In this paper, the algorithm published in [12] is utilized because this method produces few ringing artifacts and accurate restoration. Appendix II shows the complete procedure of the proposed EPI algorithm.The methods and procedures are implemented in MATLAB2023a. This software is installed in personal computer which housed 4.0 GHz Intel Core i5 processor, 8 GB RAM and Windows 11.

3. IMAGE ACQUISITION

Neutron imaging experiments are performed at MNA, which houses the Research TRIGA PUSPATI (RTP) reactor. The RTP is a swimming pool-type, light, water research reactor containing enriched uranium–zirconium–hydride fuel and a graphite reflector. This reactor has a nominal power of 1 MW and is thus categorized under low-power research reactors. RTP possesses three radial beam ports, one tangential beam port, and one thermal column. The NR imaging facility is constructed around the radial beam port. The details are described elsewhere [14]. A total of six NR images of common objects produced at this facility are investigated, three of which are illustrated in Fig. 4.


(a)

(b)

(c)

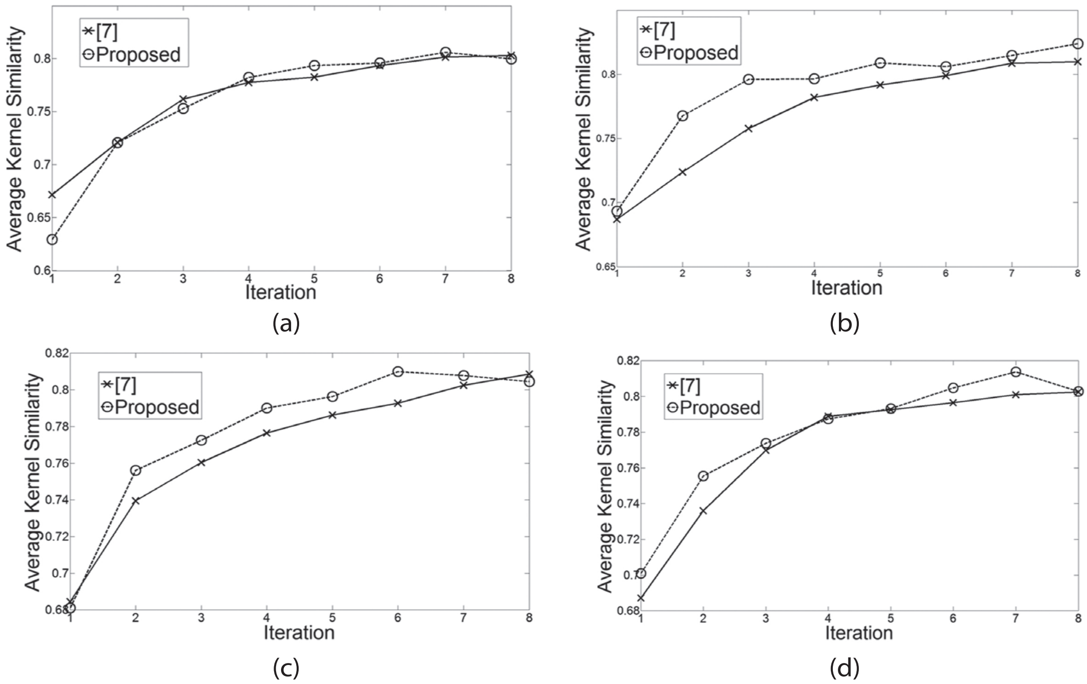
Fig. 4. Examples of neutron images produced by RTP. Regions containing vague useful features are marked with solid-line circles. (**a**) Hard-disk drive (1509 x 1248), (**b**) Honeycomb (1515 x 2322), (**c**) Lily flower (1728 x 2132)

The remaining images are presented in Appendix III. These 16-bit images are captured at an exposure time of 300 s and they are shown here after noise removal and brightness adjustment. Figs. 4(a), 4(b), and 4(c) correspond to images of a hard-disk drive, an aircraft honeycomb, and a lily flower, respectively. The size of each image is indicated in the figure. Visually, these images are blurry, which causes difficulty in the identification of important features or useful structures. For illustration purposes, regions containing useful information are encircled with solid lines. In the case of the hard-disk-drive in Fig. 4(a), five small anomalies located on the controller unit appear faded and blurry. Fig. 4(b) illustrates the loss of minute details, such as small structures of the honeycomb. Meanwhile, left petal of a flower's image in Fig. 4(c) show fine filaments that are blurry and out of focus. The proposed deblurring method is applied to enhance the images in Fig. 4. The results are presented and discussed in next section.

4. RESULTS

4.3. DETERMINATION OF PATCH SIZE

The effect of various patch sizes and the number of iterations is investigated first. Given that the ground measurements required for such an investigation and the lack of ground truth image for NR at present, optical images are the best alternative option. The popular dataset published in [15] is considered for this purpose. This dataset includes four ground truth images (255×255) and eight PSF kernels (25×25). These ground images are blurred using eight various PSF kernels, which results in 32 blurred images. The results are evaluated in terms of the following quality indices: similarity kernel ($S(k, \hat{k})$), mean peak signal ratio (PSNR), and mean structural similarity (SSIM). Figs. 5–6 show the plotted results comparing the proposed method and [7]. The former reveals variation in $S(k, \hat{k})$, and the latter depicts the trends of PSNR and SSIM with the increase in patch size and iteration. In terms of $S(k, \hat{k})$, both algorithms show no considerable variation when different patch sizes are used. Only the number of iterations exhibits a crucial effect on $S(k, \hat{k})$, which increases with the increase in the number of iterations. This trend is expected because k approaches $\hat{k}$ as the iteration increases. Moreover, $S(k, \hat{k})$ converges to almost the same value for both algorithms after eight iterations (Fig. 5). A striking difference is noted upon close examination of this figure. Although the $S(k, \hat{k})$ values calculated from the proposed approach show no significant difference from those computed in [7], the former registers slightly and consistently higher values (Figs. 5(b-c)). The same trend is achieved for SSIM, as suggested by the results in Figs. 6(a-d)(ii). The competitiveness of the proposed method is best shown in terms of the PSNR. Fig. 6 reveals the significantly higher PSNR values of the proposed method compared with those in [7], especially for smaller patch sizes. The proposed solution attains a PSNR of 29 dB compared to 26 dB in [7] for a patch size corresponding to SF=0.025. This findings translates to an approximately 11% increase in the PSNR. Results displayed in Figs. 5–6 also suggest a SF=0.025 as the best patch size for the proposed method. Hence, this size is used for the restoration of NR images in Fig. 4. The results are discussed in the following subsection. Meanwhile, Appendix IV provides results obtained using an image from Levin dataset, which prove the accuracy of the proposed method in the restoration of optical or synthetic images.



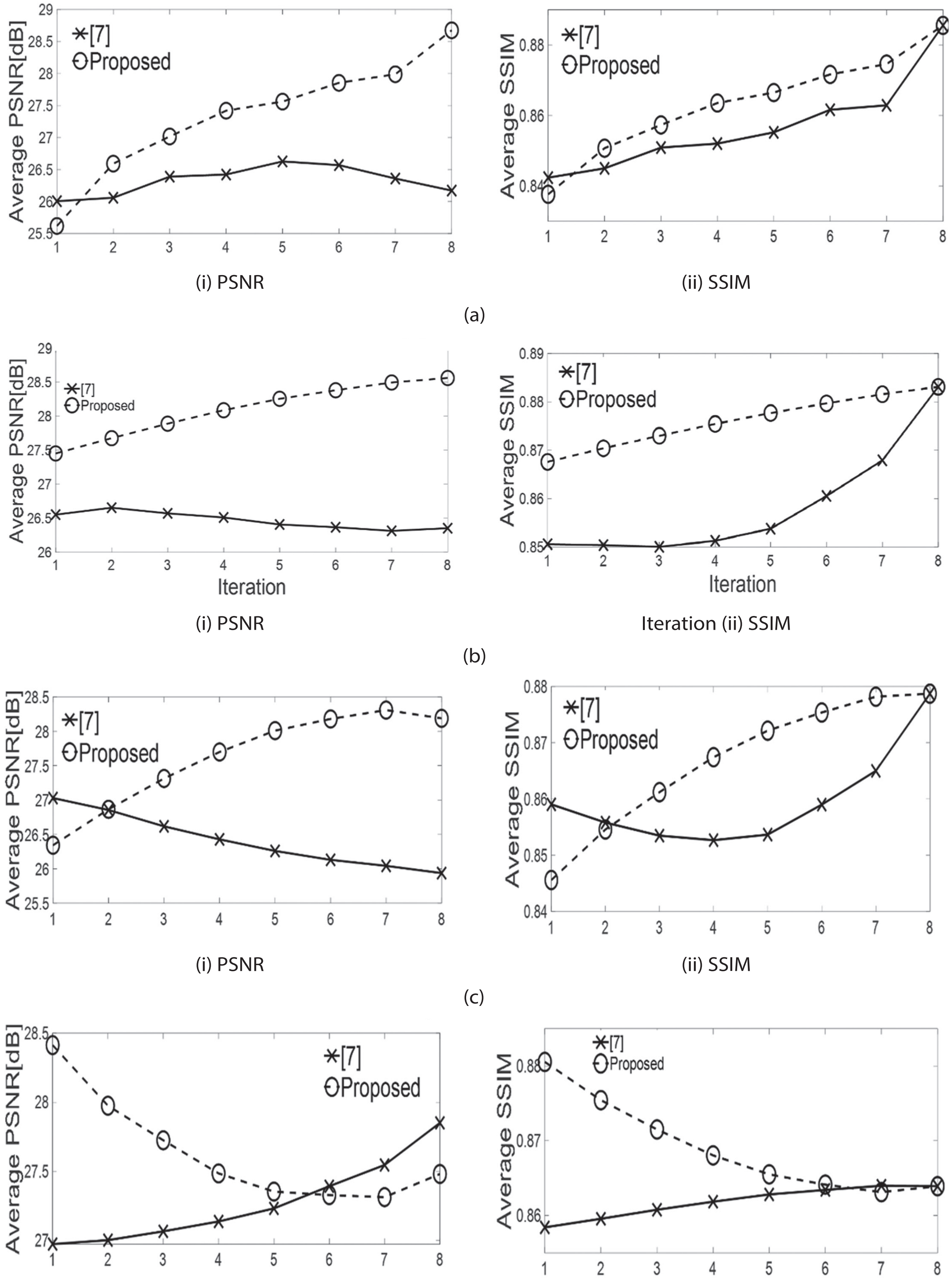
(a)

(b)

(c)

(d)

Fig. 5. Effect of varying patch size on PSNR and SSIM in the comparison of the proposed method and that of [7]. (**a**) SF=0.020, (**b**) SF=0.025, (**c**) SF=0.030, (**d**) SF=0.035



(i) PSNR

(ii) SSIM

(a)

(i) PSNR

Iteration (ii) SSIM

(b)

(i) PSNR

(ii) SSIM

(c)

(i) PSNR

(ii) SSIM

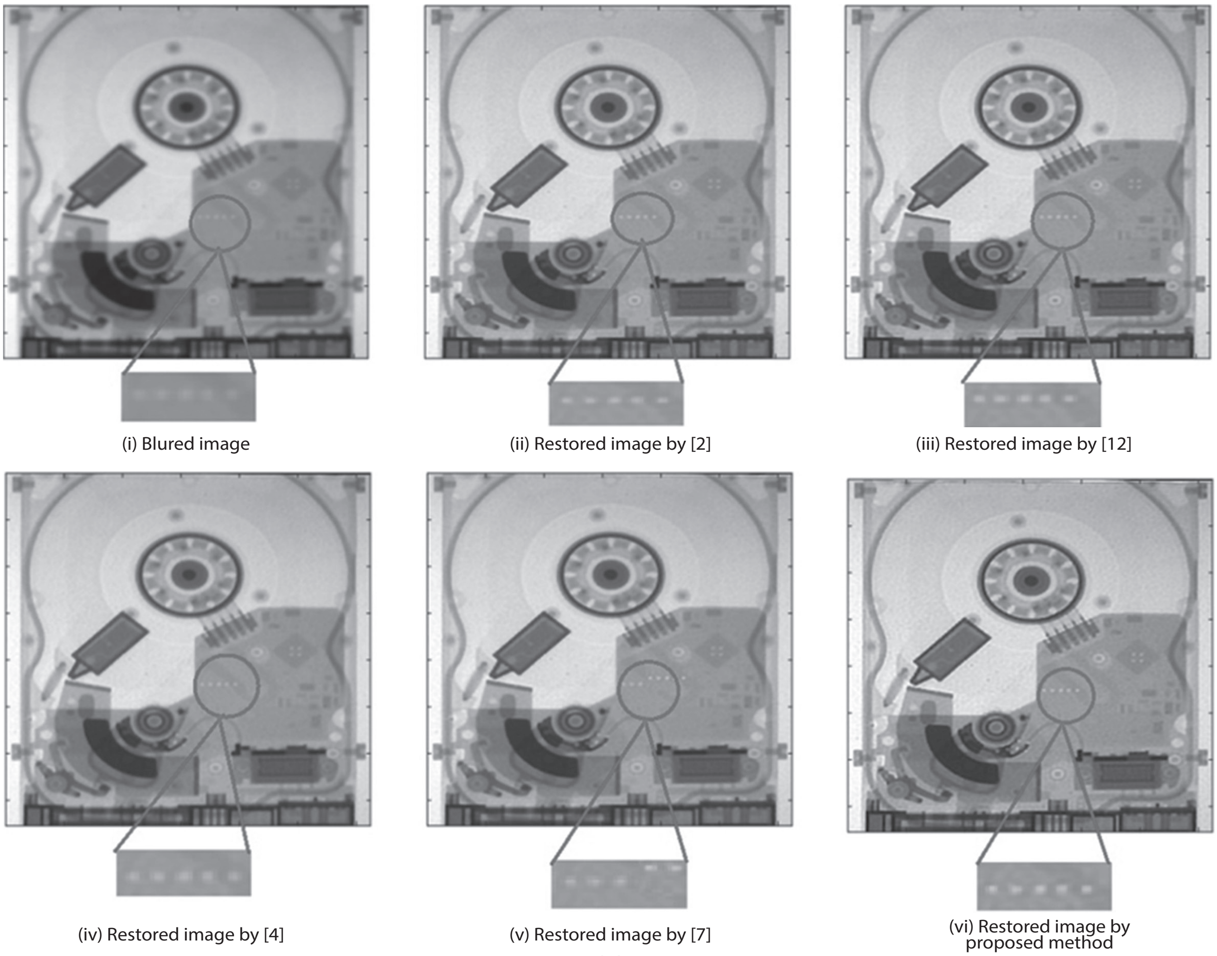
(d)

Fig. 6. Effect of varying patch sizes on PSNR and SSIM in the comparison of the proposed method and that of [7]. (**a**) SF=0.020, (**b**) SF=0.025, (**c**) SF=0.030, (**d**) SF=0.035

4.2. IMAGE RESTORATION

The settings determined from the above investigation are used in image deblurring experiments involving the real neutron images in Fig. 4. The performance of the proposed method is first evaluated visually (Fig. 7). The figure reveals the restoration findings corresponding to six neutron images of a hard-disk, a honeycomb, lily flower, aerosol spray can, rose flower, and laptop battery pack. Selected regions are zoomed-in to highlight small and fine details. Here, restorations are performed using the parameters determined heuristically through trial-and-error experiments: μ=0.004, γ=2, β_0=2μ, β_{max}=10^5, α=2, λ=0.1, and $iter_{max}$=5. The thoroughness of investigation results are compared with those of state-of-the-art methods published in [2], [4], [7], and [12]. The default settings proposed by these authors are used in restorations.

Referring to Fig. 7(a) and Fig. 7(b), the hard-disk and honeycomb resemble images captured from natural sceneries given the possible similar grayscale values of pixels in the same area and very slow gradient changes. By contrast, rapid changes can be observed in the grayscale values of pixels located in the vicinity of dominant objects. Thus, images with such pixel values exhibit a heavy-tailed distribution and local smoothing. For this reason, their intensity distributions are highly disorganized. As a result, the restoration of these images frequently results in the presence of geometrical artifacts. Despite this difficulty, overall, the proposed method produces results that are comparable to those of [7] (Figs. 7(a-b)(vi)). A close examination of Fig. 7(a)(v) reveals geometrical errors in the result produced by [7]. Visually, the locations of five small anomalies on the controller unit shifted slightly to the top. Other established methods, particularly [2], [12] and [4], resulted in blurry restoration, as evident from Figs. 7(b)(ii-iv), 7(c)(ii-iv), 7(d)(ii-iv), 7(e)(ii-iv), 7(f)(ii-iv), respectively. Although the hard-disk anomalies and fine honeycomb structures show slight improvement in their appearance, they remain blurry. Indirectly, these findings suggest that the heavy-tailed prior in [2] is an inefficient PSF estimator for textured images. Similar to the hard-disk and honeycomb, the restoration of lily flower is equally challenging because this object features a few grayscale tones during radiographical reconstruction. Such an image exhibits a low dynamic range (Fig. 7(c)(i)). As a result, the image contains very limited information that is useful for PSF estimation. Fig. 7(c)(vi) shows that the proposed method performs exceptionally well compared with other established methods. Visually, this result shows a relatively sharper and cleaner image compared with other images in the figure. In this case, the fine elements on the left side of the petal exhibit more distinct and clearer details. Other established methods also perform reasonably well, but the findings remain blurry (Figs. 7(c)(ii-iv)). The restoration of remaining images in Appendix III are shown in Appendix V. Similar trends can be observed from these results.



(i) Blured image

(ii) Restored image by [2]

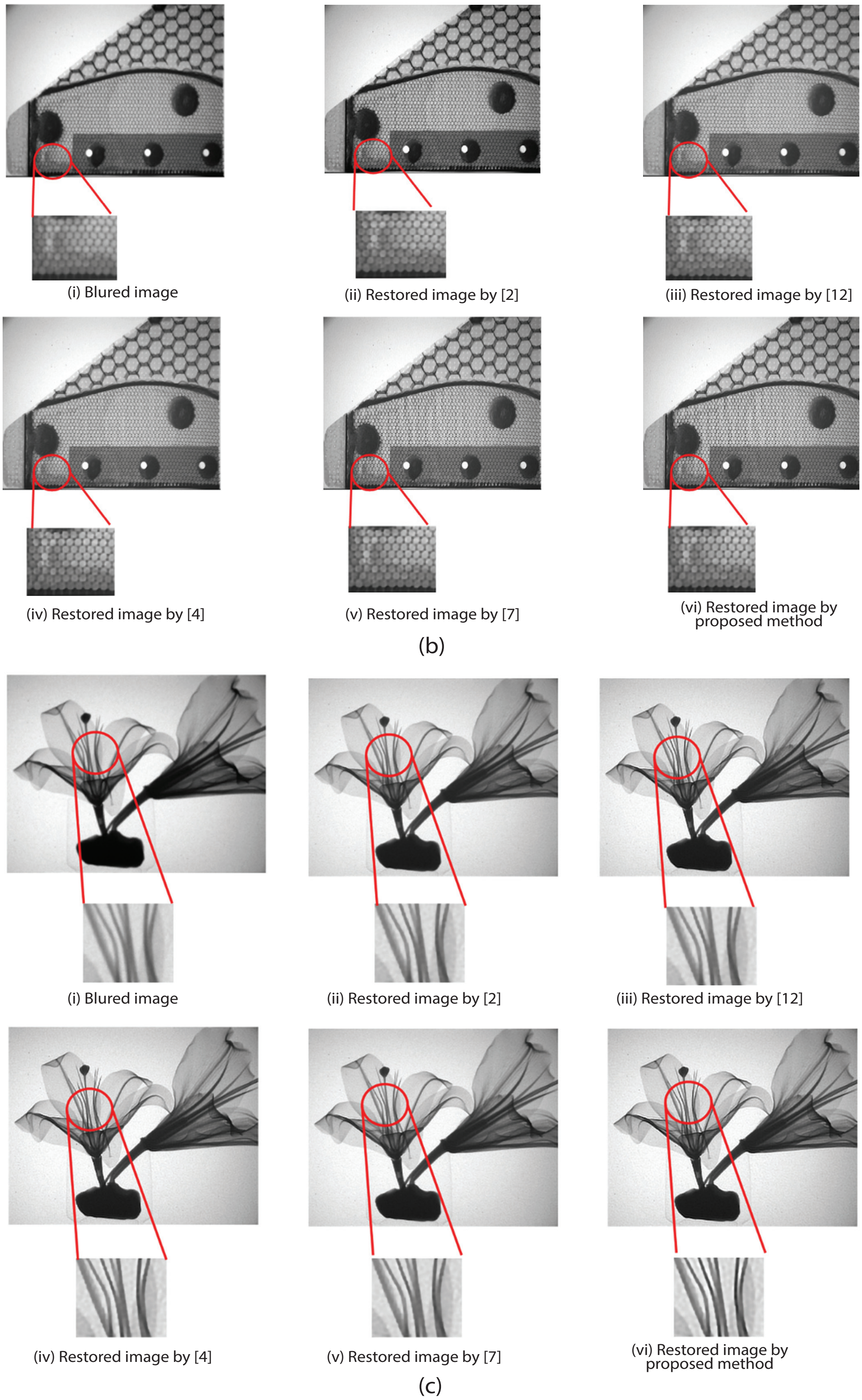
(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(vi) Restored image by proposed method

(a)



(i) Blured image

(ii) Restored image by [2]

(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(vi) Restored image by proposed method

(b)

(i) Blured image

(ii) Restored image by [2]

(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(vi) Restored image by proposed method

(c)

Fig. 7. Restoration results correspond to images in Fig. 4.
(a) Hard-disk dive, (b) Honeycomb, (c) Lily flover

In addition to visual quality, the performance of the proposed method is examined quantitatively using three evalution indices, including the Blind or Referenceless Image Spatial Quality Evaluator (BRISQUE)[16], image information entropy, and image contrast. As a unitless quality, the smaller the BRISQUE index, the better the image quality. Meanwhile the image contrast is a measure of the difference in brightness between the highest and lowest gray values in an image, directly indicating the degree of gray level variation. Essentially, the larger the image contrast, the clearer the image is. The entropy, in the other hands, indicates the richness of information or details contained in the image. In this case the larger the entropy, the more complete the image is. These metrices are calculated for all restored images used in this paper. The values are then averaged and tabulated for each method, and Table 1 summarizes the results. The BRISQUE, entropy and contrast values for each blurred image are included in the table for reference.

Referring to Table 1, overall, the proposed method offers a superior performance, which leads to the smallest and stable BRISQUE measures. On average, the proposed method attains a BRISQUE index of 46.05. The highest and hence the least accurate restoration are those from [12], with a BRISQUE index averaging 47.38. Their algorithm works well with images dominated by low-intensity pixels because it uses dark channels when enforcing the sparsity of solutions.

This condition is difficult to meet in NR because the images that produced by the algorithm usually contain many brightly illuminated pixels (Fig. 4). This finding mainly explains the reduced performance in [12]. Meanwhile, the performances of the algorithms in [4] and [7] is in between BRISQUE indices of 46.37 and 47.19, respectively. Similarly the proposed method acheived highest scores in term of entropy and contrast, averaging at 7.09 and 1.05 respectively. In comparison the algorithm of [2] resulted in the lowest entropy and contrast, averaging at 7.06 and 0.84 respectively.

The entropy and contrast values produced by other algorithms fall within this range. Like subjective evaluation, similarly, in this case the proposed method is consistently the best performing algorithm compared to established techniques.

Table 1. Quantitative evaluation comparing proposed and established methods.

				Methods			
INDEX	**Images**	**Origial**	**[2]**	**[12]**	**[4]**	**[7]**	**Proposed method**
BRISQUE	Fig. 7 (a)	48.50	42.02	48.49	43.46	43.80	43.45
	Fig. 7(b)	66.86	44.44	43.60	43.50	49.50	43.50
	Fig. 7(c)	57.30	43.60	46.00	46.10	43.60	43.50
	App.III(a)	44.76	44.68	43.67	43.79	43.92	43.98
	App.III(b)	57.88	57.88	57.88	56.99	57.29	56.94
	App.III(c)	44.90	44.90	44.90	44.42	45.05	44.93
	Ave.	55.72	46.25	47.38	46.37	47.19	46.05
ENTROPY	Fig. 7 (a)	7.39	7.43	7.44	7.43	7.43	7.51
	Fig. 7(b)	7.64	7.70	7.69	7.70	7.74	7.75
	Fig. 7(c)	7.32	7.33	7.36	7.35	7.34	7.37
	App.III(a)	7.17	7.18	7.19	7.19	7.18	7.19
	App.III(b)	5.99	6.05	6.06	6.04	6.06	6.07
	App.III(c)	6.99	6.68	6.68	6.68	6.68	6.69
	Ave.	7.03	7.06	7.07	7.06	7.07	7.09
CONTRAST	Fig. 7 (a)	0.89	0.86	1.02	1.30	1.00	1.05
	Fig. 7(b)	0.87	0.87	0.93	0.97	1.33	1.35
	Fig. 7(c)	0.88	0.90	1.15	0.99	1.06	1.24
	App.III(a)	0.98	0.98	0.93	1.01	1.02	1.03
	App.III(b)	0.53	0.52	0.65	0.61	0.64	0.71
	App.III(c)	0.88	0.88	0.96	0.91	0.94	0.95
	Ave.	0.83	0.84	0.94	0.96	0.99	1.05

4.3. RUNTIME

Finally, the runtime performance of the proposed algorithm is evaluated in comparisom with established techniques. Table 2 tabulates the results.

Table 2. Runtime comparing proposed and established methods

Images			**Runtime (s)**		
	[2]	**[12]**	**[4]**	**[7]**	**Proposed method**
Fig. 7(a)	119	293	4125	138	127
Fig. 7(b)	129	1077	4085	265	298
Fig. 7(c)	103	234	3990	284	296
App.III(a)	85	123	1501	201	136
App.III(b)	416	82	1580	112	117
App.III(c)	276	66	1364	109	111
	Mean=188 Min=85 Max=416	Mean=313 Min=66 Max=1077	Mean=2774 Min=1364 Max=4085	Mean=184 Min=109 Max=284	Mean=180 Min= 111 Max=298

Referring to Table 2 it can be seen that the proposed method is the fastest algorithm with a runtime averaging at 180 s; minimum and maximum runtimes of 111 s and 298 s respectivelly.

Clearly the soft thresholding helps speed-up the processing since only nonzero elements are used in the computation as previously explained. In comparison the algorithm of [4] resulted in the highest runtime of 2774 s, while its minimum and miximun runtimes range from 1364 s to 2774 s. Hence, [4] is the slowest, and hence, the most complex algorithm. This is mainly due to this algorithm utilizing joint prior operation which is a very time-consuming procedure. Meanwhile the performance of other algorithms lie between the proposed method and [4] as evident from Table 2.

5. CONCLUSIONS

This work presented a relatively novel, simple, and effective patch-wise enhanced prior for image restoration in a blind deconvolution framework. Inspired by the decreased intensity of high-intensity pixels due to blurring, the proposed method incorporates the EPI prior in a non-overlapping patch combined with the existing image gradient prior to regularization of the solution. A coarse-to-fine adjustment with a multilayer scaling approach is implemented within the MAP framework. Overall, these steps provide an accurate estimation of PSF and hence a superior restoration performance. Experiments using real degraded neutron images produce competitively important results, visually and quantitatively. Importantly, the restored results reveal important structures and minute details in images. Hence, the proposed method can potentially improve the visibility of blurry neutron images, which is crucial in applications, such as material characterization. It also performs reasonably well in challenging problems that involved large PSF kernels. Nevertheless, the algorithm may yield dissatisfactory results for images that are severely degraded by gamma noise. Such a degradation is inherently observed in small nuclear reactors. Thus, an effective denoising strategy may be needed prior to restoration.

6. ACKNOWLEDGEMENTS

The work is supported by Malaysian Nuclear Agency, the Director's office of MNA, the Reactor Technology Centre, and the Technical Support Division. KY acknowledges Ministry of Higher Education Malaysia for providing scholarship enabling her to pursue postgraduate study at Universiti Sains Malaysia. This project is partly financed by Universiti Sains Malaysia (1001.PELECT.8070009).

7. REFERENCES:

[1] M. Kardjilov, N. Manke, I. Woracek, R. Hilger, A. J. Banhart, "Advances in neutron imaging", Materials Today, Vol. 31, No. 6, 2018, pp. 652-672.

[2] J. Kotera, F. Sroubek, P. Milanfar, "Blind deconvolution using alternating maximum a posteriori estimation with heavy-tailed priors", Proceedings of the 15th International Conference on Computer Analysis of Images and Patterns, York, UK, 27-29 August 2013, pp. 59-66.

[3] L. Xu, S. Zheng, J. Jia, "Unnatural L0 sparse representation for natural image deblurring", Proceedings of the IEEE on Computer Vision and Pattern Recognition, Portland, OR, USA, 23-28 June 2013, pp. 1107-1114.

[4] J. Dong, J. Pan, Z. Su, M. H. Yang, "Blind image deblurring with outlier handling", Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22-29 October 2017, pp. 2497-2505.

[5] J. Pan, Z. Su, "Fast L0-regularized kernel estimation for robust motion deblurring", IEEE Signal Processing Letters, Vol. 20, No. 9, 2013, pp. 841-844.

[6] D. Hu, J. Tan, L. Zhang, X. Ge, J. Liu, "Image deblurring via enhanced local maximum intensity prior", Signal Processing: Image Communication, Vol. 96, 2021, pp. 116311.

[7] F. Wen, R. Ying, Y. Liu, T. K. Truong, "A simple local minimal intensity prior and an improved algorithm for blind image deblurring", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 31, No. 8, 2021, pp. 2923-2937.

[8] D. Yang, X. Wu, H. Yin, "Blind Image Deblurring via a Novel Sparse Channel Prior", Mathematics, Vol. 10, No. 8, 2022, pp. 1-18.

[9] D. Zhang, G. Sun, Z. Yang, J. Yu, "A high-density gamma white spots-Gaussian mixture noise removal method for neutron images denoising based on Swin Transformer UNet and Monte Carlo calculation", Nuclear Engineering and Technology, Vol. 56, No. 2, 2024, pp. 715-727.

