

Journal of
**Electrical and Computer
Engineering Innovations
(JECEI)**

Vol. 11 No. 2, Summer-Fall 2023

• Design of a Synchronous Reluctance Motor with New Hybrid Lamination Rotor Structure	243
• Model Predictive Control of Linear Induction Motor Drive with End Effect	253
• An Improved Approach to Blind Image Steganalysis Using an Overlapping Blocks Idea	263
• Development of Wind Turbine Fault Analysis Setup based on DFIG Hardware in the Loop Simulator	277
• Design Optimization of the Delta-Shape Interior Permanent Magnet Synchronous Motor for Electric Vehicle Application	291
• PAPR Reduction in OFDM UOWC System Employing Repetitive Clipping and Filtering (RCF) Method	301
• Centrality and Latent Semantic Feature Random Walk (CSRW) in Large Network Embedding	311
• Mutual Coupling Reduction in MIMO Microstrip Antenna by Designing a Novel EBG with a Genetic Algorithm	327
• Brand New Categories of Cryptographic Hash Functions: A Survey	335
• Modeling Transport in Graphene-Metal Contact and Verifying Transfer Length Method Characterization	355
• Actor Double Critic Architecture for Dialogue System	363
• Design of Miniaturized Microstrip Antenna with Semi-Fractal Structure For GPS/GLONASS/Galileo Applications	373
• GSM based Water Salinity Monitoring System for Water Gate Management in Salt Farms	383
• 1WQC Pattern Scheduling to Minimize the Number of Physical Qubits	391
• New Platform for IOT Application Management Based on Fog Computing	399
• DPRSMR: Deep learning-based Persian Road Surface Marking Recognition	409
• Determination of the Maximum Dynamic Range of Sinusoidal Frequencies in A Wireless Sensor Network with Low Sampling Rate	419
• A New Low-Stress Boost Converter with Soft-Switching and Using Coupled-Inductor Active Auxiliary Circuit	433
• A New Hybrid NMF-based Infrastructure for Community Detection in Complex Networks	443
• Revised Estimations for Cost and Success Probability of GNR-Enumeration	459

JECEI

**Electrical and Computer
Engineering Innovations (JECEI)**

Electrical and Computer Engineering Innovations

Vol. 11 No. 2, Summer-Fall 2023

Journal of

Semiannual Publication

Volume 11, Issue 2, Summer-Fall 2023



Editor-in-Chief: Prof. Reza Ebrahimpour

Faculty of Computer Engineering, Shahid Rajaee University, Iran

Associate Editors:

Prof. Muhammad Taher Abuelma'atti

Faculty of Electrical Engineering, King Fahd University of Petroleum and Minerals, Saudi Arabia

Prof. Mojtaba Agha Mirsalim

Department of Electrical Engineering, Amirkabir University of Technology, Iran

Prof. Vahid Ahmadi

Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Iran

Prof. Nasour Bagheri

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Prof. Seyed Mohammad Taghi Bathaee

Faculty of Electrical Engineering, Power Department, K. N. Toosi University of Technology, Iran

Prof. Fadi Dornaika

Universidad del Pais Vasco, Leioa, Spain

Prof. Reza Ebrahimpour

Faculty of Computer Engineering, Shahid Rajaee University, Iran

Prof. Fary Ghassemloooy

Faculty of Engineering and Environment, Northumbria University, UK

Prof. Nosrat Granpayeh

Faculty of Electrical Engineering, K. N. Toosi University of Technology, Iran

Prof. Erich Leitgeb

Institute of Microwave and Photonic Engineering, Graz University of Technology, Austria

Prof. Juan C. Olivares-Galvan

Department of Energy, Universidad Autónoma Metropolitana, Mexico

Prof. Saeed Olyaei

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Prof. Masoud Rashidinejad

Department of Electrical Engineering, Shahid Bahonar University, Iran

Prof. Raj Senani

Division of Electronics and Communication Engineering, Netaji Subhas Institute of Technology, India

Prof. Mohammad Shams Esfand Abadi

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Prof. Vahid Tabataba Vakili

School of Electrical Engineering, Iran University of Science and Technology, Iran

Prof. Ahmed F. Zobaa

Department of Electronic and Computer Engineering, Brunel University, UK

Dr. Kamran Avanaki

Department of Biomedical Engineering, University of Illinois in Chicago

Department of Dermatology School of Medicine, University of Illinois in Chicago Scientific Member, Barbara Ann Karmanos Cancer Institute

Dr. Debasis Giri

Department of Computer Science and Engineering, Haldia Institute of Technology, India

Dr. Peyman Naderi

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Dr. Masoumeh Safkhani

Faculty of Computer Engineering, Shahid Rajaee University, Iran

Dr. Mahmood Seifouri

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Dr. Shahriar Shirvani Moghaddam

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Dr. Jian-Gang Wang

Department of Computer Vision and Image Understanding, Institute for Infocomm Research, Singapore

Executive Manager: Dr. Masoumeh Safkhani

Faculty of Computer Engineering, Shahid Rajaee University, Iran

Responsible Director: Prof. Saeed Olyaei

Faculty of Electrical Engineering, Shahid Rajaee University, Iran

Assisted by: Mrs. Fahimeh Hosseini

License Holder: Shahid Rajaee Teacher Training University (SRTTU)

Address: Lavizan, 16788-15811, Tehran, Iran.

Journal of Electrical and Computer Engineering Innovations

Vol. 11; Issue 2: 2023

Contents

Design of a Synchronous Reluctance Motor with New Hybrid Lamination Rotor Structure	243
<i>R. Rouhani, S. E. Abdollahi, S. A. Gholamian</i>	
Model Predictive Control of Linear Induction Motor Drive with End Effect Consideration	253
<i>P. Hamedani, S. Sadr</i>	
An Improved Approach to Blind Image Steganalysis Using an Overlapping Blocks Idea	263
<i>V. Sabeti</i>	
Development of Wind Turbine Fault Analysis Setup based on DFIG Hardware in the Loop Simulator	277
<i>M. Kamarzarrin, M. H. Refan, P. Amiri, A. Dameshghi</i>	
Design Optimization of the Delta-Shape Interior Permanent Magnet Synchronous Motor for Electric Vehicle Application	291
<i>S. Nasr, B. Ganji, M. Moallem</i>	
PAPR Reduction in OFDM UOWC System Employing Repetitive Clipping and Filtering (RCF) Method	301
<i>B. Noursabbaghi, G. Baghersalimi, A. Pouralizadeh, O. Mohammadian</i>	
Centrality and Latent Semantic Feature Random Walk (CSRW) in Large Network Embedding	311
<i>M. Taherparvar, F. Ahmadi Abkenari, P. Bayat</i>	
Mutual Coupling Reduction in MIMO Microstrip Antenna by Designing a Novel EBG with a Genetic Algorithm	327
<i>R. Shirmohamadi Suiny, M. Bod, G. Dadashzadeh</i>	
Brand New Categories of Cryptographic Hash Functions: A Survey	335
<i>B. Sefid-Dashti, H. Daghigh, J. S. Sartakhti</i>	
Modeling Transport in Graphene-Metal Contact and Verifying Transfer Length Method Characterization	355
<i>B. Khosravi Rad, M. khaje, A. Eslami Majd</i>	
Actor Double Critic Architecture for Dialogue System	363
<i>Y. Saffari, J. S. Sartakhti</i>	

Design of Miniaturized Microstrip Antenna with Semi-Fractal Structure For GPS/GLONASS/Galileo Applications <i>S. Komeyliyan, M. Tayarani, S. H. Sedighy</i>	373
GSM based Water Salinity Monitoring System for Water Gate Management in Salt Farms <i>M. Enriquez, A. Abella</i>	383
1WQC Pattern Scheduling to Minimize the Number of Physical Qubits <i>E. Nikahd, M. Houshmand, M. Houshmand</i>	391
New Platform for IoT Application Management Based on Fog Computing <i>S. Kalantary, J. Akbari Torkestani, A. Shahidinejad</i>	399
DPRSMR: Deep learning-based Persian Road Surface Marking Recognition <i>S. H. Safavi, M. Sadeghi, M. Ebadpour</i>	409
Determination of the Maximum Dynamic Range of Sinusoidal Frequencies in A Wireless Sensor Network with Low Sampling Rate <i>A. Maroosi, H. Khaleghi Bizaki</i>	419
A New Low-Stress Boost Converter with Soft-Switching and Using Coupled-Inductor Active Auxiliary Circuit <i>M. A. Latifzadeh, P. Amiri, H. Allahyari, H. Faezi</i>	433
A New Hybrid NMF-based Infrastructure for Community Detection in Complex Networks <i>M. Ghadirian, N. Bigdeli</i>	443
Revised Estimations for Cost and Success Probability of GNR-Enumeration <i>G. R. Moghissi, A. Payandeh</i>	459



Research paper

Design of a Synchronous Reluctance Motor with New Hybrid Lamination Rotor Structure

R. Rouhani, S. E. Abdollahi*, S. A. Gholamian

Electrical and Computer Engineering Dept., Babol Noshirvani University of Technology, Babol, Iran.

Article Info

Article History:

Received 14 July 2022
Reviewed 27 August 2022
Revised 24 September 2022
Accepted 22 October 2022

Keywords:

Synchronous reluctance motor
Hybrid laminations design
Anisotropic rotor
Axial laminations
Segmental laminations
Average torque
Torque ripple

*Corresponding Author's Email
Address: e.abdollahi@nit.ac.ir

Abstract

Background and Objectives: The rotor of synchronous reluctance machines (SynRM) is conventionally designed and implemented in two types of axially-laminated anisotropic (ALA) and transversely-laminated anisotropic (TLA). Torque ripple and power factor have always been the design challenges of this machine; however, with proper design, their values can be as close as possible to the desired value. Each of these two structures has some advantages over the other, in terms of electromagnetic performance and ease of construction. For the first time, in this paper, a hybrid anisotropic rotor is presented with both radial and axial laminations, based on the theory of anisotropic rotor structure for the fundamental harmonic and isotropic rotor structure for other harmonics, so that the designed motor meets the advantages of both structures as much as possible.

Methods: To this end, the proposed design is implemented and investigated a Magnetic Equivalent Circuit (MEC) for the first slot harmonic on a machine with stator of 24-slots. To evaluate the proposed design, its electromagnetic performance is simulated using Finite Element Method.

Results: The theory-based conceptual design method is applied to a rotor with new structure and simulation results including average torque, power factor and torque ripple of the machine are presented.

Conclusion: Based on the obtained simulation results and comparing performance of the proposed design with other structures, it is shown that there will be a significant improvement in electromagnetic features including torque ripple, average torque and power factor and the proposed design has lower torque ripple than ALA rotor and higher average torque and power factor than TLA rotor.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

According to SynRM rotor geometry, anisotropic structure is typically created in two forms of ALA and TLA. This anisotropic structure is generated by an appropriate distribution of flux carriers (made of magnetic material) and flux barriers (made of non-magnetic material or air). The lack of magnets and winding on the rotor, in addition to the simplicity of the structure, leads to a reduction in the cost of material and the manufacturing and assembly process. The lack of magnets also reduces the machine

sensitivity to temperature and fault current, resulting a simpler control method, higher torque-to-current ratio, and the ability to tolerate short-term overload capability [1]- [3]. Various investigations have been carried out regarding ALA rotors during 1990s. Meanwhile, rotor laminations with high thickness [4] and low thickness (like the thickness of normal laminations applied in TLA rotors) [5] are investigated. The lamination thickness which is 1.56 mm in [6] is rather high. Although the major component of the magnetic field is constant relative to the rotor, higher-order flux harmonics in ALA rotors result

in eddy current loss in laminations, which can impose additional losses to the machine [7], [8].

Increasing the number of laminations below each pole leads to a decrease in torque ripple, as well as reduction in losses caused by harmonic flux densities [5], also better distribution of insulations and thus increasing their effect as a flux barrier in the rotor [9]. To reduce these losses, radial slits can be created in 2-4 locations on each lamination along the d-axis to increase the length of eddy current path [10]. In general, rotors with strong anisotropy structure are made possible by ALA topology using a large number of laminations [11]. The simplest structure for SynRM machine is investigated in [12], [13]. This structure was so simple in terms of magnetic and mechanical aspects. However, magnetically, this structure has torque ripple and iron rotor losses and low torque density. In [14] and [15], instead of insulation, flexible magnets and ferrites are respectively employed between laminations. In ALA, due to the absence of tangential ribs between flux carriers in the flux weakening zone, the stator flux decreases significantly and the saliency ratio increases remarkably, which eventually leads to improved power factor [14]. The rotor anisotropy of SynRM leads to high harmonics of airgap flux density, which makes its electromagnetic analytical modeling complicated. In this regard various analytical methods are employed in order to predict the d- and q- axis inductances as well as average torque and torque ripple. Commonly MEC is used to quickly evaluate machine average torque with an acceptable accuracy [16], [17]. Also, in order to increase its accuracy and capability to calculate the torque ripple, mixed MEC is coupled with conformal mapping and other analytical tools [18], [19].

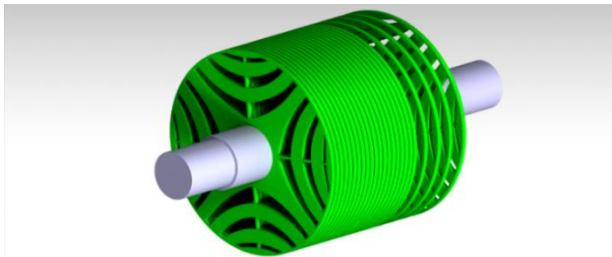


Fig. 1: Transversely-Laminated Anisotropic (TLA).

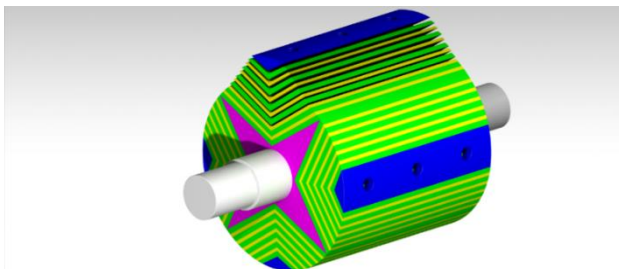


Fig. 2: Axially-Laminated Anisotropic (ALA).

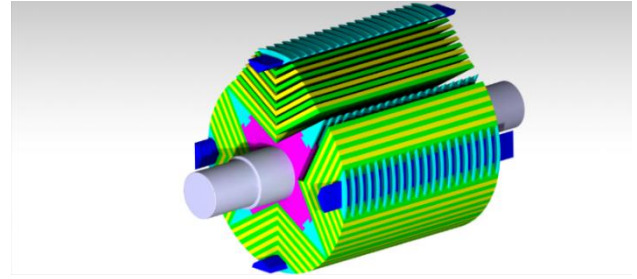


Fig. 3: Hybrid-laminated Anisotropic (HLA).

In this paper, based on the inductance theory regarding anisotropic rotors, a new structure is presented to achieve a high saliency ratio as well as the optimal torque ripple. This is the evolved version of conventional ALA and TLA structures. In this regard proposed structure is introduced first and then, analytically and numerically, shown that like ALA, it has a high saliency ratio (average torque and power factor) and, like TLA, has a desirable torque ripple. From this perspective, first, design parameters of ALA and TLA topologies are briefly reviewed. Then design fundamentals of the hybrid topology derived from the inductance theory in anisotropic rotors are presented, and finally, magnetic simulation of the proposed topology and its comparison with conventional ones are examined.

Design Parameters of Anisotropic Rotor

The operating principles of SynRM are based on inductance change of each phase of the stator with rotation of rotor. The ratio of the maximum inductance to the minimum inductance is called saliency ratio, shown by ξ . This saliency ratio is optimized by coefficient of α in ALA rotors, which is obtained by:

$$\alpha = \frac{W_i}{W_l} \quad (1)$$

in which W_i and W_l are insulation and lamination width, respectively.

The optimal value of α coefficient of ALA topology is investigated in various literatures, such that, this coefficient is 0.59 in [5]; 0.5 in [10]; in order to reach the peak torque value, approximately 0.43 in [20]; to reach the peak of constant-power speed range (CPSR) approximately 0.67, and 1 in [8], [9], [21]. A comprehensive study of effect of changing α coefficient from zero (isotropic rotor) to 1 on the inductance of d- & q-axes is carried out in [22], in which the optimal value of 0.5 is obtained. The smallness of α coefficient means that more iron is employed in the core and accordingly, lower flux density, hence, reduction in losses due to the flux harmonics in the teeth.

Design method of TLA rotor geometry is brought in [3], [11], in which the saliency ratio depends on the number

and distribution of flux carriers and flux barriers in rotor [3], [23]. In general, the number of geometry design parameters of TLA rotor is greater than the number of ALA rotor parameters [24]; however, unequal size of lamination sheets cutting, assembly and ripple improvement techniques (such as skewing, etc.) in the ALA rotor is much more difficult [2]. The saliency ratio (ξ) in TLA rotors is slightly lower than ALA rotor. Although in some literatures, saliency ratio in anisotropic rotors (especially ALA rotors) is estimated to be greater than 10 [10], [21], [25], [26], and even Kostko predicted it to be up to 25 in some structures. In loading condition, however, its saliency ratio is less than 10 [9].

The Principles of Designing Hybrid-Laminated Rotor

Anisotropic rotors (due to the presence of stator slots) have torque ripple inherently, because the fluxes existing inside and around the rotor change (in both amplitude and direction of bending of the flux) as they pass from one stator slot to another. Therefore, anisotropic rotors should be designed in a way that, in addition to approaching the highest saliency ratio, simultaneously has the highest uniformity of flux variations inside the rotor. SynRM machine inherently have high ripple, and for achieving optimal ripple, symmetric methods [3], [27] and asymmetric methods [24], [28], [29] are employed in TLA. In ALA, as mentioned earlier ripple can be slightly improved by considering the optimal α coefficient and selecting the appropriate number of laminations.

In this paper, the α coefficient for the axial laminations was considered to be 0.5, and as mentioned earlier, increasing the number of sheets in ALA, according to [5], leads to a decrease in torque ripple. Although ALA rotors have an almost smooth and circular surface, the holder section of the axial sheets around d and q axes is solid and made of non-magnetic material (Fig. 2); hence, ALA rotors have an inherent cutoff around q-axis. Although cutoff typically increases saliency ratio (hence average torque and power factor) [30], it significantly increases the torque ripple, too [31]. For this reason, TLA rotors are preferred over their longtime rival, i.e. ALA ones, because they have desired average torque, power factor, and torque ripple. Moreover, its manufacturing (cutting sheets with various dimensions and angles) and assembling process (mounting rotor sheets, assembly, and skewing the rotor) are easier than ALA.

Given in [32]-[34], d- and q-axes inductance and SynRM machine torque equation for the fundamental component are:

$$L_d(\theta) = L_{do} + \sum_{v=1}^{\infty} \Delta L_d \cos(v P N_s \theta) \quad (2)$$

$$L_q(\theta) = L_{qo} - \sum_{v=1}^{\infty} \Delta L_q \cos(v P N_s \theta) \quad (3)$$

$$L_{dq}(\theta) = - \sum_{v=1}^{\infty} \Delta L_{dq} \sin(v P N_s \theta) \quad (4)$$

$$T = \frac{m}{2} P [I_d I_q (L_{do} - L_{qo} + (\Delta L_d + \Delta L_q) \cos(P N_s \theta)) - (\Delta L_{dq} \sin(P N_s \theta) (I_d^2 - I_q^2))] - \frac{1}{2} P N_s [2 I_d I_q \Delta L_{dq} \cos(P N_s \theta) + I_d^2 \Delta L_d \sin(P N_s \theta) - I_q^2 \Delta L_q \sin(P N_s \theta)] \quad (5)$$

where v , P , N_s , θ , and m are Slot harmonic order, Machine's pole pair number, Number of stator slots per pole pair, angle of rotor reference frame and Machine number of phases respectively. In (5) there are two terms of torque ripple components. The first terms, ΔL_d and ΔL_q are the inductance changes associated with each axis due to oscillations caused by the open stator slot (Carter's coefficient) as well as the rotating flux. The second term, ΔL_{dq} is the mutual inductance changes between the stator and rotor teeth, which generates substantial torque oscillation.

In this paper, as shown in Figs. 4 & 5, these inductance variations could be reduced using proposed technique in a way that d & q axes of ALA rotor, separated by non-magnetic holders, and ultimately led to a lack of magnetic uniformity and solidity of the rotor, be modified by the radial laminated segments (TLA segmental).

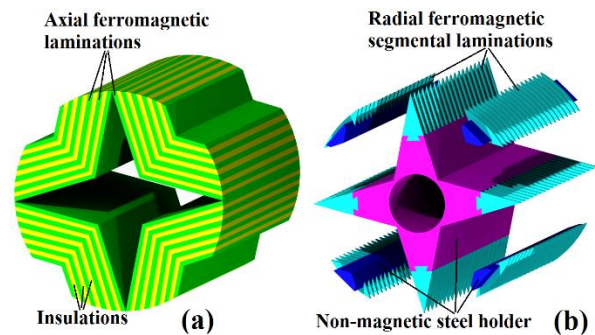


Fig. 4: The major sections of hybrid lamination rotor (a) ALA part (b) TLA part (segmental).

The design presented in Fig. 4-b, the d-axis segments are surrender and the q-axis segments surrounded on the holders.

The arrangement of d & q axes segments on the holders are not significantly different from each other magnetically, however, this is important due to the position of the segments and in order to have a better establishment and increase the mechanical strength and due to the limited space in the rotor.

Radial laminations which are segmentally established around d & q axes, ultimately have led the fluxes inside the rotor change less as they pass from one slot of the stator to another one, thus reducing the torque ripple.

Thus, the use of a rotor with two perpendicular laminations results in a structure with better peripheral uniformity while maintaining an anisotropic structure.

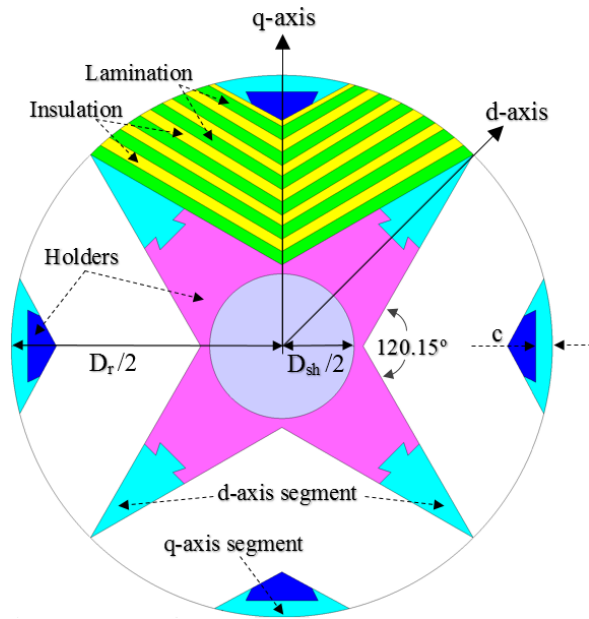


Fig. 5: Geometrical parameters of HLA rotor.

Simulation of Magnetic Performance

Improving the magnetic performance of a machine and achieving a structure that simultaneously has all the desired parameters is always one of the main design objectives, which is, in some cases, impossible to achieve, or it's necessary to reconcile a number of parameters in multi-objective functions and its optimization.

To verify that the proposed topology provides all the intended electromagnetic parameters adequately, a 1500-watt induction motor stator was used, the rotor of which was replaced by three different rotors with ALA, TLA and HLA topologies, and the electromagnetic features of all three structures were examined and compared. The geometric parameters of the stator and all three rotors are listed in Table 1. A diagram of developed electromagnetic torque related to all three rotor topologies is displayed in Fig. 9.

HLA performance features, shown in Table 2, suggest that the proposed design can be effective in performance improvement of average torque, power factor, and torque ripple.

In other words, the best design for anisotropic rotors is to have the highest saliency ratio (anisotropic feature) along with the least amount of variations in terms of stator teeth for rotor rotation as large as a polar step of the rotor, so that it leads to maximum average torque and minimum torque ripple, respectively.

Flux density distribution of TLA, ALA and HLA rotors are shown in Figs. 6, 7, and 8 respectively.

Table 1: Design characteristics of the SynRM

Common Parameter	value
Rated power	1.5 (kw)
Air gap	0.5 (mm)
Stator outer diameter	140 (mm)
Stack length	90 (mm)
S_{so}	2.5 (mm)
n_s	24
Nominal current	4.5 (A)
Based speed	1500 (rpm)
Number of turns per slot	76
D_{sh}	24 (mm)
Magnetic Sheet	M400-50A
TLA Rotor	
N_b	3
D_r	90 (mm)
R_{so}	3.3 (mm)
TR_x	1 (mm)
RR_x	1 (mm)
α_1	11.6 (deg)
k_{wq}	0.8
ALA Rotor	
W_i	0.25 (mm)
W_l	0.5 (mm)
HLA Rotor	
C	2.75 (mm)

Table 2: Comparison of the electromagnetic characteristic of SynRM with TLA, ALA AND HLA rotor

Unit	Parameter	TLA	ALA	HLA
Nm	T_{av}	11.01	13.07	12.6
%	ΔT	7.63	23.48	9.12
A	I	4.5	4.5	4.5
Nm/A	T/A	2.44	2.90	2.8
mH	L_d	329.53	317.50	326.57
mH	L_q	98.64	59.66	67.99
---	ξ	3.34	5.32	4.80
---	PF	0.539	0.684	0.655

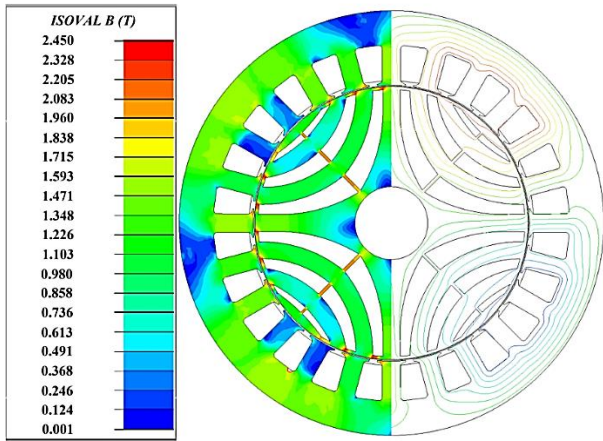


Fig. 6: Flux density distribution of TLA rotor in 4.5A current.

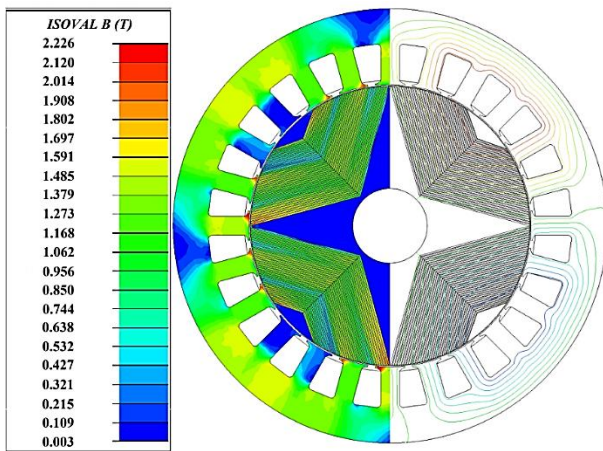


Fig. 7: Flux density distribution of ALA rotor in 4.5A current.

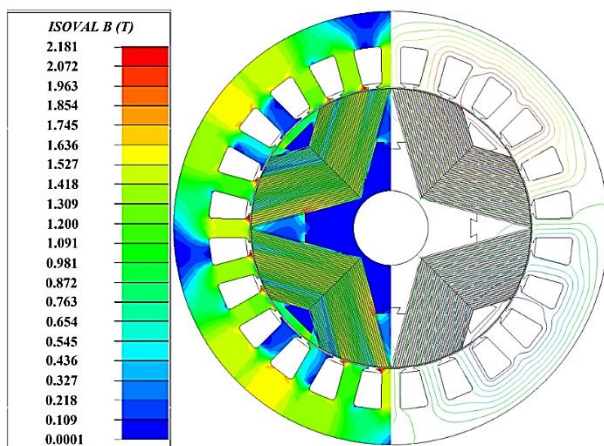


Fig. 8: Flux density distribution of HLA rotor in 4.5A current.

A. Magnetic Simulation Results

The effect of the d & q axes segments on the torque and inductance characteristics of each phase, are separately shown in Figs. 9, 10, and 11, respectively. It is observed that if the segments only affect the d-axis, since this segment is in the path of the q-axis fluxes, the loops

of the q-axis fluxes are closed in a more uniform path. According to (3), by the rotor rotation, the flux changes in the peripheral areas of the rotor decreases as it passes from one stator slot to another, resulting in a lower flux variation of q-axis and ultimately, the lower inductance variations of torque ripple generator of q-axis (ΔL_q).

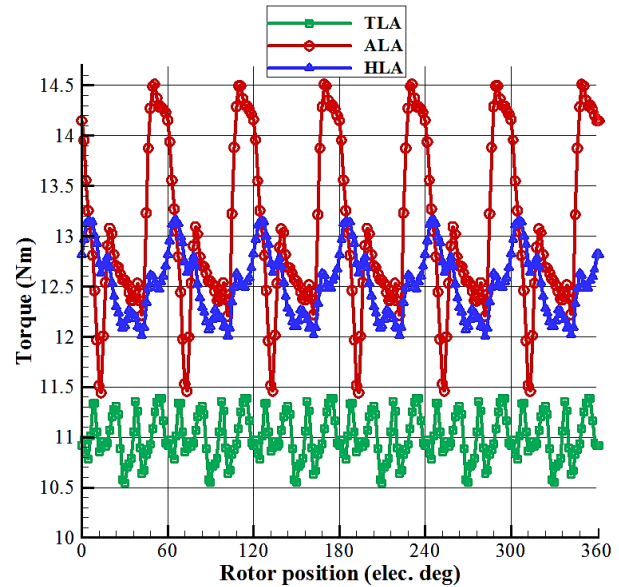


Fig. 9: Torque characteristics of SynRM with TLA, ALA & HLA rotors.

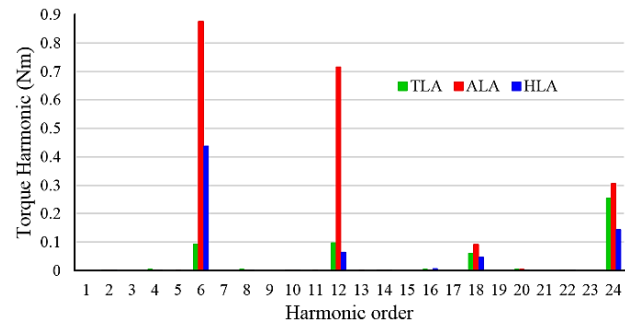


Fig. 10: Harmonic components of the torque ripple.

If the segments only affect the q-axis, because this segment is in the path of the d-axis fluxes, the loops of the d-axis flux are closed in a more uniform path. According to (2), by the rotor rotation, flux changes in the surrounding areas of the rotor decreases as it passes from one stator slot to another, resulting in a lower flux variation of d-axis and ultimately, the lower inductance variations of torque ripple generator of d-axis (ΔL_d). On the other hand, the presence of a non-magnetic holder around the q-axis in ALA topology leads to ΔL_{dq} . This quantity is produced due to simultaneous distance or approach of the holder end to the stator teeth. By placing magnetic segments around the q-axis, the rotor has more uniform external areas, resulting in reduced mutual inductance that changes between the stator and rotor

teeth. Therefore, according to Table 2 and Fig. 9, torque ripple reduction in HLA topology is greater than torque ripple reduction in ALA.

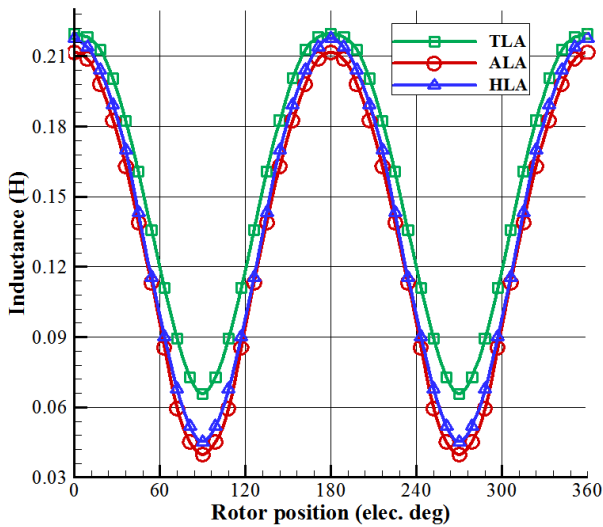


Fig. 11: Phase inductance variation of SynRM with TLA, ALA & HLA rotors.

In Fig. 11, the phase inductance variations for ALA, TLA, and HLA rotors for an electrical cycle are shown. It can be seen that the $L_{\text{phase-max}}$ of the HLA rotor is almost equal to the $L_{\text{phase-max}}$ of the TLA rotor and the $L_{\text{phase-min}}$ of the HLA rotor is almost equal to the $L_{\text{phase-min}}$ of the ALA rotor. From the Fig. 11 and Table 2, it can be found that HLA, in addition to having average torque and higher power factor, its ripple is almost the same as TLA. It also requires less magnetization current, hence its copper losses and machine torque density would improve.