[10] C. Zhao, W. Yin, T. Zhang, X. Yao, S. Qiao, "Neutron image denoising and deblurring based on generative adversarial networks", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Vol. 1055, 2023, p. 168505.

[11] Y. Khairiah, M. Z. Abdullah, I. Haidi, "Restoration of radiograpic neutron image using single-chan-

nel blind deconvolution", International Journal of Electrical and Electronic Engineering, Vol. 13, No.2, 2024, pp. 160-167.

[12] J. Pan, Z. Hu, Z. Su, M. H. Yang, "L0-regularized intensity and gradient prior for text images deblurring and beyond", Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 2, 2017, pp. 342-355.

[13] S. Cho, S. Lee, "Fast motion deblurring", ACM Transactions on Graphics, Vol. 28, No. 5, 2009, pp. 1-8.

[14] R. Jamro et al. "Monte Carlo Simulation for Designing Collimator of the Neutron Radiography Facility in Malaysia", Physic Procedia, Vol. 88, 2017, pp. 361-368.

[15] A. Levin, Y. Weiss, F. Durand, W. T. Freeman, "Understanding blind deconvolution algorithms", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 33, No. 12, 2009, pp. 2354-2367.

[16] A. Mittal, A. K. Moorthy, A. C. Bovik, "No- Reference Image Quality Assessment in the Spatial Domain", IEEE Transactions on Image Processing, Vol. 21, No. 12, 2021, pp. 4695-4708,.

APPENDIX I.

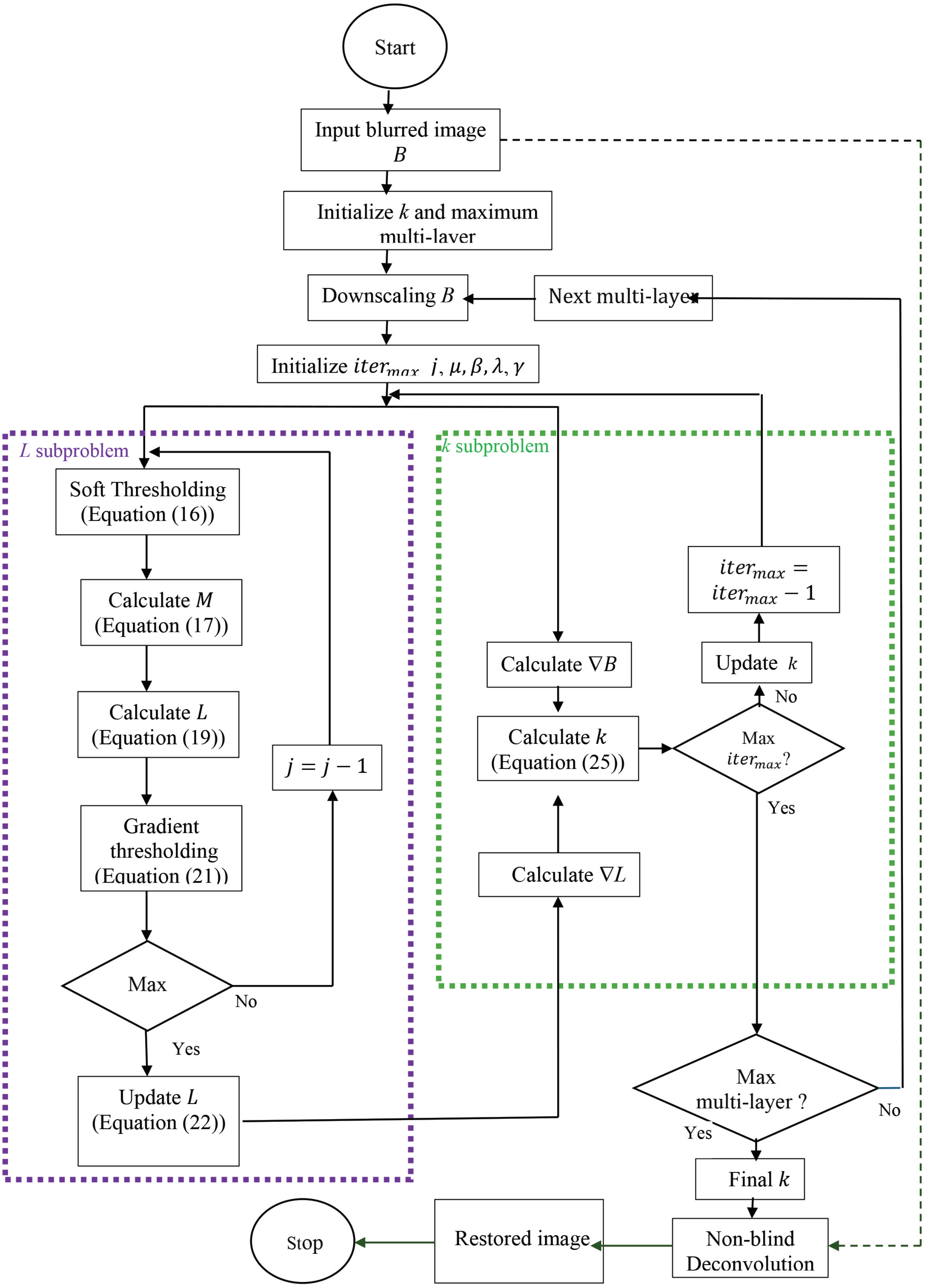
The main steps involved in solving L and k subproblems are summarized in Algorithms 1 and 2, respectively. In the latter case, the main parameter γ is set to 2.

```
Algorithm 1 $L$ subproblem
Input: Downscaled blurred image $B$, interim kernel $k^i$
$\beta \leftarrow \beta_0$, $L^0 \leftarrow B$
While $\beta \leq \beta_{max}$, do ($t$=0,1,2,…)
  $L^{t+1,0} \leftarrow L^t$
  For $j$=0:$J$-1 do
    Obtain $\tilde{L}_s^{t+1,j}$ via Equation (16)
    Compute $M^{t+1,j}$ via Equation (17)
    Update $\tilde{L}^{t+1,j}$ via Equation (19)
    Calculate gradient thresholding to obtain
    $G^{t+1,j+1}$ via Equation (21)
    Update $L^{t+1,j+1}$ via Equation (23)
  End of
  $\tilde{L}^{t+1} \leftarrow L^{t+1,j}$
  $\beta \leftarrow a\beta$
  End while
  $L^{i+1} \leftarrow L^{t+1}$
Output: Intermediate latent image estimation $L^{i+1}$
```

```
Algorithm 2 $k$ subproblem
Input: Blurred image $B$
Initialize $k^0$ from the previous layer of pyramid
For $i$=1:$iter_{max}$ do
  Estimate $L^i$ via Algorithm 1 using $k^{i-1}$
  Estimate $k^i$ via Equation (25)
End For
  $\hat{k} \leftarrow k^i$, $\hat{L} \leftarrow L^i$
Output: estimated PSF $\hat{k}$, intermediate image $\hat{L}$
```

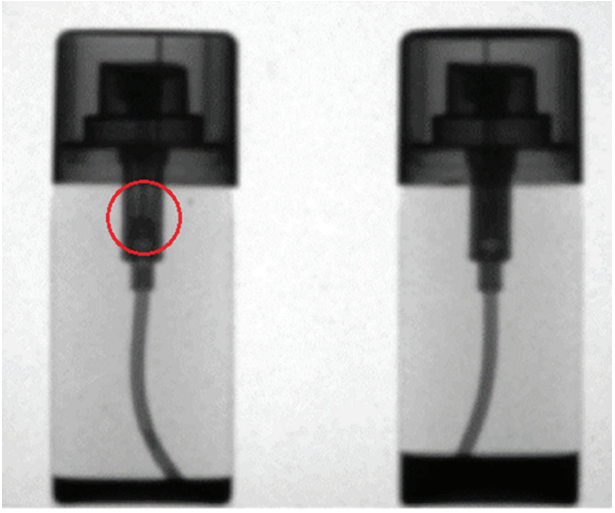
APPENDIX II.

Detailed flowchart of the algorithm for solving L and k subproblems. The dotted line indicates subsequent processing following all major calculations (solid lines).

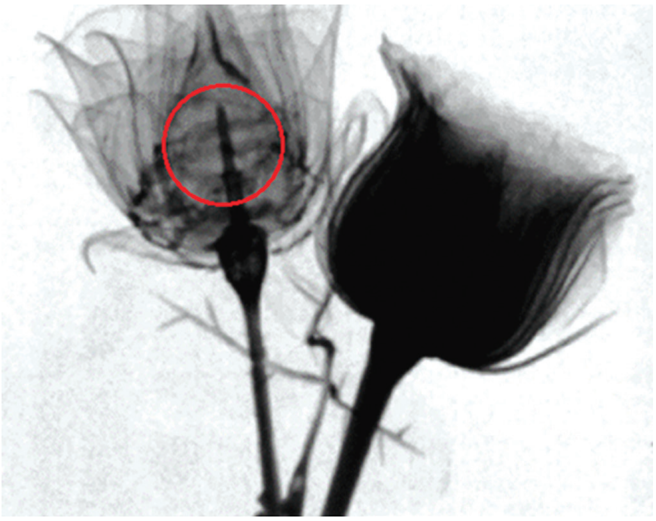

Start
Input blurred image B
Initialize k and maximum multi-laver
Downscaling B
Next multi-layer
Initialize $iter_{max}$ $j, \mu, \beta, \lambda, \gamma$
L subproblem
Soft Thresholding (Equation (16))
Calculate M (Equation (17))
Calculate L (Equation (19))
Gradient thresholding (Equation (21))
Max
No
Yes
$j = j - 1$
Update L (Equation (22))
k subproblem
Calculate ∇B
Calculate k (Equation (25))
Calculate ∇L
Max $iter_{max}$?
No
Yes
Update k
$iter_{max} = iter_{max} - 1$
Max multi-layer ?
Yes
No
Final k
Non-blind Deconvolution
Restored image
Stop

APPENDIX III.

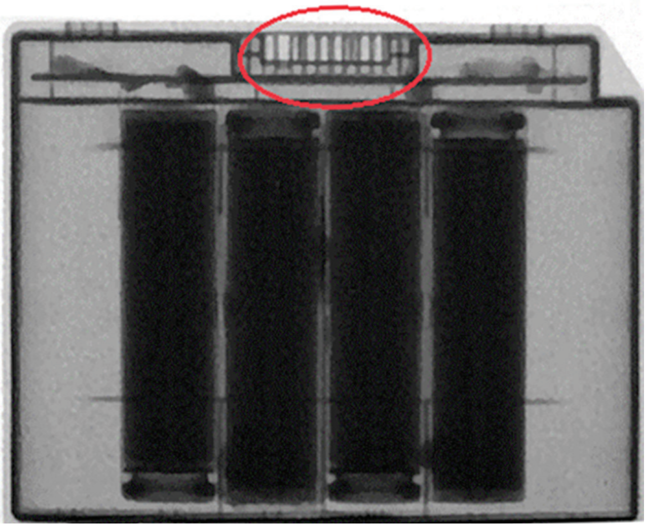
Another three examples of blurry neutron images of three common objects produced by RTP. Regions containing vague useful features are marked with solid-line circles.



(a) Aerosol spray can (1180 x 1380)



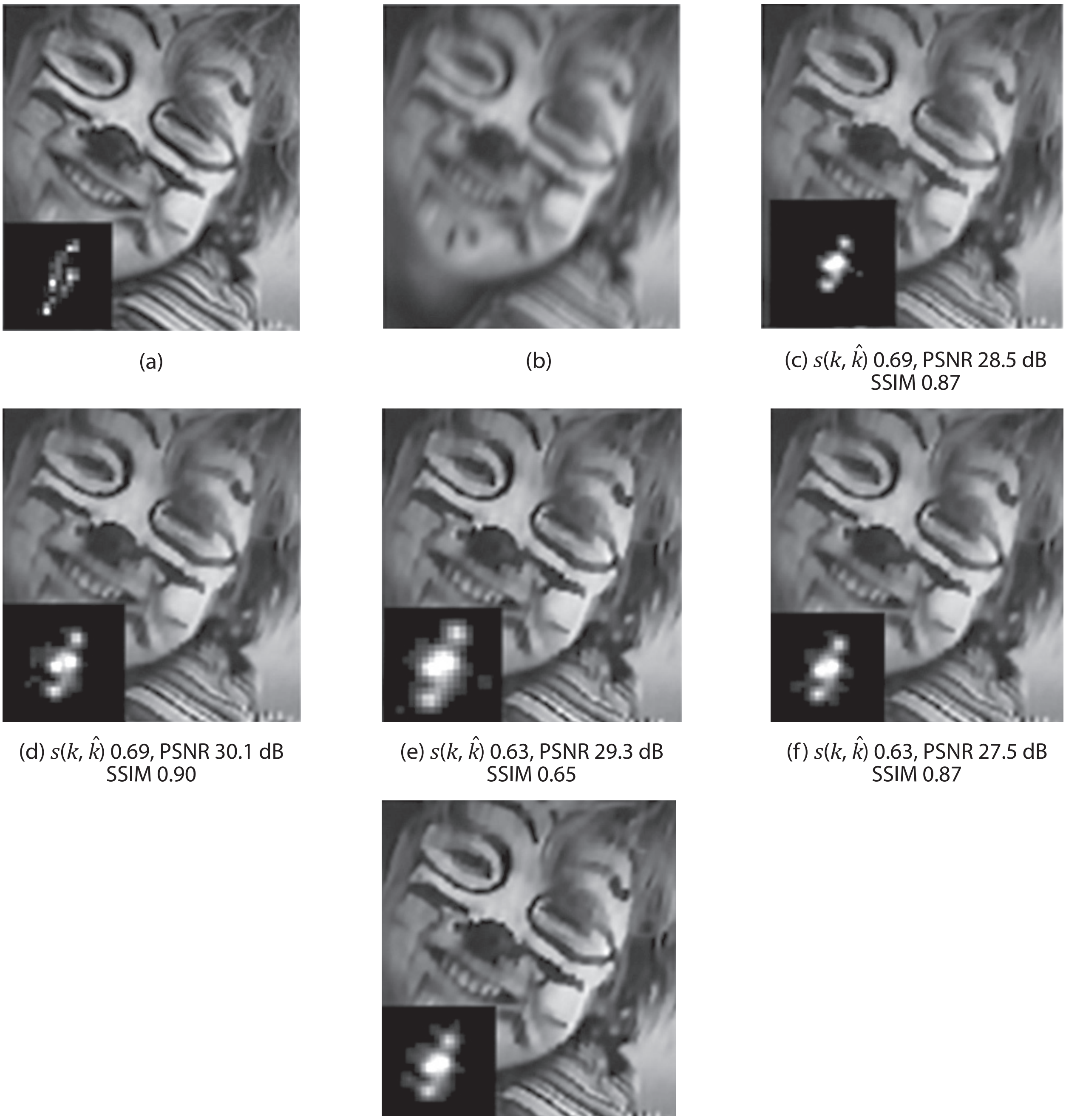
(b) Rose Flower (1143 x 1509)



(c) Laptop battery pack (966 x 1734)

APPENDIX IV.

Example of a deconvolution result from Levin's dataset showing (a) the ground truth and (b) the blurred version of (a). (c), (d), (e), (f), and (g) Restored images produced using the methods of Kotera et al. (2013), Pan et al. (2017), Dong et al. (2017), and Wen et al.(2021), and the proposed method, respectively.



(a)

(b)

(c) $s(k, \hat{k})$ 0.69, PSNR 28.5 dB
SSIM 0.87

(d) $s(k, \hat{k})$ 0.69, PSNR 30.1 dB
SSIM 0.90

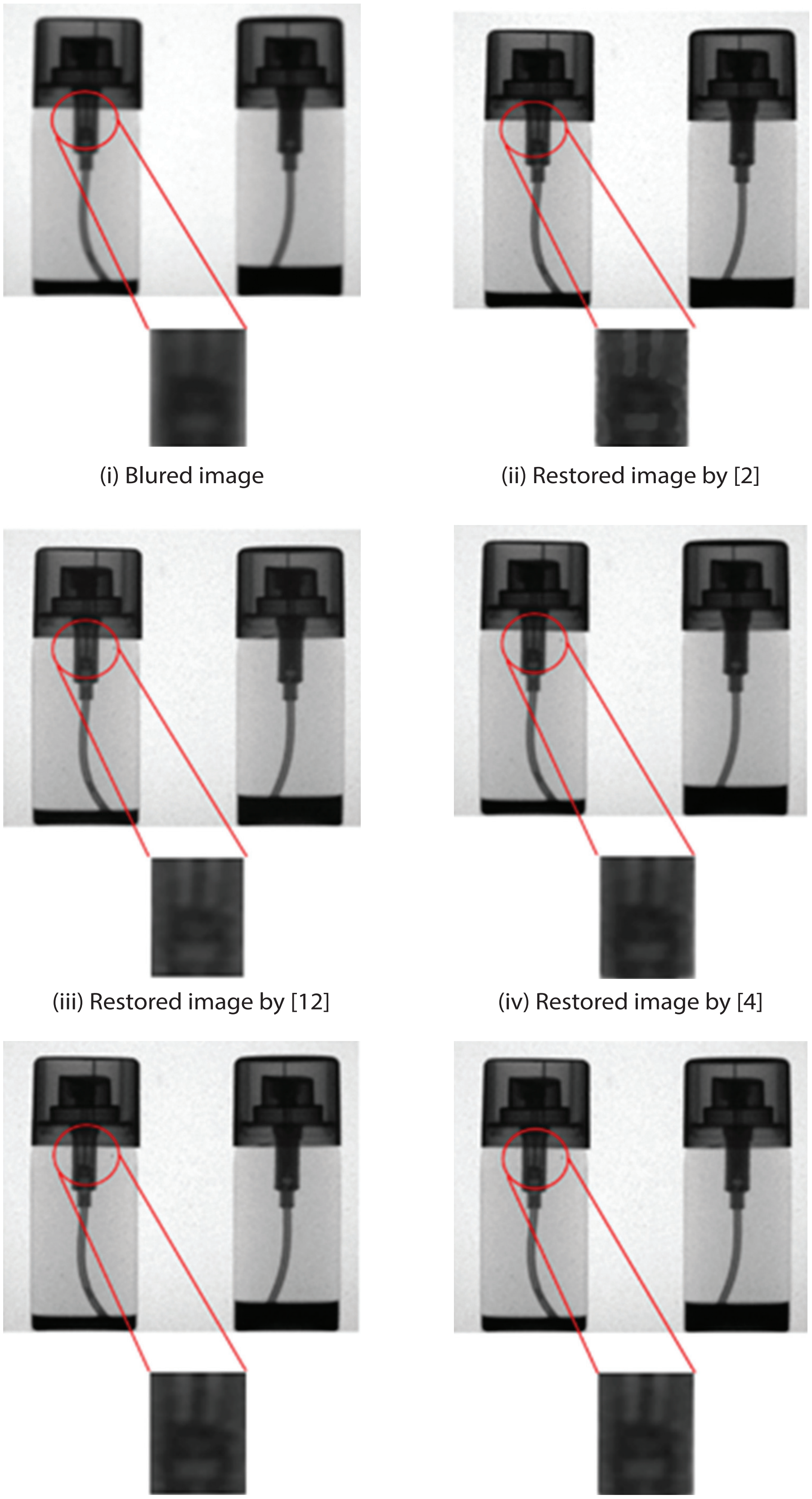
(e) $s(k, \hat{k})$ 0.63, PSNR 29.3 dB
SSIM 0.65

(f) $s(k, \hat{k})$ 0.63, PSNR 27.5 dB
SSIM 0.87

(g) $s(k, \hat{k})$ 0.73, PSNR 30.1 dB
SSIM 0.90

APPENDIX V.

Restoration results correspond to images in Appendix III.



(i) Blured image

(ii) Restored image by [2]

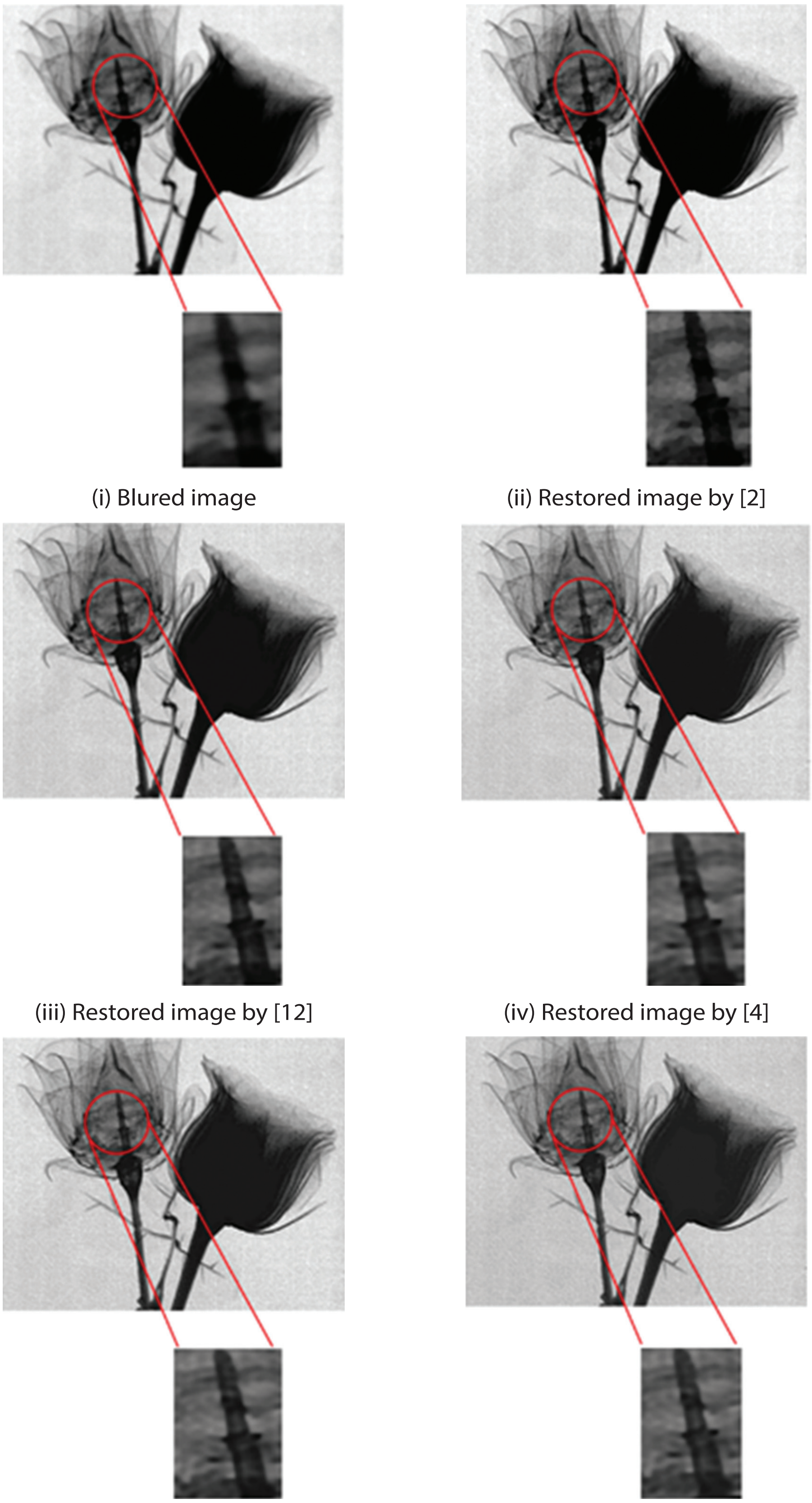
(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(iii) Restored image by proposed method

(a) Aerosol spray can



(i) Blured image

(ii) Restored image by [2]

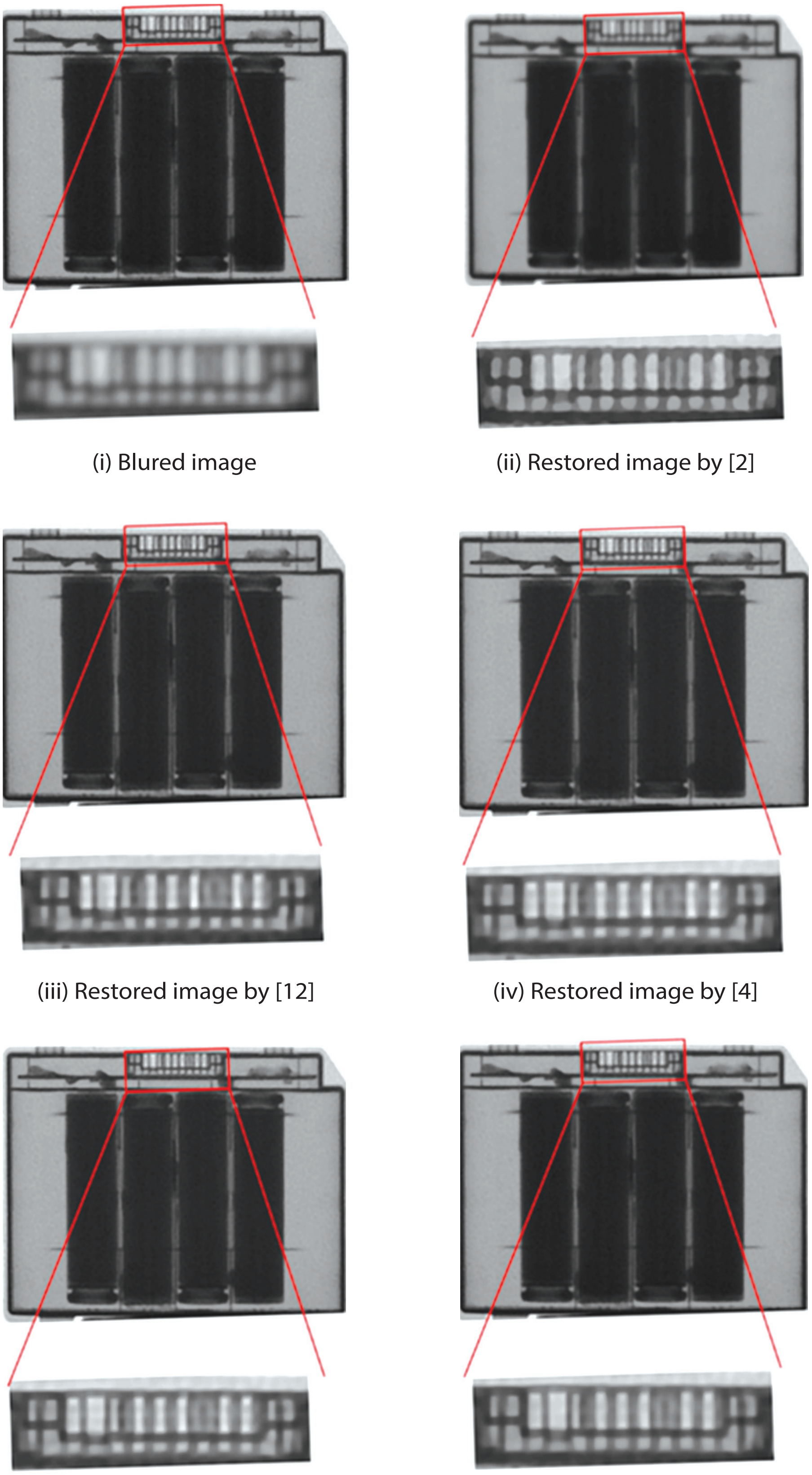
(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(iii) Restored image by proposed method

(b) Rose flover



(i) Blured image

(ii) Restored image by [2]

(iii) Restored image by [12]

(iv) Restored image by [4]

(v) Restored image by [7]

(iii) Restored image by proposed method

(c) Laptop battery pack

Breast Pathology Changes Extraction and Measurement Based on Machine Learning and DWT

Original Scientific Paper

Sahar Shakir*
Northern Technical University, Technical Engineering Collage-Kirkuk
Kirkuk, Iraq
sahar.najat23@ntu.edu.iq

Yousif A. Hamad
[1] University of Kirkuk, Department of Computer Science, Kirkuk, Iraq
[2] Siberian Federal University, Artificial Intelligence Laboratory, 660074, Krasnoyarsk, Russia
y.albayati@uokirkuk.edu.iq

Rehab Kareem
Imam Ja'afar Al-Sadiq University,
Department of Computer Technology Engineering, Collage of Information Technology
Baghdad, Iraq
rehab.hussian@ijsu.edu.iq

*Corresponding author

Abstract *– In recent years, medical image analysis has witnessed significant advancements in aiding accurate diagnosis and treatment planning. Breast tumor segmentation is a critical task in medical imaging, as it facilitates the identification and characterization of tumors for effective clinical decisions. This paper proposes a novel approach for breast tumor segmentation and analysis by integrating Fuzzy C-Means Clustering (FCM) with Discrete Wavelet Transform (DWT), called FCMDWT. This method is effective in breast diagnosis analysis, tumor size measurements, and diagnosing reports and does not require prior training in segmentation. Initially, the DWT is applied to the mammography image, decomposing it into different frequency subbands. FCM is employed on the DWT coefficients to ensure robust clustering by accommodating uncertainty and overlapping regions in the image. The experimental evaluation conducted on a comprehensive dataset and comparative analyses demonstrates the superiority of the FCMDWT approach. Furthermore, the proposed method extends beyond segmentation, incorporating tumor analysis by extracting relevant features such as size, shape, and texture. The results indicate the potential of the FCMDWT approach in not only accurate segmentation but also in providing valuable insights for clinical decision-making.*

Keywords: *DWT image processing, FCM, breast mass measurements, medical image processing, tumor segmentation*

Received: August 31, 2024; Received in revised form: November 17, 2024; Accepted: November 19, 2024

1. INTRODUCTION

Cancer is a pathological state caused by aberrant cellular alterations that result in uncontrolled proliferation. Malignant breast cells often form tumors, which are masses or lumps that are termed by the specific body location where they arise. Breast cancer is the most common form of cancer in women and the second most common cause of death globally. During the first stages of breast cancer, while it is still treatable, individuals often have minor discomfort. Therefore, screening is crucial for timely diagnosis. Timely identification of cancer and timely intervention might potentially reduce mortality rates [1].