The inductance of each phase is achieved by Finite Element Software, and for transferring this inductance to rotor reference frame, it should be multiplied by 1.5. Therefore, given in [4], [35], the d- & q-axes inductances, are obtained by:

$$L_d = \frac{3}{2} L_{\text{phase-max}} \quad (6)$$

$$L_q = \frac{3}{2} L_{\text{phase-min}} \quad (7)$$

where $L_{\text{phase-max}}$ and $L_{\text{phase-min}}$ are Inductances of each phase for the maximum rated current in the direction of the d and q rotor axis respectively. The saliency ratio is obtained:

$$\xi = \frac{L_d}{L_q} \quad (8)$$

Given [9], [22], the maximum power factor is equal to:

$$PF_{\text{max}} = \frac{\xi - 1}{\xi + 1} = \frac{L_d - L_q}{L_d + L_q} \quad (9)$$

Since in addition to the fundamental component of flux density, the spatial and slot harmonics also flow in the

rotor, this flux manifests itself as torque ripple, shown in Figs. 9 and 10.

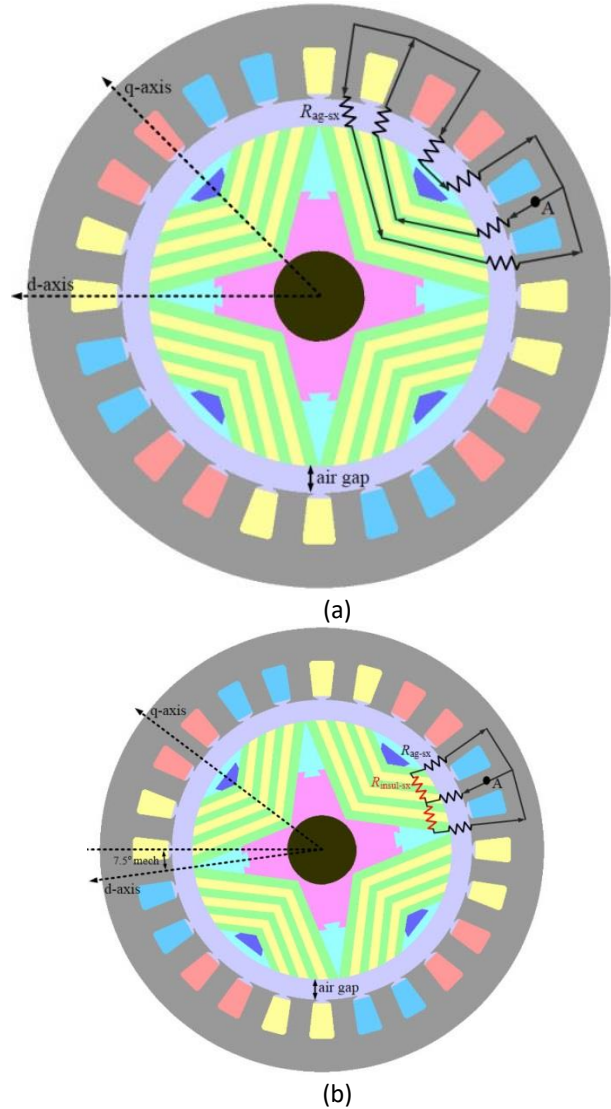


Fig. 12: The magnetic equivalent circuit of the passage of fluxes from each stator tooth to other teeth through the rotor path a) The magnetic reluctance of the d axis from the point of view of the A tooth- R_{SA-d} b) The magnetic reluctance of the q axis from the point of view of the A tooth- R_{SA-q} .

B. The Performance Theory of the Proposed Rotor

In other publications, various magnetic equation circuits are modeled for synchronous reluctance motors, usually based on the fundamental component of the machine's torque-generating magnetic field, but these circuits do not have the desired capability to describe other harmonic components and their lateral issues. If the magnetic equivalent circuit is studied and modeled for the smallest region (a stator tooth and its flux flow path through the rotor and air gap), then the harmonic behavior of the machine can be predicted and evaluated in the final results.

Designers' approach to designing the geometry of electric machines from the point of view of harmonics (except for the fundamental harmonic) can be generally based on three axes; First, the machine parameters should be designed in such a way that almost no asynchronous harmonics (space and slot harmonics) are produced (for example, with a significant increase in the number of stator slot or a machine without teeth or stator core). Second, the machine parameters should be designed to decrease the amplitude of the asynchronous harmonics (for example, using the windings method of the chorded pitch or fractional slot). Third, the machine parameters should be designed so that the most effective slot harmonic of a pole pair is opposite the adjacent pole pair to be deleted (for example, rotor with asymmetric flux barrier geometry).

However, the design of the rotor of electric machines can be such that the rotor structure is anisotropic for the fundamental component of the rotating magnetic field and isotropic for other harmonic components so that the harmonic components of the rotor and stator have no interaction with each other. In SynRM, the most effective harmonic in the torque ripple characteristic is the first order of slot harmonics. Since all torque-ripple harmonic components, both space and slot harmonics, eventually have to flow through the stator teeth and close their way through the air gap and the rotor, it seems better to be modeled the possible flux paths from the point of view of just one stator tooth.

As shown in Fig. 12, since the cycle of the slot harmonic is 15 mechanical degrees and the R_{SA-d} and R_{SA-q} are 180 electrical degrees different from each other, in Fig. 12(a), (b) the rotor is rotated for 7.5 mechanical degrees. With proper design of the rotor structure, an isotropic path can be created from the A-tooth view for different rotor positions. Considering the changes in magnetic reluctance ($dR / d\theta$) and considering the passage of fluxes from each stator tooth to other teeth through the rotor path, the magnetic equivalent circuit for each stator tooth is shown in Fig. 12. Therefore, to minimize the torque ripple ($dR / d\theta$), it is sufficient that the magnetic reluctance seen from the d and q axes for the first slot harmonic component of the stator from the point of view of each stator tooth (eg tooth A) is studied and the rotor should be designed to be $R_{SA-d} \approx R_{SA-q}$. Therefore, regardless of the magnetic resistance of the iron part of the stator and rotor:

$$R_{SA-d} = 3 R_{ag-sx} \quad (10)$$

$$R_{SA-q} = 1.5 R_{ag-sx} + 0.5 R_{insul-sx} \quad (11)$$

$$R_{SA-d} \approx R_{SA-q} \Rightarrow 3 R_{ag-sx} \approx R_{insul-sx} \quad (12)$$

where R_{SA-d} , R_{SA-q} , R_{ag-sx} , and $R_{insul-sx}$ are magnetic reluctance seen from the d and q axes, air gap magnetic reluctance and insulation magnetic reluctance from the point of view of each stator tooth respectively. This ensures that the changes in magnetic reluctance from the point of view of each stator tooth for rotation of the rotor in different positions are always minimal, and these values will be good starting points for the final design of the rotor geometry.

Therefore, for the HLA rotor, in the first place by selecting the appropriate number of laminations and insulations, also in the second place by replacing the ferromagnetic segments at the ends of the d and q axes of the rotor, an isotropic structure can be created and $dR/d\theta$ and finally the torque ripple is reduced. In other words, in the proposed structure due to the lack of radial and tangential ribs, the average torque is higher than TLA, also due to the appropriate number of insulations and laminations and ferromagnetic segments around the d and q axes of the rotor (creating an isotropic environment for slot harmonics) torque ripple are significantly better than ALA and almost equal to TLA.

C. Space Harmonic Effect

According to torque characteristics and Fig. 10, it is observed that HLA and TLA have the most desired performance. By referring to Fig. 10 and comparing the torque characteristics of the two TLA and HLA rotors, the 6th harmonic component available in the torque characteristics for the HLA rotor is significant. The 6th harmonics in the torque feature is caused by the 5th and 7th space harmonics [36]. These space harmonics can be improved by chorded winding. If the winding step is shortened by $1/n$, the winding coefficient (k_w) of (13) should be affected by the number of the winding turns:

$$k_w = \sin\left(\frac{180}{2} * \left(1 - \frac{1}{n}\right)\right) \quad (13)$$

In this relation, n is the target space harmonic order for attenuation or elimination. In this paper, n=6 was considered for attenuation of space 5th and 7th harmonics (given the number of slots and stator poles).

The torque ripple characteristics (with chorded winding) for TLA and HLA rotors are shown in Figs. 13, 14. It is obvious that ripple of HLA had a significant improvement. Fig. 14 proves that this ripple improvement is due to a significant decrease in the 6th harmonic in torque ripple. As it was expected, chorded winding had no effect on slot harmonics. Therefore, since in torque ripple characteristics of SynRM with TLA rotor there is not any remarkable space harmonic, chorded winding plays an insignificant role in performance improvement of TLA rotor. Magnetic quantities of HLA and TLA rotors with chorded and full pitch windings are listed in Table 3.

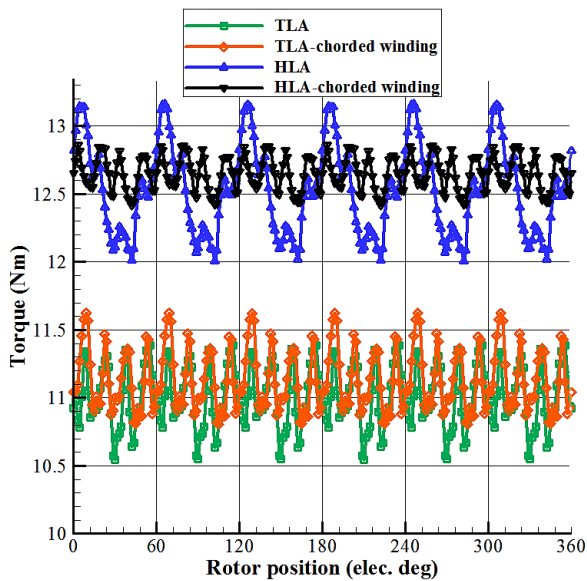


Fig. 13: A comparison of torque characteristics of HLA and TLA rotors with chorded and full pitch windings.

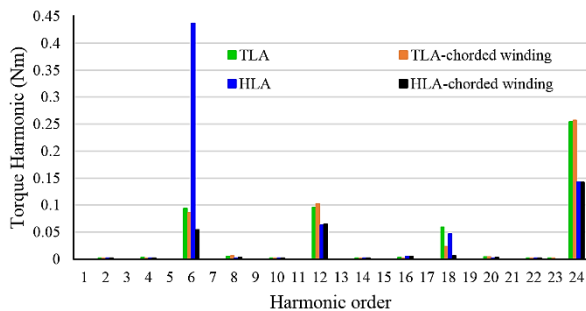


Fig. 14: Torque ripple harmonics components of SynRM with HLA and TLA rotors with chorded and full pitch windings.

Table 3: Comparison of the electromagnetic characteristic of SynRM with TLA, and proposed HLA rotor

Different Rotor					
Unit	Parameter	TLA	TLA-chorded winding	HLA	HLA-chorded winding
Nm	T_{av}	11.01	11.13	12.6	12.65
%	ΔT	7.63	7.34	9.12	3.5
A	I	4.5	4.5	4.5	4.5

Conclusion

Although ALA SynRM is not as widely recognized in the industry as TLA ones, it is further studied as hybrid lamination anisotropy in this paper and the possibility of improving its electromagnetic performance was confirmed. TLA has less ripple and simpler assembly and structure; nonetheless, ALA has some advantages such as higher power factor and average torque density than TLA. It is shown that proposed HLA exhibits moderate features

of both TLA and ALA rotors that although its torque ripple is a bit more than TLA but its average torque density and PF is modified. Also due to the magnetic characteristics of the desired torque, there is no need to any skewing to reduce torque ripple. Although HLA has more complex assembly than TLA, it has fewer design parameters, hence, it is easier for its optimization to achieve the desired electromagnetic characteristics.

Author Contributions

R. Rouhani collected the data, carried out the analysis and wrote paper, S. E. Abdollahi and S. A. Gholamian interpreted the results and supervised the research.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

W_i	Insulation Width
W_l	Lamination Width
L_d	d- axis inductance
L_q	q- axis inductance
L_{dq}	Mutual inductance
L_{do}	Constant amount of d-axis inductance
L_{qo}	Constant amount of q-axis inductance
ν	Slot harmonic order
ΔL_d	Amplitude of changes of d-axis inductance
ΔL_q	Amplitude of changes of q-axis inductance
ΔL_{dq}	Fundamental amplitude mutual inductance
P	Machine's pole pair number
N_s	Number of stator slots per pole pair
ϑ	Angle of rotor reference frame
m	Machine number of phases
I_d	d-axis current
I_q	q-axis current
N_s	Number of stator slots per pole pair
ξ	Saliency ratio
R_{SA-d}	d- axis magnetic reluctance from the point of view of each stator tooth
R_{SA-q}	q- axis magnetic reluctance from the point of view of each stator tooth

R_{ag-sx}	Air gap magnetic reluctance from the point of view of each stator tooth
$R_{insul-sx}$	Insulations magnetic reluctance from the point of view of each stator tooth
n	Target space harmonic order
k_w	Winding coefficient
ALA	Axially-laminated anisotropic
TLA	Transversely-laminated anisotropic
CPSR	Constant-power speed range

References

- [1] N. Bianchi, S. Bolognani, E. Carraro, M. Castiello, E. Fornasiero, "Electric vehicle traction based on synchronous reluctance motors," *IEEE Trans. Ind. Appl.*, 52: 4762-4769, 2016.
- [2] S. Taghavi, "Design of synchronous reluctance machines for automotive applications," Concordia University, 2015.
- [3] R. R. Moghaddam, F. Gyllensten, "Novel high-performance SynRM design method: An easy approach for a complicated rotor topology," *IEEE Trans. Ind. Electron.*, 61: 5058-5065, 2014.
- [4] Y. H. Kim, J. H. Lee, "Optimum design of ALA-SynRM for direct drive electric valve actuator," *IEEE Trans. Magn.*, 53(4): 1-4, 2017.
- [5] B. J. Chalmers, L. Musaba, "Design and field-weakening performance of a synchronous reluctance motor with axially laminated rotor," *IEEE Trans. Ind. Appl.*, 34: 1035-1041, 1998.
- [6] E. Obe, "Calculation of inductances and torque of an axially laminated synchronous reluctance motor," *IET Electr. Power Appl.*, 4: 783-792, 2010.
- [7] I. Scian, D. G. Dorrell, P. J. Holik, "Assessment of losses in a brushless doubly-fed reluctance machine," *IEEE Trans. Magn.*, 42: 3425-3427, 2006.
- [8] D. G. Dorrell, I. Scian, E. M. Schulz, R. B. Betz, M. Jovanovic, "Electromagnetic considerations in the design of doubly-fed reluctance generators for use in wind turbines," in *Proc. IECON 2006 - 32nd Annual Conference on IEEE Industrial Electronics: 4272-4277*, 2006.
- [9] D. Staton, T. Miller, S. Wood, "Maximising the saliency ratio of the synchronous reluctance motor," in *IEE Proceedings B (Electric Power Applications)*: 249-259, 1993.
- [10] I. Boldea, Z. Fu, S. Nasar, "Performance evaluation of axially-laminated anisotropic (ALA) rotor reluctance synchronous motors," *IEEE Trans. Ind. Appl.*, 30: 977-985, 1994.
- [11] Y. Wang, G. Bacco, N. Bianchi, "Geometry analysis and optimization of PM-assisted reluctance motors," *IEEE Trans. Ind. Appl.*, 53: 4338-4347, 2017.
- [12] P. Lawrenson, L. Agu, "Theory and performance of polyphase reluctance machines," in *Proc. Institution of Electrical Engineers*: 1435-1445, 1964.
- [13] P. J. Lawrenson, S. Gupta, "Developments in the performance and theory of segmental-rotor reluctance motors," in *Proc. Institution of Electrical Engineers*: 645-653, 1967.
- [14] W. L. Soong, N. Ertugrul, "Field-weakening performance of interior permanent-magnet motors," *IEEE Trans. Ind. Appl.*, 38: 1251-1258, 2002.
- [15] I. Boldea, S. Nasar, "Emerging electric machines with axially laminated anisotropic rotors: a review," *Electr. Mach. power syst.*, 19: 673-703, 1991.
- [16] C. López, T. Michalski, A. Espinosa, L. Romeral, "Rotor of synchronous reluctance motor optimization by means reluctance network and genetic algorithm," in *Proc. 2016 XXII International Conference on Electrical Machines (ICEM)*: 2052-2058, 2016.
- [17] C. López-Torres, A. G. Espinosa, J. R. Riba, L. Romeral, "c," *IEEE Trans. Veh. Technol.*, 67: 196-205, 2017.
- [18] M. Farhadian, M. Moallem, B. Fahimi, "Analytical calculation of magnetic field components in synchronous reluctance machine accounting for rotor flux barriers using combined conformal mapping and magnetic equivalent circuit methods," *J. Magn. Mater.*, 505: 166762, 2020.
- [19] A. Hanic, D. Zarko, D. Kuhinek, Z. Hanic, "On-load analysis of saturated surface permanent magnet machines using conformal mapping and magnetic equivalent circuits," *IEEE Trans. Energy Convers.*, 33: 915-924, 2018.
- [20] W. L. Soong, D. A. Staton, T. J. Miller, "Design of a new axially-laminated interior permanent magnet motor," *IEEE Trans. Ind. Appl.*, 31: 358-367, 1995.
- [21] D. Platt, "Reluctance motor with strong rotor anisotropy," *IEEE Trans. Ind. Appl.*, 28: 652-658, 1992.
- [22] T. Matsuo, T. A. Lipo, "Rotor design optimization of synchronous reluctance machine," *IEEE Trans. Energy Convers.*, 9: 359-365, 1994.
- [23] M. Ferrari, N. Bianchi, E. Fornasiero, "Analysis of rotor saturation in synchronous reluctance and PM-assisted reluctance motors," *IEEE Trans. Ind. Appl.*, 51: 169-177, 2015.
- [24] E. Howard, M. J. Kamper, S. Gerber, "Asymmetric flux barrier and skew design optimization of reluctance synchronous machines," *IEEE Trans. Ind. Appl.*, 51: 3751-3760, 2015.
- [25] E. C. F. Lovelace, "Optimization of a magnetically saturable interior permanent-magnet synchronous machine drive," *Massachusetts Institute of Technology*, 2000.
- [26] N. Bianchi, B. J. Chalmers, "Axially laminated reluctance motor: analytical and finite-element methods for magnetic analysis," *IEEE Trans. Magn.*, 38: 239-245, 2002.
- [27] S. Taghavi, P. Pillay, "A novel grain-oriented lamination rotor core assembly for a synchronous reluctance traction motor with a reduced torque ripple algorithm," *IEEE Trans. Ind. Appl.*, 52: 3729-3738, 2016.
- [28] M. Sanada, K. Hiramoto, S. Morimoto, Y. Takeda, "Torque ripple improvement for synchronous reluctance motor using an asymmetric flux barrier arrangement," *IEEE Trans. Ind. Appl.*, 40: 1076-1082, 2004.
- [29] M. Ferrari, N. Bianchi, A. Doria, E. Fornasiero, "Design of synchronous reluctance motor for hybrid electric vehicles," *IEEE Trans. Ind. Appl.*, 51: 3030-3040, 2015.
- [30] M. J. Kamper, A. Volsdhenk, "Effect of rotor dimensions and cross magnetisation on L_d and L_q inductances of reluctance synchronous machine with cageless flux barrier rotor," *IEE J. Electr. Power Appl.*, 141: 213-220, 1994.
- [31] R. Rajabi Moghaddam, "Synchronous reluctance machine (SynRM) in variable speed drives (VSD) applications," *KTH Royal Institute of Technology*, 2011.
- [32] A. Vagati, G. Franceschini, I. Marongiu, G. Troglia, "Design criteria of high performance synchronous reluctance motors," in *Proc. Conference Record of the 1992 IEEE Industry Applications Society Annual Meeting*: 66-73, 1992.
- [33] A. Fratta, G. Troglia, A. Vagati, F. Villata, "Evaluation of torque ripple in high performance synchronous reluctance machines," in *Proc. Conference Record of the 1993 IEEE Industry Applications Conference Twenty-Eighth IAS Annual Meeting*: 163-170, 1993.
- [34] S. A. Nasar, I. Boldea, *The induction machines design handbook*: CRC press, 2009.
- [35] K. C. Kim, J. S. Ahn, S. H. Won, J. P. Hong, J. Lee, "A study on the optimal design of SynRM for the high torque and power factor," *IEEE Trans. Magn.*, 43: 2543-2545, 2007.
- [36] A. Vagati, M. Pastorelli, G. Franceschini, S. C. Petrache, "Design of low-torque-ripple synchronous reluctance motors," *IEEE Trans. Ind. Appl.*, 34: 758-765, 1998.

Biographies



Rouhollah Rouhani was born in Mahmudabad, Iran, in 1989. He received the M.Sc. degree from Babol Noshirvani University of Technology, Babol, Iran, in 2017. He is interested in electric machines topics.

- Email: rouhollah.rouhani68@gmail.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Seyed Ehsan Abdollahi was born in Sari, Iran. He received the B.Sc. degree from the Amirkabir University of Technology, Tehran, Iran, in 2002, the M.Sc. degree from Iran University of Science and Technology, Tehran, Iran, in 2005, and the Ph.D. degree from University of Tehran, Tehran, Iran, in 2014, all in Electric Power Engineering. He joined the Babol Noshirvani University of Technology, Babol, Iran, as an Assistant Professor. His current research interests include electric machine design and modeling, electric vehicle and power electronics.

- Email: s.ehsan.abdollahi@gmail.com
- ORCID: [0000-0003-2277-1060](https://orcid.org/0000-0003-2277-1060)
- Web of Science Researcher ID:
- Scopus Author ID:
- Homepage: <https://ostad.nit.ac.ir/home.php?sp=390065>



Sayyed Asghar Gholamian was born in Babolsar, Iran, in 1976. He received his B.Sc. degree in electrical engineering from K.N. Toosi University of Technology, Tehran, Iran in 1999 and his M.Sc. degree in electric power engineering from the University of Mazandaran, Babol, Iran in 2001. He also received his Ph.D. in electrical engineering from K.N. Toosi University of Technology, Tehran, Iran in 2008. He is currently Assistant Professor in the Department of Electrical Engineering at the Babol University of Technology, Iran. His research interests include power electronic and design, simulation, modeling, and control of electrical machines.

- Email: gholamian@nit.ac.ir
- ORCID: [0000-0001-7654-7668](https://orcid.org/0000-0001-7654-7668)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://ostad.nit.ac.ir/home.php?sp=370509>

How to cite this paper:

R. Rouhani, S. E. Abdollahi, S. A. Gholamian, "Design of a synchronous reluctance motor with new hybrid lamination rotor structure," J. Electr. Comput. Eng. Innovations, 11(2): 243-252, 2023.

DOI: [10.22061/jecei.2022.8604.532](https://doi.org/10.22061/jecei.2022.8604.532)

URL: https://jecei.sru.ac.ir/article_1803.html





Research paper

Model Predictive Control of Linear Induction Motor Drive with End Effect Consideration

P. Hamedani^{1,*}, S. Sadr²

¹Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran.

²Department of Electrical Engineering, Tafresh University, Tafresh, Iran.

Article Info

Article History:

Received 28 June 2022
Reviewed 29 July 2022
Revised 01 August 2022
Accepted 22 October 2022

Keywords:

Linear Induction Motor (LIM)
Electrical motor drive
End Effect
Model Predictive Control (MPC)
Delay compensation
Indirect field oriented control

*Corresponding Author's Email Address:

p.hamedani@eng.ui.ac.ir

Abstract

Background and Objectives: Linear Induction Motors (LIMs) are favorite machines utilized in various industrial applications. But, due to the end effect phenomena, control of a LIM drive is more complicated than rotational machine drives. Therefore, selecting the proper control strategy for a LIM drive has been a significant challenge for the researchers.

Methods: This paper concentrates on a new Model Predictive Control (MPC) of LIM drives which considers the end effect.

Accordingly, the discrete-time model of the LIM with end effect is extracted, and the required flowchart used for the MPC of LIM drive has been presented in this paper.

Results: To study the effectiveness of the suggested strategy, simulation results of a LIM drive with MPC are presented and compared to the traditional Indirect Field Oriented Control (IFOC) of LIM drive. Simulations have been carried out using Matlab. The end effect has been considered in the LIM model and control strategies.

Conclusion: Simulation results validate that the suggested MPC of LIM drive yields excellent dynamic characteristics such as fast speed response with no overshoot. Moreover, in comparison to the traditional IFOC method, the suggested MPC strategy offers lower current ripple and lower electromagnetic force ripple, and therefore, it is suitable for industrial drive applications.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Linear electrical motors, including linear induction motors and linear synchronous motors (LSM), are popular machine types in different industries like electrical railway applications [1]-[12]. Linear induction machines, in comparison to synchronous counterparts, have a simple and robust structure, lower cost and maintenance, and self-starting thrust. These advantages make LIMs more prominent in industrial applications than SIMs [12].

However, speed control of LIMs has more difficulties than SIMs [13]-[15]. Until now, different control

strategies have been utilized for rotational induction motors, which can also be extended to LIMs, such as [16]-[17]:

- Scalar control methods, for example constant V/f method
- Field Oriented Control (FOC) methods
- Direct Torque Control (DTC) technique
- MPC

FOC and DTC have been used in many industrial and domestic applications. However, they have some issues. To overcome these issues, new variations have been proposed, which usually complicate the implementation

of the control strategy in practice [18]. In the past decade, with the development of digital signal processors (DSPs), MPC has been proposed as an interesting solution [18]-[19].

MPC approach needs the mathematical model of the system for predicting variables. A selected cost function is calculated for all possible switching states in each sampling time. Finally, the optimal switching states that minimize the cost function are chosen for firing the inverter switches in the next sampling time [18]-[21]. The main benefits of MPC are simple implementation and nonlinear solutions [18]-[20].

Although the PMC strategy has been a very popular control strategy for different electrical motor drives [22]-[23], in the case of LIMs, only a few works have been done until now [24]. In [25] and [26], MPC of LIM drive has been reported. But in these papers, the end effect has been counted in the direct axis circuit model and in the quadrature axis circuit model, the end effect is not considered. But, to accurately model a LIM, the end effect should be taken into account in both d- and q-axis equivalent circuits [27]-[28]. Moreover, in [25] and [26], the delay compensation method has not been studied in the MPC algorithm. By applying the delay compensation method, delay time that arises because of the large number of calculations will be compensated and the current ripple will be improved [30].

Consequently, this paper aims to investigate a new strategy for predictive control of the LIM drive considering the end effect. To reduce the computational time delay, a delay compensation methodology is used in the MPC of the LIM drive. Moreover, in this work, MPC and IFOC of the LIM drive are discussed from their basic theoretical concepts. The performance of these strategies is compared under transient and steady-state conditions. The same parameters and operating conditions have been considered for both approaches to guarantee a fair comparison.

The following sections of the paper will present the MPC strategy (including the discrete-time model of LIM and the MPC algorithm of LIM drive), the IFOC strategy (including the dynamic model of LIM and the vector control method of LIM drive), results, and the conclusion.

Discrete-time Model of the LIM with End Effect

In a three-phase LIM, the primary voltage equation can be written as follows:

$$\mathbf{V}_s = \mathbf{R}_s \mathbf{i}_s + \mathbf{L}_s \frac{d\mathbf{\Psi}_s}{dt} \quad (1)$$

Table 1 provides the notation for parameters and variables used in this paper.

The primary and secondary flux equations can be expressed as [21]:

$$\mathbf{\Psi}_s = \mathbf{L}_s \mathbf{i}_s + \mathbf{L}_m \mathbf{i}_r \quad (2)$$

$$\mathbf{\Psi}_r = \mathbf{L}_r \mathbf{i}_r + \mathbf{L}_m \mathbf{i}_s \quad (3)$$

Table 1: Notation for parameters and variables

symbol	Description
\mathbf{V}_s	Primary voltage vector
\mathbf{i}_s	Primary current vector
$\mathbf{\Psi}_s$	Primary flux vector
\mathbf{R}_s	Primary resistance matrix
\mathbf{L}_s	Primary inductance matrix
\mathbf{i}_r	Secondary current vector
$\mathbf{\Psi}_r$	Secondary flux vector
\mathbf{L}_r	Secondary inductance matrix
\mathbf{L}_m	Magnetizing inductance matrix
F	Electromagnetic force
τ	Motor pole pitch
k	Sampling instant
T_s	Sampling time
R_s	Primary resistance
L_s	Primary inductance
R_r	Secondary resistance
L_r	Secondary inductance
L_m	Magnetizing inductance
L_{m0}	Magnetizing inductance at zero speed
D	Motor length
V_r	Motor speed
λ_ψ	Weighting factor in cost function
ω_r	Angular velocity of LIM
ω_e	Angular velocity of reference frame
λ_{dr}	d-axis secondary flux

The electromagnetic force can be described as [28]:

$$F = \frac{3}{2} \frac{\pi}{\tau} \text{Im}\{\bar{\mathbf{\Psi}}_s \mathbf{i}_s\} \quad (4)$$

in which $\bar{\mathbf{\Psi}}_s$ is the complex conjugate value of $\mathbf{\Psi}_s$.

The discrete-time model of the LIM can be calculated from (1)-(2) using the Euler forward approximation [21]:

$$\mathbf{\Psi}_s(k+1) = \mathbf{\Psi}_s(k) + T_s \mathbf{V}_s(k) - \mathbf{R}_s T_s \mathbf{i}_s(k) \quad (5)$$

$$\mathbf{\Psi}_r(k+1) = \frac{L_r}{L_s} \mathbf{\Psi}_s(k+1) + \mathbf{i}_s(k) \left(L_m - \frac{L_r L_s}{L_m} \right) \quad (6)$$

$$\mathbf{i}_s(k+1) = \left(1 - \frac{T_s}{\tau_\sigma} \right) \mathbf{i}_s(k) + \quad (7)$$

$$\frac{T_s}{\tau_\sigma} \left\{ \frac{1}{R_\sigma} \left[\left(\frac{K_r}{\tau_r} - j K_r \omega_r \right) \mathbf{\Psi}_r(k+1) + \mathbf{V}_s(k) \right] \right\}$$

$$F(k+1) = \frac{3}{2} \frac{\pi}{\tau} \text{Im}\{\bar{\mathbf{\Psi}}_s(k+1) \mathbf{i}_s(k+1)\} \quad (8)$$

in which $R_\sigma = R_s + R_r K_r^2$, $K_r = \frac{L_m}{L_r}$, $\sigma = 1 - \frac{L_m^2}{L_s L_r}$,

$$\tau_\sigma = \frac{\sigma L_s}{R_\sigma} \quad (9)$$

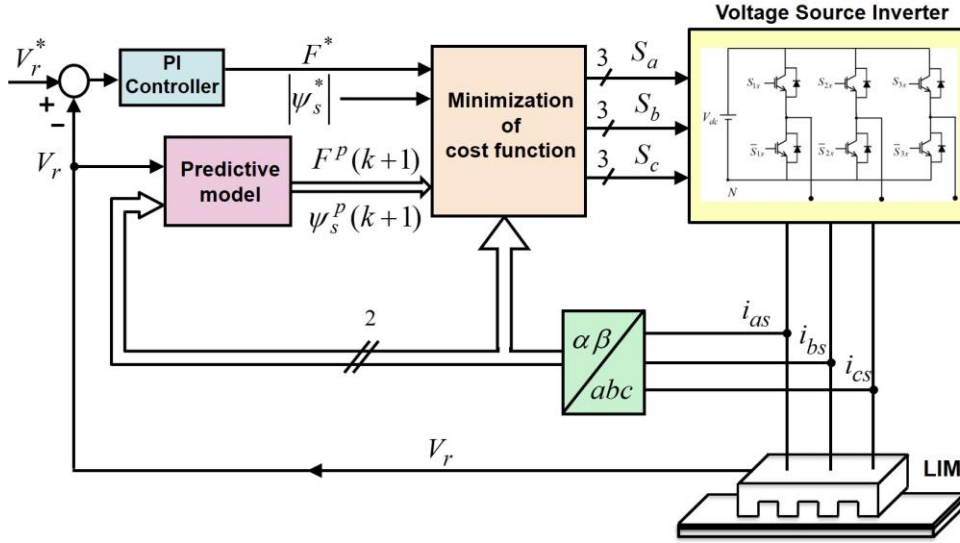


Fig. 1: Block diagram of MPC for the LIM drive.

To consider the end effect in the LIM model, the magnetizing inductance must be modified according to Duncan's model [28]:

$$L_m = L_{m0} (1 - f(Q)) \quad (10)$$

where

$$f(Q) = (1 - e^{-Q})/Q, \quad Q \cong \frac{D \cdot R_r}{L_r \cdot V_r} \quad (11)$$

Model Predictive Control of LIM drive

Model predictive control of the LIM drive is performed in the $\alpha\beta$ stationary reference frame. Therefore, Clark's transformation is utilized to convert a, b, and c primary voltage and currents to α and β primary voltage and currents [30]:

$$\begin{bmatrix} f_\alpha \\ f_\beta \end{bmatrix} = \frac{2}{3} \begin{bmatrix} 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{\sqrt{3}}{2} & -\frac{\sqrt{3}}{2} \end{bmatrix} \begin{bmatrix} f_a \\ f_b \\ f_c \end{bmatrix} \quad (12)$$

$$\mathbf{V}_s = V_\alpha + jV_\beta, \quad \mathbf{i}_s = i_\alpha + ji_\beta \quad (13)$$

in which f donates the voltage or current variables.

Fig. 1 illustrates the diagram of the MPC-based LIM drive. A discrete PI controller with anti-windup produces the reference force, F^* . The MPC diagram calculates the future values of primary flux and force utilizing (5)-(8). The predicted and command values of the primary flux and force are compared in a cost function. All possible switching conditions are considered. In a 2-level voltage source inverter, eight various switching combinations happen. The one that minimizes the cost function is selected as the next switching condition applied to the inverter.