For physicians to choose the best course of therapy and, as a consequence, save at least 40% of patient's lives, they would have to understand whether a tumor exists and what kind of malignant tumor it is [2]. Cancerous growths or lumps are assemblages of aberrant cells. They can develop from any of the trillions of cells that make up the human body. The growth and behavior of tumors vary depending on their type, with malignant (cancerous), benign (non-cancerous), and precancerous tumors all exhibiting distinct characteristics. Malignant tumors arise from cells that undergo genetic mutations and proliferate in an unregulated manner, resulting in the development of a mass or growth that can infiltrate neighboring tissues and organs. These malignant cells can detach from the primary tumor and spread to other areas of the body via the circulatory and lymphatic systems, a phenomenon referred to as metastasis. Scientists have investigated different methodologies for detecting and forecasting tumor activity [3], such as employing Conditional Random Fields (CRF) [4], analyzing volumetric white ratios of diffusion tensor (DT) in mammogram images [5], and utilizing topological imaging of human breast development through magnetic resonance imaging [6, 7]. Researchers used mammography images acquired from children in a state of regular sleep to forecast the first emergence of functional breast cancers [8]. Other research endeavors have concentrated on isolating specific histological regions within the tumor's white matter and discerning the characteristics of individual cells to get a more comprehensive understanding of tumor behavior [9, 10]. Identifying and characterizing tumors accurately is essential for effective treatment and improving patient outcomes. Imaging techniques such as mammography, ultrasound, and MRI are commonly used to detect and diagnose tumors. However, these methods are not always conclusive, and additional tests such as biopsies may be needed to confirm a diagnosis. Advancements in medical imaging and machine learning technologies have shown promise in improving tumor detection and diagnosis. For example, methods based on training and testing have been developed to analyze mammogram images and detect abnormalities that may indicate cancer. These techniques have the potential to increase the accuracy and efficiency of tumor diagnosis, reducing the need for invasive procedures and improving patient outcomes [11].

This work aims to use a combination of FCM clustering, DWT methods, and BCET image enhancement to segment and measure the size of breast tumors and non-infected areas on mammogram images. This method can reduce the workload of radiologists and confirm diagnosis analysis with high accuracy.

2. LITERATURE REVIEW

BC is defined as the development of a malignant tumor in a woman's breast. Medical practitioners use mammography pictures to identify breast cancers at different stages. Segmentation is an essential and demanding process in the categorization and interpretation of medical imaging. The FCM approach developed by Sharma and Selvakumar [12, 13] is often used for photo segmentation. In addition, [14] introduced a method for isolating breast tumors using a combined approach that used FCM.

The organization of linear and non-linear characteristics in mathematical computer vision (namely borders) may be achieved by using the conventional methodologies proposed in [15] and [16]. Identifying the specific form of cancer is a far more challenging undertaking [17]. Malignant tumors have clustered appearances, solitary ducts, and a poorly defined bulk, among other features. Effectively differentiating between cancers of the breast and other illnesses using ultrasound imaging remains tough owing to the low contrast and indistinct borders of tumors.

An innovative computational technique for locating and segmenting breast lesions in ultrasound images was also described by [18, 19]. Breast cancer cannot be cured unless it is detected early. While [20] uses discrete wavelet transform and clustering means for tumor mass delineation on mammogram images, [21] presents a technique for detecting breast tumors by fragmenting mammogram images using simple image processing algorithms that produce good results only in real time. The twofold banalization approach used by the authors of [22] was enhanced for mammography image isolation. Finally, a contour of the objects in the original image has been created using image boundary detection, making it simpler for doctors to detect cancer in various images. To overcome the shortcomings of FCN models, the authors in [23] presented the UNet model for mass segmentation based on training and testing. UNet argued that the decoder's high-level and low-level components should be combined. Skip connections were used to maintain this fusion, allowing the UNet architecture to be utilized in several medical applications, including mammography. The authors of [24] employed an UNet model for breast mass segmentation and classification.

The segmentation's F1-score for the DDSM dataset was 90%. Baccouche et al. [25] also used a UNet model to find mass lesions in full mammograms and got the same result. Their F1-score on scanned breast images was 86.91%. Using residual units to augment edge information in place of the standard neural units in the UNet architecture, researchers of [26] suggested the Residual-Dilated-Attention-Gate-UNet model. Traditional UNet scored 0.820. SegNet also performed well on the same dataset, with an overall dice coefficient of 0.817. UNet and SegNet were compared to [27] in which SegNet and UNet were the best tumor extraction models in terms of dice scores. In addition, the UNet beat the SegNet. The UNet had a dice unit of 0.883, compared to the SegNet's 0.504.

3. PROPOSED APPROACH

In this study, a BT segmentation approach is proposed that combines DWT and FCM clustering. DWT is first used to decompose breast images into different frequency sub-bands, improving the visualization of tumor boundaries. FCM is then applied to segment these enhanced features using the multi-resolution representation of DWT to improve segmentation accuracy and effectively distinguish tumor tissue from healthy regions. The outcome of the processing procedure is contingent upon the quality of the mammography images produced by the medical equipment. Acquired medical images can exhibit noticeable noise due to the technical operations of the equipment. Various approaches may be used to identify and classify breast cancers and tumors. Fig. 1 exhibits the boundary detection and FCMDWT segmentation flow charts for tumors and non-infected areas of the breast.

3.1. DATA COLLECTION AND PREPROCESSING

Medical images have been converted to the Digital Imaging and Communication in Medicine (DICOM) standard in CBIS-DDSM, resulting in a revised version of the Digital Database for Screening Mammography (DDSM) dataset. 1467 of the 2907 mammograms, which were obtained from 1555 individuals, are mass images. Two different mammography were performed on each breast. The original photographs, which have been connected to the vast locations in which they were shot [28] are around 3000 by 4800 pixels. Also, a collection of mammograms from Kirkuk City-Iraq National Medical Laboratory was collected. These images show 389 examples of breast cancer in stages 3 and 4 with mass lesions from 208 different people. Each image was preprocessed by resizing to a uniform dimension of [256 X 256] and normalized to a common intensity range. The source data consists of a mammography of the breast, which is used to diagnose breast tumors and breast cancer. If the source image is given in RGB format, it is transformed to greyscale. To preserve the image's aspect ratio, it is resized to fit the appropriate matrix. Subsequently, the scaled pictures undergo the application of the median filter, which effectively reduces random noise while preserving the borders of the image.

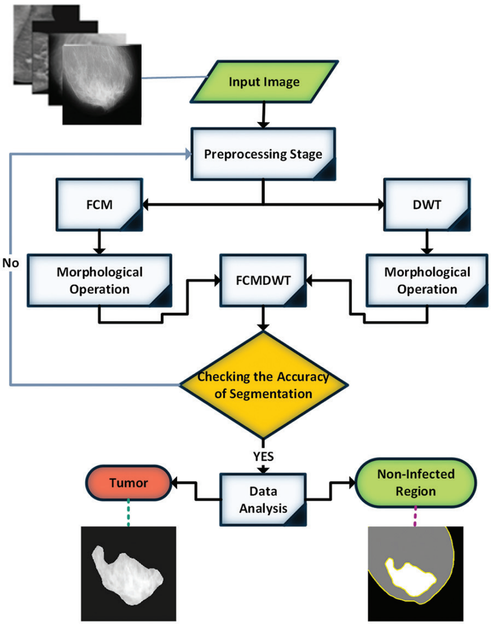


Fig. 1. FCMDWT segmentation flow chart and boundary detection scheme for tumor and non-tumor areas in the breast

During the scan enhancement phase of the source image, a noise-reducing filter is used to enhance the quality and contrast. We used the Balance Contrast Enhancement Technique (BCET) to improve and emphasize the region containing foreign entities such as tumors or nodules [29, 30]. From Fig. 2, we can recognize that the best mean value of high-quality mass segmentation is 80. Attributes of structural elements reveal an important aspect.

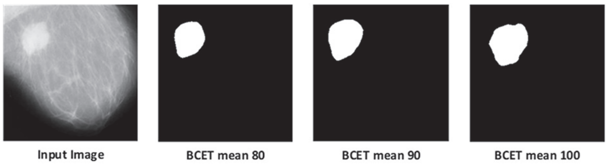


Fig. 2. Tumor segmentation using the different mean values of BCET

Mass segmentation or infection extraction refers to the process of extracting boundaries (lines, shape, size, and position) from medical scans. Decisions for breast cancer therapy heavily rely on medical image segmentation. Depending on the slide, segmenting a picture can result in a set of contours or a group of areas that make up the entire image. Based on the consistency of a tumor's features, including intensity, color, and texture, images may be utilized to classify whether the tumor class is benign or malignant. The thresholding method segments CT images by dividing the intensity of the screening mammography into two halves [31, 32].

3.2. DISCRETE WAVELET TRANSFORM (DWT)

The DWT was applied to each preprocessed image to obtain a multi-resolution representation.

The DWT technique is used in Step 1 to find the suitable threshold by dividing the source image into two sub-bands the LL subdomain and three high-frequency bands of size 83x152, which displays low-level filters, and the HH subdomain, which displays higher-level filters. Both low-pass and high-pass filters employ the thresholding technique to determine the threshold value edge enhancement. Step 2 produces a high-pass filter by imaging the remaining sub-bands with inverse wavelet transforms (horizontal, vertical, and diagonal). The LH matrix is split into a nested 3x3 matrix, the average value is calculated and assigned to the entire FCM.

The third step, which likewise divides the source image into two blocks of pixels, uses local contrast balancing to enhance an image's inherent contrast (black and white). The experimental findings demonstrate the efficiency, clearcut, and ease of understanding of the suggested approach. Smoke detection methods based on image segmentation might decrease noise interference. For this study, we utilized the DWT for its suitability in medical image analysis due to its effective edge detection.

3.3. FUZZY C-MEANS (FCM) CLUSTERING

The upgraded FCM method was utilized to do high-accuracy segmentation (normal area extraction) of the breast's uninfected tissue, and threshold segmentation was used to transform the enhanced breast mammography picture to black and white for extraction of the breast's infected tissue (size, location, and form). The primary methods used here are dilatation and erosion. The stroke area of items expands during the dilatation process and shrinks during the wear phase. The structural features served as the basis for these procedures. The stretch chooses the highest value and we find the lowest value by comparing all adjacent pixel values in the source image (CT image) given with the diagram part. Fig. 3 depicts the steps of the method proposed for segmenting breast masses.

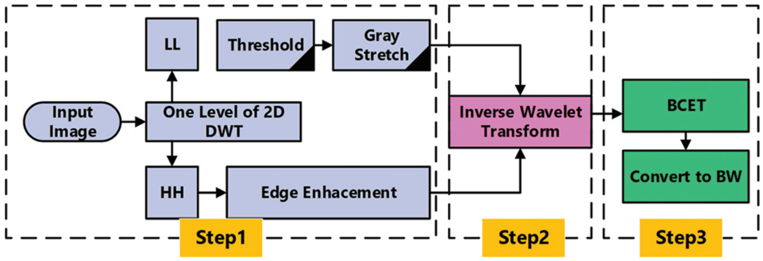


Fig. 3. Presented scheme of the breast mass segmentation method

3.4. POST-PROCESSING AND REFINEMENT

A post-processing approach was applied to refine the BT segmentation results. First, a morphological operation is performed to fill gaps within the tumor boundaries. Then, the proposed FCMDWT was used to isolate the tumor region more distinctly, removing small artifacts that FCMDWT had erroneously included. This post-processing step was essential to achieve cleaner and more precise affected area (BT) outlines and non-effected areas of the breast. The steps of applying FCMDWT segmentation on mammogram images for a complete clinical tool for mass tumor diagnosis (segmentation, measurements, and boundary detection) are presented in Fig. 4.

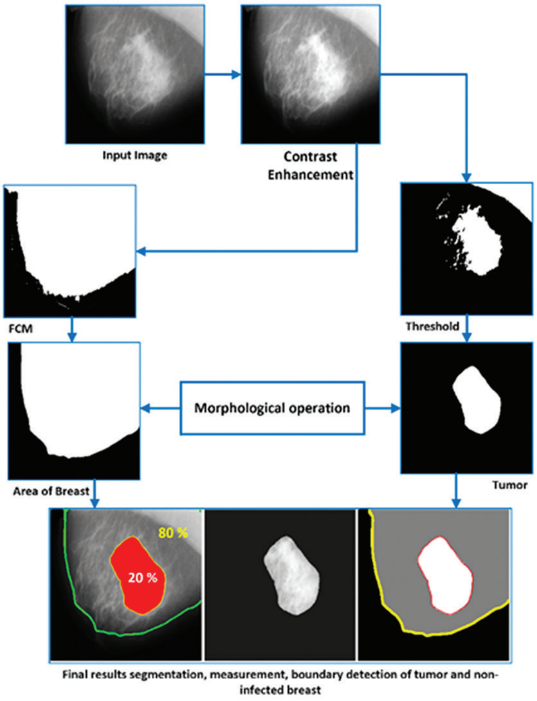


Fig. 4. The steps of the proposed FCMDWT segmentation method for breast diagnosis

3.5. EVALUATION METRICS

Breast tumor (infected region of the breast) segmentation on breast mammography images was evaluated using the Dice Coefficient (F1-score), which is defined as the geometric mean of the prediction accuracy [33]. The breast tumor, the non-infected region of the breast, and the identification of both borders are all outputs of FCMDWT segmentation. To determine whether the forecast was correct, the Jaccard coefficient of intersection over union (IoU) was used [34]. When their union splits the expected and actual bounding boxes, the result is the predicted bounding box. Predictions are labeled as "True Positive" (TP) or "False Positive" (FP) depending on whether the IOU is more than or equal to 50%.

The following formula (1) is used to determine the IOU:

$$\mathrm{IOU} = \frac{\text{area of overlap}}{\text{area of} \cup} \tag{1}$$

The IOU metric ranges from 0–1with zero signifying no overlap and one signifying perfect overlapping object segmentation.

The total area of Overlap reduced by the sum of all pixels in the two images yields the F1-score, which is calculated as follows:

$$\text{F1-score} = \frac{2\times\text{Area of Overlap}}{\text{Total combined pixels}}. \tag{2}$$

4. RESULTS AND DISCUSSIONS

In this study, we evaluated the provided computer-aided diagnostic system for mass breast analysis using the public CBIS-DDSM dataset as well as a private dataset as descripted in section 3.1. Several instances are shown in Fig. 5. The databases include tumors from a range of locations and disease types, as well as information about the shape, volume, texture, and size of the affected mass region surrounding the tumor size. We can observe the surface features and highlighted items both before and after an image is converted from one image form to another.

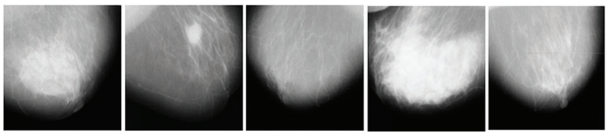


Fig. 5. Mammographic images of the breast obtained from a study's data

We experimented with a variety of breast tumor photos, all of which were 512 by 512 pixels in size, as a segmentation example. Fig. 6 and Fig. 7 show the results of various examples of breast tumor segmentation and identification using a unique segmentation approach, arranged from left to right. The original breast photos are presented in the first column, the segmented tumor results are presented in the second column, and the extracted tumor is located on the input image in the third column to help specialists better understand the location of the mass. The outlines of the breast tumor (which was extracted malignancy) and healthy breast areas are shown in the fourth column. The final row is indicated by semantic processing utilizing a color map (jet).

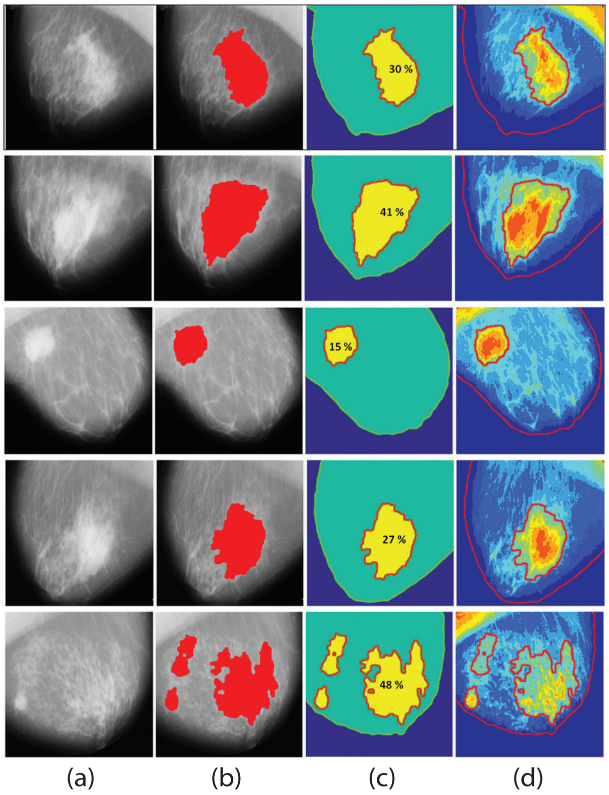


Fig. 6. Outcomes of the malignancy localization approach are provided, where (**a**) is the input images, (**b**) and (**c**) results of FCMDWT segmentation, and (**d**) mass and non-infected area is colored and localized with color-code based on FCMDWT localization on breast mammogram images (CBIS-DDSM)

Results from Fig. 6 and 7 ((b), (c), and (d)) may assist in the identification of nonspecific kinds of breast mass by the computer-aided diagnostic detection system and radiography (benign or malignant). By identifying cancers at an earlier stage, specialists may be able to save patients' lives. A survey of the scientific literature revealed that there are several issues with current molecular techniques and classifications, including their inability to identify objects, their inability to produce the same outcomes, and their lack of adequate quality control, the segmentation with high accuracy. Extracting breast tumors with the proposed approach FCMDWT in Fig. 6(b and c) and Fig. 7(b and c) enables specialists to make the right treatment decisions and provide an accurate diagnosis, which increases the chances of successful treatment as it allows for a precise examination of the tumor in terms of location, size, shape, and degree of spread.

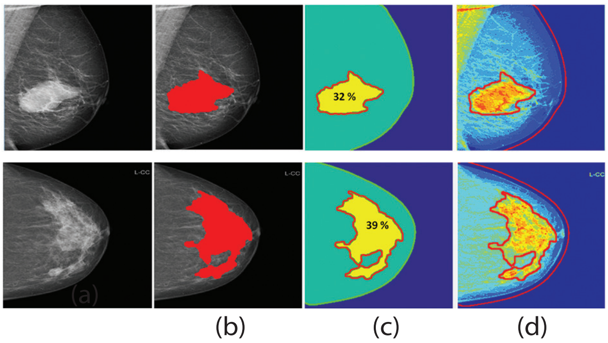


Fig. 7. Outcomes of the malignancy localization approach are provided, where (**a**) is the input images, (**b**) and (**c**) results of FCMDWT segmentation, and (**d**) mass and non-infected area is colored and localized with color-code based on FCMDWT localization on breast mammogram images (Private dataset)

The Fig. 6(d) and Fig. 7(d) help analyze and detect small tumors that may be difficult for doctors to see directly. It is also possible to predict from the shape whether the tumor is benign or malignant by analyzing patterns of stored medical data and comparing them with previous cases. The FCMDWT results help in achieving a more accurate diagnosis and providing customized treatment plans, which increases the efficiency and effectiveness of treatment and contributes to improving the quality of healthcare provided to patients.

Combining the two methods makes it possible to accurately locate tumors in medical images, as well as to accurately and quickly segment tumors in breast images. Images with edges identified and tumor and normal breast regions calculated (shown in Table 1).

Table 1. The detected concerns and their performance analysis, including the non-affected and eliminated breast regions

Data	Damaged areas (mass) %	Execution Time (s)
image 1	30 %	2
image 2	41 %	2
image 3	15 %	2
image 4	27 %	2
image 5	48 %	2

Calculating the extent of the afflicted sections by a tumor is made easier by differentiating between normal and malignant cells. The effectiveness of our suggested strategy is demonstrated by the determined area being displayed in pixel units. This paper compared our method to those in [24-27], as indicated in Table 2, for the identification and segmentation of breast tumors. Instead of depending on the empirically challenging to diagnose border lines between cancerous and non-cancerous breast sections to realize tumor location, visual analysis demonstrates that our approach outperforms the alternative in segmenting breast mass. By applying a color map to the data, our method can accurately locate tumor regions and non-cancerous breast regions on the original input image while only detecting tumor regions. This helps in displaying the current figure's color map and customizing it as shown in Fig. 6 (e) and Fig. 7 (e).

Using the FCMDWT segmentation technique, a comprehensive clinical tool for huge quantities of discovering tumors is needed, and segmentation architecture models have been developed. Radiologists can use this technology to help them find breast cancer more quickly by using it to help them identify what kind of tumor it is and how it looks. After segmentation, the edge detection of breast and tumor is applied based on the developed segmentation approach. Fig. 8 demonstrates the results of the method used to find the contours of the normal area (non-infected area) and the infected tissues (tumors).

Table 2. Performance analysis of the suggested architectures and cutting-edge techniques

Source	Methods	F1-score	IoU
Soulami et al. [24]	End-to-end UNet	90.5	--
Baccouche et al. [25]	Connected-ResUNets	86.91	90.82
Zhuang et al. [26]	Residual-Dilated-Attention-Gate-UNets (RDAU-NET)	82.00	--
Singh et al. [27]	UNet	88.30	--
Singh et al. [27]	SegNet	50.40	--
Proposed Architecture	FCMDWT Segmentation	96.47	97.34

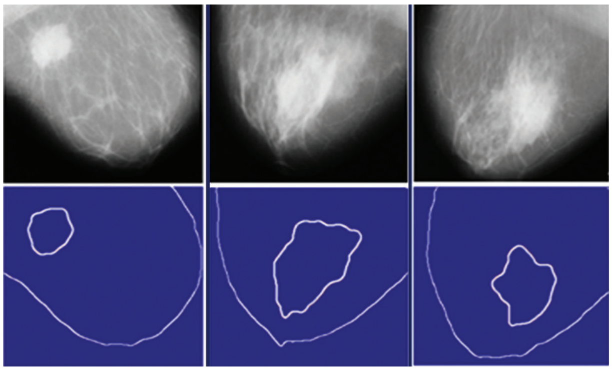


Fig. 8. Results of contour detection of normal breast and tumor, where the first line indicates the source images and the second line indicates the boundary detection of tumor and the breast based on developed FCMDWT segmentation approach

While SegNet and UNet were the best segmentation models in terms of dice scores of breast mass on mammogram images, we compared the developed method in this article to UNet and SegNet, and FCMDWT's qualitative predictions were more accurate, as seen in Fig. 9. The validation set, consisting of 30 volumes with 1325 2D slices, was used to compare the two predicted segmentations.

The Mann-Whitney test was employed to determine if there was a statistically significant difference between the predicted segmentation approach implemented by multiple methods and the ground truth labels. The predicted segmentation using FCMDWT (p-value = 0.13) was identical to the actual segmentation. According to this investigation, there was no statistically significant difference between the two approaches' projected segmentation and ground truth labels (p-value = 0.112 and 0.047, respectively). This demonstrates that the segmentations generated by these three approaches are distinct from one another. The loss, and accuracy of segmentation performance are provided in Table 3 to highlight the quantitative variations in accuracy, loss, and p-values between the actual results and those predicted by the U-test.

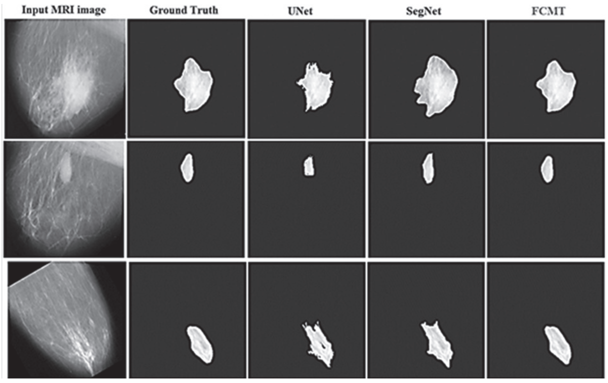


Fig. 9. The results of UNet and SegNet, FCMDWT qualitative predictions segmentation

Table 3. The accuracy iou, loss and p-value comparison between developed method fcmdwt, unet, and segnet

Methods	Accuracy (IoU)	Loss (Binary Cross-Entropy)	P-Value
FCMDWT	93.85	0.01	0.153
UNet	76.15	0.065	0.112
SegNet	68.54	0.073	0.047

The processing time of the proposed methods is not provided because some of the methods are machine learning-based while others are not (FCMDWT). Therefore, a comparison of these different methods is not applicable, since some contain a training time while others do not.

Most of the methods presented in recent years are based on machine learning, although machine learning models can produce extremely good results, they also have some limitations. The main factor affecting machine learning algorithms is data [24-27]. Class imbalance is a major problem that researchers face in almost every field. To overcome this problem, there are some methods that can be used to augment the data or there are methods that can be used to evaluate the outcomes of the algorithm. In our proposed method, the class imbalance problem does not limit the approach since it is not a machine-learning method.

Our method increases the precision of contour detection of the target object (tumor region and normal breast). Furthermore, the methods of UNet and SegNet have a lower percentage of pixels which are mistakenly thought to be the margins of breast cancers.

5. CONCLUSION

Initially, the FCM clustering algorithm is used to partition the normal area of the breast. Furthermore, the tumor zone is segmented using the thresholding approach. By integrating these methodologies, the research shows that the boundary or perimeter map of infected and non-infected parts of the breast may be determined with enhanced accuracy and precision. This method has several potential uses in clinical practice. Enhancing the precision of breast cancer extraction and localization may facilitate the identification and treatment of tumors by medical experts. Furthermore, it has the potential to decrease the need for intrusive medical procedures like biopsies and operations, which may be expensive, time-consuming, and have inherent dangers and adverse consequences. In summary, the research demonstrates that the use of machine learning and advanced image processing approaches can enhance the exactness and accuracy of breast cancer diagnosis. As these methodologies progress and improve, they possess the potential to transform clinical practice and enhance patient results. The experimental findings clearly show that the technique suggested in this work produces robust estimators that exhibit excellent picture quality for examination by medical professionals. Radiologists, who are medical specialists, assessed the edge maps, segmentation, and measurements found in cases of breast tumor pathology. The accuracy achieved by the devised segmentation and measuring approach surpasses the estimates made by similar specialists. An IoU of 98.41 and an F1-score of 96.47 were achieved. The experiment demonstrated the efficacy of the FCMDWT approach in performing edge detection, even in the presence of high levels of noise. The developed technique can also be utilized to detect lung pathology associated with COVID-19 infections with a few minor alterations. The created technology can be applied to lung segmentation, CT image pathology, and other areas where it is possible to identify cancerous cells. The authors of this study declare that the suggested method can pick up more features, making it simpler to identify the type of infected area.

6. REFERENCES:

[1] D. R. Nayak, N. Padhy, P. K. Mallick, D. K. Bagal, S. Kumar, “Brain Tumour Classification Using Noble Deep Learning Approach with Parametric Optimization through Metaheuristics Approaches”, Computers, Vol. 11, No. 1, 2022, p. 10.

[2] H. D. Cheng, J. Shan, W. Ju, Y. Guo, L. Zhang, “Automated Breast Cancer Detection and Classification Using Ultrasound Images”, Pattern Recognition, Vol. 43, 2010, pp. 299-317.

[3] L. Luo, et al., “Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions”, IEEE Reviews in Biomedical Engineering, 2024. (in press)

[4] F. Zahedi, M. K. Moridani, “Classification of Breast Cancer Tumors Using Mammography Images Processing Based on Machine Learning: Breast Cancer Tumors Using Mammography Images”, International Journal of Online and Biomedical Engineering, Vol. 18, No. 5, 2022, pp. 31-42.

[5] A. A. Alhussan, A. A. Abdelhamid, S. K. Towfek, A. Ibrahim, L. Abualigah, N. Khodadadi, A. E. Ahmed, “Classification of Breast Cancer Using Transfer Learning and Advanced Al-Biruni Earth Radius Optimization”, Biomimetics, Vol. 8, No. 3, 2023, p. 270.

[6] M. Qasim, T. Abed Mohammed, O. Bayat, “Breast Sentinel Lymph Node Cancer Detection from Mammographic Images Based on Quantum Wavelet Transform and an Atrous Pyramid Convolutional Neural Network”, Scientific Programming, Vol. 2022, 2022, pp. 1-13.

[7] A. Zotin, Y. Hamad, K. Simonov, M. Kurako, A. Kents, “Processing of CT Lung Images as a Part of Radiomics”, Proceedings of the 12th KES International Conference on Intelligent Decision Technologies, Singapore, 17-19 June 2020, pp. 243-252.

[8] Y. Hamad, O. K. J. Mohammed, K. Simonov, “Evaluating of Tissue Germination and Growth Rate of ROI on Implants of Electron Scanning Microscopy Images”, Proceedings of the 9th International Conference on Information Systems and Technologies, Cario, Egypt, 24-26 March 2019, pp. 1-7.

[9] A. Ciobotaru, D. Gota, T. Covrig, L. Miclea, “Reliable Breast Cancer Detection from Ultrasound Images

Using Image Segmentation", Proceedings of the IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, Romania, 16-18 May 2024, pp. 1-5.

[10] H. Hou, C. Zhang, F. Lu, P. Lu, "Breast Cancer Pre-Diagnosis Based on Incomplete Picture Fuzzy Multi-Granularity Three-Way Decisions", International Journal of Intelligent Computing and Cybernetics, Vol. 17, No. 3, 2024, pp. 549-576.

[11] Y. A. Hamad, K. Simonov, M. B. Naeem, "Breast Cancer Detection and Classification Using Artificial Neural Networks", Proceedings of the 1st Annual International Conference on Information and Sciences, Fallujah, Iraq, 20-21 November 2018, pp. 51-57.

[12] T. A. Kumar, G. Rajakumar, T. A. Samuel, "Analysis of Breast Cancer Using Grey Level Co-Occurrence Matrix and Random Forest Classifier", International Journal of Biomedical Engineering and Technology, Vol. 37, No. 2, 2021, pp. 176-184.