The cost function is considered as follows:

$$g = |F^* - F(k+1)| + \lambda_\psi |\psi_s^* - \psi_s(k+1)| \quad (14)$$

The weighting factor is considered as the ratio of the rated force and rated stator flux:

$$\lambda_\psi = \frac{F_n}{2|\psi_{sn}|} \quad (15)$$

To moderate the time delay that arises because of the high number of computations, the delay compensation methodology has been proposed [30]. This method calculates the predicted values in the next shifted forward sample time [30]:

$$\psi_s(k+2) = \psi_s(k+1) + T_s \mathbf{V}_s(k+1) - R_s T_s \mathbf{i}_s(k+1) \quad (16)$$

$$\psi_r(k+2) = \frac{L_r}{L_s} \psi_s(k+2) + \mathbf{i}_s(k+1) \left(L_m - \frac{L_r L_s}{L_m} \right) \quad (17)$$

$$\mathbf{i}_s(k+2) = \left(1 - \frac{T_s}{\tau_\sigma} \right) \mathbf{i}_s(k+1) + \frac{T_s}{\tau_\sigma} \left\{ \frac{1}{R_\sigma} \left[\left(\frac{K_r}{\tau_r} - jK_r \omega_r \right) \psi_r(k+2) + \mathbf{V}_s(k+1) \right] \right\} \quad (18)$$

$$F(k+2) = \frac{3}{2} \frac{\pi}{\tau} \text{Im} \{ \bar{\psi}_s(k+2) \mathbf{i}_s(k+2) \} \quad (19)$$

Consequently, the cost function can be written as:

$$g = |F^* - F(k+2)| + \lambda_\psi |\psi_s^* - \psi_s(k+2)| \quad (20)$$

Fig. 2 shows the flowchart for the MPC for LIM drive with delay compensation.

Dynamic Model of the LIM with End Effect

IFOC of the LIM drive is performed in the q-d synchronous rotational reference frame. Therefore, Park's transformation is utilized to convert a, b, and c variables to the q and d variables. Primary and secondary voltage equations are written as [29]:

$$v_{qs} = R_s i_{qs} + \omega_e \lambda_{ds} + p \lambda_{qs} \quad (21)$$

$$v_{ds} = R_s i_{ds} - \omega_e \lambda_{qs} + p \lambda_{ds} \quad (22)$$

$$v_{qr} = R_r i_{qr} + (\omega_e - \omega_r) \lambda_{dr} + p \lambda_{qr} = 0 \quad (23)$$

$$v_{dr} = R_r i_{dr} - (\omega_e - \omega_r) \lambda_{qr} + p \lambda_{dr} = 0 \quad (24)$$

Primary and secondary flux linkage equations are written as [29]:

$$\lambda_{qs} = L_{ls} i_{qs} + L_m \{1 - f(Q)\} (i_{qs} + i_{qr}) \quad (25)$$

$$\lambda_{ds} = L_{ls} i_{ds} + L_m \{1 - f(Q)\} (i_{ds} + i_{dr}) \quad (26)$$

$$\lambda_{qr} = L_{lr} i_{qr} + L_m \{1 - f(Q)\} (i_{qs} + i_{qr}) \quad (27)$$

$$\lambda_{dr} = L_{lr} i_{dr} + L_m \{1 - f(Q)\} (i_{ds} + i_{dr}) \quad (28)$$

where $p \equiv d/dt$.

The LIM thrust can be written as:

$$F = \frac{3}{2} \frac{\pi}{\tau} (\lambda_{qr} i_{dr} - \lambda_{dr} i_{qr}) \quad (29)$$

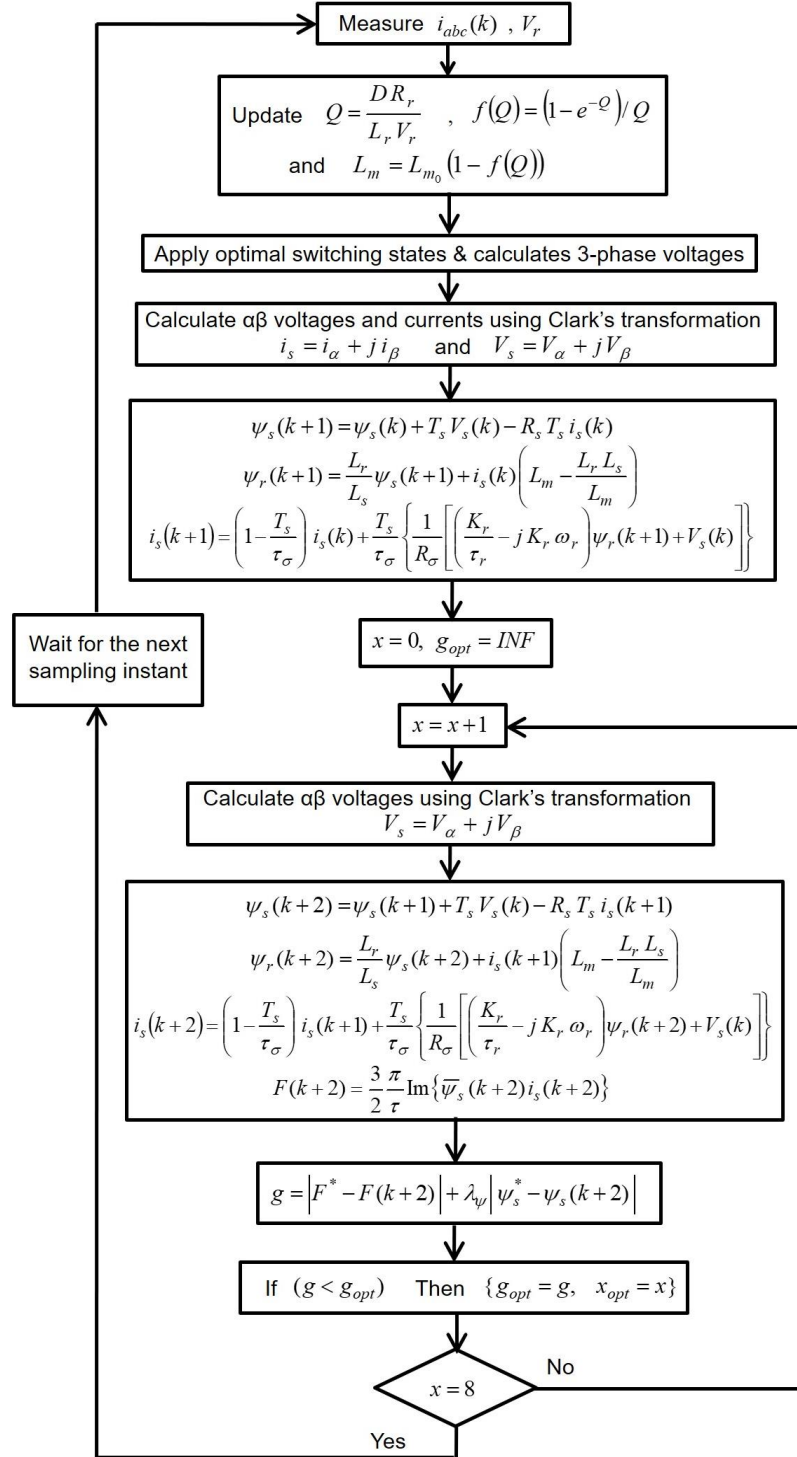


Fig. 2: Flowchart for MPC of the LIM drive.

IFOC of LIM Drive

Fig. 3 shows the IFOC diagram for the LIM drive. To decouple the flux and the LIM force, the below assumption is made in this strategy [28]:

$$\lambda_{qr} = 0 \quad , \quad \frac{d\lambda_{qr}}{dt} = 0 \quad (30)$$

As a result and by supposing $v_{qr} = v_{dr} = 0$, the slip frequency ($\omega_{sl} \equiv \omega_e - \omega_r$), λ_{dr} , and the LIM force can be computed as [28]:

$$\omega_{sl} = R_r \left[\frac{1 - f(Q)}{\frac{L_{lr}}{L_m} + (1 - f(Q))} \right] \times \frac{i_{qs}}{\lambda_{dr}} \quad (31)$$

$$\lambda_{dr} = \frac{L_m(1-f(Q))}{1 + \left\{ \frac{L_{lr} + L_m(1-f(Q))}{R_r} \right\}^p} \times i_{ds} \quad (32)$$

$$F = \frac{3}{2} \frac{\pi}{\tau} \frac{L_m(1-f(Q))}{L_{lr} + L_m(1-f(Q))} \lambda_{dr} i_{qs} \quad (33)$$

The IFOC scheme is composed of two control loops. The outer loop controls the LIM speed using a PI controller and generates the reference q-axis current (i_{qs}^*). The inner loop controls the LIM phase currents using a hysteresis controller and produces the switching pulses of the inverter.

The slip frequency (ω_{sl}) and the reference d-axis current (i_{ds}^*) are generated using (31) and (32), respectively.

As shown in Fig. 3, ω_{sl} and i_{ds}^* are calculated using gains K_1 and K_2 which depend on the end effect and machine velocity.

Results and Discussion

To investigate the effectiveness of the MPC of LIM drive with end effect, simulation results are provided in this section. The end effect is considered in the LIM model and MPC strategy. Moreover, the results are compared with the IFOC of LIM drive with the end effect. Simulations are implemented using Matlab. In both methods, the same parameters and conditions have been used for the simulations. Table 2 shows the simulation parameters. The utilized gains in the PI controller are $K_i = K_p = 50$.

Table 2: Simulation Parameters of LIM drive.

Phase voltage	220 V	R_r	0.843 Ω
Nominal current	93.65 A	L_s	4.5 mH
Power factor	0.4884	L_m	3 mH
Poles	4	L_r	3.1 mH
τ	0.1024 m	λ^*_{dr}	0.24 Wb
D	0.413 m	M	29.34 kg
R_s	0.049 Ω	Rated Load	879 N

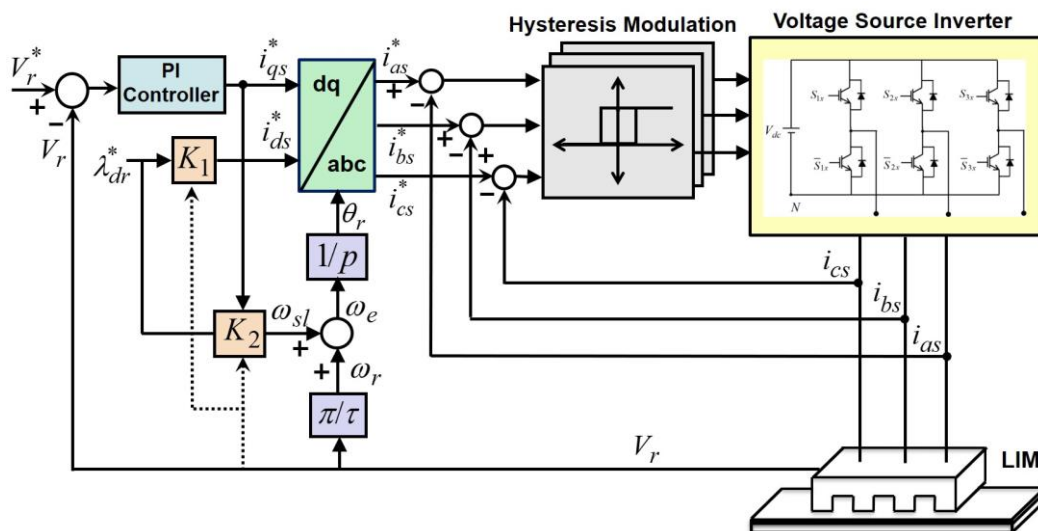


Fig. 3: Block diagram of IFOC for the LIM drive.

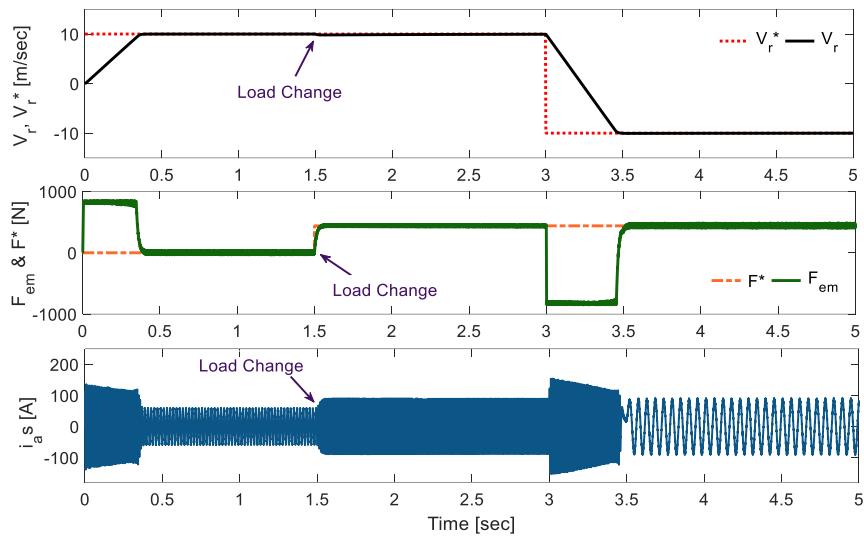


Fig. 4: Speed response, electromagnetic force response, and phase current LIM drive with MPC method.

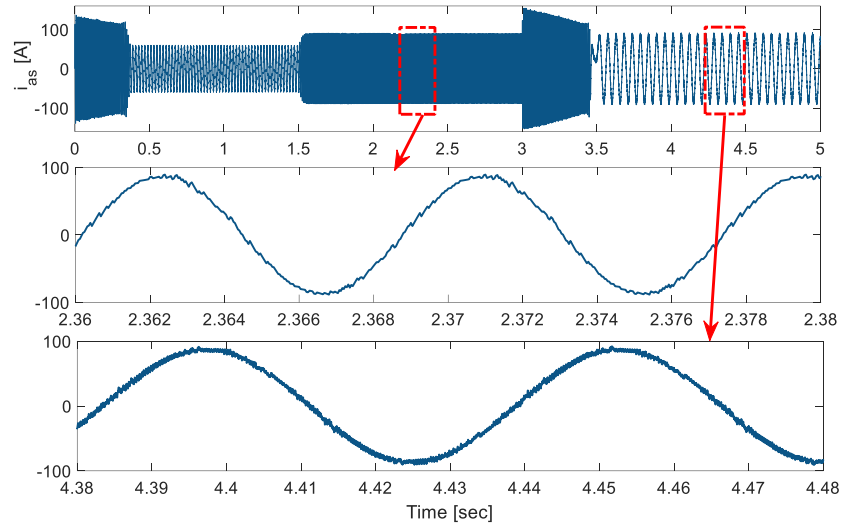


Fig. 5: Phase current ripple of LIM drive with MPC method.

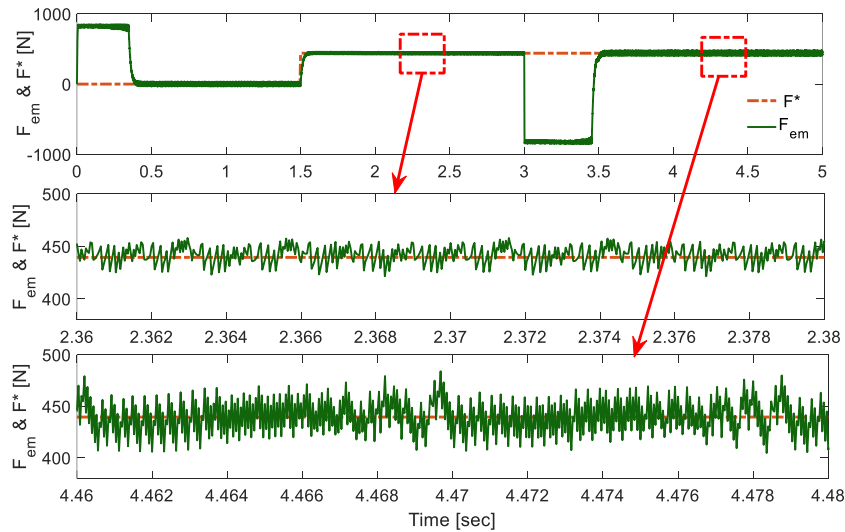


Fig. 6: Electromagnetic force ripple of LIM drive with MPC method.

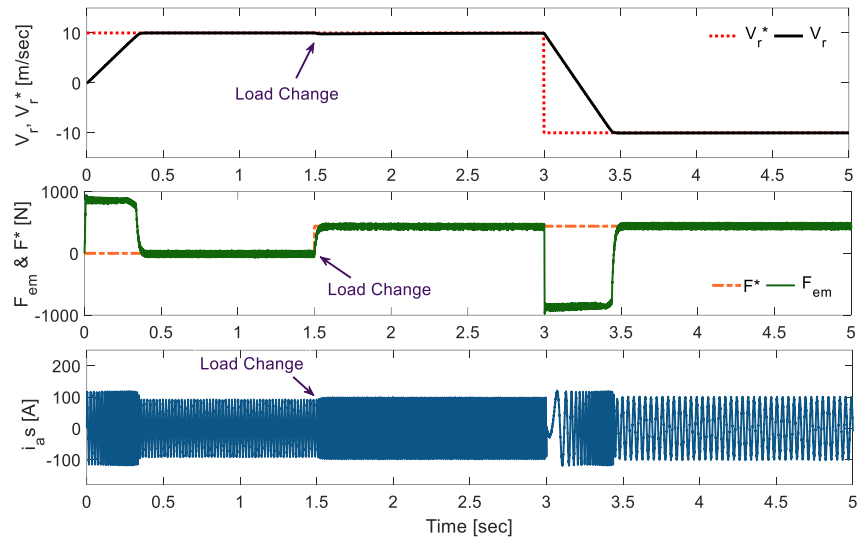


Fig. 7: Speed response, electromagnetic force response, and phase current of LIM drive with IFOC.

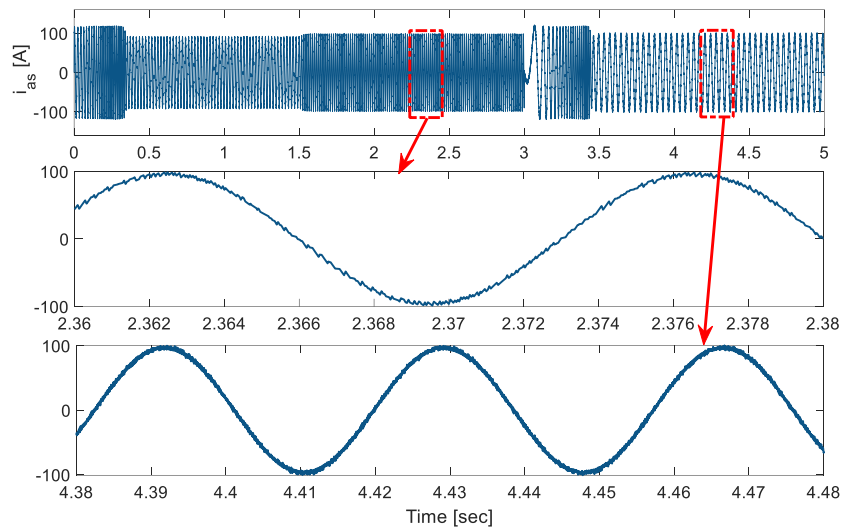


Fig. 8: Phase current ripple of LIM drive with IFOC method.

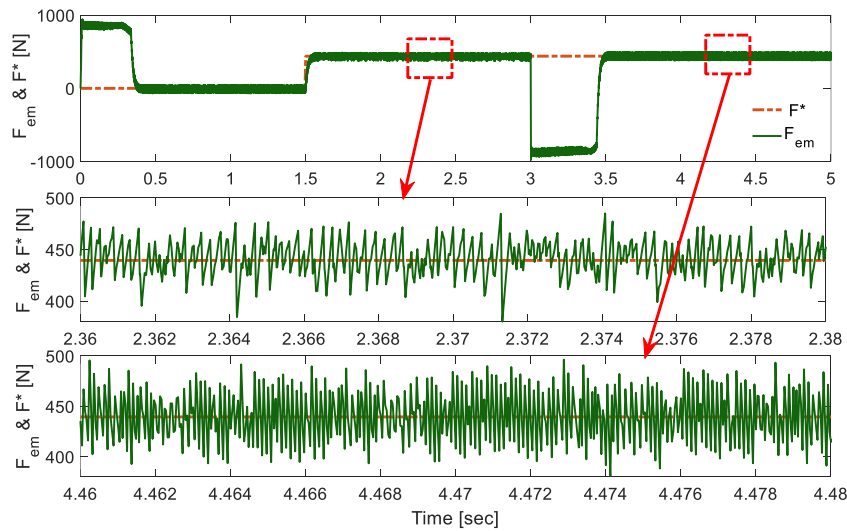


Fig. 9: Electromagnetic force ripple of LIM drive with IFOC method.

At the start, a reference speed equal to 10 m/sec is applied, and it changes to -10 m/sec at $t=3$ sec. The LIM drive starts in no-load condition, and an external load is applied to the machine at $t=1.5$ sec. For the IFOC method, the hysteresis band has been taken equal to 0.5 A.

Fig. 4 illustrates the speed, electromagnetic force, and phase current of the LIM drive with the MPC method, respectively. Clearly, the actual LIM speed follows the reference speed in motoring and braking conditions. Fig. 5 shows the phase current ripple of the LIM drive with the MPC method. Fig. 6 represents the electromagnetic force ripple of the LIM drive with MPC method.

According to Fig. 6, the LIM force tracks the external load in motoring and braking conditions.

Fig. 7 compares the speed, electromagnetic force, and phase current of the LIM drive with the IFOC method, respectively. Like the case of the MPC method, in the IFOC strategy, the actual LIM speed tracks the reference speed in motoring and braking conditions. Fig. 8 shows the phase current ripple of the LIM drive with the IFOC method. Fig. 9 represents the electromagnetic force ripple of the LIM drive with the IFOC method.

Comparison of Fig. 4 with Fig. 7 manifest that both methods yield similar dynamic performance in the speed response. However, a comparison of Fig. 5 with Fig. 8 shows that the MPC method has a lower current ripple. Moreover, a comparison of Fig. 6 with Fig. 9 demonstrates that the MPC method has a lower electromagnetic force ripple. Table 3 shows the current ripple and force ripple of the MPC and IFOC methods.

Table 3: Comparison of current ripple and force ripple in MPC and IFOC methods of LIM drive.

		IFOC	MPC
Current ripple	$V_r^* = 10$ m/sec	3 A	2.5 A
	$V_r^* = -10$ m/sec	5 A	3.5 A
Force ripple	$V_r^* = 10$ m/sec	48 A	22 A
	$V_r^* = -10$ m/sec	60 A	46 A

Conclusion

This work proposes the MPC strategy for LIM drives, considering the end effect. The discrete-time model of the LIM with end effect has been extracted, and the required flowchart utilized for the model predictive control of LIM drive has been presented. To evaluate the accuracy of the suggested strategy, MPC is compared to the traditional IFOC for the LIM drive.

Simulation results manifest that the suggested model predictive control of LIM drive achieves perfect dynamic characteristics such as fast speed response with no overshoot. In addition, compared to the traditional indirect field-oriented control, the proposed model predictive control offers lower current ripple and lower electromagnetic force ripple.

Author Contributions

P. Hamedani carried out the simulation results. S. Sadr interpreted the results. P. Hamedani and S. Sadr wrote the manuscript.

Acknowledgment

The authors thankfully appreciate the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

LIM	Linear Induction Motor
LSM	Linear Synchronous Motor
MPC	Model Predictive Control
IFOC	Indirect Field Oriented Control
DTC	Direct Torque Control
VSI	Voltage Source Inverter

References

- [1] A. Shiri, A. Shoulaie, "End effect braking force reduction in high-speed single-sided linear induction machine," *Int. J. Energy Convers. Manage.*, 61: 43-50, 2012.
- [2] X. Qiwei, S. Cui, Q. Zhang, L. Song, X. Li, "Research on a new accurate thrust control strategy for linear induction motor," *IEEE Trans. Plasma Sci.*, 43(5): 1321-1325, 2015.
- [3] R. Cao, Y. Jin, M. Lu, Z. Zhang, "Quantitative comparison of linear flux-switching permanent magnet motor with linear induction motor for electromagnetic launch system," *IEEE Trans. Ind. Electron.*, 65(9): 7569-7578, 2018.
- [4] J. Q. Li, W. L. Li, G. Q. Deng, Z. Ming, "Continuous-behavior and discrete-time combined control for linear induction motor-based urban rail transit," *IEEE Trans. Mag.*, 52(7): 1-4, 2016.
- [5] T. Wang, B. Li, B. Xie, F. Fang, "Linear induction motors for driving vehicles climbing on steel plates," *IEEE Trans. Energy Convers.*, 29(3): 788-789, 2014.
- [6] J. Lim, J. H. Jeong, C. H. Kim, C. W. Ha, D. Y. Park, "Analysis and experimental evaluation of normal force of linear induction motor for maglev vehicle," *IEEE Trans. Magn.*, 53(11): 1-4, 2017.

- [7] R. Cao, M. Lu, N. Jiang, M. Cheng, "Comparison between linear induction motor and linear flux-switching permanent-magnet motor for railway transportation," *IEEE Trans. Ind. Electron.*, 66(12): 9394-9405, 2019.
- [8] W. Y. Ji, G. Jeong, C. B. Park, I. H. Jo, H. W. Lee, "A study of non-symmetric double-sided linear induction motor for hyperloop all-in-one system (Propulsion, Levitation, and Guidance)," *IEEE Trans. Magn.*, 54(11): 1-4, 2018.
- [9] M. Shujun, C. Jianyun, S. Xudong, W. Shanming, "A variable pole pitch linear induction motor for electromagnetic aircraft launch system," *IEEE Trans. Plasma Sci.*, 43(5): 1346-1351, 2015.
- [10] H. Seo, J. Lim, G. H. Choe, J. Y. Choi, J. H. Jeong, "Algorithm of linear induction motor control for low normal force of magnetic levitation train propulsion system," *IEEE Trans. Magn.*, 54(11): 1-4, 2018.
- [11] H. Karimi, S. Vaez-Zadeh, F. Rajaei Salmasi, "Combined vector and direct thrust control of linear induction motors with end effect compensation," *IEEE Trans. Energy Convers.*, 31(1): 196-205, 2016.
- [12] P. Hamedani, S. Sadr, A. Shoulaie "Independent fuzzy logic control of two five-phase linear induction motors supplied from a single voltage source inverter," *J. Electr. Comput. Eng. Innovations (JECEI)*, 10(1): 195-208, 2021.
- [13] K. Wang, Y. Li, Q. Ge, L. Shi, "An improved indirect field-oriented control scheme for linear induction motor traction drives," *IEEE Trans. Ind. Electron.*, 65(12): 9928-9937, 2018.
- [14] D. Hu, W. Xu, R. Dian, Y. Liu, J. Zhu, "Loss minimization control of linear induction motor drive for linear metros," *IEEE Trans. Ind. Electron.*, 65(9), 6870-6880, 2018.
- [15] A. Accetta, M. Cirrincione, M. Pucci, A. Sferlazza, "State space-vector model of linear induction motors including end-effects and iron losses part i: theoretical analysis," *IEEE Trans. Ind. Appl.*, 56(1), 235-244, 2020.
- [16] A. Poorfakhraei, M. Narimani, A. Emadi, "A review of modulation and control techniques for multilevel inverters in traction applications," *IEEE Access*, 9: 24187-24204, 2021.
- [17] A. Poorfakhraei, M. Narimani, A. Emadi, "A review of multilevel inverter topologies in electric vehicles: current status and future trends," *IEEE Open J. Power Electron.*, 2: 155-170, 2021.
- [18] F. Wang, Z. Zhang, X. Mei, J. Rodríguez, R.; Kennel, "Advanced control strategies of induction machine: field oriented control, direct torque control and model predictive control," *Energies*, 11(1): 1-13, 2018.
- [19] J. Rodríguez, R. Kennel, J. R. Espinoza, M. Trincado, C. A. Silva, C. A. Rojas, "High-Performance Control Strategies for Electrical Drives: An Experimental Assessment," *IEEE Trans. Ind. Electron.*, 59(2), 812-820, 2012.
- [20] M. Rivera, J. Rodríguez, S. Vazquez, "Predictive control in power converters and electrical drives—part I," *IEEE Trans. Ind. Electron.*, 63(6): 3834-3836, 2016.
- [21] J. Rodríguez, P. Cortes, *Predictive Control of Power Converters and Electrical Drives*, vol. I. Wiley-IEEE Press: 123, 2012.
- [22] J. Rodríguez et al., "Latest advances of model predictive control in electrical drives—part i: basic concepts and advanced strategies," *IEEE Trans. Power Electron.*, 37(4): 3927-3942, 2022.
- [23] M. F. Elmorshedy, W. Xu, F. F. M. El-Sousy, M. R. Islam, A. A. Ahmed, "Recent achievements in model predictive control techniques for industrial motor: a comprehensive state-of-the-art," *IEEE Access*, 9: 58170-58191, 2021.
- [24] M. F. Elmorshedy, W. Xu, S. M. Allam, J. Rodríguez, C. Garcia, "MTPA-based finite-set model predictive control without weighting factors for linear induction machine," *IEEE Trans. Ind. Electron.*, 68(3): 2034-2047, 2021.
- [25] N. J., Merlin Mary, C. Ganguly, M. Kowsalya, "Simulation of linear induction motor using model predictive control in synchronously rotating reference frame," presented at the IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES), 2016.
- [26] S. M. Kazraji, M. B. B. Sharifian, "Model predictive control of linear induction motor drive," presented at the 43rd Annual Conference of the IEEE Industrial Electronics Society (IECON), 2017.
- [27] G. Kang, K. Nam, "Field-oriented control scheme for linear induction motor with the end effect," *IEE Proc. on Electr. Power Appl.*, 152(1): 1565-1572, 2005.
- [28] P. Hamedani, A. Shoulaie, "Utilization of CHB multilevel inverter for harmonic reduction in fuzzy logic controlled multiphase LIM drives," *J. Electr. Comput. Eng. Innovations (JECEI)*, 8(1): 19-30, 2020.
- [29] P. Hamedani, A. Shoulaie, "Modification of the field-weakening control strategy for linear induction motor drives considering the end effect," *Adv. Electr. Comput. Eng. (AECE)*, 15(3): 3-12, 2015.
- [30] P. Hamedani, C. Garcia, F. Flores-Bahamonde, S. Sadr, J. Rodriguez, "Predictive control of 4-level flying capacitor inverter for electric car applications," presented at the 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC): 224-229, 2022.

Biographies



Pegah Hamedani was born in Isfahan, Iran, in 1985. She received B.Sc. and M.Sc. degrees from University of Isfahan, Iran, in 2007 and 2009, respectively, and the Ph.D. degree from Iran University of Science and Technology, Tehran, in 2016, all in Electrical Engineering. Her research interests include power electronics, control of electrical motor drives, supply system of the electric railway (AC and DC), linear motors & MAGLEVs, and analysis of overhead contact systems. She is currently an Assistant Professor with the Department of Railway Engineering and Transportation Planning, University of Isfahan, Isfahan, Iran. Dr. Hamedani was the recipient of the IEEE 11th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC'20) best paper award in 2020.

- Email: p.hamedani@eng.ui.ac.ir
- ORCID: [0000-0002-5456-1255](https://orcid.org/0000-0002-5456-1255)
- Web of Science Researcher ID: AAN-2662-2021
- Scopus Author ID: 37118674000
- Homepage: <https://engold.ui.ac.ir/~p.hamedani/>



Sajad Sadr received the B.Sc. degree in electronic engineering from Bu-Ali Sina University, Hamedan, Iran, in 2006, and the M.Sc. degree in electrical engineering from Amirkabir University of Technology, Tehran, Iran, in 2009. Ph.D. degree in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2016. He is currently an Assistant Professor with the Department of Electrical Engineering, Tafresh University, Tafresh, Iran. His research interests include power electronics and motor drives.

- Email: sadr@tafreshu.ac.ir
- ORCID: [0000-0002-6113-8930](https://orcid.org/0000-0002-6113-8930)
- Web of Science Researcher ID: GVT-2596-2022
- Scopus Author ID: 56367494300
- Homepage: <http://faculty.tafreshu.ac.ir/sadr/fa>

How to cite this paper:

P. Hamedani, S. Sadr, "Model predictive control of linear induction motor drive with end effect consideration," J. Electr. Comput. Eng. Innovations, 11(2): 253-262, 2023.

DOI: [10.22061/jecei.2022.9191.586](https://doi.org/10.22061/jecei.2022.9191.586)

URL: https://jecei.sru.ac.ir/article_1804.html





Research paper

An Improved Approach to Blind Image Steganalysis Using an Overlapping Blocks Idea

V. Sabeti*

Department of Computer Engineering, Faculty of Engineering, Alzahra University, Tehran, Iran.

Article Info

Article History:

Received 12 August 2022
Reviewed 18 October 2022
Revised 28 October 2022
Accepted 31 October 2022

Keywords:

Steganalysis
Steganography
Blind steganalysis
Block-based steganalysis
Spatial steganography
JPEG steganography

*Corresponding Author's Email
Address: v.sabeti@alzahra.ac.ir

Abstract

Background and Objectives: Steganalysis is the study of detecting messages hidden using steganography. Most steganalysis techniques, known as blind steganalysis, focus on extracting and classifying various statistical features from images. Consequently, researchers continually seek to improve the accuracy of blind detection methods. The current study proposes a blind steganalysis technique based on overlapping blocks.

Methods: The proposed method began by decomposing the image into identically sized overlapping blocks, then extracted a feature vector from each block. Subsequently, a tree-structured hierarchical clustering technique was used to partition blocks into multiple classes based on extracted features, and a classifier was trained for each class to determine whether a block is from a cover or stego image. The block decomposition process was repeated for each test image, and a classifier was selected based on the block class to make a decision for each block. Furthermore, the majority vote rule was utilized to determine whether the test image is a cover or stego image.