[13] J. A. Basurto-Hurtado, I. A. Cruz-Albarran, M. Toledano-Ayala, M. A. Ibarra-Manzano, L. A. Morales-Hernandez, C. A. Perez-Ramirez, "Diagnostic Strategies for Breast Cancer Detection: From Image Generation to Classification Strategies Using Artificial Intelligence Algorithms", Cancers, Vol. 14, No. 14, 2022, p. 3442.

[14] E. A. Zanaty, "Determination of Gray Matter (GM) and White Matter (WM) Volume in Brain Magnetic Resonance Images (MRI)", International Journal of Computer Applications, Vol. 45, No. 3, 2012, pp. 16-22.

[15] S. F. Ameer, Z. T. Nayyef, Z. H. Fahad, I. R. Niama ALRubee, "Using Morphological Operation and Watershed Techniques for Breast Cancer Detection", International Journal of Online & Biomedical Engineering, Vol. 16, No. 5, 2020.

[16] Y. A. Hamad, M. E. Seno, M. Al-Kubaisi, A. N. Safonova, "Segmentation and Measurement of Lung Pathological Changes for COVID-19 Diagnosis Based on Computed Tomography", Periodicals of Engineering and Natural Sciences, Vol. 9, No. 3, 2021, pp. 29-41.

[17] M. G. Kanojia, S. Abraham, "Breast Cancer Detection Using RBF Neural Network", in Proceedings of the 2nd International Conference on Contemporary Computing and Informatics, Greater Noida, India, 14-17 December 2016.

[18] L. Liu, K. Li, W. Qin, T. Wen, L. Li, J. Wu, J. Gu, "Automated Breast Tumor Detection and Segmentation with a Novel Computational Framework of Whole Ultrasound Images", Medical & Biological Engineering & Computing, Vol. 56, No. 2, 2018, pp. 183-199.

[19] I. Singh, K. Sanwal, S. Praveen, "Breast Cancer Detection Using Two-Fold Genetic Evolution of Neural Network Ensembles", Proceedings of the International Conference on Data Science and Engineering, Cochin, India, 23-25 August 2016, pp. 1-6.

[20] Y. Hamad, J. Kadum, A. Rashed, A. Safonova, "A Deep Learning Model for Segmentation of COVID-19 Infections Using CT Scans", AIP Conference Proceedings, Vol. 2398, 2022.

[21] S. Dalmiya, A. Dasgupta, S. K. Datta, "Application of Wavelet-Based K-Means Algorithm in Mammogram Segmentation", International Journal of Computer Applications, Vol. 52, No. 15, 2012.

[22] S. M. Badawy, A. A. Hefnawy, H. E. Zidan, M. T. GadAllah, "Breast Cancer Detection with Mammogram Segmentation: A Qualitative Study", International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.

[23] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5-9 October 2015, pp. 234-241.

[24] K. B. Soulami, N. Kaabouch, M. N. Saidi, A. Tamtaoui, "Breast Cancer: One-Stage Automated Detection, Segmentation, and Classification of Digital Mammograms Using U-Net Model-Based Semantic Segmentation", Biomedical Signal Processing and Control, Vol. 66, 2021, p. 102481.

[25] A. Baccouche, B. Garcia-Zapirain, C. Castillo Olea, A. S. Elmaghraby, "Connected-UNets: A Deep Learning Architecture for Breast Mass Segmentation", NPJ Breast Cancer, Vol. 7, No. 1, 2021, pp. 1-12.

[26] Z. Zhuang, N. Li, A. N. Joseph Raj, V. G. Mahesh, S. Qiu, "An RDAU-NET Model for Lesion Segmentation in Breast Ultrasound Images", PloS ONE, Vol. 14, No. 8, 2019.

[27] V. K. Singh, H. Rashwan, M. Abdel-Nasser, M. Sarker, F. Akram, N. Pandey, S. Romaní, D. Puig, "An Efficient Solution for Breast Tumor Segmentation and Classification in Ultrasound Images Using Deep Adversarial Learning", arXiv.1907.00887, 2019.

[28] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, D. L. Rubin, "A Curated Mammography Data Set for Use in Computer-Aided Detection and Diagnosis Research", Scientific Data, Vol. 4, No. 1, 2017, pp. 1-9.

[29] L. J. Guo, "Balance Contrast Enhancement Technique and Its Application in Image Colour Composition", Remote Sensing, Vol. 12, No. 10, 1991, pp. 2133-2151.

[30] C. Ortiz-Toro, A. García-Pedrero, M. Lillo-Saavedra, C. Gonzalo-Martín, "Automatic Detection of Pneumonia in Chest X-Ray Images Using Textural Features", Computers in Biology and Medicine, Vol. 145, 2022, p. 105466.

[31] E. M. Kabaev, Y. A. Hamad, K. V. Simonov, A. G. Zotin, "Methods of Interpretation of Data from Isokinetic Tests and MRI Studies During Rehabilitation of Patients After Reconstructive Shoulder Joint Surgery", International Archives of the Photogrammetry, Remote Sensing and Spatial Information, 2021, pp. 91–97.

[32] Y. A. Hamad, A. Alekhina, T. Pleshkova, O. Shestakova, "Infusion Extraction and Measurement on CT Images Based on Computer Vision and Neural Network", BIO Web of Conferences, Vol. 84, 2024, p. 02006.

[33] Y. Sasaki, "The Truth of the F-Measure", https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf (accessed: 2024)

[34] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15-20 June 2019, pp. 658-666.

Asphalt Pavement Distress Detection by Transfer Learning with Multi-head Attention Technique

Original Scientific Paper

Ahmed Bahaaulddin A. Alwahhab *
Middle Technical University, Technical College of Management, Information Technology management Department
Baghdad, Iraq
ahmedbahaaulddin@mtu.edu.iq

Vian Sabeeh
Middle Technical University, Technical College of Management, Information Technology management Department
Baghdad, Iraq
viantalal@mtu.edu.iq

Ali Abdulmunim Ibrahim Al-kharaz
Middle Technical University, Technical College of Management, Information Technology management Department
Baghdad, Iraq
ali.al-kharaz@mtu.edu.iq

*Corresponding author

Abstract – *Roads and highways represent a crucial lifeline between communities in all countries. They have to be healthy enough for safe and effective transportation. The traditional ways of inspecting roads by human inspectors consume time, and the inspection results may be subjective. For this reason, researchers are motivated to automate pavement distress detection to help the road monitoring and maintenance process. Additionally, many researchers have tried to present models to detect distress on road infrastructure. However, these models face accuracy challenges and overfitting because of the nature and complications of distress images. This paper proposes a model that combines pre-trained VGG16 with a multi-head attention layer. The proposed paradigm began with smoothing as a pre-processing step to eliminate the granular effect of the asphalt gravel and make asphalt damage more distinct. Then, data augmentation was conducted to improve model generalization by adding various distress scenes to the dataset in geometric, color, and intensity cases. This work also contributes to the broader body of research by collecting a local dataset that contains three types of asphalt distress (cracks, potholes, and ruts). The proposed model was tested using three benchmarked datasets in addition to the locally collected one, and it showed efficiency in detecting asphalt distress using offline and real-time images. The model achieved an accuracy 1.00 in the Pavmentscapes dataset, outperforming the UNET model, and a fully connected network was trialed with the same dataset. With the Deep Crack dataset, our model scored an accuracy of 1.00.*

In contrast, ResNet achieved an accuracy of 0.72 on the same dataset. The NHA12D dataset was also used to test the proposed model and achieved an accuracy of 1.00, but the VGG16 without an attention layer used on that dataset scored only 0.64. All previous obvious tests prove that the proposed VGG16 and multi-head attention paradigm outperform the earlier models. Additionally, the proposed model has undergone a real-time test on local roads. The future directions are to try to make the self-attention mechanism more explainable and implement an attention layer for multi-scales.

Keywords: *Asphalt distress detection, Computer vision, Transfer learning, Multi-head attention*

Received: July 25, 2024; Received in revised form: October 26, 2024; Accepted: October 27, 2024

1. INTRODUCTION

As the asphalt pavement is the major part of the transportation infrastructure that leads to possible transportation operations, it is substantial in the long-term investment maintenance to ensure safety and prolonged useful pavement life. However, the empirically controlled system for monitoring schedules can no longer meet the demands in many global areas, such as long waiting times, unstable inspections, and low adequate verification. The asphalt pavement, which has a poor ability to resist huge impacts from climate and traffic loadings, will exhibit various distresses, such as the effects of friction conditions, climatic conditions, and traffic loading classifications [1]. Therefore, many developed countries aim to maintain roads and highways as strategic target to boost their economies and reduce poverty [2]. For example, in 2018, the Philippines set aside 11 billion dollars to maintain their national roads and bridges. In 2019, the US spent about 29 billion dollars on infrastructure, with highway and roadway infrastructure accounting for nearly 50% of federal transportation spending [3]. In comparison, China spent about 702.6 billion Yuan in 2023 on high-speed roads, a 12% increase from the year before [4].

Many factors, such as road aging, traffic loads, construction materials, lack of maintenance on time, and weather conditions, significantly impact pavement damage. As a result, Pavement distress rates can instrumentally amount to the risk of losing significant worth of pavement around the world because of restated upkeep rather than rebuilding. This information helped to strengthen the concerning asset management and to promote the reorientation of the interests to fewer resources while maintaining infrastructure by adopting life cycle production. Detecting pavement distress in good time plays a crucial role in eliminating the degradation of pavement surfaces [2]. Conventional detection primarily depends on manual techniques and is plagued by solid subjectivity, expensive implementation costs, and is time-consuming and unsuitable for quick detection. Pavement distress detection has been undertaken utilizing different image processing technologies (DIP) for nearly two decades. Almost all the aforementioned methodologies lay down some limitations without providing ways to improve them. Existing methodologies lack the capability to accurately model the entire spectrum of pavement distresses for asphalt suitability in either environmental factors or management. The DIP methodologies encompass processes such as edge detection [5], threshold segmentation [3], and morphological processing [4]. Advancements in DIP for pothole detection have been significant, yet the system still faces challenges in achieving impeccable accuracy and reliability in automatic pothole detection. Simultaneously, machine learning (ML) techniques utilized training classifiers – Naive Bayesian Classifiers (NBC), Support Vector Machines (SVM), and Artificial Neural Networks (ANN) – to identify diverse forms of pavement distress, including damages and cracks. The process involves these specific classifiers to learn how to recognize particular pavement defects by focusing on their distinctive features within images. This necessitates prior knowledge coupled with engineering expertise. However, the complexity inherent in feature extraction may occasionally impact detection accuracy, which could demand customization, such as refining an algorithm for a particular detection scenario [6]. Recent years have highlighted the significant efficiency of deep learning models in solving various computer vision problems, including object detection, image classification, and segmentation [7]. Image segmentation and pothole/crack detection tasks often employ Convolutional Neural Networks (CNNs) due to their ability to extract crucial features from asphalt images, such as texture, edges, and corners [8]. U-NET is the most common type of CNN used for pothole segmentation. It depends on the encoding-decoding feature, which uses a bounding connection inside neural architecture to save most of the information after the down-sampling process. The U-NET mechanism aided in preserving the spatial information from images, yielding an accurate segmentation process. U-NET works more accurately in pothole segmentation and has begun to be used widely by researchers in this field [9]. After the success of CNN, a new approach was developed called transfer learning (TL), which uses one performed task to implement another job. TL is a set of models already trained with a vast dataset related to the problem under investigation. These models allow researchers to build accurate models with much less training time and they need to be fine-tuned for the dataset of the specific problem.

Rangoli et al. 2018 explored a transformers-based object recognition model for distress detection purposes called (YOLOv4) which stands for You Only Look Once. The model demonstrates a noticeable performance as it was pre-trained based on the asphalt distress detection objective. The model registered a precision of about 0.87 in identifying asphalt distress from images. The crucial issue in that work is the quality of images taken by echoes, which can be attracted by various conditions like camera settings, position, speed of the vehicle, and degree of sunlight. All the mentioned factors may cause a decrease in the accuracy of any transfer model [10].

In 2024, Vinodhini et al. utilized a pre-trained AlexNet model modified by adding a final layer to detect asphalt distress from images. Despite that, the model achieved an accuracy of 0.96, but this work does not mention other metrics like precision and recall. Furthermore, the paper did not discuss the conditions that might affect the model’s efficiency, such as lighting conditions and weather effects in real-time implementations [11].

Apeagyei et al. (2023) proposed a deep convolution network (DCNN) pre-trained using TL to detect eight pavement distress classes. Seven models were used for comparison. Low false negative values specified the best models. However, despite their low general accuracy, various models performed well in detecting specific distress types. Researchers have noticed that image quality impacts model performance regarding prediction speed and precision [12].

Recently, a more efficient novel multi-head attention mechanism was proposed, which enhances the network's learning capability to focus on different locations separated in the feature space in parallel. The attention mechanism can improve the performance of machine learning tasks such as NLP, speech recognition, object detection, recommendation systems, and time series data tasks, for example, in forecasting and diagnosis [13]

. For example, Xue et al. (2021) used the multi-head attention layer with a transfer model in face expression recognition. The model relies on a multi-head layer to drop attention maps during learning. This technique makes the model focus on various local points in the face while ignoring the weak points or features [14].

Zhao et al. (2023) designed a multi-head attention based on two-stream EfficientNet. The proposed model architecture recognizes human actions. Their model consists of two streams utilizing EfficientNet-B0 to extract spatial/temporal features from the video. Then, the model incorporates multi-head attention to extract crucial key points from extracted features [15].

Hong et al. (2021) model consists of convolution extraction blocks and attention modules to detect COVID-19. The first two convolutional blocks are made up of two depth-wise separable convolution layers and a maximum pooling layer. The multi-head attention mechanism (MHAM) is used to extract effective feature information from COVID-19 X-rays and CT images. This mechanism allows the model to focus on different parts of the input image simultaneously, enhancing the ability to capture relevant features across various scales. The multi-head worked by taking various filtered CNN features and putting these features into a multi-head layer to get attention arrays for various image parts[16]. Accordingly, using the multi-head mechanism proposed in the Transformers architecture, a multi-head attention method can get the different channels that can effectively extract different features from the input. In our case, the multi-head attention mechanism can learn different features from the hidden vector to get the different features of the input as well as get the different features across the multiple hidden vectors projected in parallel as vectors at the same input and then reduce to the final number of features. Therefore, applying the multi-head attention mechanism to the detection model can enhance the feature extraction from the input data.

According to previous works, the poor performance of several deep-learning transfer models can be largely attributed to image quality. However, this sparked an ingenuity that helped formulate a new and effective strategy to enhance the distress detection of asphalt surfaces. Image degeneration results from different factors, such as being captured as a low-resolution image due to, for example, using mobile phone cameras under variable lighting and weather conditions. Moreover, the coarse texture of road surfaces and irrelevant objects (e.g., pedestrians, vehicles, or trees) may adversely influence the detection rate [14].

We conclude from all the above that the distress in asphalt pavements may cause a reduction in the service life of the road. Humans inspect the road by tradition; however, an objective and unbiased evaluation using computer vision techniques is crucial to aid human inspectors in decision-making. The main problem addressed in this paper is caused by the local variability in the surface texture, and the distress contributes to the difficulty of automatic detection. The difficulties also result from (1) the fast weather changes resulting in changes in road surface coloration or asphalt texture, (2) the various distress, and (3) the artifacts of the road expansion joints. Consequently, the following are the primary contributions of the current study:

1. Develop an accurate asphalt distress detection model using TL by improving the VGG16 model with a multi-head attention layer that consists of four heads. The multi-head attention layer focuses on crucial features extracted from the VGG16 model that formulates the pattern of asphalt distress. The attention layer with a pre-trained VGG16 model has been tested for the first time in this type of application.
2. Collect a local dataset for asphalt-damaged images from Baghdad streets. This dataset comprised three classes (cracks, potholes, and ruts). This dataset is the first national dataset, and its distinction comes from the uniqueness of crack shapes and potholes caused by the abnormally high temperature in Iraq, which may reach over 50 degrees Celsius.
3. Propose a series of pre-processing and augmentation of the dataset that are used to evaluate the proposed paradigm. These augmentation operations enlarge the dataset to contain images of the various intensity conditions.
4. Evaluate the model on real-time stream images to detect asphalt distress.

The rest of this paper discusses the following subjects. First, the related works and image pre-processing techniques are discussed, including the steps (smoothing, edge detection, and dilation). After that, the proposed model is illustrated in detail with results and discussions. The paper ends with a conclusion.

2. RELATED WORKS

In the past few years, researchers have conducted numerous computer vision-based studies with the specific aim of automatically identifying asphalt distress. These investigations employ various approaches, including Gabor filters [17], binary patterns [18], tree structure algorithms, and shape-based methods [19], among others. Although generally valuable, these methods require assistance to extract distinguishing characteristics from images to discern between non-cracked and cracked pixels. Furthermore, these techniques must enhance their ability to detect asphalt distress in real-world scenarios accurately, varying pavement textures and lighting conditions. Deep learning (DL), however, has demonstrated significant potential to address comparable problems and deliver superior accuracy results, notably through the utilization of DCNN equipped with TL – an approach that Gopalakrishnan et al. employed within the context of computer vision-based pavement distress detection [20]. After initial training with the ImageNet database, the DCNN detects pavement image cracks on Hot-mix asphalt (HMA) and Portland cement concrete (PCC) surfaces. The research achieved a significant increase in complexity by training a classifier using combined images of pavements featuring diverse surface properties - HMA and PCC. Optimal results are achieved when utilizing a single-layer neural network classifier that is pre-trained on ImageNet and trained with features from the DCNN. Employing pre-trained DCNN models for cross-domain image classification, a general approach, has proven to be efficient in computer vision-based automated pavement crack detection. However, certain drawbacks were also observed, like the inclusion of non-crack characteristics such as joints, the inhomogeneity of cracks, and diversity within surface texture, all compounded by background complexity.

In 2019, Liu and his colleagues proposed a Deep Crack CNN model. This innovative approach featured multilevel convolutional layers. Additionally, demonstrating their commitment to advancing research, they introduced an invaluable dataset termed 'Deep Crack.' The utilized model, a variation of the VGG architecture, employed its first 13 layers. Deep crack with augmented data emerged as the most exemplary tested model; it yielded unprecedented performance in experimental tests, with an F-score and precision both measuring at 0.96 and recall registering at 0.86. The sole constraint identified in this work was the need to supplement the dataset with additional non-crack images [21]. In their 2020 study [22], Fan et al. introduced a system of multiple DCNNs specifically designed for automated crack detection and measurement in pavements. These CNNs, working collectively, recognize patterns of small gaps within raw images. They combine these findings to produce not only an overfitting-reducing result but also a predictive probability map. The approach outperforms alternative methods, achieving superior precision, recall, and F1 scores in evaluations using two publicly available crack databases. The proposed algorithm also facilitates the length and width measurement for various types of cracks. However, the suggested model faced two limitations: firstly, the system failed to detect cracks from the video streaming as it necessitates a more extensive and diverse dataset on which to offer performance evaluation, and secondly, an improved functioning is required, i.e., there is a need to test the system using more data. The researchers assembled a dataset consisting of 21,000 images taken from three different nations containing four different crack types.

Mandal et al. (2020) used three pre-trained models to detect pavement distress. These models were Hourglass-104, CSPDarknet53, and EfficientNet. The CSPDarknet53 model received the highest F1 score (0.58). Hourglass-104 came in second with (0.48), and EfficientNet came in third with a score (0.43). When compared to such models, the YOLO-based CSPDarknet53 model performed quite well; however, it has some drawbacks in terms of shadow-related conditions. In addition, EfficientNet encountered difficulties when attempting to locate cracks in roads that were wet [3].

A TL method was presented by Li et al. [23] in their article from 2021. This approach was designed to solve the difficulty of varying model performance across various types of cameras as well as mounting positions in the context of pavement distress detection. The approach is comprised of two primary components: model transfer and data transfer. The use of a distress detection model in unfamiliar settings is made possible through components that significantly reduce the requirement for considerable training data by no less than 25%. Also, such an approach enhances model accuracy by an amazing 26.55% compared to traditional approaches. Yet, it is essential to keep in mind that the efficiency of the training model could be affected by differences brought about by the use of multiple cameras that capture a wide variety of data and settings. This might potentially limit the potential for the model to be generalized. It turns out that obtaining labeled data for new scenes is very necessary, but given the framework that we have suggested, this process could cost a significant amount of time and effort. The utilization of GANs in data synthesis and transfer could result in a potentially hazardous circumstance. Distress annotations that have been created could contain inaccuracies or errors, which is one of the consequences that might have adverse impacts on model performance. It is necessary to perform manual screening of synthesized images after the completion of GAN style transfer to mitigate that danger. Even though this can be time-consuming and may require the removal of some training data validities, this stage is crucial for achieving optimal results. Errors or inconsistencies in the model's initial labeling could negatively affect the quality of the synthesized images and, consequently, the model’s performance.

Smadi and Gosh (2021) used DL techniques, including YOLOv3 and Faster R-CNN, to perform the automatic categorization and identification of pavement problems from high-resolution 3D surface images. This was a powerful strategy. In terms of demonstrating robust performance, such models performed quite well, with an average precision rate for distress detection and classification reaching as high as 89.2% through Faster R-CNN. YOLO achieved an even higher level of efficiency, reaching 90.2%. A feasible alternative to manual Quality Assurance and Quality Control (QA/QC) methods is presented by the developed methodology. It reflects the outputs of QA/QC in an efficient manner, which is a big step towards streamlining operational procedures. One of the research's limitations is that its testing and training datasets are smaller than the image datasets that are typically used [24].

Abbas et al. (2021) used advanced image processing techniques to automate the detection of pavement distress like cracks and potholes. The proposed model employs various image processing techniques like mathematical morphology to identify cracks. In addition, the model used segmentation methods to improve crack detection, using dynamic segmentation techniques that relied on six segmentation algorithms. Their model outperforms ML models because of the dynamic optimization approach designed to handle noise better than traditional methods, allowing for more precise identification of crack patterns even in less-than-ideal imaging conditions. The proposed model can also detect the degree of curvature and make the model distinguish between potholes and cracks accurately. This model's limitations are associated with variability in lightning conditions, as the model's performance can be reduced under various lighting conditions. Second, environmental factors like strain lane marking may affect the view of the cracks. Third, the model's effectiveness depends significantly on the image's quality. Lastly, the model may not detect all types of pavement distress. These limitations resulted in the need for further collaboration to build a more generalized model [25].

During their research conducted in the year 2022, Zhu and colleagues [5] suggested using an Unmanned Aerial Vehicle (UAV) equipped with a high-resolution camera to collect pavement damage data. To train a dataset that contained images of pavements showcasing six different kinds of damage, they used three object-detection algorithms: YOLOv3, Faster R-CNN, and YOLOv4. The YOLOv3 algorithm produced a good performance with a mean average precision (MAP) score of 56.6%. This result considerably improves the effectiveness of non-destructive automated pavement condition evaluations. Yet, in order to have a comprehensive understanding of this study, additional information regarding dataset size throughout the training of the model is required. It is expected that the research would provide significant insights into the adaptability of trained models to a wide variety of types of pavement conditions and surroundings.

Yihan et al. (2022) introduced a new Transformer-based approach called LeViT for automatically classifying asphalt pavement images. LeViT's architecture incorporates convolutional layers, transformer stages, and two classifier heads. This method has been found to achieve excellent performance in terms of accuracy, precision, recall, and F1 score when tested on Chinese and German asphalt pavement datasets, surpassing the capabilities of existing state-of-the-art models. LeViT exhibits faster inference speed than the original Vision Transformer and other CNN-based models. Additionally, the paper proposes a visualization technique that combines Grad-CAM and Attention Rollout to enhance the interpretability of the results, while it does not provide information regarding overfitting [26].

Zheng et al. (2022) have contributed to collecting a benchmarked Pavementscapes dataset. Pavementscapes comprised 4000 images with a resolution of 1024 x 1024 pixels for each image. Several pre-trained DCNN models were examined. The CNN models used were variations from VGG16 to detect cracks, potholes, and ruts. The best model was the segmentation transform. The main limitation noticed while conducting this work is the inefficiency of detecting small damage instances [27].

Huang et al. (2022) collected a dataset of cracks called NHA12D. Their work is a comparison study between a set of state-of-the-art crack detection algorithms. The NHA12D dataset comprised 80 pavement images divided into 40 asphalt and 40 concrete images. Three models were tested using the proposed dataset: first VGG16, Deep Crack and ResNet 3. The Deep Crack model shows performance with a 90.3 recall and precision of 0.35 for asphalt cracks. Meanwhile, for concrete cracks, the recall was 0.96, and the precision was 0.25. The Huang et al. work's limitation was the failure to classify concrete joints from cracks [28].

Eslami et al. (2023) [29] examined the performance of DCNN classifiers in the context of automated pavement assessment. Among all the factors tested, using multi-scale inputs had the most significant positive impact, resulting in an average performance improvement of 20% as measured by the F-score. Interestingly, when distinguishing between road distress and non-distress classes, the CNN classifiers performed better on area-based objects (patches) than linear objects (cracks). The M-VGG19 model achieved the highest F-score and demonstrated reduced variation in classification accuracy across different class types. Additionally, adding more layers to shallow networks with fewer than four convolution layers improved classification accuracy, particularly for smaller objects. However, there are some limitations to consider in this study. Firstly, the paper should provide a more detailed explanation of the rules governing the DCNN classifiers used in the research. Secondly, it is essential to note that the study focuses exclusively on pavement assessment and does not explore the broader applications of DL algorithms. Furthermore, the paper must compare non-deep learning-based methods for classifying road objects. Lastly, the study should address the computational requirements and training time associated with the DCNN classifiers used in the research.

The model known as Crack Forest was designed by Shi et al. for crack detection using a function for specifying features relying on detecting cracks based on intensity inhomogeneity. The method is based on a random structure forest, which is an improved ML method that combines algorithms for learning patterns to make decisions without being programmed explicitly. The proposed algorithm is superior to the previous models. Crack forest based on the SVM classifier registered a precision of about 90.28%, recall 0.86, and F1 89.39%. The only limitation of this work is that video streaming was not taken into consideration [30].

All the previous researchers either implemented existing transfer models or proposed their own models. They implemented the models directly on the images without pre-processing steps. In contrast, the current study presents an innovative improvement on the pre-trained CNN VGG16 model by adding a multi-head attention layer with a vast number of augmented images.

Cano-Ortiz et al. (2024) focused on comparing various YOLOv5 variants. These models have been evaluated according to their efficiency in detecting the targeted objects. Their main contribution is a novel filtering post-processing mechanism. This filtering mechanism is used to reduce false positive detection by 20.5%. The proposed post-processing mechanism relies on a rule-based approach that includes several rules to reject overlapped detection cases. The proposed model was evaluated on two datasets and achieved a precision of about 0.56 and 0.57 for the RDD2022 and CPRI datasets, respectively. This study has limitations because it evaluated the model on the existing dataset, which may not accommodate real-world conditions. Also, this study focused on one type of distress, cracks, which means that this model may not assess all types of distress comprehensively. Lastly, the results were validated on an open dataset that raises concerns about the generalization capabilities of the proposed architecture [31].

A. Nasertork et al. (2024) designed a model to improve the detection of pavement distress inception. This work utilized a proposed image processing feature extraction with AI techniques. The proposed model combined a set of texture features such as Gray Level Co-occurrence Matrix (GLCM), Local Binary Patterns (LBP), and Histogram of Oriented Gradients (HOG). All these features are used as discriminators to detect pavement images. Various ML algorithms trained by the extracted features, including XGBoost (XGB), Logistic Regression (LR), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), and Convolutional Neural Networks (CNN) were used. The best classifiers that achieved accuracy higher than 90% were SVM, XGB, and KNN. The most crucial limitation of this work was the variation of the feature selected effectiveness that relied on the dataset itself. The datasets were limited, leading to the specific model, because the model's training and evaluation depended on the quality and diversity of the image dataset [32].

M.Guerrieri et al. (2024) employed a pre-trained DL model YOLOv3 detection algorithm, which is known because of its efficiency in real-time object detection. YOLOv3 utilized Darknet, which used 3x3 filters inspired by ResNet that efficiently detect small objects in real-time. The dataset includes a diverse range of pavement damage types. The variety in the dataset allows the model to learn and extract relevant features that distinguish between different types of distress, enhancing its classification capabilities. The limitation of YOLOv3 is its difficulty in managing scale variations, especially when detecting small or large objects. The model faced a challenge from relying on a public dataset for training and validation. While this dataset provided substantial data, it may not encompass [33].

K. Ijari et al. (2024) utilized the EfficientNetB3 architecture, one of the EfficientNet variations. This model is notable for its compound scaling method that optimally adjusts depth, width, and resolution. The EfficientNetB3 model achieved superior performance with fewer parameters than traditional models like ResNet. The model detects various types of distress, like cracks and potholes. The researchers utilized a Swin Transformer-based GAN to generate images of synthetic pavement cracks. This augmentation was crucial for improving the efficiency and accuracy of the pavement damage assessment process. The efficientNetB3 model, when combined with the SwinGAN data augmentation process, achieved impressive testing accuracy ranging from 76.7% to 78.2%. This paper addressed a set of limitations. First, data quality issues, such as poor data quality, can lead to inaccurate predictions and classifications. Second, the model's sensitivity to environmental factors: the model's performance can be adversely affected by environmental factors such as shadows, reflections, and road markings, which can introduce noise into the image data. Third, handling complex crack geometry because the study identifies the lack of existing models for complex crack topologies. Many CNN models face difficulties classifying cracks with irregular shapes, which can decrease their accuracy in real-world applications [34].

3. BACKGROUND

This research paper proposes a model for asphalt distress detection. The proposed model is based on a pre-trained model that relies on improving the VGG16 with a multi-head attention layer. Before that, a batch of processes was conducted to prepare the datasets prior to training the model. These processing procedures included a smoothing process and then data augmentation processes. The following sections illustrate the methods used in implementing the proposed system. These methods make the datasets more suitable for efficiently training the model to produce more accurate results.

3.1. SMOOTHING PROCESS

Smoothing, also known as averaging, is used to smooth any image by spatial filters to reduce sharp details in images. It is used to lessen the sharpness of irrelevant details in the image [35]. A bilateral, linear filter replaces each pixel's intensity with a nearby pixel's average weight. The bilateral smoothing can preserve edges at the same time. Each neighbor is weighted by spatial components, considering the distant pixels and the difference between pixels of various intensities. Their combination value ensures that only nearby similar pixels contribute to the final pixels of the same region. Bilateral works are based on Eq.1 for each pixel p and q, whose loop is nested within p. The equation relies on taking the central pixel p and its neighborhoods such that $|p\text{-}q| < 2\ \sigma s$, considering the contribution of pixels outside the range of σs is negligible because of the spatial kernel.

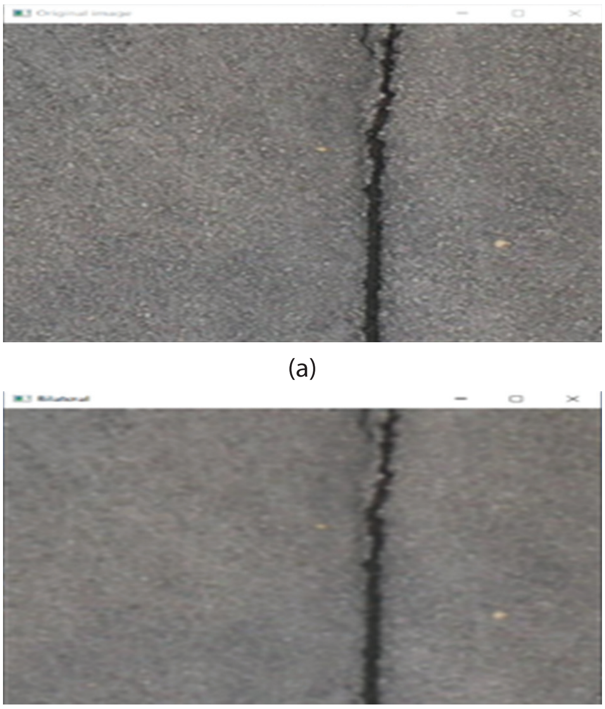
$$g\ \sigma s(||q-p||)g\sigma s(f(q)-f(p)).(f(q)) \tag{1}$$

It is used for unwanted texture removal. In our approach, the granular texture in asphalt must be removed [36]. (Fig. 1) illustrates how the smoothing preprocess has been applied to the original asphalt image.

3.2. TRANSFER LEARNING

TL models are those pre-build models trained using a specific dataset to be used later for building new models for various purposes using a different dataset. The *TL* principle relies on generating a model for one purpose and utilizing it for other activities. In these models, knowledge is gained from previously trained models on past tasks. As a result, this paradigm is beneficial when the data is limited because the limited data makes the model difficult to generalize. In addition, *TL* makes the model faster to train than developing the training model from scratch. The idiom of transfer learning comes from transferring existing knowledge to learn a new model with or without a labeled dataset [37]. (Fig. 2) summarizes the principle of *TL*.

The source dataset domain D_s and training task Tt, and target domain Dt with training model for task Tt. The target of *TL* is to use the knowledge in Ds and Tt to learn the targeted prediction model f in Dt given that $Ds \neq Dt$ and $Ts \neq Tt$ [37].



(a)

(b)

Fig. 1. Crack image **a**) Before smoothing, **b**) after smoothing

3.3. ATTENTION LAYER

The paradigm of using the attention layer was inspired by how humans penetrate a specific region of a scene. The attention layer works as a spotlight within the architecture of neural networks, specifying essential features in an image.

Therefore, the neural can adaptively adjust the neural weight according to the image features to learn from special regions in an image [38]. Special attention is widely used to focus on specific regions in any image. Generally, the operations of the attention layer consist of the following steps. First, compute addressing scores between various regions of the input image, such as pixels in an image. Second, weights are determined based on scores to indicate the importance of areas. Third, weight is used to refine the output, focusing on the most relevant features [39].

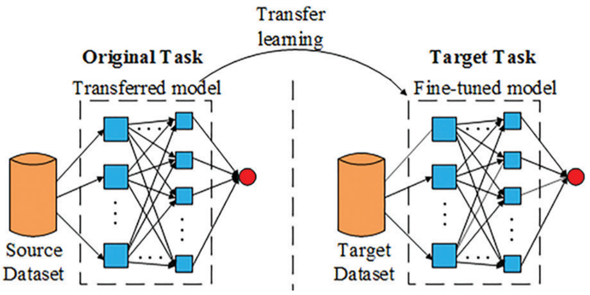


Fig. 2. TL for a new model [36]

Self-attention can be defined as an attention spatial filter applied to a single context or pixel instead of multiple contexts. So, queries, keys, and values are used to extract features from the spatial domain.

For example, in (Fig. 3) below, to extract the feature set of a pixel $xi, j \in R^{din}$.

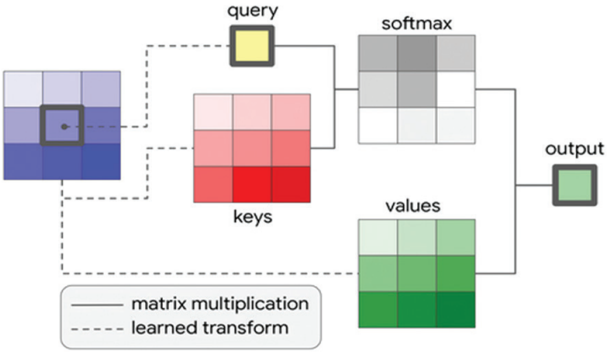


Fig. 3. The technique of attention layer [36]

As a spatial region, the attention layer has to extract the local regions of pixels in position around the specific pixel xi, j or $\in Nk\,(i, j)$. Then, the single-head attention process is computed by applying *softmax* on the query multiplied by the key to get the attention-focused features in the Eq. 2 below

$$yij = \sum_{ab\,\in Nk(i,j)} Softmax(qij^T kab)vab \tag{2}$$

Where qij^T is the query, kab is the key, and vab are values $Wvxab$, which represent the transformation of the pixel in position ij and the surrounding pixels. The softmax operation is conducted on all learned transforms. Self-attention works in a way that is similar to a spatial convolution filter by collecting information over the pixel and its neighbors, and the aggregation is done by using a convex combination of value vectors by using the softmax function [40]. This operation is repeated for every pixel in an image.

The advanced type of attention mechanism is the multi-headed attention that is illustrated in (Fig. 4) The multiple attention heads paradigm is used to gain various representatives for the input. The multi-head attention begins with partitioning the pixel features into N parts $xi, j \in R^{din/N}$ by conducting a single-head attention operation on each group separately with various transformations $W^n\ Q$, W^nK, $W^nv \in R^{dout/N\ x\ din/N}$ for each head. Then, concatenating the output from each head into one final output $yij \in R^{dout/N}$. CNN begins by extracting the features from the image, and then the attention layer maps the essential features using the weighted average of the values. Lastly, the multi-headed attention blocks output are concatenated into one feature set [41].

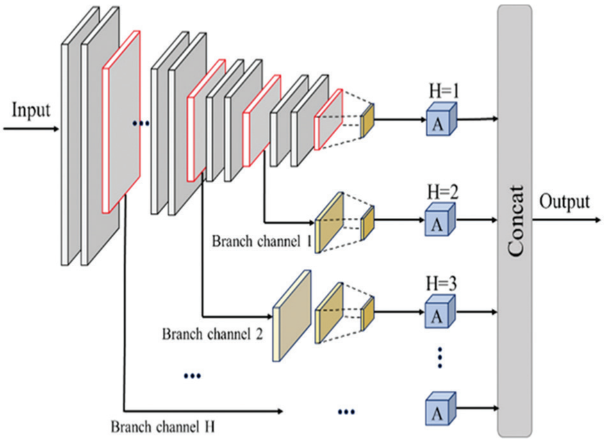


Fig. 4. The mechanism of multi-head attention layers [40]

3.4. DATA AUGMENTATION

Data augmentation is a technique used to generate new training samples from the existing seed of the dataset. This operation is like taking from the existing training samples and producing modified copies to train any classification model. There are various augmentation processes:

- Geometric transformations: These operations alter some geometric features in images like flipping images, cropping parts from an image, scaling an image (zooming in, zooming out), rotating to a specific degree, and shearing by distorting the image along an axis to rectify the perception angles.
- Color space augmentation: relying on modifying color within an image. The image is augmented by changing lights, saturation, and hue. Changing the colors within images can add realistic light variations and other color elements. This process makes the model less susceptible to overfitting.
- Noise injection is a technique in image augmentation that depends on adding a specific amount of noise to existing samples of images. This injects variations that simulate real effects or camera noise. For example, Gaussian noise, which is commonly used, is implemented by adding random values with normal distributive values to each image pixel. The intensity of noise is controlled by standard deviation [42].

4. THE PROPOSED MODEL

The proposed asphalt distress detection model has been inspired to detect three types of asphalt damage: cracks, potholes, and ruts. This system overcomes the issues noticed in previous works, as some systems have low accuracy or overfitting. The issues in accuracy came from datasets with a small number of images in each class or specific classes or poor-quality images. The proposed system consisted of stages. This work proves the system is free from overfitting because the model must be generalized to unseen data, not just memorize the training data. (Fig. 5) represents the proposed system stages.

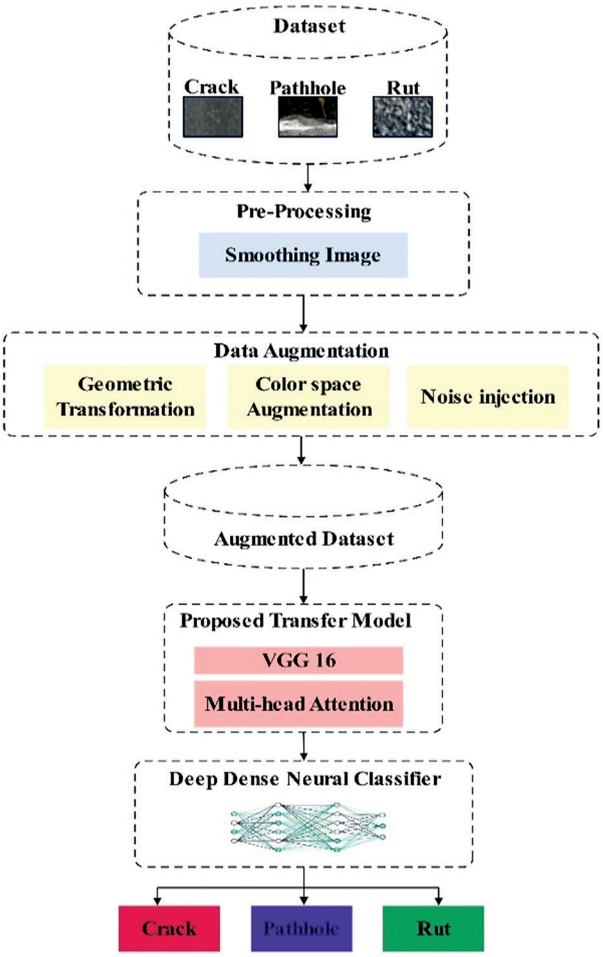


Fig. 5. General view of the proposed model

In Fig. 5, the model begins with the pre-processing stage to enhance the asphalt image quality. This enhancement is conducted by applying a smoothing operation to eliminate the granular shape of the gravel within the asphalt texture because pieces of gravel add noise to the image, especially when the system tries to detect cracks from non-cracked asphalt cases. The result of pre-processing is saved in a dataset repository.

The second stage is data augmentation to increase the number of training samples that make the system more generalizable. The augmentation process is useful in asphalt distress classification because it aids the classification model in being responsive to variations of the scenes in the real world. For instance, asphalt cracks may look different according to various lighting conditions. So, by implementing augmentation operations like cropping, flipping, blurring, and adding noise, we can make the dataset wider, containing various cases of images that will help the model learn as much distress in as many conditions as possible. (Fig. 6) Shows the steps for each of the augmentation operations

According to (Fig. 6), we begin the data augmentation with geometric transformation. The first process in geometric transformation is flipping; flipping the image from left to right horizontally helps the model to identify the distress pattern regardless of the image orientation. The second geometric transformation is cropping part of the image because the image does not always capture the entire area of interest. Consequently, cropping helps the classifier detect the distress area even when the damage does not occupy all the image scenes. The next operation is scaling the image by zooming in and zooming out. Zooming in can make the model focus on a magnified area. Zooming out lets the model learn the pattern from a more comprehensive view.

Consequently, these provide a broader context, allowing the model to learn to identify larger-scale distress features like potholes. The fourth process is the rotation of the image by a slight random angle. Rotation may simulate the variations in camera orientation. The last geometric transformation is shear, which tilts the image slightly in a specific direction, either horizontal or vertical. The next batch of operations is the color space augmentation, which begins with the flowing operations. First, Gaussian blur is an image augmentation technique widely used in computer vision tasks such as asphalt distress detection applications. Gaussian blur applies a Gaussian filter to the image, yielding a smooth image by blurring its details. The second is multiplying the image by a value greater than 1 to increase brightness.

In contrast, multiplying the image with a value less than 1 increases the darkness of the image. Third, contrast normalization is a data augmentation technique used in image processing to improve the overall contrast and visibility of features within an image. It is particularly helpful for DL tasks where models rely on extracting meaningful features from image data. The last augmentation process noise injection is implemented by adding Gaussian noise. It involves adding random noise following a Gaussian distribution to the image. This injects a controlled level of "artificial noise" that mimics real-world variations or sensor imperfections.

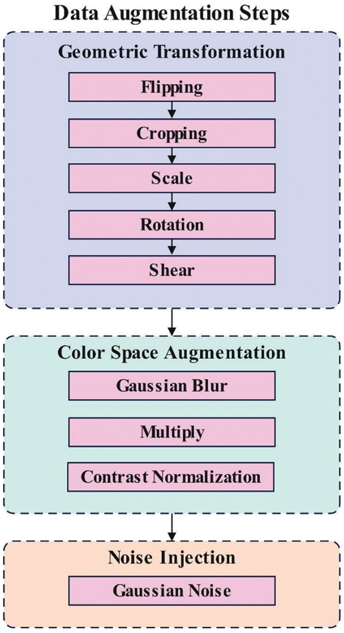


Fig. 6. Image augmentation stage process

The third stage is the proposed VGG16 model with a multi-head attention layer that has to be trained using the prepared dataset. (Fig. 7) illustrates the details of the improvement.

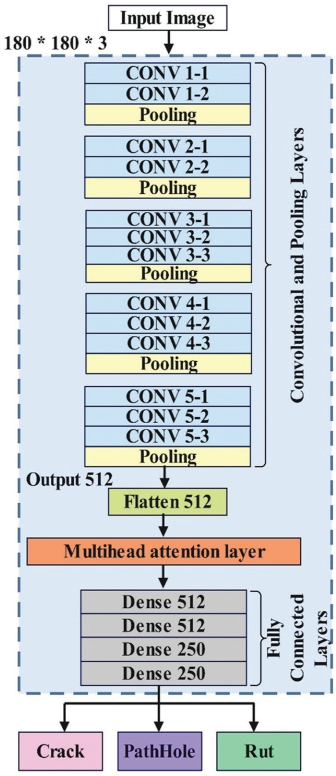


Fig. 7. VGG16 with multi-head attention layer structure

The image is size 180x180 and has three channels entering the model. The first part of the model is the VGG16, which consists of five layers. Each layer has CNN filters and one pooling layer. The output feature set from the VGG16 layer consists of a two-dimensional 512 array. Then, this feature set is flattened to be a one-dimensional array. The flattened feature set is the input to the multi-head attention layer that focuses on different parts in the feature set to extract the best representative for each image. The final output from the attention layer enters the final component of the deep dense neural network DDNN. This DDNN is used to classify the input image into one of three classes after learning from the features of each image during the training epochs.

(Fig. 8) illustrates the details of the mechanism used to extracting the features from the dataset's images. Initially the image was divided into four parts. Each of the parts entered the VGG16 transfer model. Extracting features from the image parts gives the attention process a richer understanding of the data. This can lead to more accurate and informative attention weights. In addition to that, dividing the image into parts allows the system to catch different aspects of the input image. By concatenating them, we allow the attention mechanism to consider these various aspects simultaneously, potentially leading to a more comprehensive understanding of the data. All the feature sets extracted by VGG16 are flattened into a 512 array. These feature sets enter the process of a multi-head attention layer that consists of four heads, one head for each part of the divided image. Accordingly, the number of iterations within the multi-head attention layer would be 16. Each element corresponds to a different region or aspect on the input image. Multiple heads allow the capture of various patterns. The multi-head attention layer begins the iteration 16 times by choosing arbitrary values of K and Q. Q: Represents the information you want to attend to, K Represents the information you want to attend with, and V: Represents the information you want to retrieve if there's a match between query and key. (Fig. 9) represents the details of each attention head. Fig. 9 shows the architecture of the single-head attention layer. The similarity between the query and key is scaled between +1 and -1, calculated by finding the dot product of two vectors. Multiplying the key (K) and query (Q) yields an attention filter.

$$A = QK^T \tag{3}$$

Then, scale the attention scores in the attention filter. The attention filter scores enter the softmax process to get more detailed crucial features, as in Eq. 4.

$$A_{softmax} = softmax(A) \tag{4}$$

After that, the attention filter is multiplied by the original image to remove unnecessary details. The features set is concatenated with the original image to obtain a more focused and detailed final image. The concatenated values are projected back to the original dimensionality using a projection matrix W_O:

$$Output = (X \parallel V(A\_softmax))\, W_Q \tag{5}$$

The multi-head attention allows one to focus on various parts of the image. So, each attention head outputs an attention filter that may focus on different details inside the image.

The proposed VGG16 with a multi-head attention layer has been developed using a mathematical model, and the details of the mathematical model are in the following steps:

4.1. CONVOLUTION (FEATURE EXTRACTION IN VGG16)

The VGG16 operations are repeated four times to get the feature set map. Suppose I represent the input image of a 3-dimensional tensor (height, width, channels). The filter W is the learnable weight of 3 dimensions. The convolution operation of VGG16 to extract the feature set is as follows:

$$O_{ij} = \Sigma\big(W_{khw} * I_{(i+k)(j+h)(c)}\big) + b \tag{6}$$

Where O_{ij} is the value of the output of position (*i, j*) in the feature map. W_{khw} is a single value from the filter W. $I_{(i+k)(j+h)(c)}$ represents a specific pixel value in the input image at a shifted position (*k, h*) within the kernel and channel (*c*). b is the bias term for that particular feature map.

4.2. POOLING

The pooling function is used to select more significant features by applying either average pooling or max pooling. Pooling also reduces the dimensionality of the feature set. The max pooling function is applied after conducting each CNN layer as in the equation.

$$O_{ij} = max(F_{(i+k)(j+h)}) \tag{7}$$

For all k, h within the window, where $F_{(i+k)(j+h)}$ is a single element in the features set.

4.3. MULTI-HEAD ATTENTION:

This process is used to extract the distress region pattern within the asphalt images and produce a feature set that is focused on the distressed parts in the asphalt. The attention process is done by conducting self-attention with many parts within the image to extract the more crucial part in deciding the image class. So, suppose X is the flattened feature vector extracted by the VGG16 pretrained model. Now, define three weight matrices, W_Q, W_K, and W_V, for projecting the input vector into a query (X_Q), key (X_K), and value (V) vectors, respectively.

The attention process is implemented as in the equation:

$$S = X_Q * X_K^{\ T} \tag{8}$$

Where S is the attention score.

After that, the softmax is applied to extract important features from the attention layer as in the equation:

$$A = softmax(S) \tag{9}$$

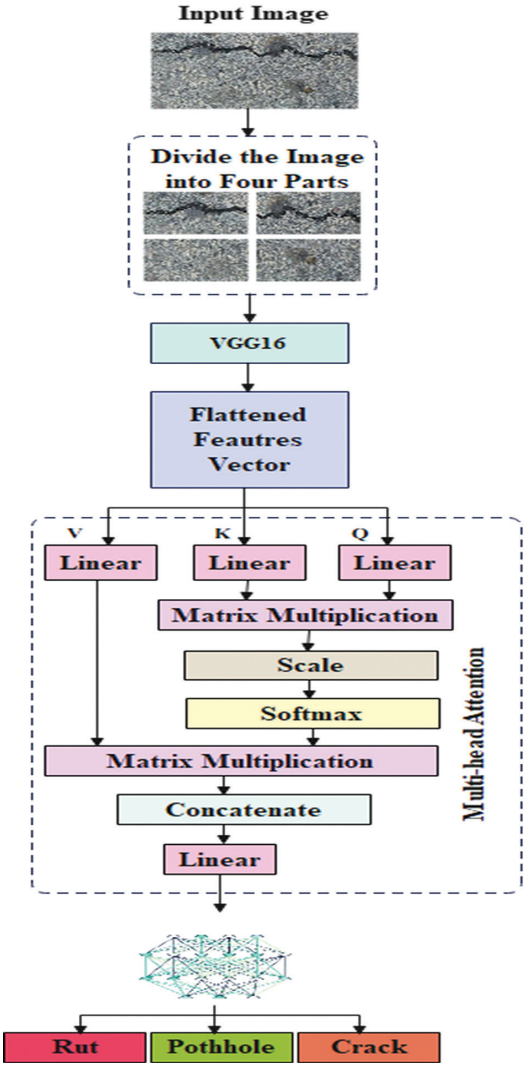


Fig. 8. Detailed architecture of the extraction features process in the proposed model

Now highlight the significant feature by multiplying the value of the image by the max-pooled feature set in the equation below:

$$\text{Context vector: } C = A * X_V \tag{10}$$

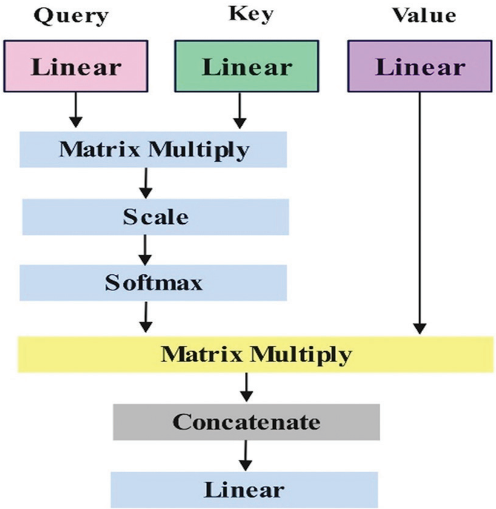


Fig. 9. Attention layer architecture

4.4. DENSE LAYERS

A deep dense classifier consists of four layers, trained over the features extracted from the multi-head attention layer to predict the input images later as cracks, potholes, or ruts. So, C represents the features vector, W is a set of weights, and B is the bias vector. The decision of an image is calculated by implementing the equation:

$$Y = ReLU(W * C + b) \tag{11}$$

So, the image is predicted by multiplying the W after training to the feature of an image after adding the bias vector.

4.5. SOFTMAX (OUTPUT LAYER)

SoftMax is implanted on Y to reach the final decision about the image class, which is either a crack, pothole, or rut. So, suppose Y is the output vector from the final dense layer. Then implement the softmax function for each class probability (i):

$$P(i) = exp(Z_i) \,/\, \Sigma(exp(Z_j)) \tag{12}$$

For all classes j, the last stage in the proposed model is the deep dense neural that accepts the feature set from the proposed VGG16 to decide the input image to which class it belongs.

5. EXPERIMENTAL RESULTS

This section presents the tests conducted on four datasets for various transfer models for computer vision problems. Initially, the experimental environment, datasets, and evaluation metrics must be explained.

5.1. EXPERIMENTAL ENVIRONMENT AND DATASET

The experiments were conducted using Python version 3.1.10 on Windows 10, CPU core I 7, and GPU Gforce 940 MX to accelerate the data training time while the transfer models are executed. Libraries like TensorFlow and Keras were used to build the proposed models. Three well-known benchmark datasets were utilized and divided into training and testing sets with a proportion of 0.8 training and 0.2 testing sets to evaluate the performance of the suggested models more precisely. The details of each dataset are as follows:

- The Pavementscapes dataset conducted by Zhang et al. contained 4000 images, each with a size of 1024*1024 pixels. The dataset was labeled with six classes related to asphalt detection. In this work, we use only the data related to cracks, potholes and ruts, which are 2300 in total. (Fig. 10) illustrates the three types of asphalt distress.
- The Deep Crack dataset contains 537 RGB color images, each of which is a fixed size of 544*304 pixels. The dataset was annotated manually and labeled into two classes: cracks and non-cracks.

- NHA12D dataset consists of 80 pavement images divided into 22 cracks and 58 normal images. Each image has a size of 1920*1080 pixels.
- Iraq asphalt dataset: This dataset was collected from Iraqi streets because there is a need for a national dataset because of the shape of the cracks and distress in this country. The asphalt distress in Iraq is produced by high temperatures over 50 degrees Celsius and unauthorized digging. The dataset images were collected using a mobile camera with 48 megapixels. Each image consisted of 3000 x 4000 pixels and was saved as a JPG file. The photos were labelled by the research group of this paper and consisted of 250 for each class (cracks, potholes and ruts). Anyone who wants the data should contact the authors.

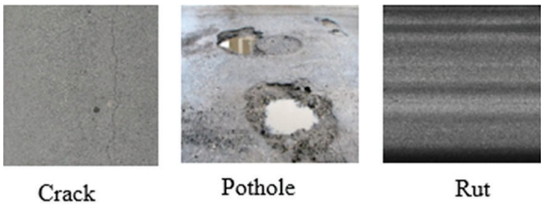


Fig. 10. Three types of asphalt distress under the study

5.2. EVALUATION METRICS

Four evaluation metrics were used to check the proposed model's efficiency. First, accuracy in Eq. (12) measures the model and predicts the outcomes correctly.

$$Acc = \frac{TP+TN}{all\ predictions} \tag{12}$$

Second, precision in Eq. (13) represents how often the model correctly predicts the positive class. Precision will be better when it is closer to 1.

$$precsion\ = \frac{TP}{TP+FP} \tag{13}$$

Third, recall in Eq. (14) measures how often classification learning correctly identifies positive instances of the positive class. Recall will be better when it reaches 1.

$$Recall = \frac{Tp}{TP+FN} \tag{14}$$

Lastly, the harmonic mean of precision and recall.

$$F1\ = 2 \times \frac{precsion\ \times recall}{precsion+recall} \tag{15}$$

We have to introduce the following idioms to understand the metrics used to evaluate the performance of the models. (Fig. 11) explains the components of the confusion matrix. True positive Tp represents when the classifier correctly predicts an instance related to a positive class. For example, when the model predicts an image holding crack damage to the crack class. True negative TN represents when the model correctly predicts an instance related to a negative class. For example, when a classifier predicts an input image without a crack as a normal image without damage. False positive FP means the classifier model incorrectly predicts a case as positive when it belongs to the negative class. For example, if a normal asphalt image is classified as a damaged case. False negative FN represents an error case. This happens when the predictor model mispredicts an instance as negative. For example, an image of damaged asphalt can be classified as normal.

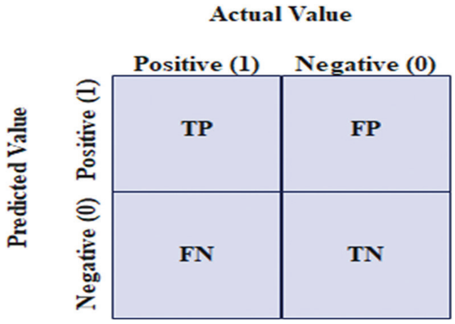


Fig. 11. Confusion Matrix Shape

6. RESULTS AND DISCUSSIONS

The performance of this proposed asphalt distress detection model was evaluated relying on matrices of accuracy, precision, recall, and F1 score. We depend on Macor's average accuracy with related matrices because the datasets are imbalanced. This research tested the possibility of overfitting by plotting the difference between training and validation accuracies and loss during the training epochs. Table 1 illustrates the performance of the proposed model.

The proposed pavement distress detection model was conducted on four benchmarked datasets, including the IRAQ asphalt dataset. The Pavementscapes dataset, consisting of three distress types (cracks, potholes, and ruts), has been used to evaluate the proposed model. The precision and recall for cracks both reached 1.0. At the same time, the general macro average precision and recall were 0.99. Pothole precision is 1.00, while the ruts precision reached 0.99. Ruts predicting achieved a true positive of about 0.99, as in the confusion matrix in (Fig. 12). The imbalanced data caused these differences between the precision values of classes.

Table 1. Performance Of Various Transfer Models

Dataset	Precision	Recall	F1	Accuracy
Pavementscapes	0.99	0.99	0.99	1.00
NHA12D	0.99	0.99	0.99	0.99
Deep Crack	1.00	1.00	1.00	1.00
Crack Forest	1.00	1.00	1.00	1.00
IRAQ dataset	0.96	0.96	0.96	0.96

The behavior of the proposed system towards the possibilities of overfitting was acceptable during the training process. (Fig. 13) registers the system's accuracy during training epochs, which refers to high training accuracy compared to the validation accuracy in the same training cycles. At the same time (Fig. 14) shows the difference between the validation loss and training loss. Validation loss was lower than the training loss in all the training epochs. An early stop mechanism was used to terminate the training process in five epochs to ensure achieved weights for the model.