Results: The proposed method was evaluated using the INRIA and BOSSbase datasets. Several parameters, including the number of block classes, feature extraction method, block size, and number and block overlapping level, affected the performance of the proposed method. The optimal block size was 64×64 by 32 steps, and the number of block classes was set to 16. WOW, S-UNIWARD, PQ, and nsF5 were the steganographic methods employed to evaluate the proposed method. Experimental results indicated that using overlapping instead of non-overlapping blocks increased the detection of data embedded in both the spatial and Joint Photographic Experts Group (JPEG) domains by an average of over 9%. In addition, the proposed method's accuracy in detecting the S-UNIWARD method was comparable to that of other deep learning-based steganalysis techniques.

Conclusion: The concept of using overlapping blocks improves the efficiency of blind steganalysis by providing the benefit of additional and larger blocks. One of the main advantages of the proposed method is comparable detection accuracy and less computational complexity than recent deep learning-based steganalysis techniques.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

With the advent of the web and its increasing use as a

platform for digital data transmission, data protection techniques are more crucial than ever. To this end, numerous solutions have been developed for data

security and safety. Individuals other than the sender and receiver are unable to understand the communication content due to data encryption before the transmission occurs through the network. Only the sender and receiver have access to the encryption key. Although the existence of encrypted communication is not hidden, steganography is the process of transmitting secret information by embedding it within a cover media. Consequently, this method conceals the existence of confidential information [1].

The secret data and cover media used in steganography may be text, image, video, or audio files. Among the various techniques, image steganography (i.e., hiding secret data in an image) is the most popular and widespread because it facilitates the transfer of large volumes of images over the internet. Digital images and videos contain a high proportion of repetitive bits, making them more suitable for data hiding [2].

In image steganography, the original image used to carry secret information is known as a cover image. The image resulting from the embedding process is known as a stego image. The success of steganography is predicated primarily on the secrecy of the embedded concealed data. Image steganography comprises three main requirements: hiding capacity, imperceptibility, and security [3].

The embedding process used to hide secret data in the cover image forms the basis of the steganographic process. Since it is possible to embed secret data in the spatial and transform domains of the cover image, existing steganographic techniques can be classified based on the cover domain employed. Due to the simplicity of embedding and extraction operations, spatial domain embedding techniques are more popular and utilized than transform domain techniques. However, they possess less robustness and reliability. Embedding in Discrete Cosine Transform (DCT) coefficients of Joint Photographic Experts Group (JPEG) images is one of the transform domain techniques [4].

Steganalysis is the opposite of steganography, the art, and science of deciphering covert communications through in-depth knowledge of steganographic techniques. Steganalysis is the science of attacking steganography in a never-ending battle with the primary objective of gathering sufficient evidence of the presence of an embedded secret message and breaching the security of the message carrier. Steganalysis methods are generally classified into three visual, structural, or statistical categories [5]. Visual steganalysis approaches, the simplest form of steganalysis, seek to detect visual anomalies within the stego image. Numerous visual steganalysis methods rely on deficiencies in embedding algorithms. Structural steganalysis detects modifications to the stego file format and reveals the presence of

embedded data by comparing the structure to its standard set. As the most prevalent available techniques, statistical steganalysis techniques uncover covert data by comparing the statistical characteristics of a stego image to a set of covers.

Alternatively, steganalysis techniques can be classified as either special or universal steganalysis [6]. Special steganalysis methods are designed for a particular steganographic algorithm. In contrast, universal or blind steganalysis is a general technique that can decipher data embedded by any steganographic algorithm, even a previously unknown one.

Recent steganalysis literature has primarily focused on blind statistical steganalysis methods (referred to as blind steganalysis in the present study). Enhancing the quality of extracted feature vectors from images is one strategy for improving the performance of blind steganalysis algorithms. The richer this vector is with informative features, the better the algorithm's performance. Therefore, the quantity and quality of image features extracted have become crucial for blind steganalysis design. Indeed, several recent studies also employed deep learning algorithms, such as Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and other algorithms in which feature extraction and selection are performed automatically [7].

Frame-based steganalysis and block-based steganalysis are two blind steganalysis approaches that use the entire image or image blocks to extract features, respectively. The complete structure of block-based steganalysis methods is presented in [8]-[10]. The present study uses image blocks instead of the whole image in the feature extraction process. The majority of existing methods employ completely distinct and non-overlapping blocks. In addition to block images, this paper proposes the concept of overlapping image blocks, which significantly improves detectability by coordinating the number and size of image blocks. The implementation results demonstrate that the proposed method significantly outperforms its predecessors. The following summarizes the study's main contributions:

- Provide block-based steganalysis methods that use overlapping blocks for feature extraction.
- Investigate the influence of parameters, including the number of block classes, the feature extraction method, the size and number of blocks, and the degree of overlapping blocks, on the accuracy of the proposed approach.
- Evaluate the probability of using the proposed method to discover steganographic techniques in the spatial and JPEG domains.

The remainder of the paper is organized as follows: Section 2 presents several block-based steganalysis techniques after introducing the structure of blind

statistical steganalysis methods. Section 3 describes the overlapping blocks-based steganalysis method. Section 4 presents the results of implementing and applying the proposed steganalysis approach in detecting several existing JPEG and spatial steganographic methods. Finally, Section 5 provides conclusions and recommendations for future research.

Related Work

Steganalysis is said to be successful only if the hidden message embedded in media is proven. Recent steganographic techniques attempt to leave cover media with minimal quantitative and statistical traces. Conversely, in response to this practice, standard steganalysis approaches attempt to broaden their analysis dimensions and employ complex and expert processes to achieve greater sensitivity. Therefore, modern steganalyzers require significantly more computing resources and power than in the past.

New approaches are required to conserve resources and simplify steganalysis, reducing computational complexity and time while increasing productivity [12]. The blind statistical steganalysis methods are comparable to pattern classification techniques. After applying some image preprocessing operations, most existing blind steganalysis methods extract a vector of features from images. Then, they select or design a suitable classifier and train it using the extracted features from the training image set. The training images consist of both cover and stego images. The output of the training phase is a classifier that can be used to determine the state of the test images. After applying image preprocessing operations and feature vector extraction in the testing phase, the trained classifier classifies the test image into one of two cover or stego categories. Blind steganalysis general steps are depicted in Fig. 1 [11].

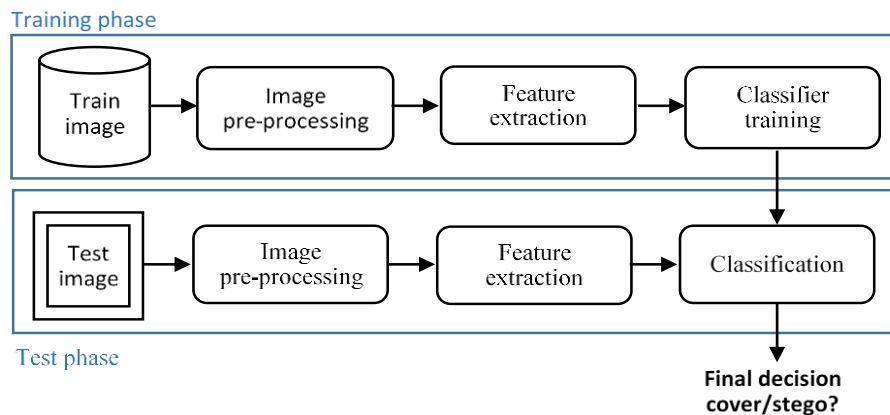


Fig. 1: The framework of blind steganalysis methods [11].

In the image preprocessing step, some operations are applied to the image before feature extraction, including converting RGB images into grayscales, cropping, JPEG compression, and DCT transformation, among others. The feature extraction step includes a set of unique statistical properties of an image, referred to as features. This step attempts to extract informative image features. The informative feature indicates that the selected feature must be embedding sensitive information. Following feature extraction, a low-dimensional feature vector is constructed, which reduces the computational complexity of the training and classification phases. Feature extraction techniques are wavelet decomposition, Markov empirical transition matrix, image quality metrics, co-occurrence matrix, and image histogram, among others [13]. Currently, existing steganalysis methods, such as special and blind methods, aim to enhance detectability and efficiency [7]. In this regard, numerous advancements have been made by researchers. Some extract more features from existing images [6], whereas others search for the optimal, highest

-quality, and smallest feature vector without increasing the number of features [12]. Another group of methods selects the desired features from the specific parts of the images [14], while others focus on improving the classifiers' performance [15].

Since the performance of blind steganalysis algorithms is highly dependent on the quality of the extracted features, a line of related research was devoted to determining the requirements for feature extraction. Several of these studies attempted to identify the most probable location for information embedding and feature extraction based on the type of cover image, image texture, and image color, among other factors [16].

Images are comprised of different decompositions with varying complexities and frequencies. When hidden information is embedded into an image, the effect is more noticeable when the image is less complex. A single image may also contain heterogeneous regions, and some of its decompositions may have greater complexity and frequency rates. Consequently, block-based steganalysis methods emerged in response to these challenges and

observations [8]. The central idea is to decompose images of comparable complexity into smaller, more uniform blocks than the original image. Then, each obtained block is considered a discrete steganalysis input image. It has been demonstrated that steganographic embedding correlates much more strongly with similar blocks. Thus,

the features of these smaller blocks are used to create a content-based classifier. Table 1 summarizes several blind steganalysis methods that employ image blocking prior to feature extraction. These methods are described in greater detail below.

Table 1: The summary of block-based steganalysis methods

Ref.	Blocking	Overlapping	Texture analysis	Feature extraction method	Domain	Tested methods
[8], [9]	✓	×	×	Pevny's method	Transform (JPEG)	OutGuess, F5, MBS
[10]	✓	✓	×	SPAM	Transform (JPEG)	PQ, MBS
[17]	✓	×	✓	Merged-274 feature set	Transform (JPEG)	F5, nsF5 MB1, PQ, JPHide
[18], [19]	✓	×	×	IQM	Transform (JPEG)	F5, Quickstego, StegHide
[14]	✓	×	✓	SPAM CC-PEV	Spatial	HUGO
[20]	✓	×	✓	SRM	Spatial	LSBR, LSBM, HUGO, S-UNIWARD

Studies [8]-[10] can be cited as being among the first to utilize the idea of block-based steganalysis. The image is first decomposed in these methods into smaller, fixed-size blocks. Based on the feature vector of each block, the blocks are then decomposed into multiple classes. For each block class, a classifier is trained based on the features extracted from each block in that class. Similar to the training phase, the block and feature extraction operations are repeated for each input image during the testing phase. Then, each image block is assigned to a particular class, and the classifier for each class determines whether the block is cover or stego. The final step is to combine the results of all classifiers regarding the type of blocks to determine whether the entire image is a cover or stego.

Wang et al. [17] proposed a block texture-based cover image method for JPEG image feature extraction and steganalysis. Their method decomposes the input images into several sub-images based on the JPEG block texture complexity. The calibrated set of features is then extracted from each of these subimages. Separate sets of subimages with the same texture complexity are used to construct and train the classifier. The end result of steganalysis is attained through a process of weighted fusing. Due to the insufficiency and limitation of the obtained feature set, this method lacks the optimal detection accuracy for detecting embedded images with a low rate or some novel and unknown methods.

Another disadvantage of this method is the significant computational complexity and computation time complexity of detecting image texture.

Suryawanshi et al. [18], [19] also presented a blind statistical method for digital image steganalysis. In this method, the image is first decomposed into identical blocks. Then, the statistical features of each block are extracted. Several sub-classifiers of a multi-class classifier are used to classify the image based on these features. The proposed scheme outperforms several existing approaches due to each block's initial image block and multidimensional feature extraction.

Mohammadi et al. [14] proposed a universal statistical steganalysis method that decomposes test images into sub-images using the Artificial Bee Colony (ABC) algorithm. Then, the optimal sub-region concerning density and energy is selected, and the desired features are extracted from this region. These two feature sets are combined to train the Support Vector Machine (SVM) classifier. Experimental results from the algorithm implementation demonstrate that the proposed method improves detection accuracy and increases True Positives (TPs) and True Negatives (TNs).

Zhu et al. [20] used image decomposition to introduce a block-based steganalysis method. This method decomposes the image into subimages with differing texture levels. Subimages are utilized for training the classifier, which aids in simulating statistical detectability.

This technique is only employed to decipher spatial steganographic techniques.

The Proposed Model

This paper proposes a block-based steganalysis framework. As mentioned earlier, other block-based steganalysis methods have also been presented. However, the advantage of the proposed method over other existing methods is that the overlapping blocks idea is used for a fixed-size image in addition to the number of

blocks with specific dimensions. Therefore, since the number of blocks is not necessarily fixed in this technique, the number and size of the blocks can be coordinated appropriately. As the size of blocks increases, their number does not decrease, and the feature vector is rich enough to train the classifier used well. Fig. 2 illustrates the block diagram of the proposed block-based steganalysis system. This system includes training and testing phases, each of which will be discussed in detail in the subsequent sections.

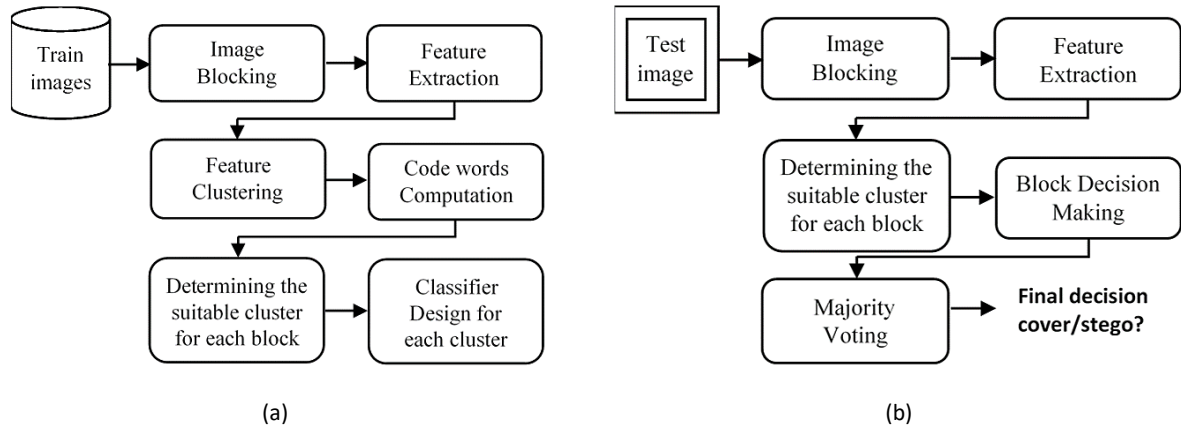


Fig. 2: The block diagram of the block-based steganalysis system (a) training phase (b) testing phase.

A. Training Phase

A set of images is required for the training process, including a combination of cover and stego images. After selecting a set of cover images, one or more steganographic algorithms are chosen to generate stego images. Then, using the steganographic algorithm, the secret data is embedded in the existing cover images to generate the corresponding stego image. This process creates the required set of cover and stego images. All the $M \times N$ cover/stego image pairs in the training set are decomposed into smaller homogeneous blocks of size $B \times B$. Then, the selected feature extraction algorithm is applied to all obtained blocks.

Block size and number affect the steganalysis's accuracy. There are some noteworthy points [10]:

- If the block size is enlarged, the standard deviation of the block features will decrease, and it will be easier to design a classifier that can distinguish between cover images and stego images. Therefore, the larger block size increases the extracted features' detectability.
- The more the number of blocks, the more accurate the results obtained from the classification stage.

Given these points, larger block sizes and more block numbers are more suitable for block-based steganalysis. However, there is an inverse relationship between the block size and the number of blocks in non-overlapping mode, and it is impossible to increase both parameters

simultaneously. Using overlapping blocks is an alternative solution to achieve this goal. In other words, if overlapping blocks are used, the number of blocks will increase, assuming the use of fixed-size blocks. Nevertheless, it is noteworthy that the overlapping level of blocks is an influential parameter that can be changed to control the increase in the number of blocks. A trade-off between the detection accuracy and the number of blocks is required, as decomposing more blocks from the training set increase the computational complexity. If the number of decomposed blocks is vast, a random sampling algorithm is used to select a subset of blocks, which reduces the classification complexity.

Up to this point, the training set contains cover images and corresponding stego images. For K sampled blocks selected by random sampling algorithm, $K/2$ sampled blocks are chosen from cover images. The remaining $K/2$ blocks are corresponding blocks decomposed from stego images. Generally, random sampling is better than linear sampling since it selects more diverse blocks.

Then, we should look for a way to partition the blocks into different C classes. This partitioning is based on the steganalysis features decomposed from the blocks. Therefore, all blocks are clustered into different classes based on the similarity of their features. After clustering the blocks into different classes, the averaged feature vector is calculated in each block class, referred to as the codeword of that block class. For example, if we use

merged Markov features [21] for feature extraction, each codeword will have 274 feature components. The more the number of C classes, the better the steganalysis performance, but the complexity is also increased. Therefore, we should seek a suitable C that balances computational complexity and performance.

The Tree-Structured Vector Quantization (TSVQ) method is used for block clustering, which converts image blocks into a binary tree structure based on similarity. Hence, the entire set of sampled blocks is divided into two subsets. This process is repeated in each subset until all blocks have similar features at a certain level. The K-means clustering algorithm is used in each partitioning step to decompose cluster S into two sub-clusters, which are denoted by S_1 and S_2 . It is done by minimizing the within-cluster energy sum, i.e., $E(S_1, S_2)$, which is given by (1):

$$E(S_1, S_2) = \sum_{X_i \in S_1} \|X_i - \mu_1\|^2 + \sum_{X_i \in S_2} \|X_i - \mu_2\|^2 \quad (1)$$

where X_1, X_2, \dots, X_n represent the feature vectors of n blocks, and μ_i denotes the averaged feature vector in S_i , which can be calculated according to (2):

$$\mu_1 = \sum_{X_i \in S_1} X_i, \quad \mu_2 = \sum_{X_i \in S_2} X_i \quad (2)$$

After dividing K blocks into different C classes, the averaged feature vector or codeword is determined for each class. Codewords are utilized to classify the blocks of a test image using the minimum distortion energy criterion in the feature space. In the implemented TSVQ, the partitioning operation stops when all the blocks within a node are homogeneous enough. Therefore, the stopping criterion in the clustering is based on $E(S_1, S_2)$ value, and the node with the most significant minimum distortion value is always split. This process is repeated until reaching the desired number of C classes.

After obtaining the C codewords representing the C classes of the sampled blocks, all the sample blocks in the training set are assigned into one of the C classes. This assignment is based on the distortion criterion $E_i(f_c, f_s)$, which is defined as the sum of two energies from a codeword for class i, according to (3):

$$E_i(f_c, f_s) = E_i(f_c) + E_i(f_s) \quad (3)$$

where f_c and f_s represent the feature vectors of a block from the cover and stego images, respectively, $E_i(f_c)$ indicates the energy between the number of features extracted from a cover image block, and $E_i(f_s)$ denotes the same value for the stego image block.

After calculating $E_i(f_c, f_s)$ for all C classes, if $E_i(f_c, f_s)$ has the minimum value among all $E_i(f_c, f_s)$ for $1 \leq i \leq C$, the block pair of the cover image and its corresponding stego image is assigned into class C_j . Using the features

of cover and stego image blocks for each class, a special classifier, such as an SVM classifier, can be trained for each of the C classes.

B. Testing Phase

Each test image, as in the training images, is blocked according to block size and blocks' overlapping level. Each block of the test image is assigned into a class using the minimum distortion energy. Depending on the class of each block, the classifier obtained from the training process is applied here. a decision is made about whether each block is a cover or a stego block. Therefore, every test image's total number of decisions equals the block number. Then, a majority voting approach is utilized to decide whether a given image is a cover or a stego. If the number of cover blocks is more/less than the number of stego blocks, the image is identified as a cover/stego.

Results and Discussion

The block-based steganalysis framework was implemented in the MATLAB software environment. All tests were performed on a computer with a Corei7 CPU with four cores and 6GB memory capacity. Therefore, the algorithm was tested on different steganographic methods in both spatial and JPEG domains to study the performance of the proposed block-based steganalysis algorithm. In each mode, 500 images were selected as cover and stego images (1000 images in total) in the training and testing phases.

The well-known and widely used BOSSbase dataset was used for the spatial domain. Some of these sample images are shown in Fig. 3. The dataset contains a thousand grayscale images with a size of 512×512 pixels in PGM format. The WOW and S-UNIWARD algorithms were used to embed messages in spatial domain.



Fig. 3: Sample images from the BOSSbase dataset.

The INRIA Holidays dataset was used for JPEG images. The target dataset contained more than 1,400 color images under JPEG compression with medium quality. One of the advantages of this dataset is that its images

include different subjects, textures, complexities, and sizes. These images are not special, and mostly they are universal images. The selected images are transformed into 512×512 grayscale images to be used in the proposed algorithm. Some of these sample images are shown in Fig. 4. Perturbed Quantization (PQ) and nsF5 algorithms are used to embed the secret messages. Notably, images are compressed once again in the PQ method.

The intended embedding data must undergo encryption and compression processes before the embedding process to remove any semantic relationships

between data bits, reduce the data size as much as possible, and turn the data into a bit string with random data properties; therefore, the different tests use random data (produced in MATLAB) with different lengths. Fig. 5 shows the resulting stego image from the 0.4bpp embedding level with different steganography methods in Lena image. Cover and stego images cannot be distinguished from one another by human eye. Therefore, visually, the output image of these methods is similar to the cover image, and the presence of data in these images can only be discovered through statistical analysis.



Fig. 4: Sample images from the INRIA dataset (right side: original images, left side: grayscale).



Fig. 5: Cover and resulting stego images from different embedding algorithms (0.4bpp).

Several parameters affect the performance of the proposed method, including the number of block classes (C), feature extraction method, block size and number, and block overlapping level. In the following, by changing these parameters, the accuracy of the proposed method for detecting steganographic methods in both spatial and JPEG domains is measured for different embedding rates or different Bit Per Pixels (BPP). Two versions of the proposed method were used in the tests performed. In the first version or NOBS, the image is decomposed to non-overlapping blocks, but in the second version or OBS, the overlapping blocks are used for the image blocks. The frame-based steganalysis method is also referred to as FS.

A. The Performance Comparison of Frame-Based and Block-Based Steganalysis Techniques

The idea of block-based steganalysis is proposed to improve the accuracy of frame-based steganalysis. Before examining the influence of the parameters on the proposed method's performance, this subsection examines the detection accuracy of steganalysis in two without block (frame-based method) and with block (non-overlapping) modes in both spatial and JPEG domains to compare and create an overview of the obtained results. Table 2 reports the steganalysis performance for different embedding rates of JPEG domain methods. In the NOBS, the number of block classes is 16, and the block size is

64×64. This test uses [22] for feature extraction and SVM with Gaussian kernel as a classifier. The results show that the NOBS steganalysis outperforms the traditional steganalysis methods. In addition, according to Table 1, the PQ has lower detection accuracy than the nsF5 methods, i.e., the PQ is more secure against steganalysis.

Table 3 shows the detection accuracy obtained for different embedding rates of the spatial domain. In the experiments, the method [23] was used as the FS method and NOBS method used the SPAM feature [23] for feature extraction phase. The number of block classes is 16, and the size of each block is 64×64 for NOBS. According to the results, NOBS's performance improvement over the FS method is evident in all embedding rates. Furthermore, the S-UNIWARD algorithm is slightly more robust against the steganalysis attacks than the WOW algorithm.

Table 2: The detection accuracy of FS and NOBS techniques for JPEG domain

BPP	PQ		nsF5	
	FS	NOBS	FS	NOBS
0.05	51.3	52.72	52.46	54.50
0.1	52.16	55.45	54.85	58.72
0.2	53.33	60.45	59.43	66.20
0.3	54.20	67.04	62.38	73.50
0.4	55.34	73.18	70.50	82.75

The results of Tables 2 and 3 demonstrate that as the embedding rate of the image increases, the NOBS exhibits higher detection accuracy. The following subsections examine the impact of the available criteria on block-based steganalysis performance. Notably, each phase selects the most robust algorithm for the experiments. Thus, for the spatial domain, the implementation results are tested on stego images created by the S-UNIWARD and images created by the PQ embedding technique for the JPEG domain.

Table 3: The detection accuracy of FS and NOBS techniques for spatial domain

BPP	WOW		S-UNIWARD	
	FS	NOBS	FS	NOBS
0.05	52.80	54.30	53.20	53.90
0.1	55.25	59.70	55	58.45
0.2	57.40	66.5	56.5	65.85
0.3	61.75	74.04	59.23	73.33
0.4	66.72	84.20	64.40	82.85

B. The Influence of Block Class Numbers

After blocking the input images, the decomposed blocks are classified into different classes based on the feature extracted through a classifier. It is expected that with the increase in the number of block classes (C) and more available codewords, the average block detection accuracy and the final detection accuracy of the algorithm will improve because of increasing the similarity of the block features belonging to each class. Table 4 reports the experimental results of the effect of block class number on the PQ steganalysis for the embedding rate of 0.6. As expected, the detection accuracy increases from 71.50 to 82.56 as the number of block classes increases from 2 to 64. However, the algorithm's performance improvement becomes saturated when the number of classes reaches 32 and more. Obviously, as the number of classes increases, more costs must be paid for computations. Therefore, experiments usually consider the middle bound for the number of classes to achieve optimal performance and balance the number of classes and computational complexity.

Table 4: The detection accuracy of the NOBS for different block classes

Number	Accuracy
2	71.05
4	73.27
8	75.31
16	78.53
32	81.70
64	82.56

C. The Influence of the Feature Extraction Method

Each feature extraction method focuses on specific image dependencies and statistical data. Since there are different feature extraction methods, the performance of available steganalysis algorithms will also vary. Therefore, this subsection examines the effect of different feature extraction methods on the proposed steganalysis performance. This study used three basic and well-known techniques to extract the desired features to measure the NOBS's performance for the JPEG domain. Other steganalysis researchers have widely used these techniques. Table 5 shows the obtained results.

This experiment used 512×512 images with 64×64 block size and eight block classes. Data were embedded into images with embedding capacity from 0.05 to 0.4. As can be seen, for low embedding capacities, the [24] method slightly outperforms the two methods [21], [22]. However, as the embedding capacity increases, [22] has

the highest detection accuracy among these three approaches. In any case, [21] is the least influential compared to other tested methods. By default, all the experiments applied [22] for the feature extraction phase to achieve the best performance and detectability.

Two different feature extraction methods were used to investigate the effect of the extracted feature set on the spatial domain steganalysis. The first method is the SPAM [23], which extracts 686 features based on second-order Markov features from each image. The second method is the spatial rich model (SRM) [25], which extracts 34671 features from each image. Table 6 reports the obtained results.

Table 5: The NOBS detection accuracy for the effect of different feature extraction methods on the PQ approach

BPP	Feature extraction method		
	[22]	[21]	[24]
0.05	51.59	50.45	52.04
0.1	55.22	53.86	55.90
0.2	57.04	55.90	58.40
0.3	65.22	57.95	61.81
0.4	72.27	68.40	70.02

Table 6: The NOBS detection accuracy for the effect of different feature extraction methods on the S-UNIWARD

BPP	Feature extraction method	
	SPAM	SRM
0.05	50.80	52.80
0.1	55.20	57.20
0.2	62.50	65.5
0.3	71.35	75.35
0.4	80.18	81.18

According to Table 6, the SRM outperforms the SPAM due to the high feature dimensions and richer feature vector. On the other hand, the steganalysis execution procedure is time-consuming due to the high computational complexity, and the performance decreases compared to the expected level. The following subsections use the SPAM feature extraction method for the spatial domain to facilitate the experimental process.

D. The Influence of Block Size and Number

As mentioned earlier, as the block size increases, the average block decision accuracy increases and improves the overall algorithm's detectability. As the number of

blocks increases, the performance also improves due to the availability of more features. However, assuming non-overlapping blocks, in practice, there is an inverse relationship between these two parameters. For a fixed size 512×512 image, increasing the block size will result in fewer blocks, negatively impacting performance. Table 7 presents the average detection accuracy of the NOBS for different embedding rates on the PQ algorithm in the JPEG domain for 512×512 images and 16 block classes.

As the block size increases, which reduces the number of blocks for four different block sizes, the algorithm performance also decreases. Alternatively, the smaller the size of the blocks, the higher the algorithm's computational complexity due to the increase in the number of blocks. Fig. 6 depicts these results. According to the results, for lower embedding capacities, the detection accuracy obtained is relatively the same and somewhat negligible. However, the performance degrades as the data embedded in the image and the block size increase.

It is also worth noting that the experiments considered the smallest block size as 32×32. For smaller sizes, it is possible to reverse the results and reduce the detection accuracy due to excessive computational complexity. Therefore, the minimum threshold for the block size will be 32×32. Both the block size and the number of blocks individually have a significant impact on block-based steganalysis performance. Hence, their relationship should be balanced to achieve optimal performance and benefit from both positive effects.

Table 7: The NOBS detection accuracy for different values of block size and number

BPP	Block size			
	32×32	64×64	128×128	256×256
0.05	55.22	52.72	50.22	49.70
0.1	57.54	55.45	52.50	50.54
0.2	63.18	60.45	55.45	53.22
0.3	72.40	67.04	62.27	56.50
0.4	78.53	73.18	60.77	57.95

E. The Influence of Overlapping Blocks

The previous subsection investigated the inverse relationship between block number and size and their impact on the proposed steganalysis performance. According to the previous results, there is a dependency between block number and size, and an average limit should be considered for optimal performance. The proposed method uses overlapping blocks to overcome this limitation. For a block with a given size, the blocks are

moved by predetermined step sizes (S) to get more. Therefore, one can take advantage of both larger blocks and more blocks. It will significantly improve the algorithm performance. Table 8 lists the number of blocks for a 512×512 image with different block sizes (four modes) and different step sizes (three modes). Note that if the step size equals the block size ($S=B$), the blocks are not overlapping. As the step size decreases, the overlapping degree of blocks and thus the number of blocks increase.

Table 9 presents the detection accuracy of the proposed method for both domains for two modes of non-overlapping blocks and overlapping blocks with a step size equal to one-half of the block size ($S = B/2$). In the experiment, the block size is 64×64 in normal mode. Consequently, by implementing the overlapping approach and considering a step size of 32 pixels for 512×512 images, we will have a fixed number of 225 blocks for each image. According to the results, the detection accuracy is improved as the number of overlapping blocks increases.

Table 9 shows the detection accuracy improvement rate for the PQ algorithm related to the JPEG domain for two modes based on blocks with a fixed number and

overlapping blocks. Considering the step size equal to one-half of the block size, the number of blocks also increases in the spatial domain, and as a result, the algorithm performance improves compared to the previous state. The results of the S-UNIWARD experiments are also shown in Table 9. On average, the block-based steganalysis detection accuracy using the idea of overlapping blocks improves by more than 9% for PQ detection and more than 9.32% for S-UNIWARD detection at various embedding rates.

Table 8: Number of blocks for a 512×512 image with different block sizes and step sizes

block size	S: Step size, B: Block size		
	$S = B$	$S = B/2$	$S = B/4$
256×256	4	9	25
128×128	16	49	169
64×64	64	225	841
32×32	256	961	3721

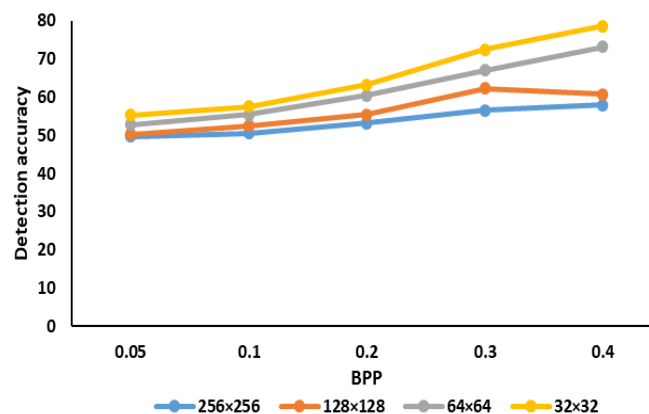


Fig. 6: Decreased NOBS detection accuracy with increasing block size.

Table 9: NOBS and OBS detection accuracy for two PQ and S-UNIWARD approaches at different embedding rates

BPP	PQ			S-UNIWARD		
	NOBS	OBS	improvement	NOBS	OBS	improvement
0.05	52.72	55.50	5.3%	53.90	56.20	4.3%
0.1	55.45	62.25	12.3%	58.45	66.70	14.1%
0.2	60.45	64	5.9%	65.85	74.25	12.8%
0.3	67.04	73	8.9%	73.33	79.5	8.4%
0.4	73.18	82.5	12.8%	82.82	88.64	7%

In general, for a given step size, the smaller the step size is, the more blocks are obtained, with fixed image size and fixed block size, and thus, the steganalysis performance is improved. However, despite the performance improvement, minimum step sizes are impractical due to the high computational complexity. By benefiting from overlapping blocks, the algorithm performance can be easily enhanced for different modes. For convenient comparison, Figs. 7 and 8 illustrate the bar charts of three steganalysis modes for the JPEG and spatial domains.

F. The Performance Comparison of the Proposed Method with State-Of-The-Art Techniques

The results of the earlier subsections demonstrate that if the steganalysis method uses the features extracted from the image to decide the state of the test image, the use of overlapping blocks (rather than the whole image or non-overlapping blocks) for feature extraction can enhance the accuracy of the steganalysis techniques. However, several novel steganalysis approaches use deep

learning algorithms in which feature extraction is performed automatically. For better performance measurement, Table 10 presents the comparison results of the proposed approach with six deep learning-based steganalysis methods for S-UNIWARD at different embedding rates. These results were extracted from [26], [27] and [28], and in cases marked with a dash (–), the desired result was not cited in the references.

Table 10 reveals that the automatic feature extraction approaches are not significantly superior to the proposed method. Another main advantage of the proposed method is the much lower computational complexity. The statistics provided in [27] indicate that the time required to train the fastest method available [28] is about 3 hours. In contrast, less time is needed for non-automatic feature extraction-based steganalysis techniques. Indeed, despite the growing research of deep learning-based steganalysis methods and the existence of potent state-of-the-art hardware, it is not unlikely that these methods will be developed very quickly.