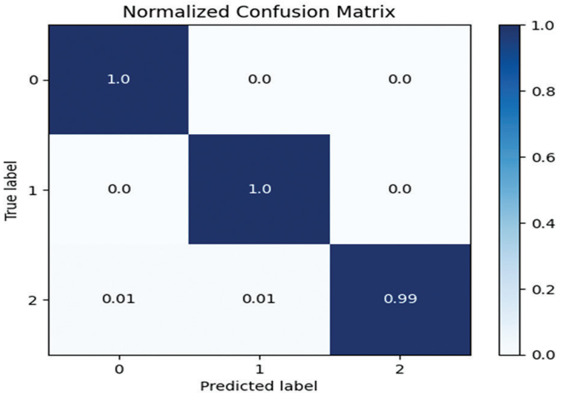


Fig. 12. Pavement scape dataset confusion matrix

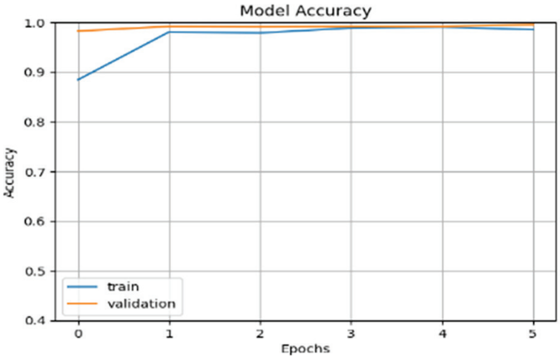


Fig. 13. Model training accuracy to validation accuracy for Pavement scape dataset

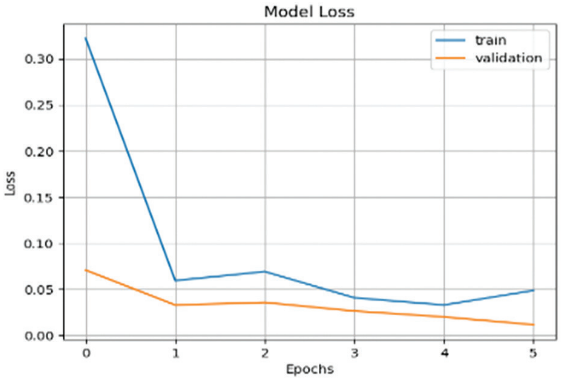


Fig. 14. Model training Loss to validation loss Pavement scape dataset

NHA12D was also used to check the validity of our paradigm, although this dataset consisted of two classes (crack, non-cracked) of asphalt images. The precision and recall were 0.99 despite the difference in precision between cracks and non-cracks classes. The model predicts crack classes with a precision reaching 0.98, while the precision for the non-cracked class was 1.00. This small difference is because of the imbalanced dataset. The confusion matrix is clear in (Fig. 15). The 0 label refers to the crack class, and the 0 class refers to the non-crack class.

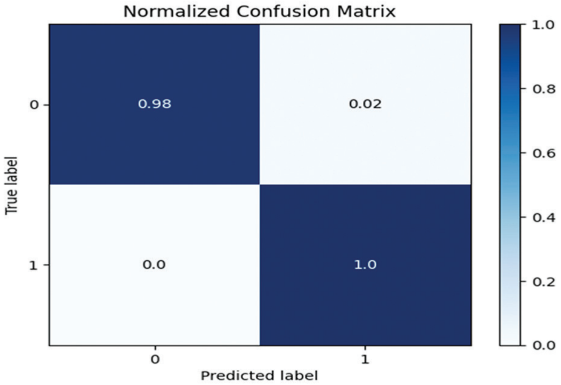


Fig. 15. Model confusion matrix on the NHA12D dataset

In this experiment, we also used an early stop mechanism; the model needed seven epochs for training. (Fig. 16) shows that the validation accuracy is close to the training accuracy during the training process. In contrast, the validation loss was lower than the training loss except for the last epoch before conducting the early stop, as shown in (Fig. 17). Model training accuracy. Early stop is used to prevent any possibility of overfitting or overtraining in the process.

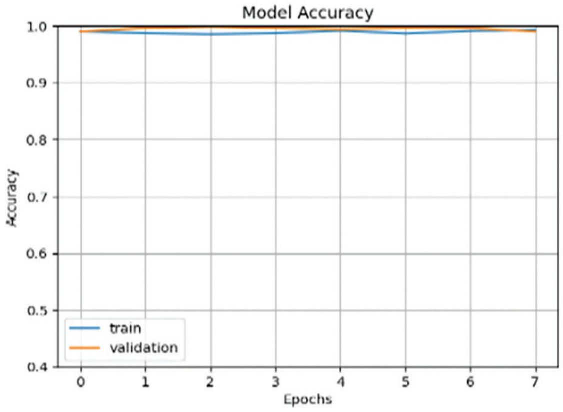


Fig. 16. Model training accuracy relate to validation for NHA12 Ddataset

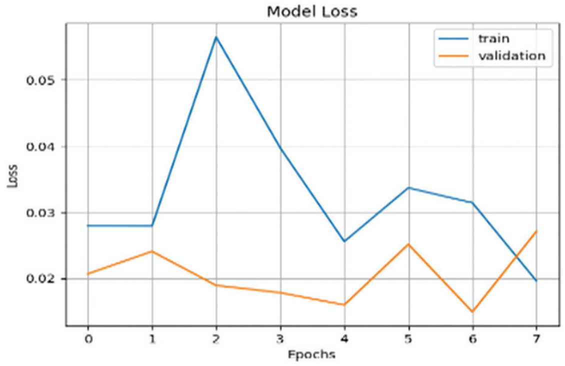


Fig. 17. Model training loss relate to validation loss for NHA12 Ddataset

The Deep Crack dataset also contains two classes (cracks and non-cracks). The cracks of multiple scales and scenes make this dataset a crucial benchmarked dataset to evaluate crack detection models. The dataset is relatively balanced, resulting in an efficient model for detecting cracks and normal asphalt without damage. The model achieved high precision and recall in this dataset, reaching 1.00 in the prediction of both classes. The confusion matrix is clear in (Fig. 18). Label (zero) represents the crack class, while Label (one) represents the non-cracks class in the confusion matrix.

The model's behaviour during the training was also investigated by plotting accuracy and loss. (Fig. 19) shows the accuracy during eight epochs of the training process. The X-axis in the figure represents the number of epochs, while the y-axis represents the accuracy. In all the training epochs, the validation accuracy was close to the training accuracy in a consistent trend between the two groups.

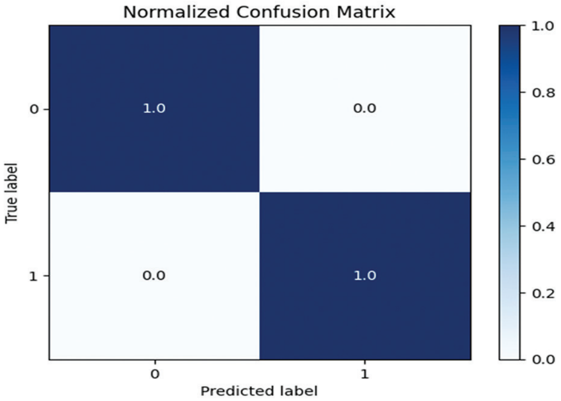


Fig. 18. Model confusion matrix for the deep crack dataset

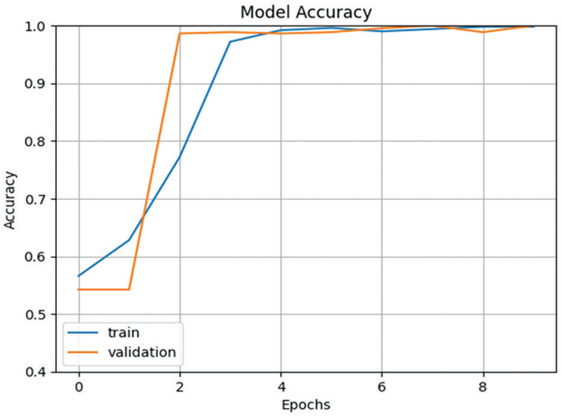


Fig. 19. Model training accuracy towards validation in deep crack dataset

In (Fig. 20), the validation loss is less than the training loss during the training process. Both accuracy and loss plots refer to the efficient behavior of the model in detecting new instances of cracks and non-cracks during the training and the trend of the model to stay away from overfitting in the working process.

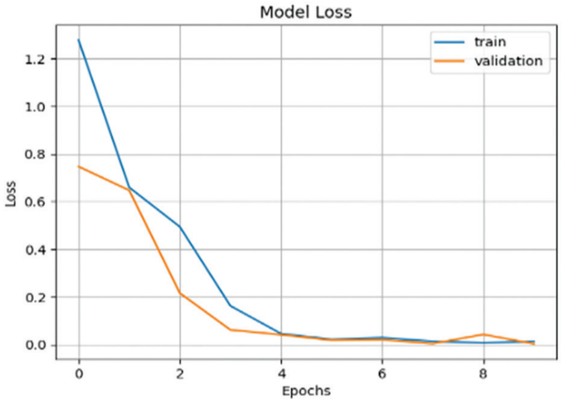


Fig. 20. Model training loss towards validation loss in deep crack dataset

The fourth test dataset was a Crack Forest that consisted of two classes (crack class and non_cracked class). The precision and recall for the cracked asphalt class were consistent and registered 1.0. The non-cracks class precision was 1.0; in contrast, the recall was 0.96. The low recall of the non-cracks class is due to the small number of images in the test set. The model's accuracy was 1.00, according to the confusion matrix in (Fig. 21).

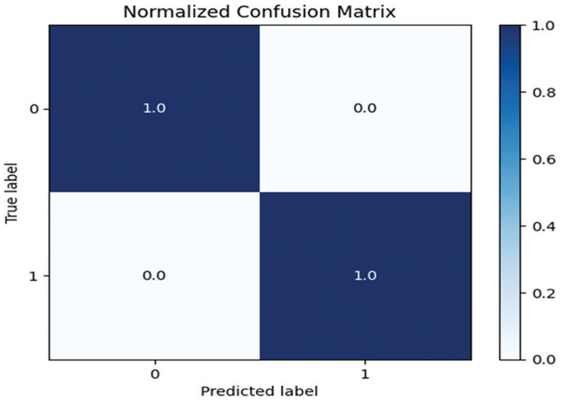


Fig. 21. Model confusion matrix on Crack Forest dataset

The model's behavior towards overfitting was apparent in the trend of increasing accuracy within training epochs, as in (Fig. 22).

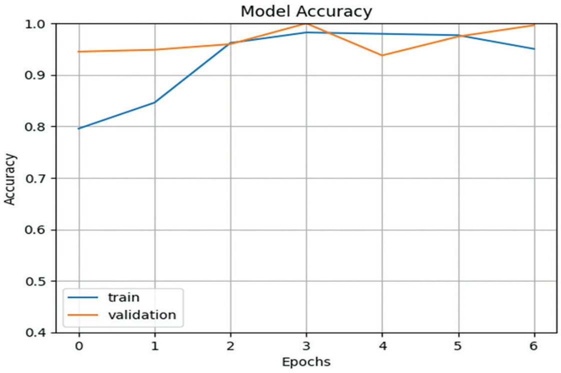


Fig. 22. Accuracy of the training model for crack forest dataset

Validation accuracy increases gradually during the model training. The loss of the validation also went lower than the loss of the training set except for the fourth epoch, which scored lower in the last two epochs, as in (Fig. 23).

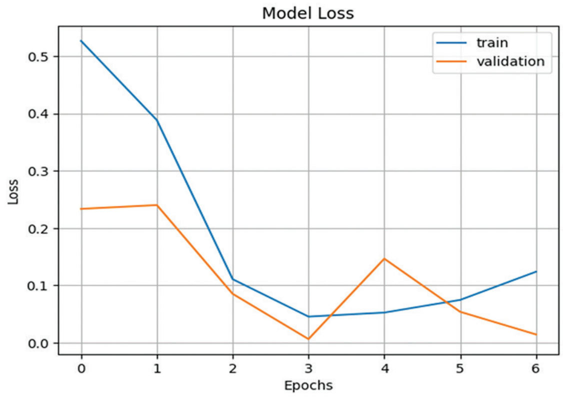


Fig. 23. loss of training model for crack forest dataset

The system has been tested on the Iraqi distress dataset. This dataset contains images taken under bright sunlight. Additionally, this dataset is distinguished by unique crack shapes due to the high temperature and drilling works, as mentioned. The system achieved an accuracy of 0.96. According to the confusion matrix in (Fig. 24), cracks were detected with an accuracy of 0.97. In contrast, potholes were detected with an accuracy of 0.91, and ruts were detected accurately in 1.0. The reason behind the low accuracy of pothole detection is that some overlap with cracks, as some images contain cracks and potholes at the same time. (Fig. 25). tracks the difference in training versus validation accuracy during the training process. In all 18 training epochs, the validation accuracy was close to the training accuracy.

In (Fig. 26), by comparing the loss validation to the loss of training in the 18 epochs, we can notice that the validation loss was generally less than the training loss, especially from epoch ten forward. Fig. 25 and Fig. 26. Refer to the fact that the system was far from entering the overfitting case within the training epochs.

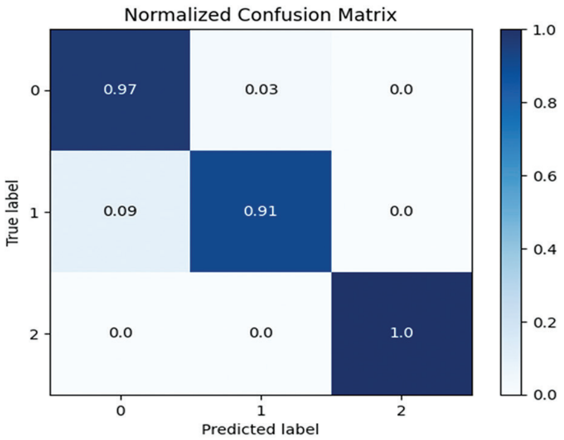


Fig. 24. Iraqi dataset confusion matrix

The system shows efficient behavior for both asphalt concrete distresses. The dataset Deep Crack contains the cracked concrete images used to train the proposed model in this research. (Fig. 27) represents a crack in concrete.

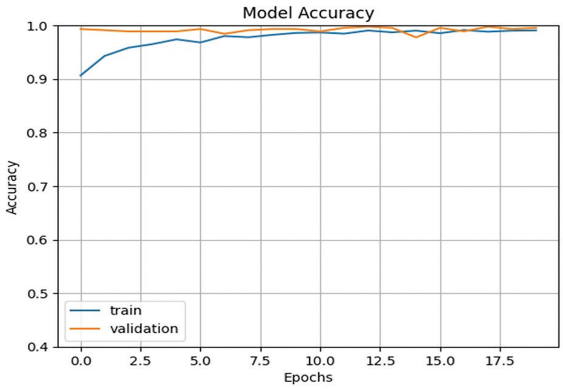


Fig. 25. Training accuracy of Iraqi dataset

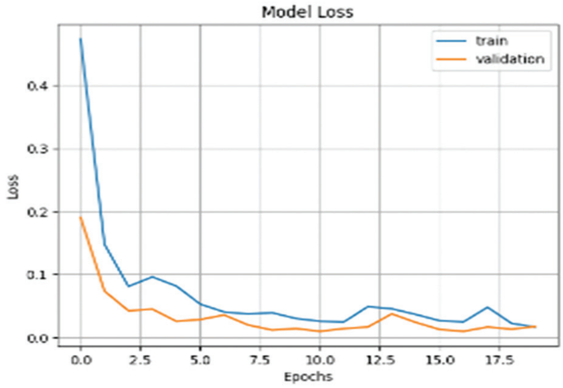


Fig. 26. Training loos of Iraqi dataset

Fig. 27. Crack in concrete surface

The proposed model efficiently detects asphalt distress under various circumstances like high lighting intensity, shadows, and the existence of traffic signs. This efficiency of the proposed model is due to training relying on augmented datasets and the successes in covering most scene cases like rotating image, flipping and shear, providing a variety of scenes and angles of the same image. Color augmentation provides the system with high or low-intensity images by multiplying processes and contrast normalization. Additionally, Gaussian blur generates images with noise. (Fig. 28) shows samples of the augmented images for both concrete and asphalt cracked areas. The augmentation process resulted in building a generalized system responsive to a wide range of crack, pothole and rut scenes. In the real world, the system was tested on Iraqi streets.

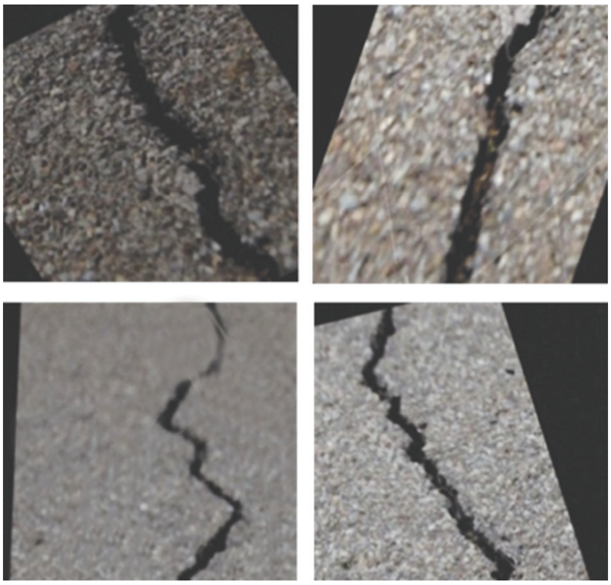


Fig. 28. samples of augmented images

The photos were captured using a web camera and sent directly to the system operated on a laptop. The system worked efficiently in the high-intensity light at noon and low-intensity light intensity near sunset. The time consumed was considerable, and the system took about 5 to 6 seconds to detect distress in the asphalt. The model is connected to a web cam of 1080p with 4k.

The execution time to detect the cracks increases gradually as the number of frames increases. (Fig. 29) represents the time of execution while the number of frames increased.

Additionally, the model was trained to detect special cases in Iraqi streets caused by insurgent drilling processes from people to establish water pipes through the streets, representing a crucial public problem in this country. (Fig. 30) presents one of the illegal drilling operations on Asphalt Street. We also compared the results of the proposed model with previous works that used the same benchmarked dataset adopted to evaluate this work. Table 2 compares those results with our proposed model's results.

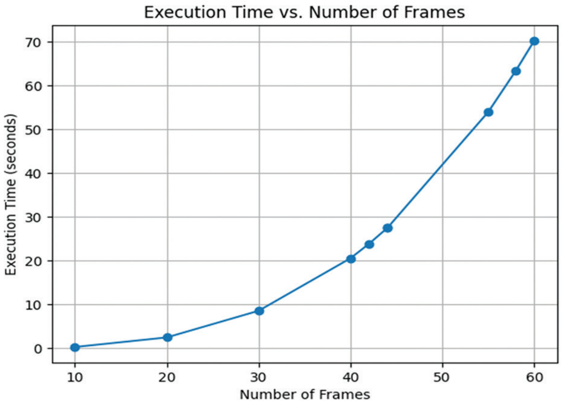


Fig. 29. proposed system execution time relate to the number of frames

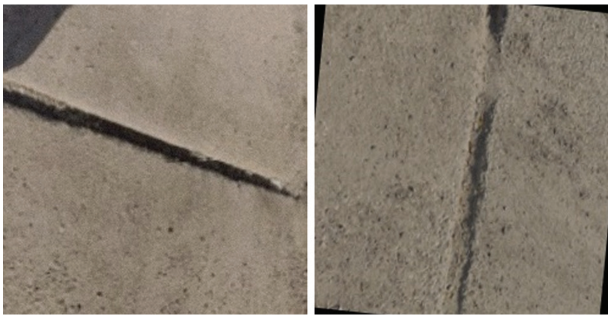


Fig. 30. cracks from illegal drilling from Iraqi streets

Zhang et al. registered 0.6 accuracy using the pavmentscapes dataset in their work. This low accuracy is caused by a wide variety of distress taken in their study model in addition to the noisy images that contain trees or other obstacles, such as environmental reasons like rain, snow, and sunlight, which may affect the quality of the image.

In contrast, our proposed model shows a performance accuracy of 1.00. The precision and recall in our model were higher than Zhang's work because we took three damage types to be predicted: cracks, potholes, and ruts. The second model used for comparison is Liu's Deep Crack model. Our model outperforms the Deep Crack model regarding precision, recall, and F-score metrics. Liu's model has about 0.87 precision, 0.85 recall, and 85.7 as the F-score, while our model got 1.00 for both precision and recall.

Table 2. compare the proposed model results with previous works

Paper	Dataset	Model	Accuracy	Precision	Recall	F-score
Zhang Tong	Pavementscapes	Segmentation transformer	0.60	0.4	0.73	8.42
Yahui Liu	Deep crack	Proposed model	1.00	0.99	0.99	0.99
Zhening Huang	NHA12D	Deep crack	-----	0.86	0.84	85.7
		Proposed model	1.00	1.00	1.00	1.00
		VGG16	----	0.35	0.90	0.5
		Proposed model	0.99	0.99	0.99	0.99
Shi et al.	Crack forest	Crack forest model	-----	0.82	0.89	95.68
		Proposed model	1.00	1.00	1.00	1.00

This difference in accuracy between the proposed model and the Deep Crack model is caused by a small number of non-cracked images in addition to the complexity of the surface that is coming from different asphalt textures and colors. The third work for comparison with the proposed system is for Huang et al. and the NHA12D dataset. Our model registered an average accuracy of 0.99 for the same dataset.In contrast, Huang used the straight VGG16 and achieved a precision of 0.35 and a recall of 0.90. The first reason behind the recall being higher than the precision in Huang's research is that the model is more biased toward the crack class than the non-crack class.

The second reason is that detecting concrete joints as cracks increases the false positive number, which leads to decreased precision. The Shi et al. model, called Crack Forest, achieved a precision of about 82.28 and a recall of 89.44, referring to the high number of false negative cases or high accuracy in detecting non-cracks in their model rather than cracks. Our model was implemented with the same dataset and achieved consistent precision and recall of 1.00, referring to balanced and efficient work predicting cracks and cracks cases. Furthermore, leveraging ImageNet pre-trained weights with the VGG16 model reduced the training time. In addition to this, VGG16 has efficient low-level feature crafting, helping the attention layer focus on high-level pattern understanding.

Our results are compared with a bench of baseline models such as (VGG16, ResNet, Unet, FCN, self-attention network, YOLOv8, YOLOv7, and RCNN). Table 3 presents a comparison of the main models used with the dataset used in this paper. These models were tested with the same datasets to test the proposed model. In [21], Liu et al. (2019) tried two models, VGG16 and ResNet. VGG16 achieved an accuracy of 0.30, while ResNet achieved an accuracy of 0.72 for the same dataset because ResNet has a residual connection that can collect deeper features without the gradient vanishing. Tong et al. [27] experimented with three models in their paper. The first model was UNET, which scored an accuracy of about 69.56.

Table 3. Compression table between a set of essential models

Model	Dataset	Accuracy	IOU
VGG16 [20]	Deep crack NHA12D	0.30 0.64	0.54
ResNet [20]	Deep crack	0.72	0.77
UNET [26]	Pavementscapes	69.56	54
FCN[27]	Pavementscapes	67	52
Self-attention network [27]	Pavementscapes	73.07	58.71
Yolov8 [43] Yolo v7 RCNN	RDD2022	78.4 57.8 49.4	
Texture feature extraction + Machine learning [28]	RDD2022	90.00	

The second model, the fully connected network FCN, was better, with 67% accuracy. The third model was the best, with the same dataset of Pavementscapes and an accuracy of 0.73. UNET shows low-performance returns due to noisy images in the Pavementscapes dataset containing shadows, traffic marks, etc. Therefore, FCN might work better than UNET in such cases. We noticed that a self-attention network was more effective because it can capture long-term dependencies. YOLOv8, YOLOv7, and RCNN were tested by Dong et al.(2024) [43] on the dataset, RDD2022. YOLOv8 outperforms YOLOv7 and RCNN with an accuracy of 78.4. YOLOv8 sometimes integrates label smoothing to regularize training and prevent overfitting.

In [29], their model combined a set of texture features such as GLCM, LBP, and HOG. Just the SVM, XGBOOST, and KNN classifiers gain an accuracy of over 0.90. The proposed model outperforms all mentioned models because the model incorporates VGG16 feature extracting with a multi-head attention layer that can understand long-range dependencies as one integrated feature set.

7. CONCLUSION

The infrastructure of roads and highways plays a vital role in the economy by connecting producers to markets and enabling more accessible transportation across regions and countries. Because of that, many countries are trying to enhance their transportation networks.

Detecting cracks and other damaged areas is essential because asphalt distress creates unseen surfaces that may increase the risk of accidents for drivers. Consequently, catching small cracks early is much cheaper than waiting for the damage to turn into major repairs.

As a result, the early detection of these issues allows repairs to be undertaken before damage worsens. Previous researchers have worked on designing CNNs to detect distress and damaged parts, while others have experimented with pre-trained models. However, their efforts have faced issues with accuracy because of an imbalanced dataset or the nature of the images.

The proposed system leverages the strength of VGG16 and the multi-head attention approach to focus on asphalt distress parts. VGG 16, a pre-trained CNN model on a massive dataset, extracts general features from images. Then, adding a multi-head attention layer makes the system focus on specific relationships between different parts of images. The proposed paradigm is beneficial for asphalt distress detection where the spatial context or how the cracks or potholes are internally connected is essential for accurate classification. For future work, a large, diverse dataset encompassing various climates and pavement types is needed to enhance model generalization. The models also require a lightweight architecture that enables real-time deployment on mobile devices for on-the-go road inspections. Last, the researchers must work on an explainable model that produces reasoning behind the classification or decision.

8. REFERENCES

[1] F. R. Bruinsma, S. A. Rienstra, P. Rietveld, "Economic Impacts of the Construction of a Transport Corridor: A Multi-level and Multiapproach Case Study for the Construction of the A1 Highway in the Netherlands", Regional Studies, Vol. 31, No. 4, 1997, pp. 391-402.

[2] J.-A. R. Sarmiento, "Pavement Distress Detection and Segmentation using YOLOv4 and DeepLabv3 on Pavements in the Philippines", arXiv:2103.06467, 2021.

[3] V. Mandal, A. R. Mussah, Y. Adu-Gyamfi, "Deep Learning Frameworks for Pavement Distress Classification: A Comparative Analysis", Proceedings of the IEEE International Conference on Big Data, Atlanta, GA, USA, 10-13 December 2020, pp. 5577-5583.

[4] L. Jia, S. Yang, W. Wang, X. Zhang, "Impact analysis of highways in China under future extreme precipitation", Natural Hazards, Vol. 110, No. 2, 2022, pp. 1097-1113.

[5] J. Zhu, J. Zhong, T. Ma, X. Huang, W. Zhang, Y. Zhou, "Pavement distress detection using convolutional neural networks with images captured via UAV", Automation in Construction, Vol. 133, 2022, p. 103991.

[6] N. G. Sorum, T. Guite, N. Martina, "Pavement Distress: A Case Study", International Journal of Innovative Research in Science, Engineering and Technology, Vol. 3, 2014, pp. 274-284.

[7] X. Chen, S. Yongchareon, M. Knoche, "A review on computer vision and machine learning techniques for automated road surface defect and distress detection", Journal of Smart Cities and Society, Vol. 1, No. 4, 2022, pp. 259-275.

[8] W. Tang, S. Huang, X. Zhang, L. Huangfu, "PicT: A Slim Weakly Supervised Vision Transformer for Pavement Distress Classification", Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10-14 October 2022, pp. 3076-3084.

[9] V. Pereira, S. Tamura, S. Hayamizu, H. Fukai, "Semantic Segmentation of Paved Road and Pothole Image Using U-Net Architecture", Proceedings of the International Conference of Advanced Informatics: Concepts, Theory and Applications, Yogyakarta, Indonesia, 20-21 September 2019, pp. 1-4.

[10] A. Ragnoli, M. R. De Blasiis, A. Di Benedetto, "Pavement Distress Detection Methods: A Review", Infrastructures, Vol. 3, No. 4, 2018, p. 58.

[11] K. A. Vinodhini, K. R. A. Sidhaarth, "Pothole detection in bituminous road using CNN with transfer learning", Measurement: Sensors, Vol. 31, 2024, p. 100940.

[12] A. Apeagyei, T. E. Ademolake, M. Adom-Asamoah, "Evaluation of deep learning models for classification of asphalt pavement distresses", International Journal of Pavement Engineering, Vol. 24, No. 1, 2023, p. 2180641.

[13] H. S. Sharaf Al-deen, Z. Zeng, R. Al-sabri, A. Hekmat, "An Improved Model for Analyzing Textual Sentiment Based on a Deep Neural Network Using Multi-Head Attention Mechanism", Applied System Innovation, Vol. 4, No. 4, 2021, p. 85.

[14] F. Xue, Q. Wang, G. Guo, "TransFER: Learning Relation-aware Facial Expression Representations with Transformers", Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada,11-17 October 2021, pp. 3581-3590.

[15] A. Zhou et al. "Multi-head attention-based two-stream EfficientNet for action recognition", Multimedia Systems, Vol. 29, No. 2, 2023, pp. 487-498.

[16] G. Hong, X. Chen, J. Chen, M. Zhang, Y. Ren, X. Zhang, "A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19", Scientific Reports, Vol. 11, No. 1, 2021, p. 18048.

[17] R. Shahabian, A. M. Moghaddam, S. A. Sahaf, H. reza Pourreza, "Second-Order Statistical Texture Representation of Asphalt Pavement Distress Images Based on Local Binary Pattern in Spatial and Wavelet Domain", Rehabilitation in Civil, Vol. 7, No. 3, 2019, pp. 48-67.

[18] M. Salman, S. Mathavan, K. Kamal, M. Rahman, "Pavement crack detection using the Gabor filter", Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems, Hague, Netherlands, 6-9 October 2013, pp. 2039-2044.

[19] A. Cubero-Fernandez, F. J. Rodriguez-Lozano, R. Villatoro, J. Olivares, J. M. Palomares, "Efficient pavement crack detection and classification", EURASIP Journal on Image and Video Processing, No. 2017, 2017, pp. 1-11.

[20] K. Gopalakrishnan, S. K. Khaitan, A. Choudhary, A. Agrawal, "Deep Convolutional Neural Networks with transfer learning for computer vision-based data-

driven pavement distress detection", Construction and Building Materials, Vol. 157, 2017, pp. 322-330.

[21] Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation", Neurocomputing, Vol. 338, 2019, pp. 139-153.

[22] Z. Fan et al. "Ensemble of Deep Convolutional Neural Networks for Automatic Pavement Crack Detection and Measurement", Coatings, Vol. 10, No. 2, 2020, p. 152.

[23] Y. Li, P. Che, C. Liu, D. Wu, Y. Du, "Cross-scene pavement distress detection by a novel transfer learning framework", Computer-Aided Civil and Infrastructure Engineering, Vol. 36, No. 11, 2021, pp. 1398-1415.