Table 10: The comparison of the OBS detection accuracy and deep learning-based methods at different embedding rates

BPP	Xu-Net [32]	Ye-Net [31]	SN-Net [30]	ReST-Net [29]	CIS-Net [28]	SFR-Net [27]	AG-Net [26]	OBS
0.1	59.43	59.17	64.79	64.15	64.72	-	-	66.7
0.2	66.67	66.49	73.18	68.73	73.79	76.8	-	74.25
0.3	73.68	74.38	79.29	76.44	76.36	-	80.66	79.5
0.4	80.12	77.36	83.47	84.28	85.38	87.9	85.49	88.64

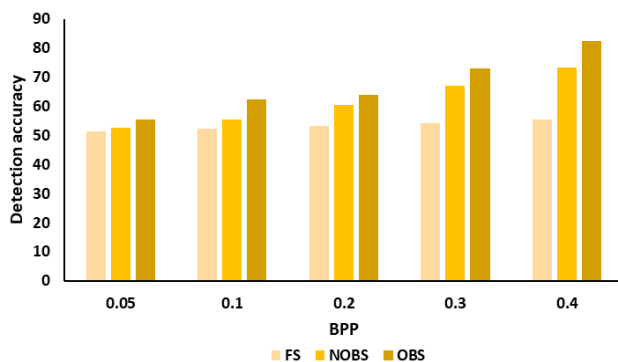


Fig. 7: The detection accuracy of FS, NOBS, and OBS approaches for PQ.

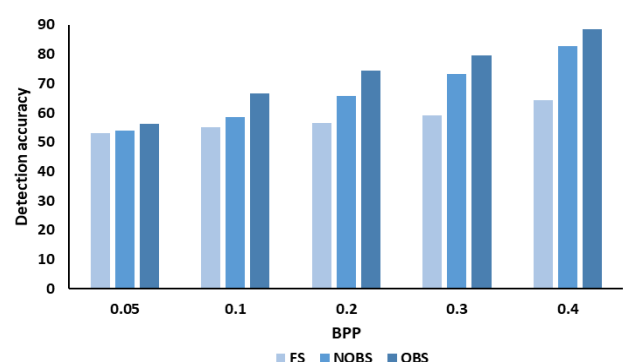


Fig. 8: The detection accuracy of FS, NOBS, and OBS approaches for S-UNIWARD.

Conclusion

Most steganalysis techniques require image feature extraction to detect stego images significantly altered by data embedding. Some steganalysis methods use the

entire image to extract features (frame-based steganalysis), while others decompose the image into blocks and perform feature extraction for each block separately (block-based steganalysis). Block-based

steganalysis has two major advantages over frame-based steganalysis. (1) the accuracy of block-based steganalysis is improved without increasing the number of features, and (2) block-based steganalysis yields more reliable results than frame-based schemes because the block decomposition process produces multiple samples. The research results indicate that increasing the number and size of blocks plays a crucial role in improving block-based steganalysis performance. However, if non-overlapping blocks are used, the relationship between these two parameters is inverted. This study proposed using overlapping blocks to resolve this contradiction, i.e., by increasing the overlapping level of the blocks, the number of blocks can be increased proportionally to the desired size.

The experimental results revealed that several parameters affected the performance of the proposed method and that there was a trade-off between these parameters and the complexity of the proposed method, making selecting these parameters a complex and challenging process. The outcomes showed that the concept of overlapping blocks improved the detection accuracy of techniques in the spatial and JPEG domains by more than 9%. In addition, one of the advantages of the proposed method is its comparable accuracy and lower computational complexity compared to state-of-the-art deep learning-based steganalysis.

Given the impact of the concept of overlapping blocks on the more precise discovery of steganographic methods and the growing popularity of CNN networks, these two concepts can be combined to create a more effective model. In this new model, rather than CNN using the entire image as input, overlapping blocks of the image that is more suitable based on the image texture can be selected and used as CNN input in a preprocessing step. Furthermore, we can expect the CNN network to succeed because the smaller input size reduces its complexity.

Author Contributions

All the authors participated in the conceptualization, implementation, and writing.

Acknowledgment

This work is completely self-supporting, thereby no any financial agency's role is available.

Conflict of Interest

The author declares no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>DCT</i>	Discrete Cosine Transform
JPEG	Joint Photographic Experts Group
<i>DNN</i>	deep neural networks
<i>CNN</i>	convolutional neural networks
<i>ABC</i>	Artificial Bee Colony
<i>SVM</i>	Support Vector Machine
TPs	True Positives
TNs	True Negatives
<i>IQM</i>	Image Quality Metrics
LSBR	LSB Replacement
LSBM	LSB Matching
<i>TSVQ</i>	tree-structured vector quantization
<i>PQ</i>	Perturbed Quantization
<i>WOW</i>	Wavelet Obtained Weights
S-UNIWARD	Spatial Universal Wavelet Relative Distortion
<i>BPP</i>	Bit Per Pixels
<i>NOBS</i>	Non-Overlapping Blocks-based Steganalysis
<i>OBS</i>	Overlapping Blocks-based Steganalysis
<i>FS</i>	Frame-based Steganalysis
<i>SPAM</i>	Subtractive Pixel Adjacency Matrix
<i>SRM</i>	Spatial Rich Model

References

- [1] S. Dhawan, R. Gupta, "Analysis of various data security techniques of steganography: A survey," *Inf. Secur. J.*, 30(2): 63–87, 2021.
- [2] I. J. Kadhim, P. Premaratne, P. J. Vial, B. Halloran, "Comprehensive survey of image steganography: Techniques, evaluations, and trends in future research," *Neurocomputing*, 335: 299–326, 2019.
- [3] V. Sabeti, M. Sobhani, S. M. H. Hasheminejad, "An adaptive image steganography method based on integer wavelet transform using genetic algorithm," *Comput. Electr. Eng.*, 99: 107809, 2020.
- [4] R. Atta, M. Ghanbari, "A high payload data hiding scheme based on dual tree complex wavelet transform," *Optik*, 226: 165786, 2021.
- [5] G. Paul, I. Davidson, I. Mukherjee, S. S. Ravi, "Keyless dynamic optimal multi-bit image steganography using energetic pixels," *Multimed. Tools Appl.*, 76(5): 7445–7471, 2017.
- [6] B. T. Hammad, I. T. Ahmed, N. Jamil, "A steganalysis classification algorithm based on distinctive texture features," *Symmetry*, 14(2): 236, 2022.
- [7] M. Dalal, M. Juneja, "Steganography and steganalysis (in digital forensics): a cybersecurity guide," *Multimed. Tools Appl.*, 80(4): 5723–5771, 2021.
- [8] S. Cho, J. Wang, C. C. J. Kuo, B. H. Cha, "Block-based image steganalysis for a multi-classifier," in *Proc. 2010 IEEE International Conference on Multimedia and Expo*: 1457–1462, 2010.
- [9] S. Cho, B. H. Cha, J. Wang, C. C. J. Kuo, "Performance study on block-based image steganalysis," in *Proc. 2011 IEEE International Symposium of Circuits and Systems (ISCAS)*: 2649–2652.
- [10] S. Cho, B. H. Cha, M. Gawecki, J. Kuo, "Block-based image steganalysis: Algorithm and performance evaluation," *J. Vis. Commun. Image Represent.*, 24: 846–856, 2013.
- [11] X. Lin, "Steganography and steganalysis," in *Introductory Computer Forensics*, Springer, Cham, 557–577, 2018.
- [12] M. Broda, V. Hajduk, D. Levický, "Universal statistical steganalytic method," *J. Electr. Eng.*, 68(2): 117–124, 2017.
- [13] X. Y. Luo, D. S. Wang, P. Wang, F. L. Liu, "A review on blind detection for image steganography," *Signal Processing*, 88(9): 2138–2157, 2008.
- [14] F. G. Mohammadi, H. Sajedi, "Region based Image Steganalysis using Artificial Bee Colony," *J. Vis. Commun. Image Represent.*, 44: 214–226, 2017.
- [15] A. Dehdar, A. Keshavarz, N. Parhizgar, "Image steganalysis using modified graph clustering based ant colony optimization and random forest," *Multimed. Tools Appl.*, 1–18, 2022.
- [16] J. Chen, W. Lu, Y. Fang, X. Liu, Y. Yeung, Y. Xue, "Binary image steganalysis based on local texture pattern," *J. Vis. Commun. Image Represent.*, 55: 149–156, 2018.
- [17] R. Wang, M. Xu, X. Ping, T. Zhang, "Steganalysis of JPEG images by block texture based segmentation," *Multimed. Tools Appl.*, 74(15): 5725–5746, 2015.
- [18] G. R. Suryawanshi, S. N. Mali, "Study of effect of DCT domain steganography techniques in spatial domain for JPEG Images steganalysis," *Int. J. Comput. Appl.*, 127(6): 16–20, 2015.
- [19] G. R. Suryawanshi, S. N. Mali, "Universal steganalysis using IQM and multiclass discriminator for digital images," *Int. Conf. Signal Process. Commun. Power Embed. Syst.*: 877–881, 2016.
- [20] J. Zhu, X. Zhao, Q. Guan, "Detecting and Distinguishing Adaptive and Non-Adaptive Steganography by Image Segmentation," *Int. J. Digit. Crime Forensics*, 11(1): 62–77, 2019, doi: 10.4018/IJDCF.2019010105
- [21] T. Pevný, J. Fridrich, "Merging markov and DCT features for multi-class JPEG steganalysis," *Secur. Steganography, Watermarking Multimed. Contents IX*, 6505: 650503, 2007.
- [22] C. Chen, Y. Q. Shi, "JPEG image steganalysis utilizing both intrablock and interblock correlations," in *Proc. IEEE Int. Symp. Circuits Syst.*: 3029–3032, 2008.
- [23] T. Pevný, P. Bas, J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inf. Forensics Secur.*, 5(2): 215–224, 2010.
- [24] J. Kodovský, S. Binghamton, J. Fridrich, "Calibration revisited," in *Proc. 11th ACM workshop on Multimedia and security*: 63–74, 2009.
- [25] J. Fridrich, J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Secur.*, 7(3): 868–882, 2012.
- [26] H. Zhang, F. Liu, Z. Song, X. Zhang, Y. Zhao, "AG-Net: An advanced general CNN model for steganalysis," *IEEE Access*, 10: 44116–44122, 2022.
- [27] G. Xu, Y. Xu, S. Zhang, X. Xie, "SFRNet: Feature extraction-fusion steganalysis network based on squeeze-and-excitation block and RepVgg Block," *Secur. Commun. Networks*, 2021: 1–10, 2021.
- [28] S. Wu, S. Zhong, Y. Liu, M. Liu, "CIS-Net: A novel CNN model for spatial image steganalysis via cover image suppression," *arXiv preprint arXiv:1912.06540*, 2019.
- [29] B. Li, W. Wei, A. Ferreira, S. Tan, "ReST-Net: Diverse activation modules and parallel subnets-based CNN for spatial image steganalysis," *IEEE Signal Process. Lett.*, 25(5): 650–654, 2018.
- [30] S. Wu, S. H. Zhong, Y. Liu, "A Novel Convolutional Neural Network for Image Steganalysis with Shared Normalization," *IEEE Trans. Multimed.*, 22(1): 256–270, 2017, doi: 10.48550/arxiv.1711.07306.
- [31] J. Ye, J. Ni, Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Trans. Inf. Forensics Secur.*, 12(11): 2545–2557, 2017.
- [32] G. Xu, H. Z. Wu, Y. Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Process. Lett.*, 23(5): 708–712, 2016.

Biographies



Vajiheh Sabeti is an Assistant Professor of Engineering and Technology department at Alzahra University. She received her B.S. degree in Software Engineering in 2004 and her M.Sc. degree in Computer Architecture in 2007 and her Ph.D. degree in Computer Engineering in 2012 from the Electrical and Computer Engineering department of Isfahan University of Technology (IUT), Isfahan, Iran. Her research interests are softcomputing,

image processing, and information hiding (steganography, watermarking).

- Email: v.sabeti@alzahra.ac.ir
- ORCID: [0000-0002-9985-9143](https://orcid.org/0000-0002-9985-9143)
- Web of Science Researcher ID: AAD-1661-2022
- Scopus Author ID: 24470278600
- Homepage: <https://staff.alzahra.ac.ir/vajiheh-sabeti/en/>

How to cite this paper:

V. Sabeti, "An improved approach to blind image steganalysis using an overlapping blocks idea" J. Electr. Comput. Eng. Innovations, 11(2): 563-276, 2023.

DOI: [10.22061/jecei.2022.9241.592](https://doi.org/10.22061/jecei.2022.9241.592)

URL: https://jecei.sru.ac.ir/article_1811.html





Research paper

Development of Wind Turbine Fault Analysis Setup based on DFIG Hardware in the Loop Simulator

M. Kamarzarrin¹, M. H. Refan^{1,2,*}, P. Amiri¹, A. Dameshghi^{1,2}

¹Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

²MAPNA Electric & Control, Engineering & Manufacturing Co. (MECO), Alborz, Iran.

Article Info

Article History:

Received 16 July 2022

Reviewed 15 October 2022

Revised 29 October 2022

Accepted 07 November 2022

Keywords:

Wind turbine

HIL

DFIG

Rotor electrical asymmetry

Inter-turn short circuit

IGBT open-circuit

Abstract

Background and Objectives: Renewable energy, like wind turbines, is growing rapidly in the world today due to environmental pollution, so their maintenance plans are very important. Fault diagnosis and fault-tolerant approaches are typical methods to reduce the cost of energy production and downtime of Wind Turbines (WTs).

Methods: In this paper, a new Hardware In the Loop (HIL) simulator based on Double Feed Induction Generator (DFIG) for fault diagnosis and fault-tolerant control is proposed. The system developed as a laboratory bed uses a generator with a power of about 90 kW, which is connected from two sides to a back-to-back power converter with a power of one-third of the generator power. The generator is connected to a motor as a propulsion and wind energy replacement with a power of about 110 kW, and this connection is established through a gearbox with a gear ratio of more than three.

Results: The effectiveness of the proposed simulator is evaluated based on different fault representations back-to-back converter and generator.

Conclusion: The experiment shows that the Condition Based Maintenance (CBM) is improved by the proposed simulator and the fault is modeled before serious damage occurs. This setup is effective for the development of wind turbine fault analysis software. As the testing on real WTs is very expensive, to improve and develop the research fields of condition monitoring and WT control, this low-cost setup is effective.

*Corresponding Author's Email
Address: refan@sru.ac.ir

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Among renewable energies, wind energy is paramount and this energy now plays a significant role in the world economic equations. Regarding the high number of Wind Turbine (WT) components, it has many faults [1], [2]. Considering the development of WT, due to the rotating equipment, fault detection, and prediction are very important in these systems. Due to the installation locations of WTs, their repairs are challenging and this equipment has a variety of electrical and mechanical failures. due to their installation place [3]-[5]. Condition-

Based Maintenance (CBM) is including the CM module, Fault Detection (FD) module, fault isolation module, fault prognosis module, and fault-tolerant module (Based on Fig. 1) [3]. A study based on fault analysis is performed on different turbines. Based on this study the turbine power increase is directly related to the increase in the fault rate over a year [6]. Fig. 2 shows the failure rates of different WT components in separate studies [7]. From this figure, it is clear that the electrical failure rate is high, and of course, the highest damage and shutdown is in the gearbox, generator, blade, and propulsion. Electrical components have the highest fault rates and the lowest

shutdown rates. That is, considering that the failure rate is high in the electrical sector, but the damage and shutdown are greater in other areas [8]. Mechanical subsystems including gearboxes, blades, and generator components have low fault rates but high shutdown rates [9].

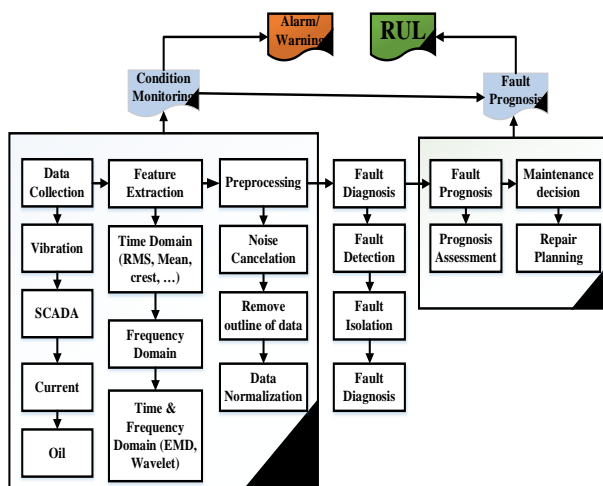


Fig. 1: Different modules in CBM.

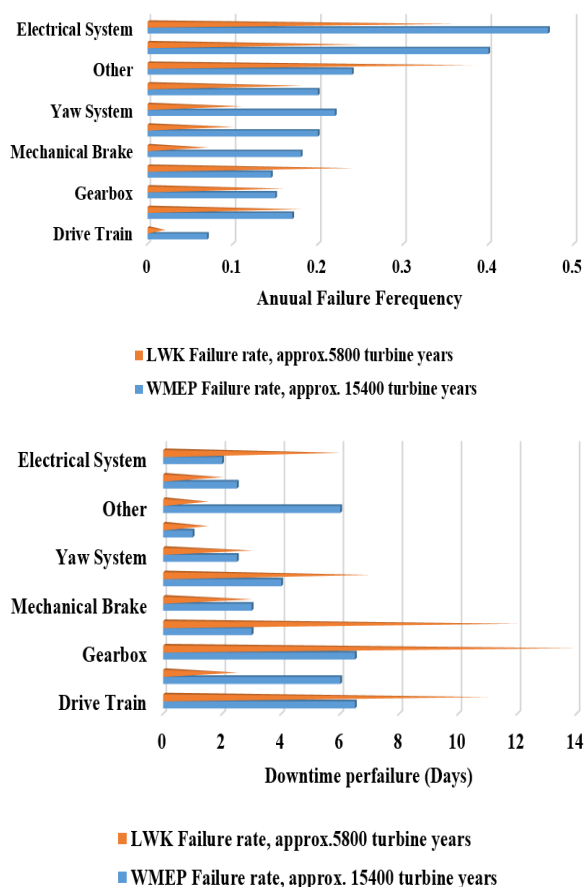


Fig. 2: Relationship between fault frequency and shutdown results from two studies on European WTs during 13 years [13].

This does not mean poor design of these parts, but due to the functional complexity and hardware of these parts [10]. The two main components, the generator, and the gearbox are very important in the reliability of the entire WT system [11]. 75% of faults cause only 5% of shutdowns, while only 25% of faults cause 95% of shutdowns [12]-[14]. It is known that in small and medium-sized turbines, the faults related to the rotor and the slip ring are significant compared to the total faults of the generator, while the bearing fault is more limited [13], [14]. Fig. 3 shows this in an analytical and statistical study of WT fault [15]. A study is conducted in 2018 that analyzes the collected data related to the ReliaWind project, the results of which are shown in Fig. 4 [16]. This result is based on information from more than 4,000 turbines. In new variable speed turbines, most of the faults are related to the rotor and power subsystems. In Fig. 4, the vertical axis on the left of the graph represents the percentage of the total fault rate per (fault/turbine/year) while the horizontal axis on the right shows the cumulative fault rate for each section marked with an orange line [17].

The purpose of this article is to provide a platform to test and improvement of these fault diagnoses and fault-tolerant modules. The availability and reliability of WTs are under the influence of components failure including gearbox, generator, back-to-back converter, main bearing, blades, and tower [18]. In this paper, to identify a different symptom of WTs fault, a HIL test rig for simulating turbine faults is provided; this test rig is a DFIG with a prime mover motor. The topology of the WT system is the same as Fig. 5. Similar to this test rig different setups have been developed in various universities and research centers [13]-[15]. Different faults including Rotor Electrical Asymmetry (REA), Stator Inter-Turn Short Circuit (ITSC), and an open circuit switch fault are applied to display the performance of the HIL setup. In stator WT generators, the failure rate is 30%, the rotor has a failure rate of 40% and the bearing has a failure rate of 12%. In this article, the defects of the rotor winding are investigated. The fault of the rotor screw system is due to the manufacturing method, mechanical and thermal stresses, and current leakage [18], [19]. In [20], the rotor winding fault detection is performed based on the analysis of current spectra. In [22], the rotor winding error is performed based on the analysis of current spectra. In [21], the rotor winding fault is investigated based on the frequency spectrum of the power and current signals, the power signal has a better detection power. The use of model-based methods based on rotor winding modeling is described in [22]. Using frequency analysis and frequency-based rotor fault detection is the most common method for REA fault diagnosis [23].

Percentage Contribution to turbine failures

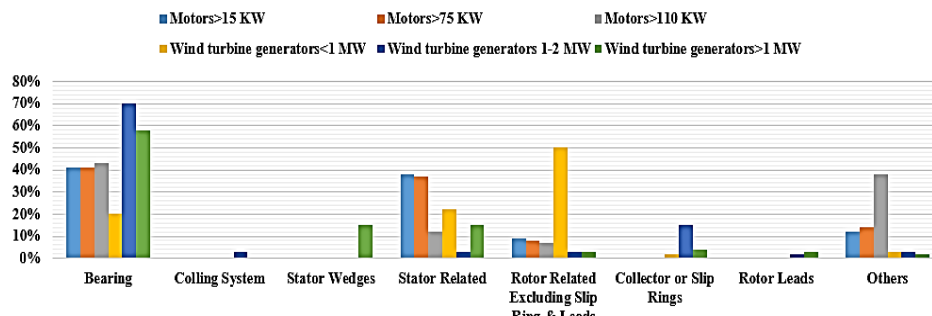


Fig. 3: Relationship between turbine power and separately component fault percentage [15].

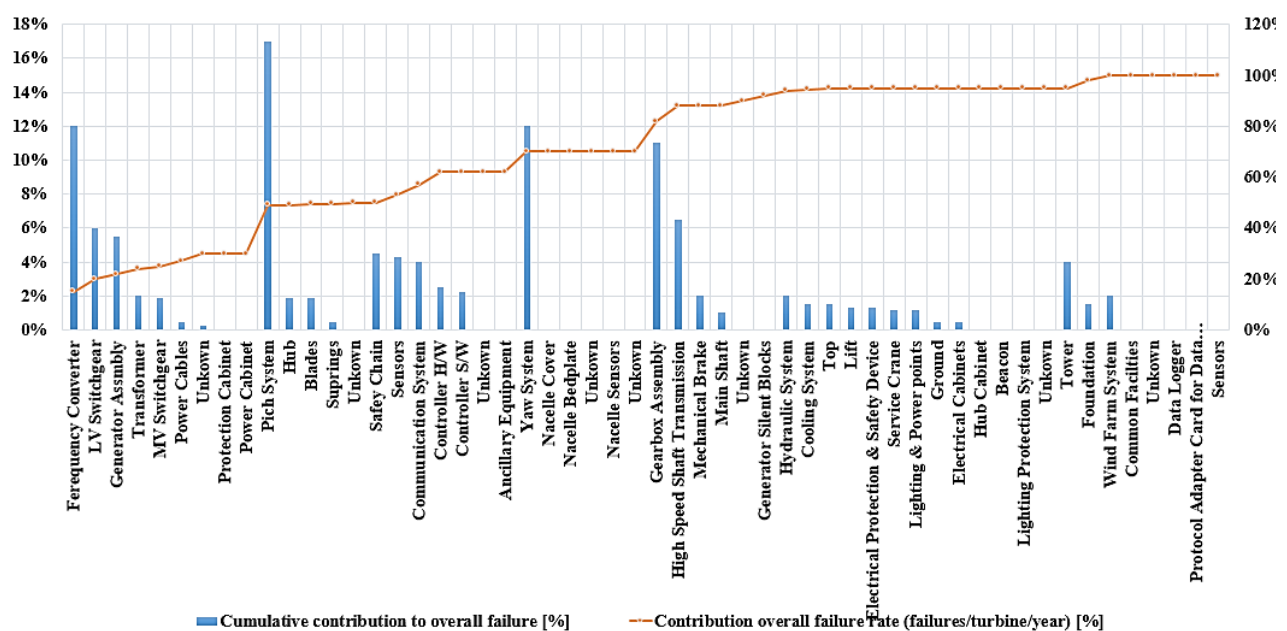


Fig. 4: Expansion of normalized outages related to WT subdivisions and different manufacturers in the analysis of the WT project carried out in 2017 [15].

These two signals are used in [24], [25] based on two generators for the error studied in the paper. But this REA fault causes certain changes in the frequency spectrum of current and power signals. These frequencies are a ratio of the main frequency of the source and depend on the fraction slip of the generator.

The most important frequency is $2sf_s$ [24], which is very efficient for REA fault detection. These frequencies are proportional to the different types of rotor and stator current signals and the control current signals are reflected in Fig. 6 [23], [26], [27]. The second case study fault is related to the stator. Generally, in induction machines including the DFIG, the ITSC fault, which damages the insulation of windings, is common [28]-[30]. Detecting this fault prevents the expansion of the fault. Faults that are related to the winding insulation have weak signatures in signals and their detection is more

difficult than other faults [30]-[32]. Moreover, the variable speed of the WT and its unstable dynamics cause disturbances in current signals and make the FD very difficult. Few types of research related to ITSC fault in DFIG-based WTs have been conducted so far. However, many types of research have been conducted about this fault in induction motors [9]-[13]. Forty percent of the electrical faults in WTs are related to the generator, and 40% of these faults are due to faults in the stator [28]. Investigating the faults related to stators shows that the fault of windings is the most important fault in stators. Damage to the insulator occurs due to ITSC, and partial discharge takes place between stator turns. The latter leads to an inter-turn fault and emerges largely in the form of some kind of faults, such as coil-to-coil, phase-to-phase, open-circuit phase, and phase-to-ground, which may result in WT shutdown [32].

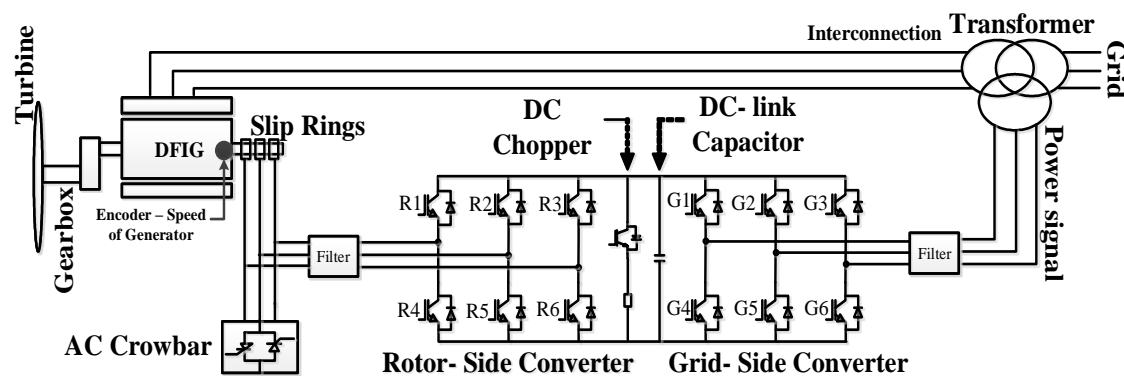


Fig. 5. WT DFIG topology

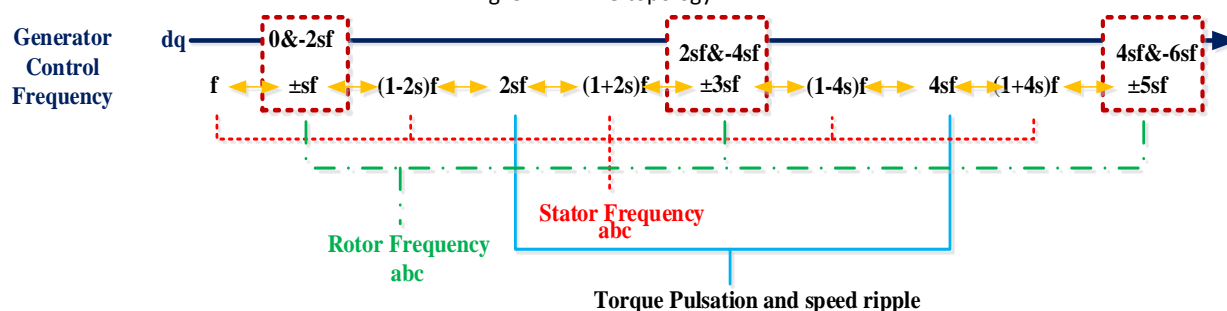


Fig. 6: The REA fault symptoms in frequency-domain.

This fault has little effect on the performance of the machine, and therefore, it can hardly be observed through the performance parameters of the WT; however, its development is very costly and destroys the generator. In Table 1, an appropriate description of various conditions of fault in the stator windings is presented [30]–[32].

Table 1: Various conditions of fault in the stator windings

State	Action
inter-turn short circuit	The generator will continue to operate, but for how long?
Shorts between coils of the same phase	The generator can continue to operate but for how long?
Phase to phase short	The generator fails and protection equipment disconnects the supply
Phase to earth short	The generator fails and protection equipment disconnects the supply
Open circuit in one phase	The generator may continue to operate, depending on the load conditions

Third case study fault related to the back-to-back converter. Power converter or back-to-back converter fault rates in DFIG and PMSG WT are high, in power converters nearly 35% of faults are due to IGBT switches, these faults are often due to mechanical stress, gate

control circuit faults, thermal stress, wiring, and current problems [34]. There are two faults in the WT converter, an open circuit fault, and a short circuit fault. Unlike open circuit faults, which are often solved by software, the short circuit fault of the IGBT switch is often controlled by hardware, the short circuit fault is controlled by the WT protection system, and the WT is turned off, but The fault of the open circuit of IGBT WT converter does not cause the WT to down, but it is effective in the operating conditions of the WT, such as power quality, disturbance of balance and balance between phases [34]. Open circuit failure will cause offset current leakage in the phases, especially the phase related to the same switch. Except for offset and field, it produces jumps in torque and stator current frequency, which reduces the maximum average available for the drive, and of course, this affects power and other performance factors. Of course, the offset current also causes problems in the IGBT characteristics, and this may cause secondary defects and damage to the entire converter structure [34].

The reliability and availability concept is one of the most important issues in the WT industry, where unforeseen downtime can lead to significant economic losses. Fault diagnosis and fault-tolerant approaches are typical methods to reduce the cost of energy production and downtime of WTs. The doubly-fed induction generator converter Fault-Tolerant Control (FTC) plays a significant role in improving the reliability and availability of modern WTs. This article deals with the development of a 90 kW wind turbine simulator based on the DFIG

generator to be used in wind turbine fault diagnosis applications. In this context, intermittent Different faults and failures with different domains can be implemented and controlled in a single and combination format. To check the results and performance of the developed simulator, three categories of errors were checked. The first category of faults is related to defects in the rotor winding. Different faults including REA, Stator ITSC, and an open circuit switch fault are applied to display the performance of the HIL setup. The second case study fault is related to the stator. And finally, the Third case study fault is related to an open circuit fault in the back-to-back converter. The experiment shows that the CBM is improved by the proposed simulator and the fault is modeled before serious damage occurs. This setup is effective for the development of wind turbine fault analysis software. As the testing on real WTs is very expensive, to improve and develop the research fields of condition monitoring and WT control, this low-cost setup is effective.

The contributions of the article are categorized into four main parts.

- For fault diagnosis and Fault-tolerant control, hardware in the loop setup is proposed in this paper.
- The developed structure is quite similar to a real wind turbine; this makes it possible to examine fault and control issues.

- Simulate the actual behavior and dynamics of the WT based on the HIL test setup and perform experiments based on collecting real signals.
- The developed laboratory setup is prepared to emulate three faults; REA, IGBT open circuit fault, Stator ITSC.

The best parts of this paper are organized as follows: the second section is the system design for fault analysis. The third section has introduced the simulator. In section 4, fault representation is described. HIL performance analysis is presented in section 5 and finally, the last section is the conclusion.

HIL (Test Rig) Design for Fault Analysis

The WT hardware simulator platform is used in the developed loop to emulate faults and collect WT signals. In this context, intermittent faults and failures with different domains can be implemented and controlled in single and combination formats.

The fault prediction and detection and fault tolerant control algorithms can be implemented in this hardware. The general structure of the developed hardware simulator is shown in Fig. 7. As can be seen, this structure includes mechanical measurement sections, electrical measurement sections, and a control section.

The important components of the implemented structure are as follows:

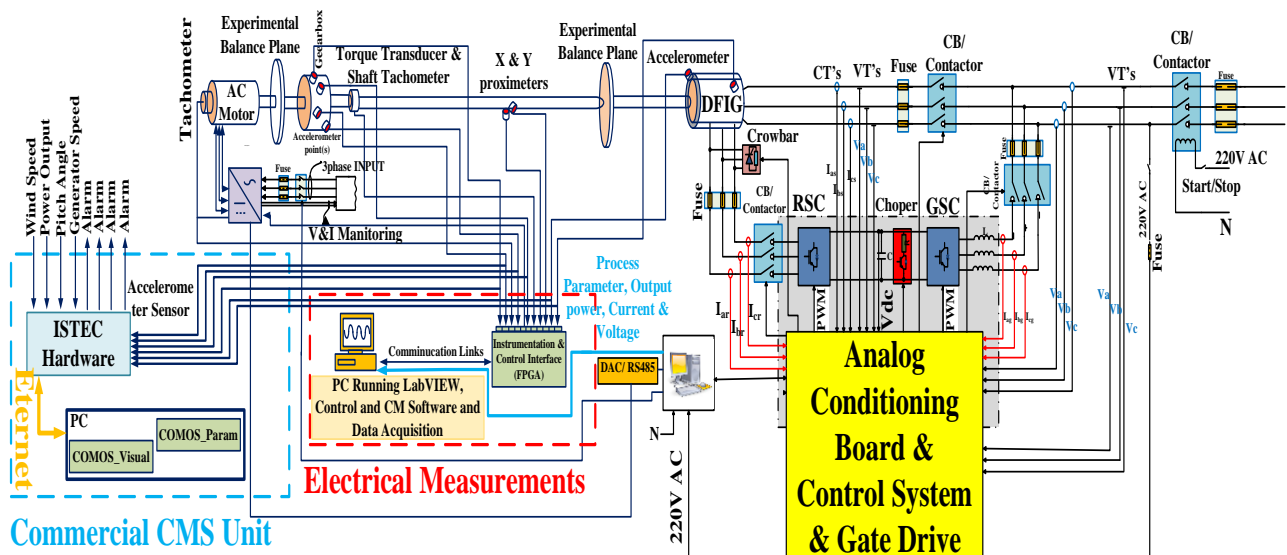


Fig. 7: Schematic diagram of the test rig.

ISTEC hardware for vibration monitoring (2) Accelerometer sensors to measure vibration intensity (3) A PC to install CMS software (Param and Visual) (4) Tachometer to measure generator speed (5) AC motor as the prim mover of the system (6) Shaft balancing plate to detect imbalance and failure in communication shafts between motor and gearbox (7) Torque measurement based on transducer (8) Variable speed driver with programmable ability to drive the motor (9) Drive input

information monitoring system (10) Shaft balancing plate to detect imbalance and failure in communication shafts between generator and gearbox (11) Circuit breaker (crowbar) to protect against high leakage current (12) Current and voltage measurement sensors (13) Fuses and contactors in different directions (14) Power converter with stator side voltage source drive and rotor side voltage source drive (15) Analog and digital measuring boards (16) Central control board (17) Transformers (18)

DC link capacitor and related inductors. The MAPNA HIL system is included a generator from the VEM brand, this generator is operate with 90 kW power. The one gearbox is between the generator and primmover ABB motor (power: 110 kW). The system can be used to simulate converter and generator faults. An overview of these three interconnected devices is shown in Fig. 8. The system is capable of operating at constant and variable speeds. The primary drive motor is designed so that the DFIG generator operates with 4 pairs of poles in synchronous, sub-synchronous, and super-synchronous operating areas. For this purpose, the motor speed is 750 rpm with a grid frequency of 50 Hz 8 poles are selected. The gearbox conversion ratio activates the DFIG generator at a synchronous speed of about 1500 rpm in all three operating areas. The hardware simulator generator is 8 poles and can operate at 50 Hz and 400 volts. The HIL proposed in the paper is designed to simulate WT behaviors and simulates real similar signals and situations. Defects similar to WTs can be created in it or a defective condition can be removed from it, its prominent feature is the modeling of both faults in the electrical, mechanical and electro-mechanical parts.

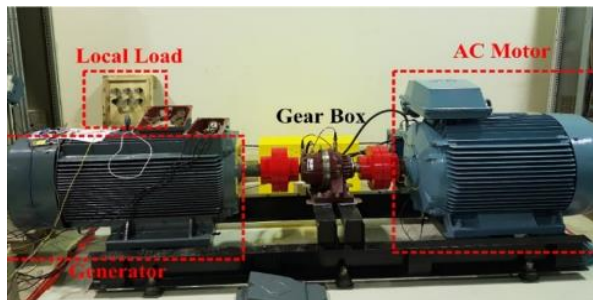


Fig. 8: DFIG WT test rig.

Table 2 shows the hardware simulator parameters.

Test Rig Mechanical and Electrical Components

A. Generator

In DEFA-based WTs, the generator power is often more than 2 MW; In this setup, a 90 kW DFIG is used to simulate a WT. This simulator requires upstream control and control at the converter and generator levels.

To simulate the fault, the winding rotor model must be considered, so a model based on the resistance bank for the rotor is used. In Fig. 9, this model is for the rotor. speed and sub-synchronous and super-synchronous. An external resistor is used to emulate the fault instead of the short circuit.

In the healthy state, the resistance of the three phases of the rotor is Balanced, this balance is disturbed by the addition of resistance. In balanced mode, the slip changes are small, and with the addition of this resistance, these changes are intensified. In this paper, the rotor phase

resistance at NON- asymmetries IS 1.3 Ω . This amount is due to the WT SETUP being in synchronous.

Table 2: Parameters of test-rig components

Drivetrain	
Power	90 kW
Speed	1488RIV/MIN
Stator Voltage	400V
DC Link Voltage	700V
Current	199 A
Torque	578 Nm
Number of pole pairs	4
Magnetizing inductance	120.4 mH
Cos Φ	0.88
Stator resistance	24.8m Ω
Stator inductance	44mH
Rotor resistance	16.6 m Ω
Rotor inductance	33mH
Gearbox	Gear with 1 stage by 1:3.32
Prime mover motor	8 pole / 110 kW/ 400 V/ 742 rpm
Converter	
Filter resistance	5m Ω
Filter inductance	800 μ H
Sampling frequency	100 kHz
Switching frequency	2.5 kHz
Converter control	FOC
DC-Link Capacitor	7 mF

B. Gearbox

The gearbox in the designed setup can be detached and a new gearbox replaced. This gearbox has a solar structure in the form of a parallel shaft. The output of the gearbox is a high speed and its input is low speed therefore the gearbox increases the distance. his gearbox is used to emulate various gear and bearing faults. The gearbox is coupled to the motor and generator in such a way as to prevent intrinsic vibration and achieve the main fault by measuring the signals. In [4], the presented method has been compared with other methods.

The HIL simulator receives wind speed maps and signal samples using a motor drive; the corresponding drive can receive a fixed map or a constant wind speed. This drive also can convert wind speed data to torque as a reference for variable speed motor performance and thus variable speed performance of WT hardware simulator.

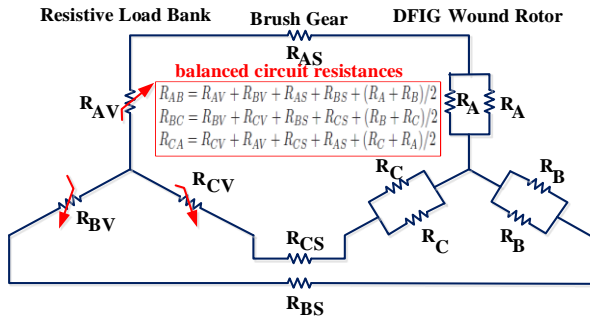


Fig. 9: Generator rotor circuit diagram.

C. Power Converter

In proportion to this input, the drive applies a certain frequency to each wind speed data to the motor. This drive capability allows the implementation of the plan to change the operating areas of the WT and transient conditions. The corresponding control cabinet is shown in Fig. 10 and the upper part of this cabinet contains the relevant drive. The lower part of this cabinet is ISTEK CMS. The back-to-back converter cabinet used in this structure is in Fig. 11. The system consists of two drive-based voltage sources in a back-to-back form, its power is 30 kW, a capacitor bank, and an output filter of 800μH-5mΩ, receptively. The DC link voltage of the converter is 700 volts. In the design of the converter power board, 6 SEMIKRON switches with two IGBTs connected are used. There is a crowbar circuit to prevent a short circuit. The ARM/FPGA-based control with a sampling frequency of 100 kHz for A/D channels has been implemented using conventional PWM (six-leg control), the switching signals are generated at 2.5 kHz frequencies and transmitted via fiber optics to the gate drive and applied to the IGBT, 12 gate drives are used for driving of 12 SEMIKRON switches. The vector control strategy is used for the converters, the control strategy is field vector control that which is based on space vector modulation for rotor and grid side converters.

D. Measurement Devices

To measure the signals required in this paper, a structure similar to Fig. 12 is used. A tachometer is used for measurement of the rotational speed of the generator and it is measured as input to the CMS. The control signals required on the control board are stored in the FPGA data archiving unit. Hioki power analyzer measures current/voltage information of three-phase, power information is measured through ISTEK CMS system. To achieve the research objectives, various measurements are designed in the hardware simulator.

Based on this structure, the rotational speed of the generator is used in the status monitoring and diagnosis process. This signal is measured using a tachometer. CMS measures the vibration signals from this hardware simulator as follows with 8 sensors.

Vibration sensors are accelerometers. The sensors are of the accelerometer type with a measurement accuracy specified by a magnet connector to a specific location on the hardware simulator and measure the vibration of the hardware simulator.

1. Two sensors are located at the end of the drive motor to indicate that the sensors are located in the main bearing of the WT before the gearbox low-speed shaft.

2. Four sensors are located at the gearbox similar to the location of the real gears.

3. Two accelerometer sensors are located on the generator at the beginning and end of the generator similar to the real situation in the WT.

Table 3 shows the general information of the hardware simulator measuring equipment.

E. CMS

MAPNA Kahak wind farm currently uses ISTEK CMS. This system is only for collecting information and calculating some frequency functions and time functions. It is the situation to do this offline, performance information including power, wind speed, generator speed, and step angle is received as the primary parameters from the control unit in the CMS system. Analysis should be performed taking into account the operating condition of the WT.

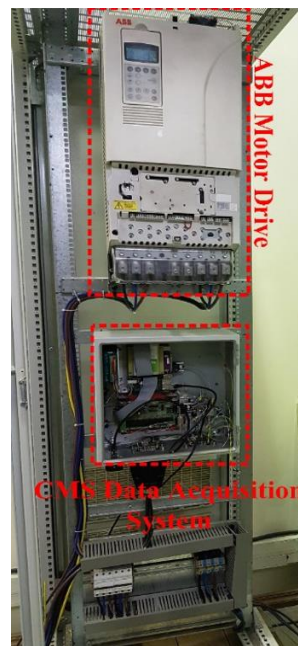


Fig. 10: Test rig instruments: prime mover motor drive and CMS.

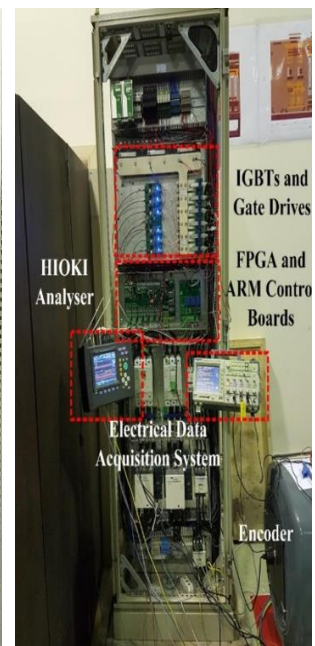



Fig. 11: Back-to-back converter.

Fig. 13 shows the condition monitoring hardware. The same system is also installed on the rig test. This system uses 8 accelerometer sensors, which are located according to Fig 14. 8 sensors are installed in the components of the gearbox (4), generator (2), and main bearing (2).

Table 3: The HIL test rig measuring devices

Hioki Power Quality Analyzer		WT CMS	
Input specifications			
Measurement line types	Single-phase 2-wire, Single-phase 3-wire, Three-phase 3- wire (3P3W2M, 3P3W3M), or Three-phase 4-wire, plus one extra input channel	Input channels	Time processing command variables up to a maximum of 4 channels (power, pitch angle, wind speed, generator speed)/ 8 acceleration sensors (Vibration channels)
Input channels	Voltage: 4 channels (U1 to U4) (channel U4 can be switched between AC and DC) Current: 4 channels (I1 to I4)		
Measurement specifications			
Current	calculated continuously every 10 or 12 cycles at 50 or 60 Hz respectively)	System architecture	
Measurement	2 MHz sampling	Measurement	32 kHz
Sensors	Clamp on Sensors: 1000 A AC, 1000 A continuous	Sensors	8 ICP acceleration sensors 10 mV/ms-2

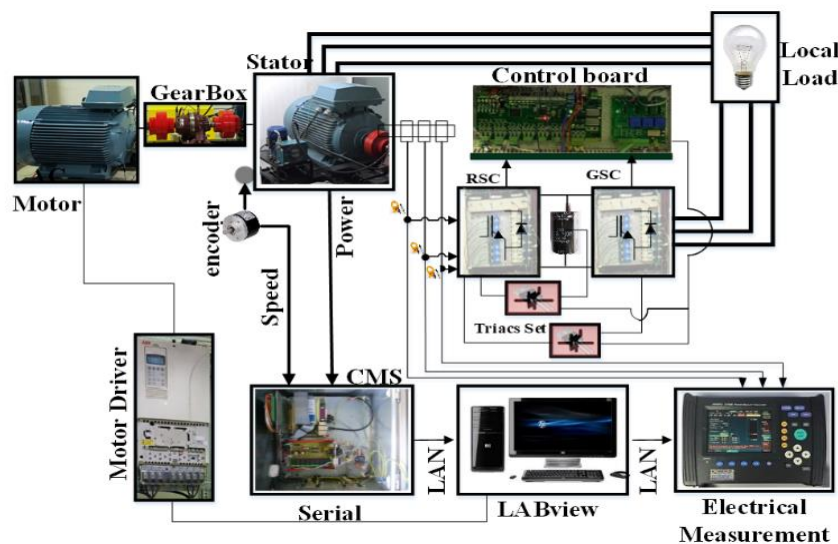


Fig. 12: The data collection system and HIL setup communication.

The system uses time-frequency and frequency domain analysis and statistical analysis performed inside the hardware to process the received signals. This product has proprietary software for hardware settings. This product has other software for statistical and graphical analysis. This system was launched about two years ago with the participation of the research author on the MAPNA Kahak site, and now the data collection of vibration signals from WTs on the Kahak site is underway. Also, this system is launched in the MAPNA WT laboratory and is installed on the WT test rig. The hardware of this system is a PC that operates upstream and has a DSP unit in which time and frequency domain functions are implemented. This system calculates the functions of maximum value, minimum value, crest, kurtosis, mean, and histogram from the time domain. Also, in the frequency domain, FFT is the main function and its output in the software can be seen at different time intervals.



Fig. 13: CMS hardware.

Fault Representation

A. Rotor Electrical Asymmetries (REA)

This fault is caused by an increase in the resistance that is series connected to the rotor winding [18]. Each of the winding and brush sections is modeled with a resistor, and the added resistor forms three series resistors for each phase.



Fig. 14: Location of vibration sensors of MAPNA WT condition monitoring system.

The corresponding schematic is shown in Fig. 9. It is also shown in the figure of total resistors [19]. Resistors R_{ea} , R_{eb} and R_{ec} are the same external resistors added, when the balance is disturbed in three-phase when one of the resistors is larger than the other two resistors. At balances, (1) is established, and (2) indicates an imbalance:

$$R_{ec} = R_{eb} = R_{ea} \quad (1)$$

$$R_{ea} = R_{eb} + \delta R = R_{ec} + \delta R \quad (2)$$

where δR is the amount of resistance added due to the asymmetry fault. An asymmetry is created in the same way in the desired simulator. The electrical asymmetry of the rotor can be defined as follows:

$$\delta R = |R_{1af}e^{i\theta_1} + R_{2bh}e^{i\theta_2} + R_{3ch}e^{i\theta_3}| \quad (3)$$

Where $i = \sqrt{-1}$, $\theta_1 = ?$, $\theta_2 = \frac{2\pi}{3}$, $\theta_3 = \frac{4\pi}{3}$.

in (3), h indicates a healthy state and f is a faulty condition. On the other hand, the degree of electrical asymmetry is determined by the following equation:

$$\Delta R(\%) = \frac{\delta R}{R_{1ah}} \times 100 = \frac{\delta R}{R_{2bh}} \times 100 = \frac{\delta R}{R_{3ch}} \times 100 \quad (4)$$

in (4) with resistance changing at various levels, different faults can be emulated. Based on (4), this fault percentage is determined [6]-[8]. In [5], [19] the presented method has been compared with other methods.

B. Inter-Turn Short Circuit (ITSC)

The ITSC fault causes damage to the stator winding insulation in the generator in the area between the coils. Condition monitoring and FD methods in induction devices are based on current analysis. In the stator

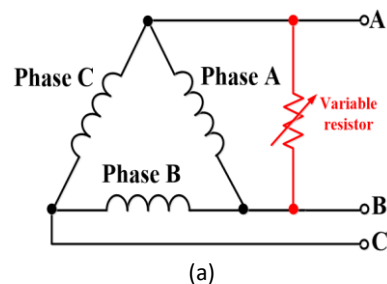
current signal, (5) shows the fault frequency component in the corresponding spectrum.

$$f_{st} = f_1 \left[\frac{n}{p} (1 - s) \pm k \right] \quad (5)$$

where f_{st} is the frequency component, f_1 is the supply frequency, $n = 1, 2, 3, \dots$, $k = 1, 3, 5$, and p is the pole pair and frictional slip is denoted by s . According to this equation, the frequency of the fault depends on the slip. In transient conditions, s is changing, therefore, it is difficult to detect a fault in the s variable. When there is an imbalance on the DFIG stator side, a negative harmonic component at the $-f_1$ of the stator produces a magnetic field. This magnetic field produces a harmonic component at the frequency $(2-s)f_1$. If in the frequency domain this fault is checked in the rotor current signal, the symptom of this fault component is the frequency $(2K \pm s)f_1$. The stator winding has a constant amplitude voltage. This voltage is connected directly to the grid. This voltage has a frequency equal to f_1 . The DFIG rotor winding is connected to the generator via a back-to-back converter. After the stator ITSC fault occurs, a regular periodicity occurs in the positive and negative directions, with the corresponding frequencies at f_1 and $-f_1$. Therefore, the grid voltage or stator voltage will create a new frequency component $2f_1$ in the active and reactive power. This frequency has an amplitude of more than twice the power control loop and is well reflected in the signal. Because the ITSC fault occurs due to insulation damage between the loops of a coil, it is modeled as a short circuit in the coil, which affects the current stability. This modeling is based on Fig. 15.a. For this purpose, it is assumed that in modeling, the impedance of the shorted stator winding decreases. The amount of this reduction is related to the severity of the fault. Hence, in the simulator, a variable resistor, similar to Fig. 15.b, is placed parallel to one of the legs. The number of short-circuit loops is determined by the changes in this resistance.

C. Converter IGBT Open-circuit

To open the switch circuit, the gate cut of each switch is used [34]. For each side of the converter, there are six switches similar to Fig. 16. The manual switches shown in Fig. 16 are a solution for testing gate IGBT Open-circuit fault. To implement the hardware structural fault-tolerant scenario, a TRIAC circuit is used; this board receives commands from the main control board of the converter (similar to Fig. 12).



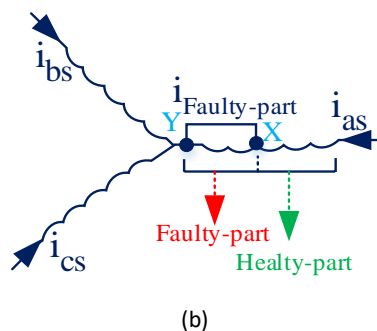


Fig. 15: Stator winding structure despite a fault in a winding phase.

Experimental Results and Discussion

A. Performance Analysis

To show the condition of the converter, Fig. 17 shows the thermal status of the switches and their switching condition. The general outputs of the HIL simulator, including the ratio of the generator speed diagram to the motor speed, and the generator speed diagram to the output power, are shown in Fig. 18 and Fig. 19, receptively. In Fig. 18, the gearbox conversion ratio will determine the generator speed. Generator speed is measured as one of the inputs of the CMS equipment channel.

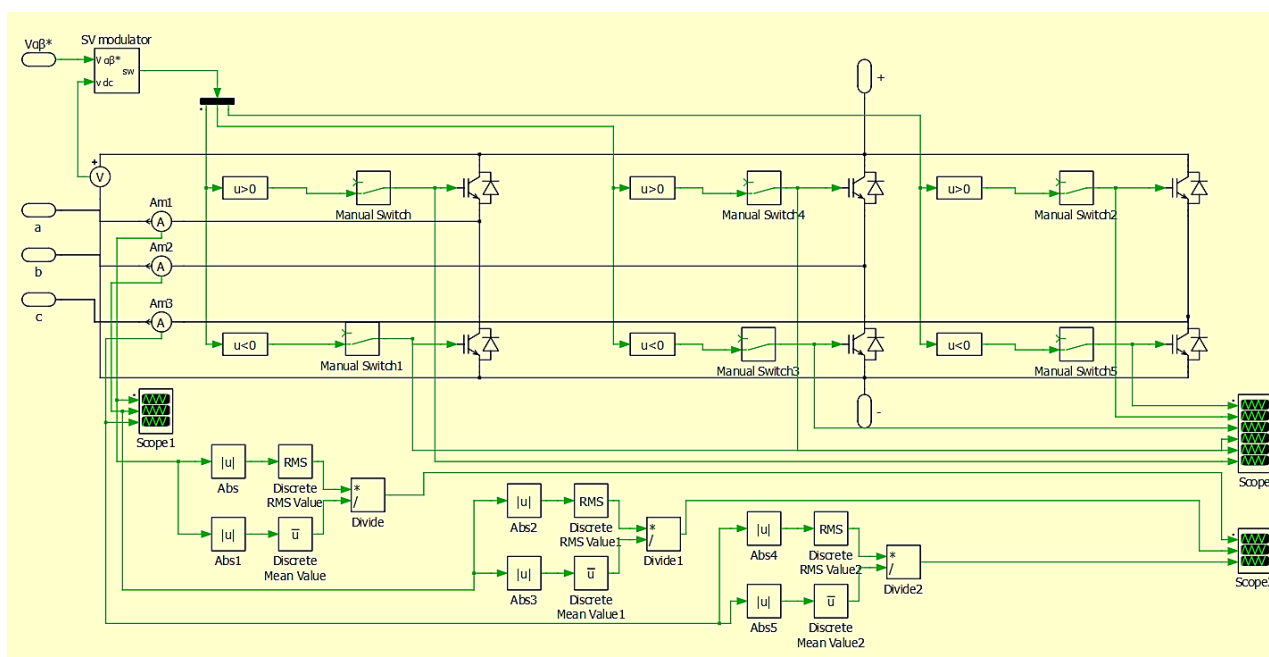


Fig. 16. schematic representation of converter IGBT open-circuit hardware implementation in WT test rig.

The motor speed is read from the drive and the generator speed is read from the CMS. Fig. 19 shows the WT HIL power generation performance. This diagram is based on the collected information about the power and speed of the generator, which is measured through the CMS.

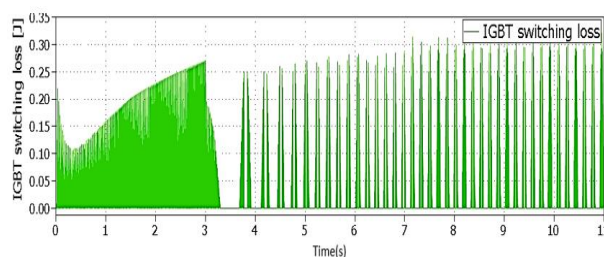
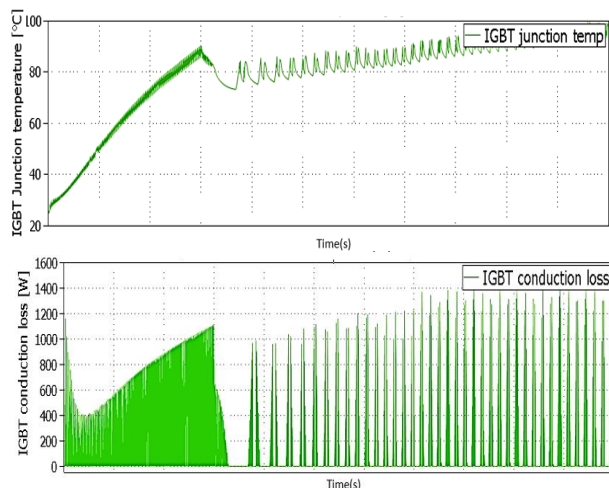


Fig. 17: IGBT status output in the back-to-back converter.

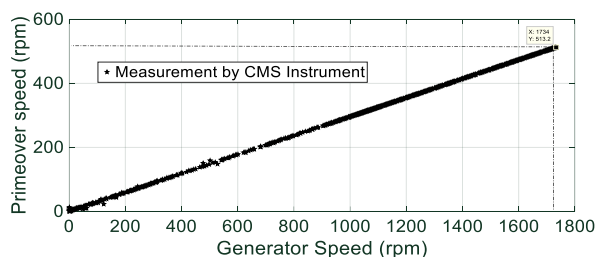


Fig. 18: Diagram of the motor prim mover speed of the motor than the generator speed.

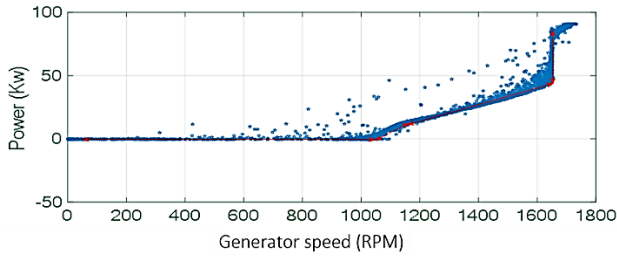


Fig. 19: WT HIL simulator power generation diagram based on generator speed.

B. Rotor Fault

Rotor asymmetries mean a fault in the rotor winding is simulated, which includes the effect of a fault in the winding, a brush imbalance, or an air gap imbalance on the rotor of a DFIG generator. Different levels of asymmetries are analyzed to measure the effect of the fault. These levels are shown in Table 4. The test is performed at different speeds from the WT HIL operating areas. The generator speed is from 1220 to 1600. This generator speed will be extracted through the wind speed pattern generated by the drive motor. To implement the experiments, some parameters must be set, for the power signal ω_c is equal to $2sf$ and for the stator current signal this value is equal to $((1-2s)f)$.

The time-domain signals are extracted in this experiment based on Fig. 20. A. From this figure, due to generator variable speed, It is clear that the fault symptom cannot be detected from the generator current (Fig. 20.a) or the power output (Fig. 20.d) of the generator. Generator rotation speed (Fig. 20.b), mechanical torque (Fig. 20.c), single-phase stator current (Fig. 20.a) and total power signal (Fig. 20.d) are measured from the simulator. The stator current frequency $((1-2s)f)$ at faulty condition (Fig. 20.e) and the fault frequency $(2sf)$ (Fig. 20. f) for the power output are extracted. The variation of power and current signals is like the behavior of the generator rotation speed signal. Degradation of the generator fraction slip causes a change in the corresponding fault frequency.

Table 4: Rotor electrical asymmetries applied to DFIG generator

Time (s)	Condition	$\Delta R(\%)$	Rotor resistance (m Ω)
0-50	Normal	0	16.6
50-150	Low asymmetries	9	18.09
150-200	Normal	0	16.6
200-300	Medium asymmetries	20	19.92
300-350	Normal	0	16.6
350-450	High asymmetries	46	24.23
450-500	Normal	0	16.6

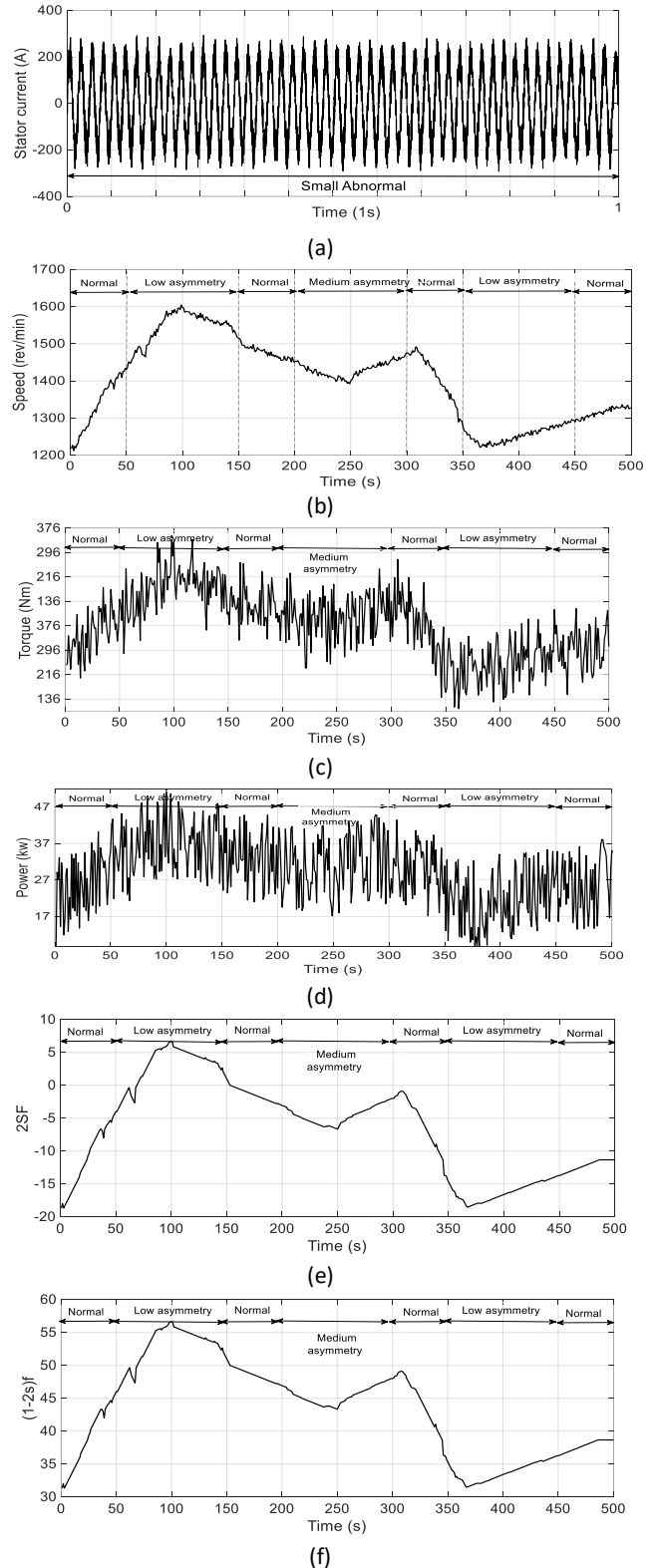


Fig. 20: The experimental result from WT HIL in the rotor fault condition.

C. Stator Fault

Table 5 shows the phase difference between the three stator phases for different fault values in the phase. For example, the failure rate in phase A increases the difference between phases A and B. While the phase

difference between the two healthy phases is very little (close to 120 degrees). Table 6 shows the complete harmonic distortion change. In this table, it is clear that the fault percentage increases with THD, but in the other two steps, the fault percentage changes are small and close to THD in a healthy state.

Table 5: Phase difference between the three phases of the stator in the presence of the fault

Fault Percentage	AB	BC	CA
Normal	120.012	1120.063	119.358
5	122.747	120.087	116.996
10	125.536	120.997	113.467
25	135.096	119.667	105.237
40	142.198	119.558	98.244

Table 6: THD changes in different fault percentages

Fault Percentage	C (%)	B (%)	A (%)
Normal	3.20	3.54	3.39
5	3.22	3.57	3.62
10	3.18	3.65	3.92
25	3.23	3.52	5.59
40	3.39	3.48	7.98

D. Converter Fault

By disconnecting the gate signal from the gate drive and removing the signal related to the gate, an open-circuit defect is created. Fig. 21 shows the experimental results of generator rotor currents. First, by disconnecting the driving signals of the IGBT gates S1 and S4 in 2.441 and 2.521 seconds, respectively, an open circuit fault occurs in one phase. With the IGBT open circuit fault, the current in the phases is damaged and distorted. This increases heat loss and vibration in the stator shaft. These behaviors and problems are exacerbated by the second open circuit fault. Fig. 22 shows the results obtained for the open circuit fault of the two IGBT switches S3 and S2 in 4.322 and 4.486 seconds on the grid side, respectively.

In [3], [19], [35], the presented method has been compared with other methods.

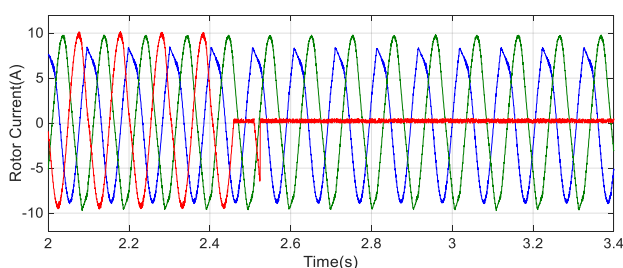


Fig. 21: Rotor current in presence converter fault.

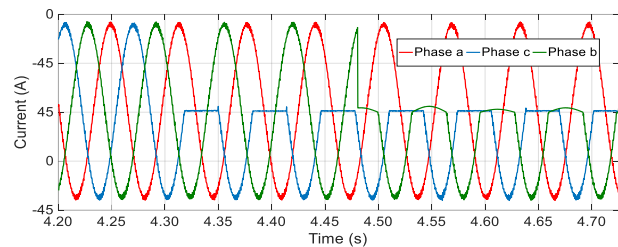


Fig. 22: Stator current in presence converter fault.

To provide more examples for verification and review of the presented results, you can refer to [36], [37].

More details of the implemented method, including relations, simulations, implementations, results, and their review are presented in [3]-[5], [19], [35]-[37].

Conclusion

In this paper, a hardware simulator for WTs is presented. This hardware in the loop is designed and launched in the research and development laboratory of MAPNA Group. The need to develop a simulator to investigate faults in WTs and their monitoring methods are first considered, an analysis of the types and percentages of faults in WTs is presented, the test rig structure is described and its subsystems are investigated. Fault generation and scenario-building solutions in the test bench are analyzed for the rotor, stator, and converter faults, and finally, an analysis of the performance of the test rig in the presence of a fault and healthy condition is presented. This structure can be used to develop new fault diagnosis software, fault prediction, and fault-tolerant control methods. Testing on real WTs is very expensive, but using this low-cost setup is effective to develop and improve the field of condition monitoring and WT control.