[24] R. Ghosh, O. Smadi, "Automated Detection and Classification of Pavement Distresses using 3D Pavement Surface Images and Deep Learning", Transportation Research Record, Vol. 2675, No. 9, 2021, p. 1359.

[25] H. Abbas, M. Q. Ismael, "Automated Pavement Distress Detection Using Image Processing Techniques", Engineering, Technology & Applied Science Research, Vol. 11, No. 5, 2021, pp. 7702-7708.

[26] Y. Chen, X. Gu, Z. Liu, J. Liang, "A Fast Inference Vision Transformer for Automatic Pavement Image Classification and Its Visual Interpretation Method", Remote Sensing, Vol. 14, No. 8, 2022, p. 1877.

[27] Z. Tong, T. Ma, J. Huyan, W. Zhang, "Pavementscapes: a large-scale hierarchical image dataset for asphalt pavement damage segmentation", arXiv:2208.00775, 2022.

[28] Z. Huang, W. Chen, A. Al-Tabbaa, I. Brilakis, "NHA12D: A New Pavement Crack Dataset and a Comparison Study Of Crack Detection Algorithms", arXiv:2205.01198, 2022.

[29] E. Eslami, H.-B. Yun, "Comparison of deep convolutional neural network classifiers and the effect of scale encoding for automated pavement assessment", Journal of Traffic and Transportation Engineering, Vol. 10, No. 2, 2023, pp. 258-275.

[30] Y Y. Shi, L. Cui, Z. Qi, F. Meng, Z. Chen, "Automatic Road Crack Detection Using Random Structured Forests", IEEE Transactions on Intelligent Transportation Systems, Vol. 17, No. 12, 2016, pp. 3434-3445.

[31] S. Cano-Ortiz, L. L. Iglesias, P. M. R. del Árbol, P. Lastra-González, D. Castro-Fresno, "An end-to-end computer vision system based on deep learning for pavement distress detection and quantification", Construction and Building Materials, Vol. 416, 2024, p. 135036.

[32] A. Nasertork, S. Ranjbar, M. Rahai, F. M. Nejad, "Pavement raveling inspection using a new image texture-based feature set and artificial intelligence", Advanced Engineering Informatics, Vol. 62, 2024, p. 102665.

[33] M. Guerrieri, G. Parla, M. Khanmohamadi, L. Neduzha, "Asphalt Pavement Damage Detection through Deep Learning Technique and Cost-Effective Equipment: A Case Study in Urban Roads Crossed by Tramway Lines", Infrastructures, Vol. 9, No. 2, 2024.

[34] K. Ijari, C. D. Paternina-Arboleda, "Sustainable Pavement Management: Harnessing Advanced Machine Learning for Enhanced Road Maintenance", Applied Sciences, Vol. 14, No. 15, 2024, p. 6640.

[35] R. C. Gonzalez, R. E. Woods, "Digital Image Processing", 3th Edition, Prentice-Hall, 2006.

[36] R. G. Gavaskar, K. N. Chaudhury, "Fast Adaptive Bilateral Filtering", IEEE Transactions on Image Processing, Vol. 28, No. 2, 2019, pp. 779-790.

[37] Q. He, Z. Xiang, P. Ren, "A CLSTM and transfer learning based CFDAMA strategy in satellite communication networks", Plos One, Vol. 16, No. 3, 2021, p. e0248271.

[38] M.-H. Guo et al. "Attention Mechanisms in Computer Vision: A Survey", Computational visual media, Vol. 8, No. 3, 2022, pp. 331-368.

[39] J. Gupta, S. Pathak, G. Kumar, "Deep Learning (CNN) and Transfer Learning: A Review", Proceedings of the International Conference on Applications of Intelligent Computing in Engineering and Science, Raipur, India, 12-13 Febreuary 2022, p. 012029.

[40] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, "Stand-Alone Self-Attention in Vision Models", Advances in Neural Information Processing Systems, Vol. 32, 2019.

[41] G. Hong, X. Chen, J. Chen, M. Zhang, Y. Ren, X. Zhang, "A multi-scale gated multi-head attention depthwise separable CNN model for recognizing COVID-19", Scientific Reports, Vol. 11, No. 1, 2021, p. 18048.

[42] C. Shorten, T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning", Journal of Big Data, Vol. 6, No. 1, 2019, p. 60.

[43] X. Dong, Y. Liu, J. Dai, "Concrete Surface Crack Detection Algorithm Based on Improved YOLOv8", Sensors, Vol. 24, No. 16, 2024, p. 5252.

Applying Artificial Intelligence Techniques For Resource Management in the Internet of Things (IoT)

Original Scientific Paper

Salwa Othmen*
Computers and Information Technology Department, College of Science and Arts Turaif, Northern Border University, Arar, Saudi Arabia
e-mail: Salwa.Othmen@nbu.edu.sa

Wahida Mansouri
Computers and Information Technology Department, College of Science and Arts Turaif, Northern Border University, Arar, Saudi Arabia

Radhia Khdhir
Department of Computer Science, College of Science and Arts in Qurayyat, Jouf University, Saudi Arabia

*Corresponding author

Abstract – *Internet of Things (IoT) applications in smart cities (SCs) rely on free-flow services streamlined by artificial intelligence (AI) paradigms. However, the nature of resource constraint prevails due to external infrastructure costs and energy-based allocations. Existing approaches to smart city resource distribution rely on static thresholds or reactive responses, which are not always sufficient. These approaches may limit system performance and scalability in dynamic IoT environments owing to increased energy consumption, postponed resource allocation, and frequent device failures. This article introduces a Concerted Resource Management (CRM) using the Leveled Reinforcement Training (LRT) method. The proposed method accurately identifies cost-complex and high energy-consuming sharing intervals based on service response time and device failure. The reinforcement learning and training concerts both energy and device incorporations for SC applications based on its demand. This process requires leveled training in resource management, from energy depletion to device activeness. The interrupted sessions are identified using resource allocation failures, and the active resources with optimal energy expenses are selected to pursue resource management. The training method thus identifies the demands based on independent or concerted resource allocations to mitigate the management constraints in an SC environment. This proposed method reduces the resource constraint-based waiting for allocations and allocation failures in any SC application services. Under the varying devices, the following is observed: Improvements: 9.1% (Allocation Rate), 10% (Device Detection), 11.88% (Constraint Mitigation—Energy), 9.06% (Constraint Mitigation—Resource Allocation); Reduced: 8.01% (Allocation Failure), 9.64% (Waiting Time).*

***Keywords**:* *Power flow control, three port converter, high frequency transformer, phase shift control technique*

Received: August 23, 2024; Received in revised form: November 24, 2024; Accepted: November 25, 2024

1. INTRODUCTION

Resource management is a process that ensures the network has all the required resources to complete or perform a task. RM is widely used in organizations to reduce unwanted difficulties in performing user tasks [1, 2]. IoT-based smart city environments are commonly implemented to enhance the overall performance level of the network [3, 4]. The analyzed data produces optimal features for resource allocation in smart cities. The management framework improves the services' lifespan, maximizing the performance range of smart cities [5, 6].

IoT devices are advanced technology that increases the capability level of smart cities. IoT-enabled devices reduce smart cities' latency and energy consumption [7, 8]. IoT devices provide various heterogeneous services to users, which enhances the feasibility and robustness of smart cities. IoT devices use wireless sensors, which reduces the energy consumption ratio when providing services to users. Wireless sensors collect the data from the base station, which minimizes the complexity of analyzing the necessity of the request [9, 10].

Many cities are embracing smart technology to build data-centric settings that can instantly analyze massive amounts of data. City planners, traffic engineers, and resource allocators may all benefit from the insights provided by machine learning algorithms that study this data [11, 12]. The study's effectiveness in an ever-changing Internet of Things (IoT) setting depended on Leveled Reinforcement Training's (LRT) hierarchical framework for smart city resource management. Learning rate tuning (LRT) allows for the progressive learning and adaptation of resource allocation choices in response to changing demand and energy usage [13, 14]. The system may progressively enhance its decision-making capabilities using LRT's multi-tiered learning technique. It starts with basic scenarios with low energy consumption and moves on to more complicated ones with high device activity and resource constraints. By adopting this tier-wise method, the system can pinpoint crucial stress points, including spikes in energy usage or device failures, and then prioritize managing available resources accordingly [15].

The RM process uses machine learning (ML) methods and techniques in smart cities. ML methods are mainly used to reduce the computational cost level in the RM process [16, 17]. The examined data produces optimal information for the RM process. The dataset analyzed minimizes the management process's latency and energy consumption ratio [18, 19]. This article introduces CRM using Leveled Reinforcement Training (LRT) to address resource constraint issues due to flexible device features. LRT helps smart city resource management adapt to demand and energy consumption variations. This method proactively plans resources in response to energy availability and real-time device manipulation, reduces waiting times, and increases system efficiency, making it appropriate for the recommended system.

The paper’s novelty is that this article discusses smart city resource optimization utilizing CRM based on leveled Reinforcement Training. Our method dynamically distributes resources based on service demand and device performance to identify energy-intensive periods and prices. CRM optimizes IoT-based smart city services by monitoring energy usage and device failures. CRM is faster and eliminates resource allocation errors than previous methods.

Therefore, the research contributions are as follows:

- Design of a CRM method aiding less complex constraint mitigation for ease of sharing and allocation.
- The modification and implication of interrupt considered application services for SC environments.
- A comparative study will be conducted to validate the method’s performance using allocation, constraint, and time-based metrics.

This diagram shows one possible smart city resource management system based on reinforcement learning. A Smart City (SC) 's fundamental considerations are resource management and allocation.

2. RELATED WORKS

Researchers in [20] created AACF, an application-centric framework for IoT powered smart city applications. AACF prioritizes user-friendly and adaptable IoT architectures, emphasizing robust quality of experience in real-time smart city applications. The method focuses on the flexible distribution of application services, considering both application state and longevity. The proposed method enhanced the overall user experience in smart city applications. Alam et al. [21] devised an access control system for IoT devices in smart cities using blockchain and big data. Su et al. [22] suggested that smart cities use cloud computing and the IoT for information processing. The method provides helpful insights for future smart city development in the same context. To ensure uninterrupted smart services by balancing energy trade-offs between utility cost and consumption, Xiaoyi et al. [23] proposed a multi-objective distributed dispatching algorithm (MODDA) for efficient green energy management in IoT-driven smart cities. Liu et al. [24] proposed a method for efficiently allocating edge computing resources in IoT-based smart cities (MJOM). Zhang et al. [25] created an IoT green energy system for smart cities. The suggested method plans and assesses the development of smart power systems, considering various on-site and off-site resources. Zhong et al. [26] proposed a method for environmentally friendly smart cities using Green IoT. CRM with LRT is a better and faster alternative to conventional resource management methods in smart cities. This method improves the system's overall speed and scalability, making it better manage the intricate IoT settings in smart cities today. Fawzy et al. [27] introduced TPRUDF, a data-fusion framework for efficient resource utilization in smart environments leveraging the IoT. This method demonstrates a reduction in energy consumption and an improvement in throughput.

3. CONCERTED RESOURCE MANAGEMENT (CRM) USING THE LEVELED REINFORCEMENT TRAINING (LRT)

Real-world applications of the IoT include smart city traffic management systems that employ LRT to distribute resources optimally. Interconnected sensors and cameras monitor traffic. LRT examines energy consumption, traffic patterns, and traffic signal performance when detecting congestion. The preliminary step is identifying the cost complex derived above expressed in Equation (1).

$$\alpha = \frac{1}{d_n} + \sum_{s_c}^{s_v}(i_t * d_0) + \left(\frac{\left(\frac{i_t * s_c}{s_v}\right)}{\sum \frac{1}{d_n}}\right) * \left(\frac{i_t + d_0/s_v}{d_n * s_c}\right) + \left[\left(\prod_{s_c}(d_0 * s_v)\right) * (s_c + i_t)\right] * \left\{\left((d_0 + \cdots + d_n) * \left(i_t + \frac{s_c}{s_v}\right)\right)\right\}. \tag{1}$$

Calculating the complicated cost of resource management in a smart city setting. The number of devices $\{d_0, \ldots d_n\}$. and services s_v, s_c that are engaged in distributing resources is one of the factors included in the calculation. The cost complex in IoT provides reliable smart city device sharing, and it is expressed as $[(\prod_{s_c}(d_0 * s_v)) * (s_c + i_t)]$. IoT is initialized for the service-oriented computation in the environment study. The scope of this is to forward the resources to the requested devices in IoT, for this cost complexity is identified.

$$\theta = \prod_{s_c}(d_0 + e_y) * c_r + \left\{ \left(\frac{\overset{\vdots\cdots\vdots}{i_t + s_c}}{\sum_{d_0}^{d_n}(s_v + w')} \Big/ {e_y} \Big/ {g'} \right) \right\} * \\ [(e_y - g') + (i_t + w')] + \int_{d_0}^{d_n}(c_r(e_y) + v') - \left(\frac{v'}{w' + d_0} \right) \tag{2}$$

The high energy-consuming resources are observed, and the intervals are detected above (2). They depict an Equation (3) that deals with device failure inside an IoT framework and distributes resources to devices according to their requirements. Before allocating resources, the system examines energy use and service demands to determine the most intricate cost possibilities. Identifying resources with high energy consumption and closely tracking them at certain points can improve the efficiency and reliability of smart city applications. Equation (3) detects the IoT services handling where the device management forwards the resources, which is θ. The symbol represents this management function θ, and the highlighted Equation (3) is concerned with identifying IoT services that handle resource allocation via device management. It is essential for the effective administration of IoT services and optimizing resource sharing by deciding when devices send resources. The equation considers important variables such as reaction time and device failures to determine the period and cash needed for resource sharing. Improved service delivery in smart cities, managing resource limits, and navigating complicated IoT settings rely on the following equation.

$$\mu = \begin{cases} \left({s_v * s_c} \big/ {w' + d_0} \right) + (o_c - q_e) - \left(v' * \frac{1}{d_n} \right), \in r_m \\ \prod_{i_t}(e_y - g') + q_e(s_v) - w' + \sum_{v'}(e_y - c_r), \in d_f \end{cases} \cdot \tag{3}$$

The complexity occurring intervals are decided as presented in Fig. 1. This complexity is considered based on device cost (replacement and functional) and energy drain (for *g*') such that *w*' is nevertheless defaced (Fig. 1). Here, the processing step relies on the response time to analyse a particular method where concurrent detection is required. The second derivation states device failure, where resource management defines the constraints of energy devices. Energy-based device management provides artificial intelligence in a better manner. By stating the same device request for the identical resource, then, the device failure occurs, and it is $\prod_{i_t}(e_y - g') + q_e(s_v)$. Based on this approach, the device failure is addressed in this derivation.

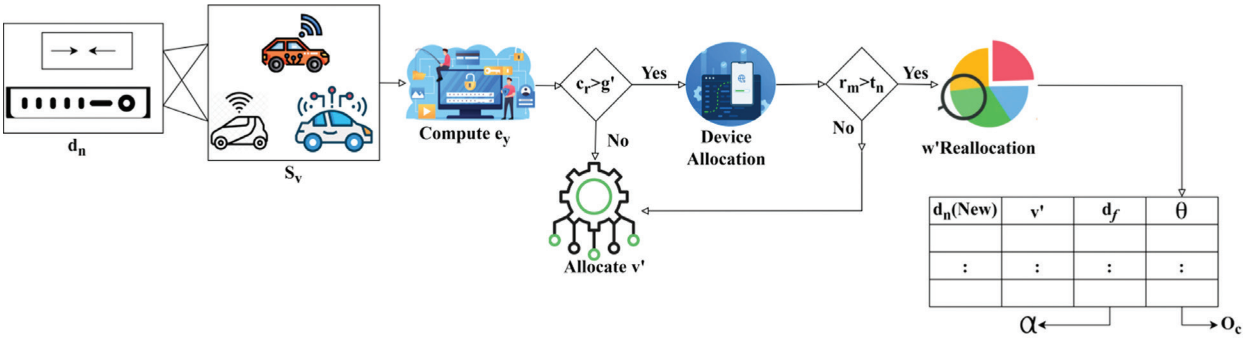


Fig. 1. Complexity Occurring Sharing Intervals

$$p_a = (r_a + d_0) * \left(\frac{\prod_{s_v}^{w'}(v' - r_m) + c_r}{\frac{{e_y}/{m_d}}{a_p + i_t}} \right) + (a_p * d_f) + \sum_\theta[(v' - i_t) * \\ (q_e + m_d)] + \sum\left[(r_m - d_f) + \frac{w'}{s_v} \right] \tag{4}$$

Equation (4) expresses that incorporating energy and devices for the smart city application is based on demand. Equation (4) combines smart city energy and device management for distributing resources based on demand. The real-time device failure handling and resource distribution are effective. This dynamic method boosts smart city reliability.

3.1. REINFORCEMENT LEARNING FOR CONSTRAINT ANALYSIS

It is a decision-making approach suitable for the specific situation and improves the reward function.

Three phases of computation are pragmatic in this learning; the first is the input, which is the initial stage of acquiring the request from the device.

In Equation (5), energy depletion is derived. Equation (5) calculates smart city energy depletion using resource allocation and forwarding time intervals. Demand and energy use are measured. This equation optimizes resource management in smart cities by reducing energy waste.

$$l_p = \left(\frac{{\prod_{d_0}(i_t + m_d)} \big/ {r_a}}{{w' + m_d} \big/ {\sum_{r_a}(e_y - d_f)}} \right) * [(o_c + a_p) + (v' + \mu)] * \\ \left(\frac{{r_a + d_0} \big/ {q_e}}{\prod_{g'}(r_m + d_f)} \right) + \{[(c_r + e_y) + (w' + d_0)] * (\sum_{m_d} o_c + v') + \alpha\} \cdot \tag{5}$$

The level-based constraint representation is given below. The reinforcement learning states the active response where the device-based service is defined for the demand. The allocation failure is equated below.

$$a_f = (q_e + s_v) + d_0 * \left(\frac{\theta + c_r}{\sum_{w'}(l_p + s_c)}\right) - v' \tag{6}$$

In Equation (6), the allocation failure is examined. Allocation failure (a_f) in Equation (6) happens when demand exceeds supply, and a system cannot allocate resources to a device. Resource depletion (v') is added to service and queuing times. The resource management system targets this failure to deliver continuous service at desired intervals. In Algorithm 1, the pseudo-code for a_f is presented.

Algorithm 1 a_f Detection

```
Step 1: for all $d_n$ do {
Step 2: Allocate $s_v$ and $\mu$
Step 3: compute $p_a$ using equation (2b)
Step 4: if {$p_a > \mu/s_v$} then
Step 5: Estimate $l_p$ using equation (3)
Step 6: while {$d_f$!=1} do
Step 7: Assign $s_v$ to $d_n$ $\forall(d_n * \mu) = r_m/v_e$
Step 8: compute $c_r$ and $g'$
Step 9: if {$c_r < g'$} then
Step 10: Assign a new $d_n$ in $\mu$ and Repeat
         from Step 5 until $r_m/q_e = 1$
Step 11: else if {$c_r > g'$} then
Step 12: Repeat from Step 4
Step 13: Update $q_e$
Step 14: $a_f = (q_e + s_v)$
Step 15: End if
Step 16: End while
Step 17: $l_p = l_v(e_y)$
Step 18: End if
Step 19: $a_f = a_f$-v^'
Step 20: End for
```

Hence, the allocation failure is observed in this part, and the session-based interruption is derived in the section below (Fig. 2):

$$\mu(i_p) = |w' + q_e| * \{a_f - l_p\} + \left(\frac{a_p * g'}{\alpha * o_c}\right) - i_t. \tag{7}$$

As found in Equation (7), the allocation failure is observed, and session-based interruption has been examined. In Equation (7), session-based interruption $\mu(i_p)$ refers to service disruption caused by allocation failure a_f. When session resources are misallocated, this disturbance occurs. The Equation considers interruption time (i_p) and resource depletion parameters to evaluate service effects in IoT applications, especially smart cities. Thus, incorporating the service along with the energy and device faces the interruption of services to avoid this identification of resource allocation failure and active resource is equated below.

$$\alpha(r_a) = \left[\left(a_p + (i_p * p_a)\right)\right] * \left(\frac{d_0 + w'}{\sum(v' + m_d)}\right) - o_c. \tag{8}$$

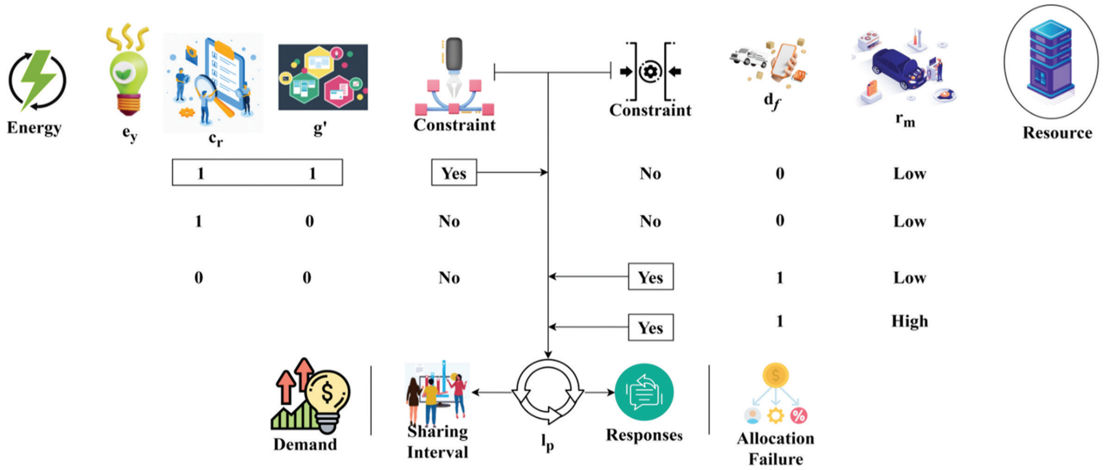


Fig. 2. Level-based Constraint Representation

As deliberated in Equation (8), resource allocation failure and active resources have been described. Consider active resource equations and allocation failure to maximize resource distribution in dynamic situations like smart cities. They assist proactive management in sustaining service operations by predicting resource shortages and breakdowns. These equations effectively detect active devices and energy use to scale systems and reduce waste. The training is introduced in this learning method, where demands are identified based on the independent or concerted resource allocation. The derivation is expressed as follows.

$$t_n = \prod_{d_n}(s_v * m_d) + \left(\frac{a_f + l_p/r_a}{w' + i_t}\right) * (i_p + c_r) - p_a \tag{9}$$

As examined in Equation (9), concerted resource allocation has been explored. The effective and coordinated distribution of resources across various devices is crucial in dynamic contexts, such as smart cities, and this can only be achieved by concerted resource allocation. The system optimizes energy utilization, minimizes delays, and eliminates competition for scarce resources by aligning allocation with real-time demand. Overall, this method prevents allocation errors and enhances system performance and scalability. This interconnection of levels is discussed and equated to the below equation.

$$l_v(e_y) = [(w' + d_0) + (r_a - v')] * \left(\frac{(d_f+a_p)}{\sum_{m_d} t_n}\right) + (o_c - i_p) \quad (10)$$

As discussed in Equation (10), interconnection levels are explained. A high degree of connectivity is crucial for the smooth operation of the many IoT devices in smart cities. With their help, extending resource management becomes simpler, allowing the system to react quicker to changes in real-time demand.

The level of interconnection is based on this approach where the cost complex is involved for the service forwarding for the allocation failure, and it is described as l_v. In this case, an interruption of service is associated with device failure, where training is involved for the processing levels. The learning process for $l_v(e_y)$ is illustrated in Fig. 3.

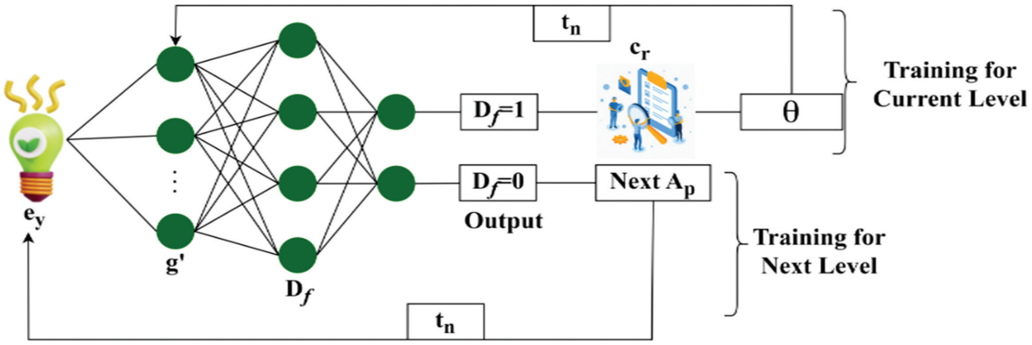


Fig. 3. $l_v(e_y)$ Learning Process

The equation below states the device failure:

$$l_v(d_f) = \prod_{d_0}(a_p * w') + \left(\frac{l_p * i_p}{m_d}\right) - i_t(d_0) \quad (11)$$

As explored in Equation (10), device failure was calculated. After this, resource allocation is performed, and training levels are observed in the equations below.

$$l_v(r_a) = \left(\frac{\alpha + l_p}{a_f}\right) * (s_v - \theta) + \left(\frac{{}^{o_c}/_{g'}}{t_n + a_f}\right) * (d_f + t_n) \quad (12)$$

$$t_n(l_v) = (e_y + d_f + r_a) + \left(\theta * \frac{\mu}{v\prime}\right) * r_m. \quad (13)$$

Training level observation was derived as deliberated in Equation (12) and Equation (13) performed for resource allocation. Before allocating resources, it is essential to conduct training-level observations to identify and respond to evolving needs and performance measures.

This procedure makes the system more adaptable to new circumstances, reduces delays, and guarantees efficient use of resources in smart city settings. The learning process for $l_v(r_a)$ is illustrated in Fig. 4.

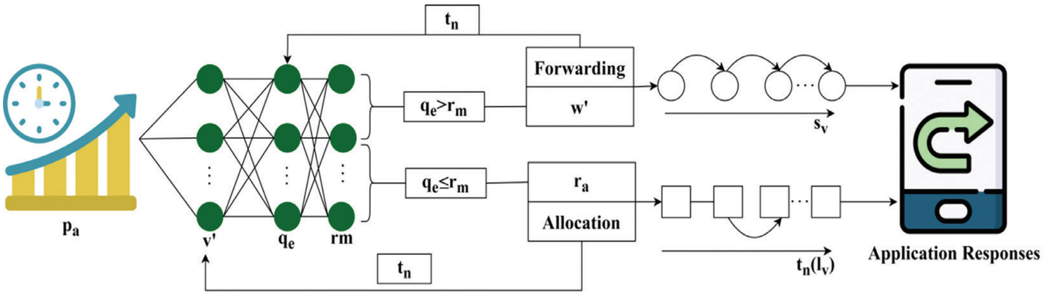


Fig. 4. Learning Process for $l_v(r_a)$

The learning process connected to levels is presented as a pseudo-code in Algorithm 2.

Algorithm 2 Learning Process for Connected Levels

```
Input: $p_a, m_d$
Step 1: for all $p_a$ $m_d$ ∀ $d_n$ do
Step 2: compute $\mu(ip)$ using equation (4)
        and perform $\alpha$ $(r_a)$
Step 3: if {$\alpha(r_a)$!=($a_p$*$r_a$/($t_n$($l_v$)))} then
Step 4: Estimate o using equation (1b)
        and $a_f$ using equation (3b)
Step 5: if {$t_n$=$e_y$} then
Step 6: Allocate $m_d$ with $d_n$ such that ($d_f$+$a_p$)=$\sum t_n$
Step 7: Update ($e_y$, $d_f$)|($c_r$, $o$) ∀ $d_f$=1 in Step 6
Step 8: Perform $t_n$ for $o$ and ap∈($c_r$,$o$)
        until $\alpha(r_a)$=$a_p$*$r_a$/($t_n$($l_v$))
Step 9: Estimate $r_m$ ∀ $p_a$ under $q_e$
Step 10: if {$r_m$≥[$q_e$-$t_n$($l_v$)]} then
Step 11: Include $l_v(d_f)$ ∀ $i_t$ ($d_o$), $d_o$ ∈ $m_d$=Allocation
Step 12: Update ($p_a$=$p_a$-v'/($t_n$ ($l_v$)))
Step 13: else
Step 14: $l_v$ ($r_a$)=($d_f$+$t_n$) until $t_n$=($e_y$+1) ∈ $d_n$
Step 15: Goto Step 3 for all $p_a$=0 in $t_n(l_v)$
Step 16: End if
Step 17: Perform $l_p$ for $l_v$ $(e_y)$ and $l_v$ $(r_a)$ until $r_m$=0
Step 18: End if
Step 19: End if
Step 20: End for
```

4. PERFORMANCE ASSESSMENT

This article discusses layered reinforcement learning and its potential applications in coordinated resource management to ease SC service constraints. With limited energy and resource allocation, this concept took the big picture into account while managing SC resources. The approach first uncovered the intricate processing limitations based on energy to guarantee high device availability. Device availability and resource allocation concerns are associated with excessive energy usage and depletion. This section describes the performance of the proposed method through a comparative study. This section discusses comparative analysis based on resource allocation rate, active device detection, constraint mitigation, allocation failure, and allocation wait time. The number of devices (20 to 240) and the sharing interval time (30s to 300s) are the X-variants considered. The proposed method is compared with MODDA [23], MJOM [24], and TPRUDF [27] methods discussed in the related works section.

Table 1 shows the experimental setup.

Table 1. Experimental Setup

Parameters	Description
Processor	Intel core i7;3.5
Memory	16GB RAM
Storage	1 TB SSD
IoT Devices	Raspberry Pi 4
Sensors	Temperature, humidity, energy sensors
Network	5G
Operating System	Ubuntu 20.04
Programming Language	Python 3.8
Machine Learning Library	TensorFlow
Reinforcement Learning	OpenAI, stable baselines 3
Simulation environment	MATLAB
Monitoring Tool	Grafana
Blockchain Integration	Hyperledger Fabric

Dataset Description: Turning the city into a "smart city" is the government's goal. The plan is to turn it into an intelligent and digital city to make services more efficient for the people. The administration is dealing with traffic as one of its challenges. Data scientists are contributing to improved municipal traffic management and future infrastructure development. So, that we can plan accordingly for the next four months' worth of traffic at each of these intersections, the traffic data from many periods since the sensors at each junction were taking readings at various times. Some intersections have supplied little or incomplete data, which further complicates matters and necessitates care in developing future predictions. According to data collected over the last twenty months, the government depends on reliable traffic forecasts for the next four months. A bigger change is coming to the city, and machine learning algorithms will be the cornerstone of it. It will become smart and intelligent.

4.1. RESOURCE ALLOCATION RATE COMPARISON

The resource allocation rate for the proposed work increases for the smart city where the demands are satisfied by the number of devices and constraints used to provide energy based on the levels of observation. The identification of the cost-complex is developed for the consuming sharing intervals, and it is represented as $((((i_t*s_c)/s_v))/(\sum 1/d_n))*(((i_t+d_0)/s_v)/(d_n*s_c))$. This processing step allocates resource management to the appropriate device based on reinforcement learning. Here, the high energy-consuming sharing intervals detect device failure in smart city applications. In this case, the incorporation of energy and device is examined for the interrupt sessions, and it is equated in Equation (2a). Device failure provides energy depletion and allocation failure, whereas the cost complex provides the response time analysis. The incorporation rate in this work increases where the energy drop is detected, and based on this, resource allocation is performed. The device failure is identified for the levels of observation where the output is trained in this session using reinforcement learning as shown in Fig. 5 (a) and (b).

Fig. 5 (a) and (b) show that the proposed CRM method with LRT regulates the resource allocation rate by adapting the distribution of resources in real-time to patterns in energy consumption and device demand. It discovers energy-consuming periods and optimizes the sharing time by assessing service response times and likely device faults. This approach minimizes power consumption during sharing times, adjusts to the requirements of various devices, continuously educates the system to deal with allocation mistakes, and allows for the efficient and timely allocation of resources to running devices.

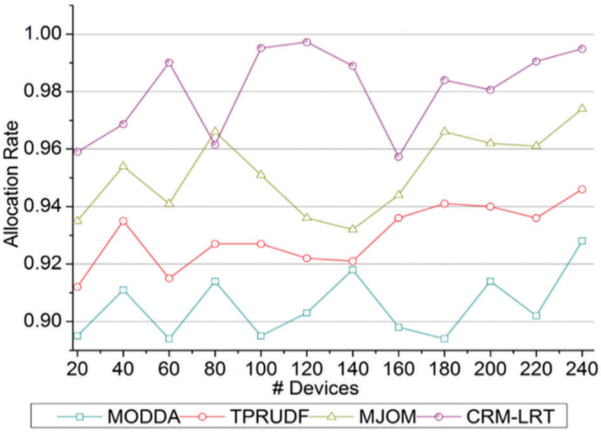


Fig. 5 (a). Resource Allocation Rate for #Devices

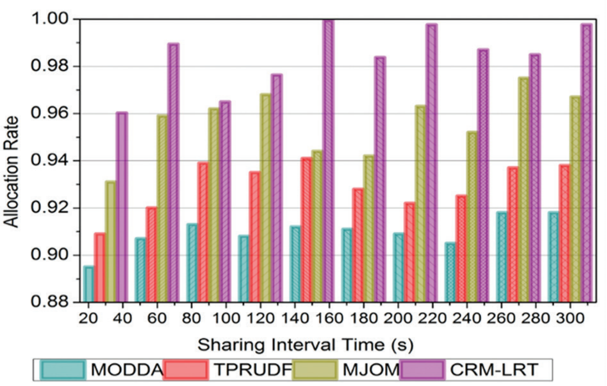


Fig. 5 (b). Resource Allocation Rate for Sharing Interval Time(s)

4.2. ACTIVE DEVICE DETECTION COMPARISON

In Fig. 6 (a) and (b), active device detection improved for the varying energy drop and resource allocation. Based on the request, the response time is observed and forwards the resource to the appropriate device on time. It states the alternative allocation where it provides the consuming and sharing intervals. Energy and devices are incorporated to state smart applications and provide resource management. The interruption is detected based on the forwarding interval and allocates the resource to the device. In this case, an active device is detected to forward the resource in this process, and the failure is reduced.

Fig. 6 (a) and (b) express how tracking power utilization and service response with the LRT during sharing times enhances active device recognition. It accurately detects active devices using real-time performance data, including energy utilization and demand fulfillment. The technique ensures that only active session participants acquire resources by monitoring and reacting to each device's operating state. With dynamic detection, smart city resources may be dispersed more effectively between devices.

4.3. CONSTRAINT MITIGATION COMPARISON

The constraint mitigation is enhanced in this work; if device failure is detected, then the computation process is developed to detect whether the device is active or not. Resource management is developed in this IoT based on device activation, which provides service-based demands. The device failure occurs and observes whether any resource allocation is occurring. In terms of this method, the levels of observation take place for the constraint mitigation, and it is equated as $d_0*((\theta+c_r)/(\sum(w')(l_p+s_c)))$ and shown in Fig. 7 (a) and (b).

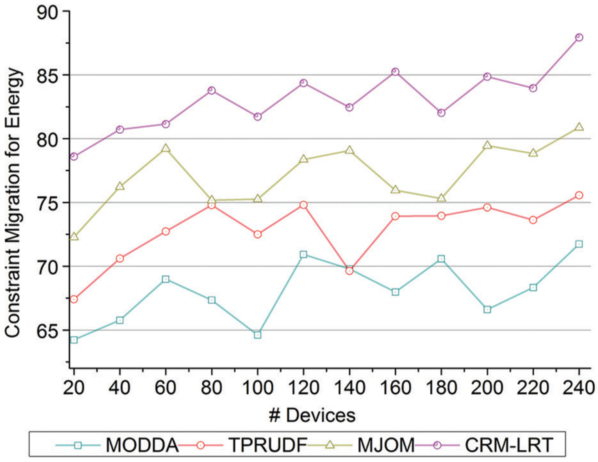


Fig. 6 (a). Active Device Detection for #Devices

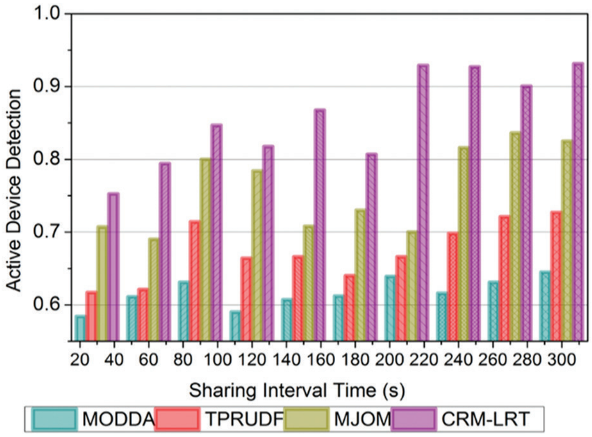


Fig. 6 (b). Active Device Detection for Sharing Interval Time(s)

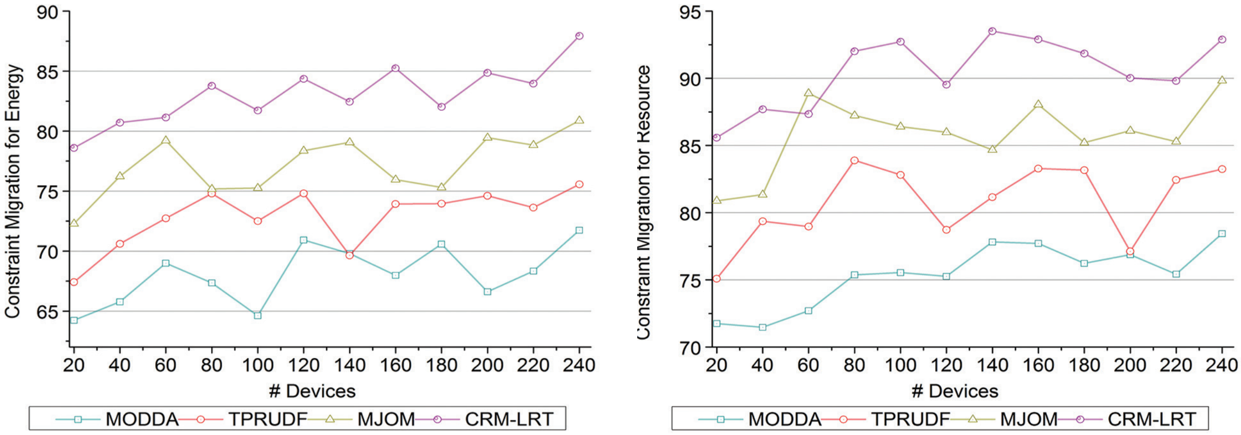


Fig. 7 (a). Constraint Mitigation for Energy and Resources against #Devices

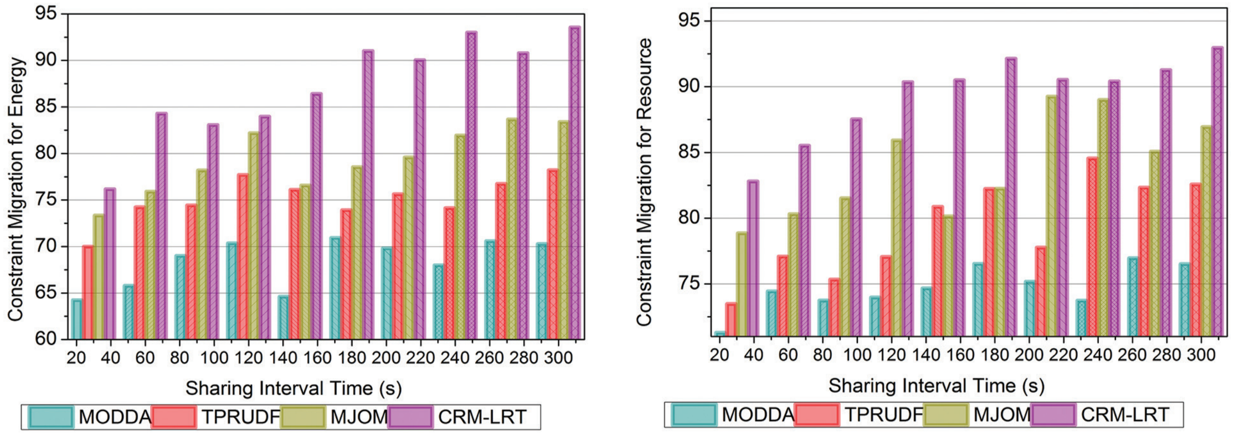


Fig 7 (b). Constraint Mitigation for Energy and Resources against Sharing Interval Time(s)

Fig. 7 (a) and (b) explain how to reduce limitations caused by various devices and sharing intervals. The suggested solution employs LRT to adapt resource management on the flight.

4.4. ALLOCATION FAILURE COMPARISON

The allocation failure is due to the device failure, and a response is given to the identification method. This processing case was developed for resource allocation and is based on the levels of observation. From these levels, the training is processed and meets the demands of examination in resource management. From this management system, the n-number of devices is associated with energy consumption, and depletion is detected in this reinforcement learning. The learning estimates and delivers the smart city application for the high energy in this processing step. The active resource is forwarded to the end device that requests the resource, and it is termed concerted resource allocation or independent resource management. The level of interconnection is formulated as $\{a_f\text{-}l_p\}+((a_p*g')/(\alpha*o_c))\text{-}i_t$, from the processing intervals and examines the device failure Fig. 8 (a) and (b).

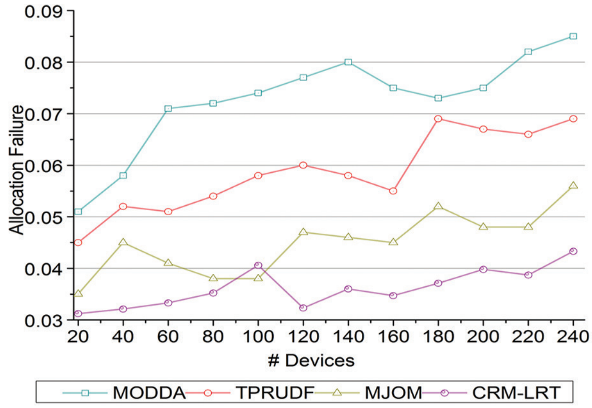


Fig. 8 (a). Allocation Failure for #Devices

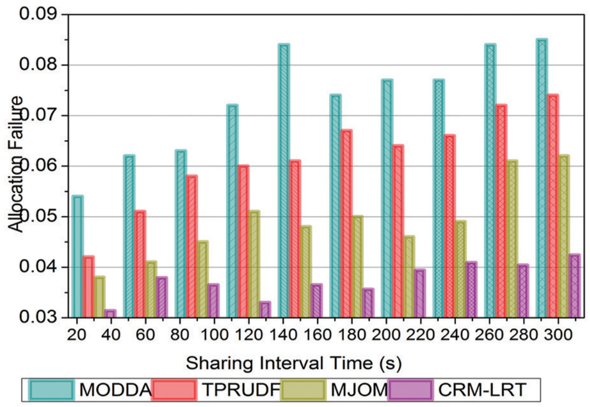


Fig. 8 (b). Allocation Failure for Sharing Interval Time

Fig. 8 (a) and (b) explored the proposed LRT to solve allocation failure by anticipating and controlling resource requirements across devices and sharing durations. In the event of a failure caused by inadequate resources or high energy use, it identifies this immediately and modifies allocations appropriately.

4.5. ALLOCATION WAIT TIME COMPARISON

In Fig. 9 (a) and (b), allocation wait time is reduced in this process, where energy depletion and allocation failure occur. Here, demand estimation is performed to detect the interruption of the session. The levels of interconnection are processed for resource allocation where the satisfaction of the demands. The service to the smart city is deployed in the IoT device and detects depletion and allocation failure. In this processing step, the incorporative state is used to identify the consuming sharing intervals, and it is represented as [(a_p+(i_p*p_a))]. The training in the levels of interconnection is used to state the independent or concerted resource allocation in which the demands are satisfied. This allocation wait time is decreased if the active resource is forwarded to the requested device. Incorporating energy consumption provides an alternative allocation where demands are identified based on this computation process. Here, it states the device failure and reduces the waiting time. The wait time is observed, and the smart city application is provided with better resource forwarding.

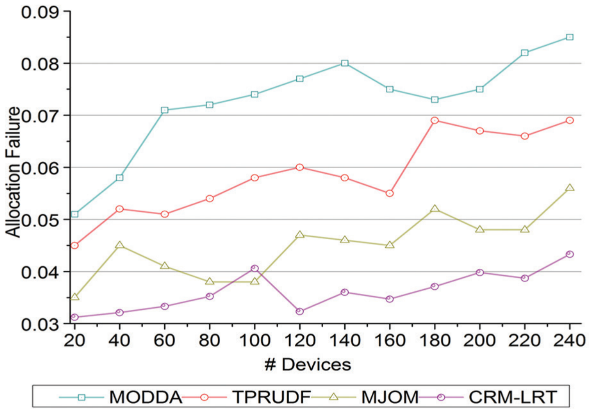


Fig. 9 (a). Allocation Failure for # Devices

Fig. 9 (a) and (b) deliberated that the suggested technique uses patterns of device activity and sharing intervals to forecast resource demands and allocate them using LRT, reducing allocation wait time.

Table 2. Execution Time for Different Devices

	MODDA	TPRUDF	MJOM	CRM-LRT
20	0.418	0.324	0.554	0.326
40	0.581	0.538	0.464	0.526
60	0.524	0.479	0.656	0.313
80	0.289	0.335	0.614	0.346
100	0.656	0.417	0.452	0.323
120	0.549	0.584	0.413	0.419
140	0.442	0.611	0.804	0.333
160	0.534	0.643	0.721	0.526
180	0.554	0.698	0.821	0.418
200	0.587	0.624	0.522	0.581
220	0.602	0.587	0.513	0.524
240	0.624	0.673	0.818	0.326

The Leveled Reinforcement Training (LRT) approach, which the suggested CRM method employs to optimize resource allocation in response to changing service needs and device statuses in real-time, improves execution time. In contrast to more conventional approaches, CRM anticipates and resolves device failures and high-cost, energy-intensive sharing periods. This proactive method expedites service execution by minimizing delays caused by resource restrictions.

Table 3. Memory Usage for Different Devices

Number of Devices	MODDA	TPRUDF	MJOM	CRM-LRT
20	0.503	0.718	0.599	0.478
40	0.51	0.704	0.587	0.464
60	0.529	0.69	0.566	0.45
80	0.532	0.68	0.559	0.44
100	0.548	0.678	0.543	0.438
120	0.555	0.658	0.532	0.428
140	0.562	0.647	0.522	0.417
160	0.577	0.639	0.50	0.409
180	0.587	0.625	0.491	0.395
200	0.593	0.605	0.481	0.386
220	0.553	0.625	0.491	0.395
240	0.544	0.605	0.481	0.386

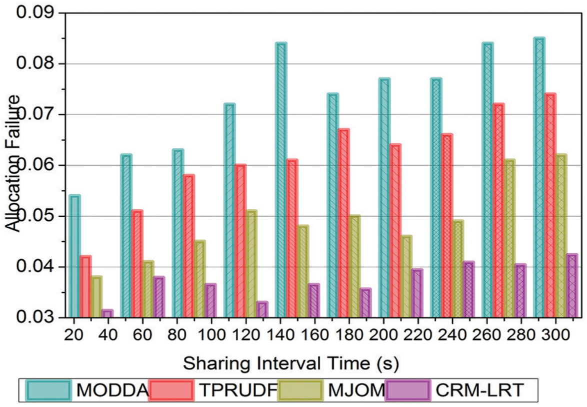


Fig. 9 (b). Allocation Failure Sharing Interval Time

The suggested CRM technique decreases memory utilization by enhancing resource management via reinforcement learning and selective allocation. Instead of continuously storing and processing data for all smart city devices and services, the CRM technique prioritizes data points with significant energy and cost consumption. By recognizing and handling these crucial periods, the system may reduce the processing and storage of unnecessary or redundant data. Leveled Reinforcement Training (LRT) also allows the approach to adjust resource allocation in real-time in reaction to demand, which might reduce the need to retain vast volumes of past data. Smart city applications benefit greatly from the system's enhanced speed and efficiency through memory management and selective data processing.

Table 4. Performance of the Study

Number of Devices	Energy Consumption	Resource Management	Waste Management	Decision Making
20	80.2	80.9	70.8	87.6
40	81.3	81.6	71.7	88.7
60	84.6	82.5	72.4	89.1
80	86.9	83.2	73.5	90.3
100	88.5	84.4	74.7	87.7
120	90.4	85.1	75.2	88.9
140	91.6	86.2	76.3	89.5
160	93.2	87	77.6	90.9
180	94.4	88.1	86.8	91.2
200	95.7	88.9	86.8	93.1
220	96.4	88.1	88.5	94.4
240	96.7	90.3	89.8	95.7

Table 4 shows the performance of the proposed study. Maintaining unrestricted service flow in a smart city (SC) setting is achieved by carefully managing resources, including energy consumption and device performance. These services are often interrupted when resources are scarce, leading to increased energy consumption, slowed resource allocation, and broken devices. Attempts to control resources in SC systems using static thresholds or reactive strategies have failed miserably. An intermediate-sized smart city employs CRM-LRT to address these issues. Intelligent transportation, smart waste management, and public safety applications are just a few of the urban systems enabled by IoT. These systems have difficulties allocating resources because of fluctuations in device performance and energy consumption patterns, especially during high service demand. Issues with managing resources in the city's smart services were successfully managed by the CRM solution that was based on LRT.

Implementation of CRM-LRT:

I. Following these phases, the CRM approach is implemented throughout several city sectors:

II. The city's IoT sensors track energy use, response times to service requests, and device malfunctions. This data is analyzed using AI algorithms to comprehend cost-complex intervals and energy consumption patterns.

III. LRT uses the data to determine when resources are most needed during critical service periods. The training is centered on allocating resources from energy depletion stages to device activity degrees. The predictions of the LRT model are used to pick devices with the best combination of energy efficiency and availability.

IV. The LRT method employs a coordinated resource allocation, considering energy needs and device preparedness. By working together, we can decrease allocation failures and energy waste by letting the system dynamically assign resources based on real-time requests.

V. Management of Interruptions and Failures: The CRM system can identify instances when services are interrupted due to problems with devices or the distribution of resources. They were redistributing resources to devices actively using less energy, which guarantees that services will continue uninterrupted.

The approach enhanced system performance, reduced energy consumption, and expanded scalability of SC applications by dynamically distributing resources according to real-time data. Collaborative decision-making for smart city resource management is shown in this scenario.

The performance assessment in Section 4 shows how effectively the suggested CRM system works, but a complete investigation of device counts and sharing intervals may reveal its true scalability. Given the growing number of connected devices, comparing results across different IoT device sizes is one approach to evaluate the strategy. This investigation may reveal the CRM technique's resilience under more rigorous resource allocation and energy consumption testing. Testing the technique with different sharing intervals will reveal its flexibility to shift demand and performance in real-time, dynamic environments. Comparing the two shows the strategy's practical scalability; this may help determine its feasibility for larger smart city networks.

5. CONCLUSION

This article introduced and discussed the performance of concerted resource management using leveled reinforcement learning for SC service constraint mitigation. This proposal considered the energy and resource allocation constraints that are chained together in retarding SC resource management. First, the method identified the complex processing constraints based on energy to ensure high device availability. Computed with energy depletion and high energy utilization features, device availability and resource allocation failures are connected. This connectivity provides shared or re-allocated/failure-less resource management for different SC application services. The shared sessions are validated using complex constraints that remain unaddressed post the allocation. The concerted reinforcement learning was used to identify the allocation failures using the connectivity between different levels in the resource management process. Both constraints rely on different training inputs per energy and resource allocation conditions to improve the allocation rate. Under the varying devices, the following is observed: Improvements: 9.1% (Allocation Rate), 10% (Device Detection), 11.88% (Constraint Mitigation—Energy), 9.06% (Constraint Mitigation—Resource Allocation); Reduced: 8.01% (Allocation Failure), 9.64% (Waiting Time). Potentially impassable difficulties with the proposed method include very dense IoT network scalability and unexpected, rapid shifts in resource requirements. The scalability of the CRM technique might be studied further by testing it in various real-world IoT scenarios and incorporating deep reinforcement learning for increased adaptation. Improved performance optimization and resilience in ever-expanding smart city applications can only be achieved using edge computing and beefing up security measures.

6. ACKNOWLEDGMENT

The author gratefully acknowledges the approval and the support of this research study by the grant no CSAT-2023-12-2159 from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A

7. REFERENCES

[1] T. Shafique, R. Gantassi, A. H. Soliman, A. Amjad, Z. Q. Hui, Y. Choi, “A review of Energy Hole mitigating techniques in multi-hop many to one communication and its significance in IoT oriented Smart City infrastructure”, IEEE Access, Vol. 11, 2023, pp. 12345-12367.

[2] V. M. Kuthadi, R. Selvaraj, S. Baskar, P. M. Shakeel, A. Ranjan, “Optimized energy management model on data distributing framework of wireless sensor network in IoT system”, Wireless Personal Communications, Vol. 127, No. 2, 2022, pp. 1377-1403.

[3] I. Hussain, A. Elomri, L. Kerbache, A. El Omri, "Smart City Solutions: Comparative Analysis of Waste Management Models in IoT-Enabled Environments Using Multiagent Simulation", Sustainable Cities and Society, Vol. 103, 2024, pp. 1-14.

[4] A. Nauman, N. Alruwais, E. Alabdulkreem, N. Nemri, N. O. Aljehane, A. K. Dutta, W. U. Khan, "Empowering smart cities: High-altitude platforms based Mobile Edge Computing and Wireless Power Transfer for efficient IoT data processing", Internet of Things, Vol. 22, 2023, pp. 1-26.

[5] R. Selvaraj, V. M. Kuthadi, S. Baskar, P. M. Shakeel, A. Ranjan, "Creating security modelling framework analysing in internet of things using EC-GSM-IoT", Arabian Journal for Science and Engineering, Vol. 48, 2023, pp. 2607-2620.

[6] F. Beştepe, S. Ö. Yildirim, "Acceptance of IoT-based and sustainability-oriented smart city services: A mixed methods study", Sustainable Cities and Society, Vol. 80, 2022, p. 103794.

[7] J. Duque, "The IoT to Smart Cities - A design science research approach", Procedia Computer Science, Vol. 219, 2023, pp. 279-285.

[8] C. Li, "Development of IoT Smart Cities and Optimization of English Education Systems Based on 5G Networks", Soft Computing, Vol. 27, No. 12, 2023, pp. 1-19.

[9] M. Hosseinzadeh, A. Hemmati, A. M. Rahmani, "Clustering for smart cities in the internet of things: a review", Cluster Computing, Vol. 25, No. 6, 2022, pp. 4097-4127.

[10] A. Ullah, S. M. Anwar, J. Li, L. Nadeem, T. Mahmood, A. Rehman, T. Saba, "Smart cities: The role of Internet of Things and machine learning in realizing a data-centric smart environment", Complex and Intelligent Systems, Vol. 10, No. 1, 2023, pp. 1607-1637.

[11] K. Peng, H. Huang, P. Liu, X. Xu, V. C. Leung, "Joint optimization of energy conservation and privacy preservation for intelligent task offloading in MEC-enabled smart cities", IEEE Transactions on Green Communications and Networking, Vol. 6, No. 3, 2022, pp. 1671-1682.

[12] L. A. Ajao, S. T. Apeh, "Secure edge computing vulnerabilities in smart cities sustainability using Petri net and genetic algorithm-based reinforcement learning", Intelligent Systems with Applications, Vol. 18, 2023, p. 200216.

[13] A. Ullah, S. M. Anwar, J. Li, L. Nadeem, T. Mahmood, A. Rehman, T. Saba, "Smart cities: The role of Internet of Things and machine learning in realizing a data-centric smart environment", Complex & Intelligent Systems, Vol. 10, No. 1, 2024, pp. 1607-1637.

[14] X. Li, D. Zhang, Y. Zheng, W. Hong, W. Wang, J. Xia, Z. Lv, "Evolutionary computation-based machine learning for smart city high-dimensional big data analytics", Applied Soft Computing, Vol. 133, 2023, p. 109955.

[15] A. Qadeer, M. J. Lee, "HRL-edge-cloud: Multi-resource allocation in edge-cloud based smart-streetscape system using heuristic reinforcement learning", Information Systems Frontiers, Vol. 26, 2023, pp. 1399-1415.

[16] Y. Kim, B. C. Jung, Y. Song, "Online learning for joint energy harvesting and information decoding optimization in IoT-enabled smart city", IEEE Internet of Things Journal, Vol. 10, No. 12, 2023, pp. 10675-10686.

[17] Z. Huang, G. Jin, "Navigating urban day-ahead energy management considering climate change toward using IoT-enabled machine learning technique: Toward future sustainable urban", Sustainable Cities and Society, Vol. 101, 2024, p. 105162.

[18] O. M. Prabowo, E. Mulyana, I. G. B. B. Nugraha, S. H. Supangkat, "Cognitive city platform as digital public infrastructure for developing a smart, sustainable and resilient city in Indonesia", IEEE Access, Vol. 11, 2023, pp. 120157-120178.

[19] Y. Y. Liu, Y. Zhang, Y. Wu, M. Feng, "Healthcare and fitness services: A comprehensive assessment of blockchain, IoT, and edge computing in smart cities", Journal of Grid Computing, Vol. 21, No. 4, 2023, p. 82.

[20] Y. Cui, X. Song, J. Liu, K. Chen, G. Shi, J. Zhou, G. S. Tamizharasi, "AACF—Accessible application-centric framework for the Internet of Things backhauled smart city applications", IEEE Transactions on Network Science and Engineering, Vol. 9, No. 3, 2021, pp. 980-989.

[21] T. Alam, “Blockchain and Big Data-based access control for communication among IoT devices in smart cities”, Wireless Personal Communications, Vol. 132, No. 1, 2023, pp. 433-456.

[22] P. Su, Y. Chen, M. Lu, “Smart city information processing under Internet of Things and cloud computing”, The Journal of Supercomputing, Vol. 78, No. 3, 2022, pp. 3676-3695.

[23] Z. Xiaoyi, W. Dongling, Z. Yuming, K. B. Manokaran, A. B. Antony, “IoT driven framework based efficient green energy management in smart cities using multi-objective distributed dispatching algorithm”, Environmental Impact Assessment Review, Vol. 88, 2021, p. 106567.

[24] Z. Liu, “A multi-joint optimization method for distributed edge computing resources in IoT-based smart cities”, Journal of Grid Computing, Vol. 21, No. 4, 2023, pp. 1-11.

[25] X. Zhang, G. Manogaran, B. Muthu, “IoT enabled integrated system for green energy into smart cities”, Sustainable Energy Technologies and Assessments, Vol. 46, 2021, p. 101208.

[26] Y. Zhong, Z. Qin, A. Alqhatani, A. S. M. Metwally, A. K. Dutta, J. J. Rodrigues, “Sustainable environmental design using green IoT with hybrid deep learning and building algorithm for smart city", Journal of Grid Computing, Vol. 21, No. 4, 2023, p. 72.

[27] D. Fawzy, S. M. Moussa, N. L. Badr, “An IoT-based resource utilization framework using data fusion for smart environments”, Internet of Things, Vol. 21, 2023, pp. 1-24.

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia.

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Authors of the International Journal of Electrical and Computer Engineering Systems must transfer copyright to the publisher in written form.

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as “the Work”):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author’s personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

______________________________ ______________________________

Author/Authorized Agent **Date**

CONTACT
International Journal of Electrical and Computer Engineering Systems (IJECES)
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Josip Juraj Strossmayer University of Osijek
Kneza Trpimira 2b
31000 Osijek, Croatia
Phone: +38531224600,
Fax: +38531224605,
e-mail: ijeces@ferit.hr