Abbreviations

CBM	: Condition Based Maintenance
DFIG	: Double Feed Induction Generator
FD	: Fault Detection
FTC	: Fault-Tolerant Control
HIL	: Hardware In the Loop
ITSC	: Inter-Turn Short Circuit
REA	: Rotor Electrical Asymmetry
WTs	: Wind Turbines

Acknowledgment

This article was prepared with the support of MAPNA Electric & Control, Engineering & Manufacturing Co. (MECO). Our MAPNA Group colleagues at various companies have assisted in this research. Thanks to them, of course, does not mean that they agree with all the results of the article and the authors are responsible for the results of the article.

Author Contributions

M. Kamarzarrin, M.H. Refan, P. Amiri, A. Dameshghi had role in Conception and Design. M. Kamarzarrin, A.

Dameshghi did Data Acquisition, Analysis and interpretation of data, Manuscript Drafting, and Statistical Analysis. M.H. Refan, P. Amiri, and A. Dameshghi had a role in Administrative, Technical, or Material Support. Obtaining Funding was with M.H. Refan.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

References

- [1] M. Kamarzarrin, M. H. Refan, "Intelligent sliding mode adaptive controller design for wind turbine pitch control system using PSO-SVM in presence of disturbance," *J. Control Autom. Electr. Syst.*, 31(4): 912-925, 2020.
- [2] M. H. Refan, A. Dameshghi, "A new strategy for short-term power-curve prediction of wind turbine based on PSO-LS-WSVM," *Iran. J. Electr. Electron. Eng.*, 4: 392-403, 2018.
- [3] m. kamarzarrin, m. h. refan, p. amiri, a. dameshghi, "Fault diagnosis of wind turbine double-fed induction generator based on multi-level fusion and measurement of back-to-back converter current signal," *Iran. J. Electr. Electron. Eng.*, 18(2): 2074-2074, 2022.
- [4] A. Dameshghi, M. H. Refan, "Wind turbine gearbox condition monitoring and fault diagnosis based on multi-sensor information fusion of SCADA and DSER-PSO-WRVM method," *Int. J. Model. Simul.*, 39: 48-72, 2019.
- [5] A. Dameshghi, M. H. Refan, "Combination of condition monitoring and prognosis systems based on the current measurement and PSO-LS-SVM method for wind turbine DFIGs with rotor electrical asymmetry," *Energy Syst.*, 2: 1-30, 2021.
- [6] J. M. P. Pérez, F. P. G. Márquez, A. Tobias, M. Papaalias, "Wind turbine reliability analysis," *Renewable Sustainable Energy Rev.*, 23: 463-472, 2013.
- [7] T. A. Kawady, A. A. Afify, A. M. Osheiba, A. I. Taalab, "Modeling and experimental investigation of stator winding faults in induction motors," *Electr. Power Compon. Syst.*, 37: 599-611, 2009.
- [8] D. Zappalá, N. Sarma, S. Djurović, C. J. Crabtree, A. Mohammad, P. J. Tavner, "Electrical & mechanical diagnostic indicators of wind turbine induction generator rotor faults," *Renewable energy*, 131: 14-24, 2019.
- [9] M. Wilkinson, B. Darnell, T. Van Delft, K. Harman, "Comparison of methods for wind turbine condition monitoring with SCADA data," *IET Renewable Power Gener.*, 8: 390-397, 2014.
- [10] F. Haces-Fernandez, H. Li, D. Ramirez, "Improving wind farm power output through deactivating selected wind turbines," *Energy Convers. Manage.*, 187: 407-422, 2019.
- [11] K. B. Abdusamad, Condition Monitoring System of Wind Turbine Generators. PHD thesis, University of Durham, 2014.
- [12] K. Alewine, W. Chen, "A review of electrical winding failures in wind turbine generators," *IEEE Electr. Insul. Mag.*, 28:8-13, 2012.
- [13] J. M. P. Pérez, F. P. G. Márquez, A. Tobias, M. Papaalias, "Wind turbine reliability analysis," *Renewable Sustainable Energy Rev.*, 23: 463-472, 2013.
- [14] L. Alhmod, B. Wang, "A review of the state-of-the-art in wind-energy reliability analysis," *Renewable Sustainable Energy Rev.*, 81: 1643-1651, 2018.
- [15] C.J. Crabtree, Condition Monitoring Techniques for Wind Turbines. PHD Thesis, Durham University, 2011.
- [16] W. Hu, "Reliability-based design optimization of wind turbine systems," *Advanced Wind Turbine Technology*, Springer, Cham, 1-45, 2018.
- [17] M. Yadav, P. Prakash, R. C. Jha, "Reliability analysis and energy benefit analysis of distribution system incorporating wind turbine generator," in *Proc. the International Conference on Nano-electronics, Circuits & Communication Systems*: 265-273, 2017.
- [18] M. Kamarzarrin, M. H. Refan, P. Amiri, A. Dameshghi, "A new intelligent fault diagnosis and prognosis method for wind turbine doubly-fed induction generator," *Wind Eng.*, 10: 210-221, 2021.
- [19] A. Dameshghi, M. H. Refan, P. Amiri, "Wind turbine doubly fed induction generator rotor electrical asymmetry detection based on an adaptive least mean squares filtering of wavelet transform," *Wind Eng.*, 8: 1-21, 2019.
- [20] M. Kamarzarrin, M. H. Refan, P. Amiri, A. Dameshghi, "A new fault-tolerant control of wind turbine pitch system based on ANN model and robust and optimal development of MRAC method," *Tabriz Electr. Eng. J.*, 51: 83-95, 2021.
- [21] W. T. Thomson, M. Fenger, "Current signature analysis to detect induction motor faults," *IEEE Ind. Appl. Mag.*, 7: 26-34, 2001.
- [22] C. J. Crabtree, S. Djurović, P. J. Tavner, A. C. Smith, "Fault frequency tracking during transient operation of wind turbine generators," in *Proc. XIX International Conference on Electrical Machines, ICEM 2010*, Rome, Italy, 2010.
- [23] S. Djurovic, C. J. Crabtree, P. J. Tavner, A. C. Smith, "Condition monitoring of wind turbine induction generators with rotor electrical asymmetry," *IET Renew. Power Gen.*, 6: 207-216, 2012.
- [24] X. Gong, Online Nonintrusive Condition Monitoring and Fault Detection for Wind Turbines. PHD Thesis, University of Nebraska, 2012.
- [25] M. Yousefi kia, M. Khedri, H. R. Najafi, M. A. Shamsi Nejad, "Hybrid modeling of doubly fed induction generators with inter-turn stator fault and its detection method using wavelet analysis," *IET Gener. Transm. Distrib.*, 7: 982-990, 2013.
- [26] R. Roshanfekar, A. Jalilian, "Analysis of rotor and stator winding inter-turn faults in WRIM using simulated MEC model and experimental results," *Electr. Power Syst. Res.*, 119: 418-424, 2015.
- [27] R. Roshanfekar, A. Jalilian, "Wavelet-based index to discriminate between minor inter-turn short-circuit and resistive asymmetrical faults in stator windings of doubly fed induction generators: a simulation study," *IET Gener. Transm. Distrib.*, 10: 1-8, 2016.
- [28] P. V. Rodríguez, A. Arkio, "Detection of the stator winding fault in induction motor using fuzzy logic," *Appl. Soft Comput.*, 8: 1112-1120, 2008.
- [29] R. Sharifi, M. Ebrahimi, "Detection of the stator winding faults in induction motors using three-phase current monitoring," *ISA Trans.*, 50: 14-20, 2011.
- [30] A. Bechkaouia, A. Ameurb, S. Bourasc, A. Hadjadjd, "Open-circuit and inter-turn short-circuit detection in PMSG for wind turbine applications using fuzzy logic," *Energy Procedia*, 74: 1323-1336, 2015.
- [31] M. S. Ballal, Z. J. Khan, H. M. Suryawanshi, R. L. Sonolikar, "Adaptive neural fuzzy inference system for the detection of inter-turn insulation and bearing wear faults in induction motor," *IEEE Trans. Ind. Electron.*, 54: 250-258, 2007.
- [32] A. Djerdira, S. S. Moosavia, Y. Ait-Amiratb, D. A. Khaburica, "ANN based fault diagnosis of permanent magnet synchronous motor under stator winding shorted turn," *Electr. Power Syst. Res.*, 125: 67-82, 2015.
- [33] R. M. Tallam, S. B. Lee, G. C. Stone, "A survey of methods for detection of stator-related faults in induction machines," *IEEE Trans. Ind. Appl.*, 43: 920-933, 2007.
- [34] A. Dameshghi, M. H. Refan, "The fault diagnosis of open-circuit of back-to-back converters in dfig wind turbines of variable speed using combination signal-based and model-based methods," *J. Energy Manage.*, 9: 18-33, 2019.

- [35] M. Kamarzarrin, M. H. Refan, P. Amiri, "Open-circuit faults diagnosis and Fault-Tolerant Control scheme based on Sliding-Mode Observer for DFIG back-to-back converters: Wind turbine applications," *Control Eng. Prac.*, 126: 105235, 2023.
- [36] A. Dameshghi, "Wind turbine generator and converter maintenance based on smart monitoring based on data fusion strategy", Ph.D. thesis, Shahid Rajaee Teacher Training University, 2019.
- [37] M. Kamarzarrin, "Enhanced Fault Tolerant of Dual Open-circuit Faults with Back-to-Back converter of DFIG Wind Turbine", Ph.D. thesis, Shahid Rajaee Teacher Training University, 2022.

Biographies



Mehrnoosh Kamarzarrin was born in 1991 and is receiving currently her B.S. degree (with the highest honors) in Electronic Engineering from Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran, in 2013. She received her M.Sc. in Control Engineering from Shahid Beheshti University, Tehran, Iran in 2015. Now she is a Ph.D. in Electronic Engineering at Shahid Rajaee Teacher Training University. Her research interests include GPS, wind turbine, fault detection & tolerant control, Adaptive control, wireless communications, and networking with a focus on cognitive radios, Analog electronics, and Boolean Function. Now she is the wind turbine process expert in MAPNA Electric & Control, Engineering & Manufacturing Co. (MECO).

- Email: kamarzarrin.mehrnoosh@sru.ac.ir
- ORCID: [0000-0003-2292-3861](https://orcid.org/0000-0003-2292-3861)
- Web of Science Researcher ID: NA
- Scopus Author ID: 55975874300
- Homepage: <https://ir.linkedin.com/in/mehrnoosh-kamarzarrin-aa353b58>



Mohammad Hossein Refan received his B.Sc. in Electronics Engineering from the Iran University of Science and Technology, Tehran, Iran in 1972. After 12 years of working and experience in the industry, he started studying again in 1989 and received his M.Sc. and Ph.D. in the same field and at the same University in 1992 and 1999 respectively. He is currently a Professor of Electrical and Computer Engineering Faculty, Shahid Rajaee Teacher Training University, Tehran, Iran. He is the author of about 50 scientific publications in journals and international conferences. His research interests include GPS, DCS, and Automation systems, wind turbines, fault detection & tolerant control, and Adaptive control.

Email: Refan@sru.ac.ir

- ORCID: [0000-0001-5266-0586](https://orcid.org/0000-0001-5266-0586)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/mohammad-hossein-refan/>



Parviz Amiri was born in 1970. He received his B.Sc. degree from the University of Mazandaran in 1994, his M.Sc. from Khajeh Nasir Toosi University (KNTU Tehran, Iran) in 1997, and his Ph.D. from Tarbiat Modares University (TMU, Tehran, Iran) in 2010, all degrees in Electrical Engineering (Electronics). His main research interest includes electronic circuit design in industries. His primary research interest is in RF and power electronic circuits, with a focus on highly efficient and high linear power circuit design. He is currently an Associate Professor of Electrical and Computer Engineering Faculty, Shahid Rajaee Teacher Training University, Tehran, Iran.

- Email: pamiri@sru.ac.ir
- ORCID: [0000-0001-5764-0912](https://orcid.org/0000-0001-5764-0912)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://www.sru.ac.ir/en/faculty/school-of-electrical-engineering/parviz-amiri/>



Adel Dameshghi was born in 1986 and received his B.S., M.S. Ph.D. degrees in Electronic Engineering from the Department of Electrical Engineering, of Electrical and Computer Engineering, Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran, in 2011, 2013 and 2020 respectively. His research interests include Boolean Function, GPS, wind turbine, fault detection & tolerant control, Adaptive control, Electric, and Hybrid vehicles and now he is the manager of the EV & infrastructure Development Center of MAPNA Electric & Control, Engineering & Manufacturing Co. (MECO).

- Email: a.dameshghi@sru.ac.ir
- ORCID: [0000-0001-8764-4287](https://orcid.org/0000-0001-8764-4287)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://ir.linkedin.com/in/adel-dameshghi-20617941>

How to cite this paper:

M. Kamarzarrin, M. H. Refan, P. Amiri, A. Dameshghi, "Development of wind turbine fault analysis setup based on dfig hardware in the loop simulator," *J. Electr. Comput. Eng. Innovations*, 11(2): 277-290, 2023.

DOI: [10.22061/jecei.2022.8849.556](https://doi.org/10.22061/jecei.2022.8849.556)

URL: https://jecei.sru.ac.ir/article_1809.html





Research paper

Design Optimization of the Delta-Shape Interior Permanent Magnet Synchronous Motor for Electric Vehicle Application

S. Nasr¹, B. Ganji^{1,*}, M. Moallem²

¹Faculty of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

²Faculty of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran.

Article Info

Article History:

Received 25 July 2022

Reviewed 16 October 2022

Revised 28 October 2022

Accepted 07 November 2022

Keywords:

Interior permanent magnet synchronous motor

Electromagnetic modeling

Design optimization

Torque ripple reduction

Finite element method

*Corresponding Author's Email
Address: bganji@kashanu.ac.ir

Abstract

Background and Objectives: Due to exclusive advantages of the permanent magnet synchronous motors (PMSMs) such as large torque/power density, high efficiency and wide speed range in constant power region, special attention has been paid to these motors especially for electric vehicle (EV) application. A conventional type of PMSMs which is more suitable for EV application is the interior permanent magnet synchronous motors (IPMSM). The main objective of the present paper is design optimization of this type of PMSM to increase efficiency and reduce torque ripple which are important for EV application.

Methods: Using different shape design optimization methods including rotor notch, flux barrier and skewed rotor, design optimization of the delta-shape IPMSM is done and an optimized design is suggested first. One of the most important factors affecting the performance of the IPMSM is the magnet arrangement in the rotor structure. Based on the design of experiments (DOE) algorithm, optimal values of some design parameters related to magnet are then determined to improve more the motor performance of the suggested structure.

Results: The simulation results based on finite element method (FEM) are provided for a typical high-power IPMSM to evaluate the effectiveness of the proposed technique. In comparison to the initial design, 7% increase of average torque, 50% reduction of torque ripple and 1.4% increase of efficiency are resulted for the optimized motor.

Conclusion: Using the proposed hybrid design optimization procedure (shape design optimization with optimum design parameters), significant improvement of some characteristics related to the delta-shape IPMSM including efficiency, average torque and torque ripple is resulted and this conclusion is desirable for EV application.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

In order to solve the problems related to the conventional vehicles such as environment and noise pollutions, special attention has been paid in recent decades to use of the EVs. In addition, the cost of charging an EV is less than the petrol or diesel cars and this can be considered as energy saving due to the limitations of fossil fuels. One of the essential requirements of EVs is to select an appropriate

electric motor. In comparison to other types of electric motors, the PMSMs have larger torque/power density due to the presence of magnetic in the motor structure. In addition, efficiency and saving of battery capacity are high in these motors. Therefore, this type of the motor is an appropriate candidate for EV applications [1]-[5]. Nowadays, these motors are used more effectively than the past because of advance permanent magnet (PM)

materials and modern control methods. There are different rotor topologies for the PMSMs and one of them is IPMSM in which the rare-earth PMs are used to achieve high performance.

This type of PMSM produces high power density due to its utilization of both reluctance torque and electromagnetic torque. In addition, it has a more robust structure that is desirable for EV application. In comparison to the surface-type PMSM in which PMs are located on the surface of rotor, the magnetic flux of the air-gap for the IPMSM contains more harmonics and consequently it has higher torque ripple. Therefore, multi-objective design optimization of the delta-shape IPMSM is considered in the present paper with the aim of reducing torque ripple and improving average torque and efficiency simultaneously which are essential for EV application.

Significant research has been conducted on different design aspects of the IPMSM. A design optimization method is introduced in [6] to minimize torque ripple of the IPMSM. The effect of a rotor skew on performance of IPMSM has been studied in [7]. The most important factors affecting the performance of IPMSM are the arrangement and shape of PMs used in the rotor structure. In addition, some design parameters such as pole-arc to pole-pitch ratio [8], flux barrier topology [9] and number of layers [10] can impact the performance significantly. The multi-layer IPMSM is introduced in [11] for EV application and it is showed that the three-layered motor has lower torque ripple and core loss than two-layered structure due to reduction of the harmonics of air-gap flux density. An IPMSM with ∇ +U shape rotor topology is introduced in [12] for EV applications and it is demonstrated that its efficiency is better than V and ∇ shape machines. The impact of PM topology on core loss and efficiency has been also evaluated in this reference. The shapes of the rotor notch and flux barriers could significantly change the electromagnetic characteristics of the IPMSM. In [13], torque ripple of IPMSM is reduced using putting notches on the rotor surface. Optimal design of the PMSM for hybrid electric vehicles (HEVs) is considered in [14] to maximize the energy efficiency. An analytical model is also developed in this reference to determine the geometrical parameters and predict quickly the efficiency.

When the analysis of motor is carried out with finite element method (FEM), use of the optimization algorithms such as the particle swarm optimization (PSO) and genetic algorithm (GA) are usually time-consuming methods. In this case, the DOE method can be used properly for the optimization due to the high computational speed [15]–[17]. Using the response surface method (RSM) and the Taguchi method, design

optimization of the IPMSM for electric compressors of air conditioners used in EVs is done in [18] to maximize efficiency and minimize cogging torque. Based on the combination of DOE method and Taguchi's method, an optimization procedure for a submersible PMSM is introduced in [19] to have the maximum efficiency and minimum cogging torque.

As indicated above, a multi-objective design optimization of the IPMSM is proposed here to reduce torque ripple while efficiency and average torque are improved. Hence, the main contributions of the paper can be summarized as developing a multi-objective optimization of an IPMSM by using shape design, type of winding and DOE technique for optimum magnet design. The rest of the paper is organized as follows: The proposed design optimization method is described clearly in next section.

To evaluate the effectiveness of this design optimization method, it is applied to a typical IPMSM suitable for EV application and related simulation results are given in the third section of the paper. Finally, the last section highlights the main contributions and conclusions of the paper.

Design Optimization Methods

Various shape design optimization methods have been already introduced in the literatures by which performance IPMSM can be improved. Some of them are described briefly in the following. Using the Hairpin winding and the DOE method, a new design optimization procedure is also proposed and it is introduced in this section.

A. Shape Design Optimization Methods

This paper focuses on the shape design optimization of the delta-shape IPMSM for high-speed traction application and comparison of different rotor topologies is considered.

As shown in Fig. 1, four different rotor structures with the same stator are designed. Use of notch on rotor is a conventional method for reducing torque ripple of PMSMs as done in [20]. According to the position of the notch on the rotor, different topologies have been chosen for the IPMSM in this reference and an average value of 37% was illustrated for torque ripple reduction. When a notch is located on the rotor surface, it leads to stepping the air-gap and consequently the cogging torque and torque ripple could be reduced.

This approach can be also suggested for the delta-shape IPMSM as observed from comparison between Fig. 1a and Fig. 1b.

Flux barrier in motor structure is also an effective method for average torque improvement and torque ripple reduction [9].

To improve torque waveform of the IPMSM, different

symmetric flux barrier shapes are introduced in [21]. The Taguchi method is also used to optimize the design parameters. Compared to the initial design, torque ripple is decreased by 50% and average torque is increased by 8.2% for the introduced model.

Also, the shape and number of barriers can affect the value of cogging torque and torque ripple [22]. Due to reduction of magnetic flux-leakage, flux barrier can also improve the produced power.

Fig. 1c shows how the flux barrier is used for the discussed IPMSM. When the barrier is included in the structure of the motor, magnetic flux in the air-gap is increased and therefore the average torque can be improved. Moreover, the cogging torque/torque ripple can be also reduced using the skewed stator/rotor as done in [23]-[25]. The impact of different rotor skew patterns on torque ripple and average torque for an IPMSM is evaluated in [7]. Significant reduction of torque ripple (about 68%) is resulted in this research while average torque is also reduced a little (2%). The manner of skewing the rotor for the discussed delta-shape IPMSM is shown in Fig. 1d. As illustrated from this figure, 4 layers are considered here that there is a specific displacement between them.

B. The Proposed Design Optimization Method

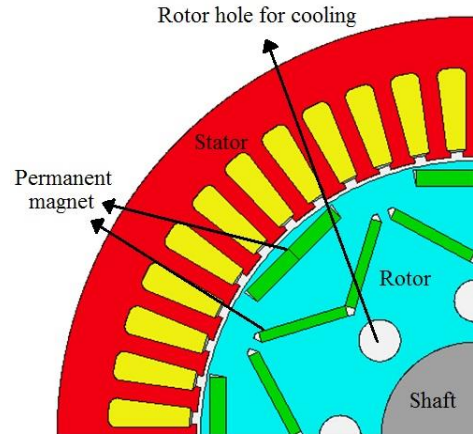
The various shape design optimization methods described above are applied to a typical delta-shape IPMSM and their impacts on the motor performance (average torque, torque ripple and efficiency) are evaluated. Then, the most effective method is selected. For this selection, two other changes are also considered in design of motor to improve more its performance. This hybrid approach is defined here as the proposed design optimization method. The two above-mentioned changes are use of the Hairpin winding and determination of the optimal values of some important design parameters explained in the following.

1. The Hairpin winding

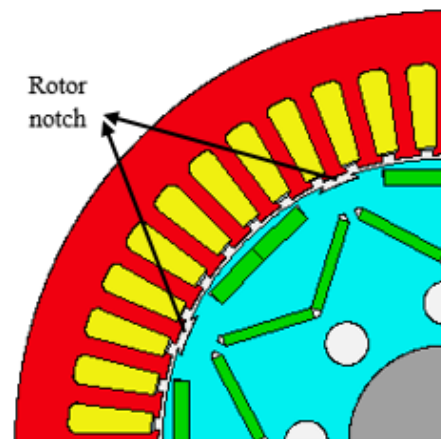
Similar to other types of electric motors, thermal issue of the IPMSM needs to be considered in design optimization of the motor especially for traction application due to the high-speed operation. It must be noted that energy saving achievement and cooling system volume reduction are resulted by improvement of the efficiency and reduction of power losses in electrical vehicle. The Hairpin winding has been introduced in [26] to improve the efficiency through the reduction of copper losses. It can also decrease the spatial harmonics because variation of the air-gap flux is lower. These variations lead to reduction of torque ripple, iron loss and vibration/noise [27].

The Hairpin winding compared with the stranded winding can also result in a higher fill factor up to 0.75.

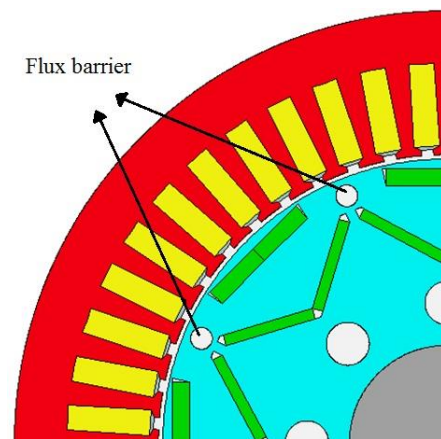
Using the Hairpin winding is desirable when high efficiency/power density is required [28]. Therefore, the hairpin winding is considered here to improve the performance of the delta-shape IPMSM. Fig. 2 shows obviously the difference between the stranded winding and the Hairpin winding.



(a)



(b)



(c)

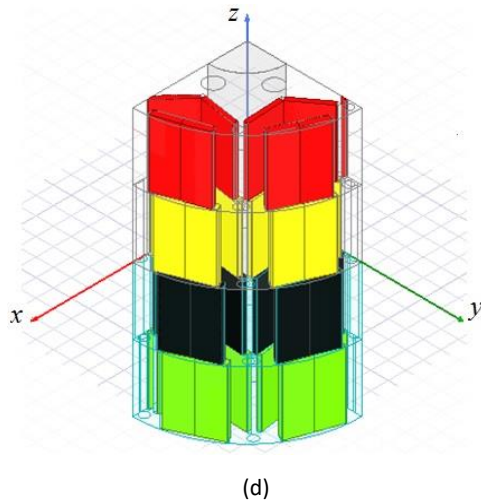


Fig. 1: Different topologies considered for the discussed IPMSM: (a) conventional structure, (b) using notch in the rotor, (c) using flux barrier in the rotor and (d) skewed rotor.

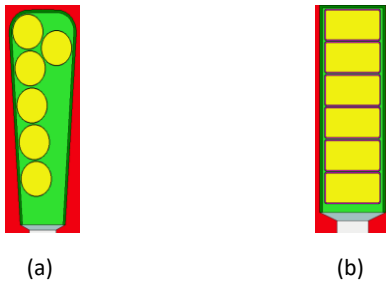


Fig. 2: Different windings: (a) stranded winding, (b) Hairpin winding.

II. Optimum design parameters

Three important design parameters for PM including the magnet thickness, the magnet bar width, and the pole V angle in two layers depicted in Fig. 3 are selected here and their optimal values are determined to have the best performance. Based on the DOE method, this optimization is done and the Minitab software is used for implementation of this method. Since the design optimization of the delta-shape IPMSM (Fig. 1a) is considered here, it must be noted that the optimum values of six design parameters must be determined in the process of optimization. In order to find the optimum values, the Taguchi method is employed here as one of the best DOE methodologies. Since a minimum number of experiments are required for the Taguchi method, it is an appropriate optimization method when the optimal design of the motor is done using FEM. In recent years, this method has been applied to the design optimization of electric motors such as PMSMs [17]. In this method, the design factors are selected first and every factor takes its value. The orthogonal table is established and the experiments are then designed to obtain influence of

factors and their different levels on optimal output. Finally, the mean value and signal-to-noise ratio (SNR) is utilized to obtain the best combination of levels [29]. As shown in Table 1, three different levels are considered for every parameter. The ranges of variables for each factor are shown in this table. It must be noted that these ranges are derived using the Nissan Leaf IPM motor described in [30].

According to the number of factors and the levels of each factor, the number of required experiments and how to combine the levels of factors in each experiment are specified using the Taguchi orthogonal arrays. The obtained results are summarized in Table 2. For three selected levels, the number of possible combinations is 729 (36). As seen obviously from this table, the Taguchi method has significantly reduced the number of tests required from 729 to 27 tests. Due to the long time of the simulations done with FEM, this reduction saves significant time to achieve the optimal point. The Taguchi method determines the optimal point according to the results of this limited number of experiments and using statistical calculations. For the experiments listed in Table 2, analysis of the motor with FEM should be carried out to calculate the average torque, torque ripple and efficiency.

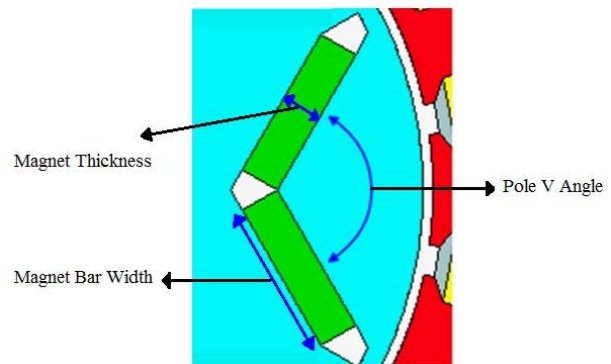


Fig. 3: Design parameters selected for main orthogonal array.

Table 1: Controllable factors and values

	Level 1	Level 2	Level 3
A = pole v angle 1 [°]	124	150	180
B = pole v angle 2 [°]	118	124	130
C = magnet bar width 1 [mm]	13.5	13.9	14.3
D = magnet bar width 2 [mm]	20.6	21.3	21.5
E = magnet thickness 1 [mm]	3.4	3.9	4.4
F = magnet thickness 2 [mm]	2.1	2.6	3.1

Using the experiments designed by the Taguchi method and after analyzing them, the optimal combination of levels for factors and values of average torque, torque ripple and efficiency at the optimal point are obtained. For instance, the mean value of the effect of level 2 of the factor B on average torque is calculated using (1).

Table 2: Orthogonal array

	A	B	C	D	E	F
1	1	1	1	1	1	1
2	1	1	1	1	2	2
3	1	1	1	1	3	3
4	1	2	2	2	1	1
5	1	2	2	2	2	2
6	1	2	2	2	3	3
7	1	3	3	3	1	1
8	1	3	3	3	2	2
9	1	3	3	3	3	3
10	2	1	2	3	1	2
11	2	1	2	3	2	3
12	2	1	2	3	3	1
13	2	2	3	1	1	2
14	2	2	3	1	2	3
15	2	2	3	1	3	1
16	2	3	1	2	1	2
17	2	3	1	2	2	3
18	2	3	1	2	3	1
19	3	1	3	2	1	3
20	3	1	3	2	2	1
21	3	1	3	2	3	2
22	3	2	1	3	1	3
23	3	2	1	3	2	1
24	3	2	1	3	3	2
25	3	3	2	1	1	3
26	3	3	2	1	2	1
27	3	3	2	1	3	2

According to Table 2, the average effect related to level 2 of the factor B is obtained from nine experiments including 4, 5, 6, 13, 14, 15, 22, 23 and 24 where the factor B is set on level 2. For the other factors, the mean value of the effect of the levels on the mean torque, torque

ripple and efficiency are obtained in a same way.

$$T_{B2} = (T_{avg}(4) + T_{avg}(5) + T_{avg}(6) + T_{avg}(13) + T_{avg}(14) + T_{avg}(15) + T_{avg}(22) + T_{avg}(23) + T_{avg}(24)) / 9 \quad (1)$$

The Taguchi experiments use the SNR to identify control factors that reduce variability. In general, the term of signal refers to the mean value of output and the term of noise indicates the undesirable value. Therefore, higher values of the SNR identify setting of the control factors that minimize the effects of the noise factors. In addition, analysis of variance (ANOVA) can be useful to determine the influence of any given input. The ANOVA analysis can also be utilized to demonstrate the mean response magnitudes of controllable process parameters. To perform the ANOVA, the sum of squares must be calculated. The goal of variance analysis is to optimize two or more factors simultaneously. The sum of squares of each of the factors (SSF) should be obtained in the first step of the variance analysis to choose the optimum levels.

Simulation Results

The proposed design optimization method is applied to a typical delta-shape IPMSM (Fig. 1a) whose specifications are given in Table 3 and simulation results are presented in this section.

Based on FEM using Maxwell software, analysis of this motor for speed of 4000 rpm and the maximum current 323.5 A is carried out and average torque, torque ripple and efficiency are obtained 175.8 Nm, 17.2% and 96.3 %, respectively.

Table 3: Motor specifications

Number of phases	3
Number of poles	8
Diameter of stator	198 mm
Diameter of rotor	130 mm
Air-gap	1 mm
Stack length	150 mm
Maximum speed	10000 RPM
RMS phase current	323.5 A
Rated torque	185 N.m
Permanent magnet material	N30UH
Core material	M250

A. The Results Related to the Shape Design Optimization Methods

Using the different shape design optimization methods

described in the second section of the paper, some design optimizations are done for the discussed IPMSM and simulation results obtained for the considered operating point are summarized in Table 4. Compared with the initial design (average torque=175.9 Nm, torque ripple=17.3% and efficiency=96.3 %), this table shows that the models 1 and 2 reduce both average torque and torque ripple.

To have the best performance, it must be explained that the number of rotor layers and the mechanical degree for the skewed rotor are selected 4 and 1.875°, respectively.

With regard to the values obtained for the initial design, it is illustrated from Table 4 that the model 2 has better performance (higher average torque, lower torque ripple and the same efficiency). Since these improvements are desirable for EV application, the model 2 is considered and its performance is improved more using the proposed design optimization method. The related simulation results are presented in next subsection.

Table 4: Comparison between the shape design optimization methods

	Average torque [Nm]	Torque ripple [%]	Efficiency [%]
Initial design	175.9	17.3	96.3
Model 1 (Initial design with notch)	173.5	13.7	96.4
Model 2 (Initial design with flux barrier)	179.9	11.5	96.3
Model 3 (Model 2 + skewed rotor)	173.5	9.7	96.2

B. Performance Improvement Using the Proposed Design Optimization Method

As indicated above, the model 2 defined in Table 4 is selected as an appropriate candidate and its performance is improved more using the design optimization method proposed in the second section of the paper. The simulation results related to this improvement are presented in the following.

1. Results related to the Hairpin winding

Using the Hairpin winding, both thermal design and efficiency of IPMSM can be improved as discussed at above.

Since these improvements are very important for EV application, the model 2 with this type of winding is

considered here and it is defined as model 4. With analysis of the model 4 for the considered operating point, the average torque, torque ripple and efficiency are obtained and they are 182.3 Nm, 8.5% and 97.5 %, respectively. Compared with the values for the model 2 (average torque=179.9 Nm, torque ripple=11.5% and efficiency=96.3 %), it is seen that all characteristics are improved using the model 4. It must be also indicated that slot area of the Hairpin winding for the model 4 is similar to that for the model 2.

As indicated above, the arrangement of the conductors for the Hairpin winding could increase the fill factor significantly. As an example, the fill factor is 0.4 for the stranded winding and 0.67 for the Hairpin winding when the areas of slots in the discussed IPMSM are 115.3 mm² for the two windings.

The efficiency maps of the two different models are shown in Fig. 4. The winding temperature and total losses including copper loss, core loss and magnet eddy current loss are predicted for the model 2 and the model 4 and they are shown in Fig. 5.

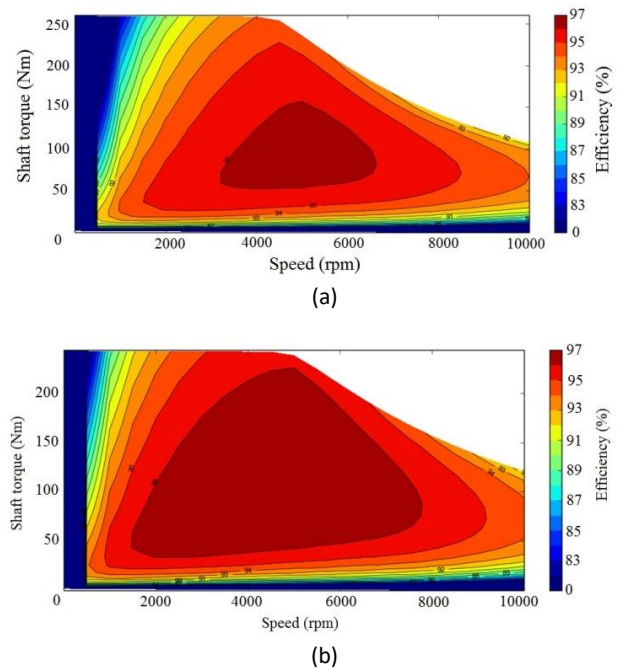
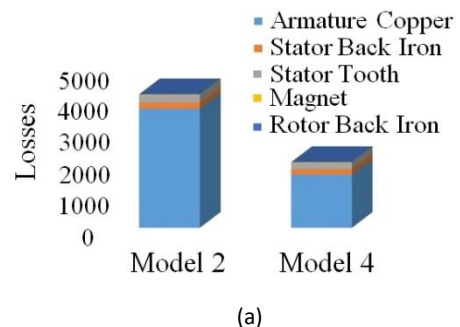


Fig. 4: The efficiency map of model 2: (a) the stranded winding, (b) the hairpin winding.



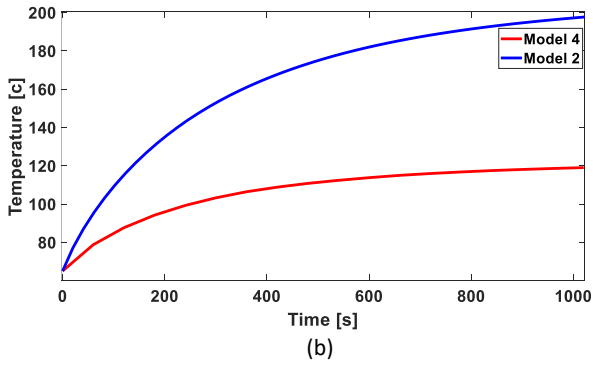


Fig. 5: The predicted winding temperature and total losses: (a) losses, (b) temperature.

II. Results with consideration of the optimum design parameters

Since the model 4 has showed the best level of performance, it is selected and its performance is improved more with determining the optimum parameters of PMs as indicated above. This optimized motor is called the model 5 and the related simulation results are presented here.

For the model 5, the mean of SNR related to average torque, torque ripple and efficiency are shown in Fig. 6. According to this figure, combination of A3-B3-C3-D3-E3-F1 leads to the highest average torque. In addition, this figure shows that combination of A1-B1-C3-D2-E3-F3 is related to the lowest torque ripple and combination of A3-B3-C3-D3-E3-F2 is for the highest efficiency. In the first stage, only the third level of the factor C and E can be selected.

Table 5 is used to find the best level for other factors. The sum of squares of the factors is calculated and it is also given in this table. To obtain the effect of factors, the value of each factor must be divided by the total. As illustrated from Table 5, the effect of factor A on average torque is 87.61%, on ripple torque is 48.82% and on efficiency is 86.1%. Therefore, level 3 is selected between the two levels 1 and 3 of factor A. Finally, it can be seen that the best optimization combination is A3-B1-C3-D2-E3-F3.

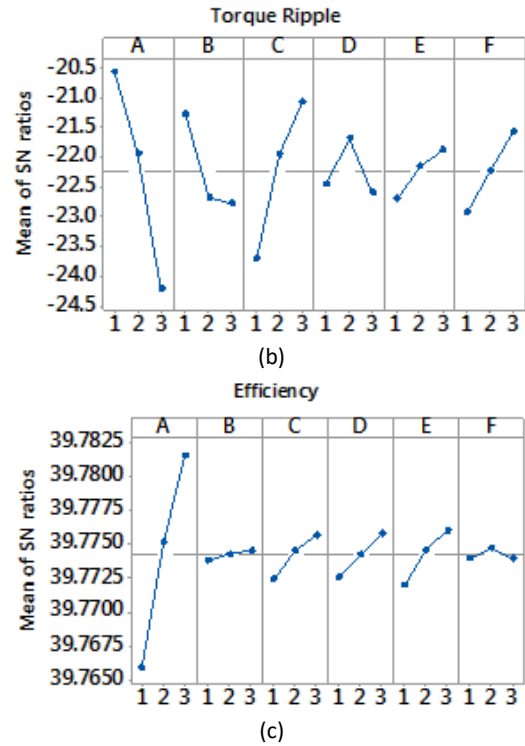
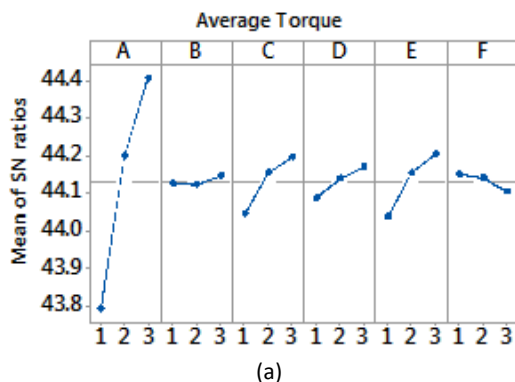


Fig. 6: The Taguchi SNR plot for: (a) average torque, (b) torque ripple, (c) efficiency.

Table 5: Impact of the selected design parameters on motor performance

	Average torque		Torque ripple		Efficiency	
	SSF	Effect [%]	SSF	Effect [%]	SSF	Effect [%]
A	198.5	86.1	55.1	49.2	0.047	86.1
B	0.3	0.1	13.4	11.9	0.0001	0.2
C	12.1	5.2	31	27.7	0.002	3.9
D	3.2	1.4	5.8	5.2	0.002	3.7
E	15	6.5	1.4	1.3	0.003	5.8
F	1.5	0.7	5.3	4.7	0.0002	0.3

Regarding the optimized IPMSM (the model 5), the average torque, torque ripple and efficiency for the considered operating point are 187.7 Nm, 6.7% and 97.6 %, respectively. Compared with the values related to the model 4 (average torque=182.3 Nm, torque ripple=8.5% and efficiency=97.5 %), it is illustrated that motor performance has been improved more using the model 5. This improvement is more evident when the optimized IPMSM (the model 5) is compared with the initial design (average torque=175.9 Nm, torque ripple=17.3% and

efficiency=96.3 %).

It must be noted that these improvements are very valuable for EV applications. The instantaneous torque waveforms predicted for two different designs (initial and the optimized motor) are also compared in Fig. 7. This figure shows obviously increase of average torque and reduction of torque ripple for the model 5 (optimal design).

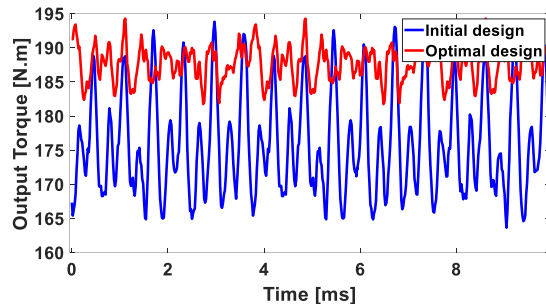


Fig. 7: The predicted instantaneous torque.

Conclusion

Using the MAXWELL, a simulation model based on FEM was developed for the delta-shape IPMSM to predict the important electromagnetic characteristics. Three different design optimization methods including consideration of notch on rotor, inserting flux barrier on rotor and having a skewed rotor were applied to a typical delta-shape IPMSM and their impacts on average torque, torque ripple and efficiency were evaluated first using FEM. Compared with initial design, it was seen that the design with flux-barrier had better performance. This design was then selected for more optimization and the optimum values of some design parameters were determined using the DOE method. In addition, the Hairpin winding was used for it instead of the conventional stranded winding. Using these two changes, significant improvement of motor performance was resulted when it was compared to the initial design (7% increase of average torque, 50% reduction of torque ripple and 1.4% increase of efficiency for the considered operating point).

Since these improvements are so valuable for an EV, the proposed design optimization method can be considered appropriately when the delta-shape IPMSM is used for this application.

Author Contributions

S. Nasr developed the model and provided the simulation results under supervision of Dr. Ganji and Prof. Moallem. The paper was written by S. Nasr and it is improved by Dr. Ganji and Prof. Moallem.

Acknowledgment

The authors gratefully acknowledge the ISKRA

company managers for their support to develop the model and provide the simulation results.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>PMSM</i>	Permanent Magnet Synchronous Motor
<i>EV</i>	Electric Vehicle
<i>IPMSM</i>	Interior Permanent Magnet Synchronous Motor
<i>DOE</i>	Design of Experiments
<i>FEM</i>	Finite Element Method
<i>PM</i>	Permanent Magnet
<i>HEV</i>	Hybrid Electric Vehicle
<i>PSO</i>	Particle Swarm Optimization
<i>GA</i>	Genetic Algorithm
<i>RSM</i>	Response Surface Method
<i>SNR</i>	Signal-to-Noise Ratio
<i>ANOVA</i>	Analysis of Variance
<i>SSF</i>	Sum of Squares of Factors

References

- [1] F. Momen, K. Rahman, Y. Son, "Electrical propulsion system design of Chevrolet Bolt battery electric vehicle," *IEEE Trans. Ind. Appl.*, 55(1): 376-84, 2018.
- [2] X. Sun, Z. Shi, Y. Cai, G. Lei, Y. Guo, J. Zhu, "Driving-cycle-oriented design optimization of a permanent magnet Hub motor drive system for a four-wheel-drive electric vehicle," *IEEE Trans. Transp. Electr.*, 6(3): 1115-25, 2020.
- [3] J. W. Chin, K. S. Cha, M. R. Park, S. H. Park, E. C. Lee, M. S. Lim, "High efficiency PMSM with high slot fill factor coil for heavy-duty EV traction considering AC resistance," *IEEE Trans. Energy Convers.*, 36(2): 883-94, 2021.
- [4] H. Dhulipati, S. Mukundan, Z. Li, E. Ghosh, J. Tjong, N. C. Kar, "Torque performance enhancement in consequent pole PMSM based on magnet pole shape optimization for direct-drive EV," *IEEE Trans. Magn.*, 57(2): 8103407, 2021.
- [5] A. Balamurali, A. Kundu, Z. Li, N. C. Kar, "Improved harmonic iron loss and stator current vector determination for maximum

- efficiency control of PMSM in EV applications," *IEEE Trans. Ind. Appl.*, 57(1): 363-73, 2021.
- [6] D. W. Kim, G. J. Park, J. H. Lee, J. W. Kim, Y. J. Kim, S. Y. Jung, "Hybridization algorithm of fireworks optimization and generating set search for optimal design of IPMSM," *IEEE Trans. Magn.*, 53(6): 16914097, 2017.
- [7] J. W. Jiang, B. Bilgin, Y. Yang, A. Sathyan, H. Dadkhah, A. Emadi, "Rotor skew pattern design and optimisation for cogging torque reduction," *IET Electr. Syst. Transp.*, 6(2): 126-35, 2016.
- [8] F. Chai, P. Liang, Y. Pei, S. Cheng, "Analytical method for iron losses reduction in interior permanent magnet synchronous motor," *IEEE Trans. Magn.*, 51(11): 15552769, 2015.
- [9] E. Sayed, Y. Yang, B. Bilgin, M. H. Bakr, A. Emadi, "A comprehensive review of flux barriers in interior permanent magnet synchronous machines," *IEEE Access*, 7: 149168-149181, 2019.
- [10] K. Yamazaki, M. Kumagai, T. Ikemi, S. Ohki, "A novel rotor design of interior permanent-magnet synchronous motors to cope with both maximum torque and iron-loss reduction," *IEEE Trans. Ind. Appl.*, 49(6): 2478-2486, 2013.
- [11] S. Zhu, W. Chen, M. Xie, C. Liu, K. Wang, "Electromagnetic performance comparison of multi-layered interior permanent magnet machines for EV traction applications," *IEEE Trans. Magn.*, 54(11): 18164250, 2018.
- [12] S. Zhu, Y. Hu, C. Liu, K. Wang, "Iron loss and efficiency analysis of interior PM machines for electric vehicle applications," *IEEE Trans. Ind. Electron.*, 65(1): 114-124, 2018.
- [13] G. H. Kang, Y. D. Son, G. T. Kim, J. Hur, "A novel cogging torque reduction method for interior-type permanent-magnet motor," *IEEE Trans. Ind. Appl.*, 45(1): 161-167, 2009.
- [14] D. Wei, H. He, J. Cao, "Hybrid electric vehicle electric motors for optimum energy efficiency: A computationally efficient design," *Energy*, 203: 117779, 2020.
- [15] C. Ma, L. Qu, "Multiobjective optimization of switched reluctance motors based on design of experiments and particle swarm optimization," *IEEE Trans. Energy Convers.*, 30(3): 1144-53, 2015.
- [16] K. Li, X. Zhang, H. Chen, "Design optimization of a tubular permanent magnet machine for cryocoolers," *IEEE Trans. Magn.*, 51(5): 8202708, 2015.
- [17] J. Si, S. Zhao, H. Feng, R. Cao, Y. Hu, "Multi-objective optimization of surface-mounted and interior permanent magnet synchronous motor based on Taguchi method and response surface method," *Chin. J. Electr. Eng.*, 4(1): 67-73, 2018.
- [18] S. K. Cho, K. H. Jung, J. Y. Choi, "Design optimization of interior permanent magnet synchronous motor for electric compressors of air-conditioning systems mounted on EVs and HEVs," *IEEE Trans. Magn.*, 54(11): 18164285, 2018.
- [19] J. Cui, W. Xiao, W. Zou, S. Liu, Q. Liu, "Design optimisation of submersible permanent magnet synchronous motor by combined DOE and Taguchi approach," *IET Electr. Power Appl.*, 14(6): 1060-1066, 2020.
- [20] M. H. Hwang, H. S. Lee, H. R. Cha, "Analysis of torque ripple and cogging torque reduction in electric vehicle traction platform applying rotor notched design," *Energies*, 11(11): 3053, 2018.
- [21] Z. Pan, K. Yang, X. Wang, "Optimal design of flux-barrier to improve torque performance of IPMSM for electric spindle," in *Proc. 18th International Conference on Electrical Machines and System*: 773-8, 2015.
- [22] T. Zhou, J.-X. Shen, "Cogging torque and operation torque ripple reduction of interior permanent magnet synchronous machines by using asymmetric flux-barriers," in *Proc. 20th International Conference on Electrical Machines and System*: 1-6, 2017.
- [23] X. Ge, Z. Zhu, G. Kemp, D. Moule, C. Williams, "Optimal step-skew methods for cogging torque reduction accounting for three-dimensional effect of interior permanent magnet machines," *IEEE Trans. Energy Convers.*, 32(1): 222-32, 2017.
- [24] Z. Shi, X. Sun, Y. Cai, Z. Yang, G. Lei, Y. Guo, J. Zhu, "Torque analysis and dynamic performance improvement of a PMSM for EVs by skew angle optimization," *IEEE Trans. Appl. Supercond.*, 29(2): 18352259, 2018.
- [25] R. Islam, I. Husain, A. Fardoun, K. McLaughlin, "Permanent magnet synchronous motor magnet designs with skewing for torque ripple and cogging torque reduction," *IEEE Ind. Appl. Ann. Meet.*, 45(1): 1552-59, 2007.
- [26] D. Deurell, V. Josefsson, "FEA study of proximity effect in hairpin windings of a PMSM for automotive applications," *MSc thesis, Chalmers University of Technology*, 2019.
- [27] D. S. Jung, Y. H. Kim, U. H. Lee, H. D. Lee, "Optimum design of the electric vehicle traction motor using the hairpin winding," in *Proc. 75th Vehicular Technology Conference*: 1-4, 2012.
- [28] G. Berardi, N. Bianchi, "Design guideline of an AC hairpin winding," in *Proc. XIII International Conference on Electrical Machines*: 2444-50, 2018.
- [29] X. Sun, Z. Shi, J. Zhu, "Multiobjective design optimization of an IPMSM for EVs based on fuzzy method and sequential Taguchi method," *IEEE Trans. Ind. Electron.*, 68(11): 10592-600, 2021.
- [30] E. Sayed, R. Yang, J. Liang, M. H. Bakr, B. Bilgin, A. Emadi, "Design of unskewed interior permanent magnet traction motor with asymmetric flux barriers and shifted magnets for electric vehicles," *Electr. Power Compon. Syst.*, 48(6): 652-666, 2020.

Biographies



Sepideh Nasr was born in Isfahan, Iran, in 1990. She received her M.Sc. in Electrical Engineering from the Isfahan University of Technology, Isfahan, Iran, in 2016. She is currently pursuing the Ph.D. degree in the Electrical Engineering at University of Kashan in Iran. Her main research interest includes electrical and thermal design of electrical machines especially PMSM for EV application.

- Email: s.nasr@alumni.iut.ac.ir
- ORCID: 0000-0003-0974-4228
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Babak Ganji received his B.Sc. degree from Esfahan University of Technology, Iran in 2000, and M.Sc. and Ph.D. from University of Tehran, Iran in 2002 and 2009, respectively, all in major Electrical Engineering-power. He was granted DAAD scholarship in 2006 from Germany and worked in institute of Power Electronics and Electrical Drives at RWTH Aachen University as a visiting researcher for 6 months. He has been working at University of Kashan in Iran since 2009 as an Associate Professor and his research interest is modeling and design of advanced electric machine especially switched reluctance motor.

- Email: bganji@kashanu.ac.ir
- ORCID: 0000-0003-2310-215X
- Web of Science Researcher ID: NA
- Scopus Author ID: 24724058600
- Homepage: <https://faculty.kashanu.ac.ir/bgjanji/en>



Mehdi Moallem (SM'95) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1989. He has been with the Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran since 1991. His research interests include design and optimization of electromagnetic devices, application of advance numerical techniques

and intelligent systems to analysis and design of electrical machines, and power quality.

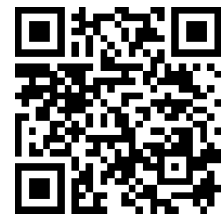
- Email: moallem@cc.iut.ac.ir
- ORCID: [0000-0002-5965-0978](https://orcid.org/0000-0002-5965-0978)
- Web of Science Researcher ID: NA
- Scopus Author ID: 7005252934
- Homepage: <https://moallem.iut.ac.ir/>

How to cite this paper:

S. Nasr, B. Ganji, M. Moallem, "Design optimization of the delta-shape interior permanent magnet synchronous motor for electric vehicle application," J. Electr. Comput. Eng. Innovations, 11(2): 291-300, 2023.

DOI: [10.22061/jecei.2022.9207.587](https://doi.org/10.22061/jecei.2022.9207.587)

URL: https://jecei.sru.ac.ir/article_1810.html





Research paper

PAPR Reduction in OFDM UOWC System Employing Repetitive Clipping and Filtering (RCF) Method

B. Noursabbaghi, G. Baghersalimi*, A. Pouralizadeh, O. Mohammadian

Department of Electrical Engineering, University of Guilan, Rasht, Iran.

Article Info

Article History:

Received 01 August 2022
Reviewed 07 September 2022
Revised 07 October 2022
Accepted 03 December 2022

Keywords:

OFDM
UOWC
PAPR
RCF method

*Corresponding Author's Email
Address: bsalimi@guilan.ac.ir

Abstract

Background and Objectives: High peak-to-average power ratio (PAPR) in Orthogonal Frequency Division Multiplexing (OFDM)-based Underwater Optical Wireless Communication (UOWC) systems is one main reason for out-of-band power and in-band distortion, leading to the degradation of system performance. Therefore, different approaches have been proposed, implemented, and examined for reducing the high PAPR of OFDM signals in such systems.

Methods: In this research, the performance of an OFDM-based UOWC system is investigated by employing the Repetitive Clipping and Filtering (RCF) technique in clear open ocean water. The Monte Carlo Modeling of Light (MCML) approach with the Henyey-Greenstein (HG) model of the scattering phase function is used to simulate the UOWC channel.

Results: The CCDF performance of the proposed system with the RCF method for various CR values is investigated. Also, the proposed system performance is examined in terms of bit error rate (BER) and error vector magnitude (EVM) at two different depths for link lengths of 1 m and 5 m.

Conclusion: The results show that the system performance is limited by increasing the link length, the number of subcarriers, and depth. Also, it is shown that the RCF method significantly leads to a reduction of the PAPR in the DCO-OFDM UWOC system and improves BER performance up to 10 dB.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Recently, ocean exploration has attracted researchers' attention due to developments in military, industry, and scientific issues. Some of the most important applications of underwater wireless communications (UWC) include oceanography study, underwater surveillance, seafloor exploration, and monitoring, underwater oil pipe inspection, remotely operated vehicles (ROV), and sensor networks [1]-[5]. Although the dominant communications schemes used in underwater are still based mainly on the wireless technologies of radio frequency (RF) and acoustic (sonar), they suffer from low transmission data rate,

limited bandwidth, high latency, and high attenuation which need to be addressed [6], [7].

One promising alternative complementary technology to aforementioned traditional schemes is optical waves. The underwater optical communications (UOWC) offer high speed (i.e., > 200 Mbps over a transmission range of up to 100 m), low attenuation (at the blue-green light wavelength of 450-550 nm), low latency, reliable and secure communication link which have become very attractive in recent years. However, UOWC link performance is affected by temperature fluctuations, scattering, dispersion, and beam steering leading to shorter transmission distance compared with other communication schemes [8].

Blue light-emitting diodes (LEDs) with phosphor coating are commonly used as a transmitter (Tx) in the UOWC systems which offer a long lifetime, energy efficiency, uniform illumination, and also simple fabrication with low cost. However, due to the slow response of phosphor, these LEDs have limited bandwidth (typically a few MHz). Moreover, white LEDs are the major source of nonlinearity in OWC systems which in turn causes signal distortion and intersymbol interference (ISI) [8], [9]. Some common solutions to mitigate these limitations are to employ multiple input-multiple output (MIMO) schemes, various equalization methods, and advanced modulation techniques [9], [10].

The simplest modulation scheme used in the UOWC systems is on-off keying (OOK), but it is not suitable for high data rate transmission due to its low spectral efficiency [10], [11]. As a result, there has been increasing interest in utilizing multi-carrier modulations such as orthogonal frequency division multiplexing (OFDM). The combination of OFDM with high-order quadrature amplitude modulation (QAM) can achieve higher data rate, higher spectral efficiency, simple one-tap equalization as well as inherent resistance to the ISI [12]. Two commonly OFDM schemes used in optical communications are asymmetrically clipped optical OFDM (ACO-OFDM) and DC-biased optical OFDM (DCO-OFDM). In this work, we use DCO-OFDM due to its higher spectral efficiency [13].

In [11] the authors evaluated the impact of different modulation orders of the QAM scheme in the OFDM-UOWC system. It was shown for a 2-m underwater channel that the best achievable bit rates are 161.36 Mb/s using 16-QAM at a BER of 2.5×10^{-3} , 156.31 Mb/s using 32-QAM at a BER of 7.42×10^{-4} , and 127.07 Mb/s using 64-QAM at a BER of 3.17×10^{-3} , respectively. In our previous work [14], the performance of an OFDM UOWC link for different depths was investigated in the clear open ocean. Results confirmed that the system BER performance degraded from 5.25×10^{-6} to 1.21×10^{-2} by

increasing depth from 1 m to 30 m.

OFDM-based UOWC system with a high peak-to-average power ratio (PAPR) is sensitive to the nonlinearity of the LEDs, which is the main reason for out-of-band power and in-band distortion leading to the system performance deterioration. Different approaches have been proposed and implemented for reducing the high PAPR of OFDM signals such as amplitude clipping and filtering, peak windowing, peak cancellation, peak reduction carrier, envelope scaling, decision-aided reconstruction (DAR), coding, partial transmit sequence (PTS), selective mapping (SLM), interleaving, tone reservation (TR), tone injection (TI), active constellation extension (ACE), clustered OFDM, and pilot symbol assisted (PA) modulation [15], [16]. For instance, in [17] a pilot-aided technique was proposed for PAPR reduction of the optical OFDM system, and results showed a 1 dB improvement in the energy-to-noise ratio ($E_{b(opt)}/N_0$) compared to the basic DCO-OFDM at a target BER of 10^{-3} . In [18] the combination of the signal-to-clipping noise ratio (SCR) and the least-squares approximation (LSA) method as a tone reservation scheme was suggested to reduce the PAPR of the DCO-OFDM in an indoor visible light communication (VLC) system. The proposed scheme exhibited PAPR improvement of about 4 dB compared with the original OFDM signal. In [19], a PAPR analysis was carried out for different types of optical OFDM schemes including DCO-OFDM, ACO-OFDM, pulse amplitude modulated discrete multitone (PAM-DMT), and Flip-OFDM. In [20], an OFDM PAPR reduction scheme based on time-frequency domain interleaved was utilized in a UOWC system, which showed a reduction of 8.4 dB in PAPR compared with the original OFDM system. In [21] the repeated clipping and filtering (RCF) method in the frequency domain was used for PAPR reduction and the results confirmed that the RCF method could lead to a reduced overall peak regrowth.

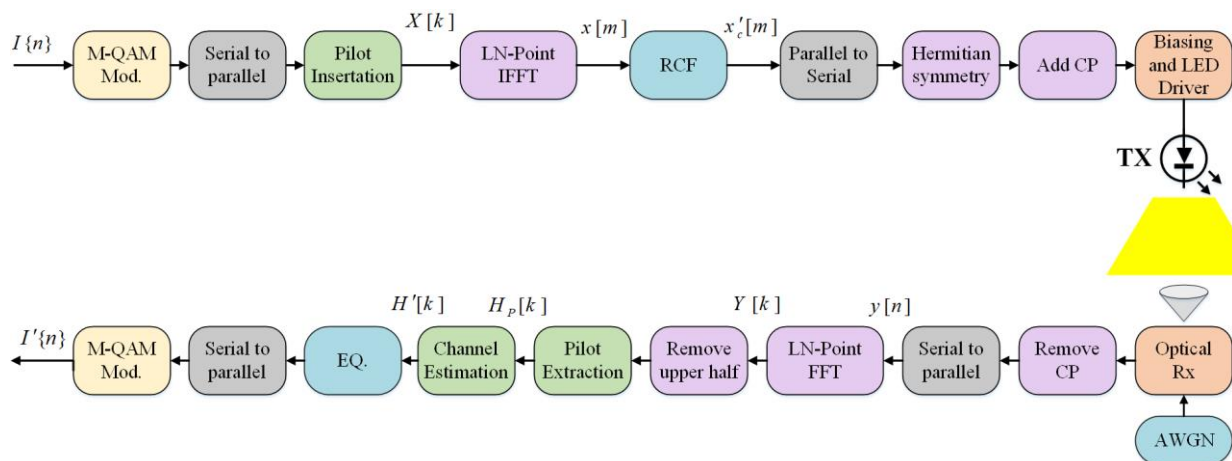


Fig. 1: Schematic block diagram of the proposed DCO-OFDM UOWC system with RCF method.

To the best of the authors' knowledge, the RCF method has not been deeply investigated in the area of UOWC. In this paper, performance of the DCO- OFDM UOWC system is evaluated for two different depths. The minimum mean squared error frequency domain equalizer (MMSE-FDE) is used for channel compensation along with the RCF method to reduce the PAPR. The rest of the paper is organized as follows. In Section 2, an overview of the DCO-OFDM UOWC system with the RCF method including the transmitter (Tx), channel, and receiver (Rx) is explained. In Section 3, the most important results are presented and discussed. Finally, Section 4 concludes the paper.

System Overview

Fig. 1 illustrates the schematic block diagram of the proposed DCO-OFDM UOWC system the with RCF method which is explained comprehensively as follows. The goal is to compare its performance with the DCO-OFDM system without the RCF method under the same conditions.

A. Transmitter

At the Tx, a random data bit stream $\{I_n\}$ is mapped onto the QAM format. Then, the modulated symbols are applied to the serial to parallel (S/P) converter. The pilot insertion block is considered to provide channel estimation at the Rx. Then, the signal is extended by inserting $N(L - 1)$ zeros in its middle, resulting in the trigonometric interpolation of the time domain signal, where L is the oversampling factor [22]. The over-sampled signal is passed through an LN -point inverse fast transform (IFFT) block generating the oversampled time-domain OFDM signal followed by the parallel to serial (P/S) converter, which is represented as:

$$x[n] = \frac{1}{\sqrt{LN}} \sum_{k=0}^{LN-1} X[k] e^{j \frac{2\pi k \Delta f n}{LN}} \quad n = 0, 1, \dots, LN-1, \quad (1)$$

As a result of intensity modulation/direct detection (IM/DD), the transmitted signal must be non-negative and real. Therefore, the Hermitian symmetry constraint is imposed to acquire a real-valued signal, as given by [23]:

$$\bar{x}_{HS} = [0, x_{CF}[1], x_{CF}[2], \dots, x_{CF}[\frac{LN}{2} - 1], 0, x_{CF}^*[\frac{LN}{2} - 1], \dots, x_{CF}^*[2], x_{CF}^*[1]] \quad (2)$$

where $(.)^*$ is the complex conjugate.

A cyclic prefix (CP) is then prepended to each OFDM signal for eliminating the ISI and inter-block interactions (IBI) [12]. Following scaling of the CP added signal $x_{CP}[n]$, we have:

$$I_s[n] = \lambda_s \times x_{CP}[n], \quad (3)$$

where the scaling factor λ_s due to the limited dynamic range of the LED is defined as [24]:

$$\lambda_s = \frac{MI \times I_{max}}{(MI+1) \times \max(x_{CP}[n])}, \quad (4)$$

Here $I_{max} = I_b + 0.5 I_{PP}$ is the maximum value of I_{in} , I_b is the DC-bias current, I_{PP} is the peak-to-peak current, and MI is the modulation index, which is defined as:

$$MI = \frac{I_{PP} - I_{PP}}{I_b}, \quad (5)$$

Then, following the digital to analog converter, a direct current (DC)-bias I_b is added to the time domain (TD) discrete scaled signal $I_s[n]$ to make it non-negative prior to IM of the light source and transmission over the underwater channel as follows:

$$I_{in}(t) = I_s(t) + I_b, \quad (6)$$

At last, it is fed to drive the LED by convolving with the LED impulse response which is characterized as a first-order low pass Butterworth filter, given by [24]:

$$h_{LED}[n] = \exp[-2\pi f_{LED} n], \quad (7)$$

where f_{LED} is the 3-dB cut-off frequency of the LED. Note, the PAPR is calculated from the L -times oversampled TD signal samples $x[m]$ as [19]:

$$PAPR\{x[m]\} = \frac{\max_{0 \leq t \leq NL-1} |x[m]|^2}{E[|x[m]|^2]}, \quad (8)$$

where $E\{\cdot\}$ is the expectation operator. The complementary cumulative distribution function (CCDF), defined as $CCDF(PAPR) = P\{PAPR > PAPR_0\}$, is utilized to appraise the PAPR reduction performance. The target value of PAPR is designated as $PAPR_0$ [20].

B. Channel

Channel modeling plays an important role in the quality evaluation of communication links. A practical and widely used method for modeling the underwater channel is Monte Carlo (MC) simulation to solve the *radiative transfer equation* (RTE). The MC numerical method evaluates the channel characteristics by generating N photons through the water simultaneously and then tracking the interactions of each photon with the medium and its trajectory through the channel from the Tx to the Rx. This method is a more flexible and simpler approach without restrictions on the scattering angles compared with the other analytical methods [25].

Three basic attributes of emitted photons include photon's weight, position in Cartesian coordinates (x , y , z), and transmission direction (characterized by polar θ and azimuthal angle φ). Photons interacting with particles in the water experience a change of direction and loss due to both scattering and absorption which can be evaluated by absorption coefficient, $a(\lambda)$, and scattering coefficient, $b(\lambda)$, respectively. The basic rules of this approach are summarized as follows [26], [27]. The initial photon's position is equal to the Tx position which is considered (0, 0, 0) in this study. The angle of θ and φ are chosen randomly between $[-\theta_{max}, \theta_{max}]$ and

$[0, 2\pi]$ with uniform distributions, respectively, where θ_{max} is the maximum initial divergence angle. The direction vector (μ_x, μ_y, μ_z) for each photon is equal to $(\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta)$.

The scattering coefficient is considered to study the effect of multiple scattering which is defined by the integral of the spectral volume scattering function (VSF) over all directions as:

$$b(\lambda) = \int_0^{4\pi} \beta(\theta, \lambda) d\Omega = 2\pi \int_0^\pi \beta(\theta, \lambda) \sin \theta d\theta, \quad (9)$$

where $\beta(\theta, \lambda)$ is the VSF and $d\Omega$ is the solid angle centered on θ . Finally, the whole attenuation coefficient of spectral light beam is defined as, $c(\lambda) = a(\lambda) + b(\lambda)$. Note, a, b , and c are in units of m^{-1} . In addition, the angular probability distribution of the scattered photons at a given wavelength, known as the scattering phase function (SPF), is given by:

$$\tilde{\beta}(\theta, \lambda) = \frac{\beta(\theta, \lambda)}{b(\lambda)}, \quad (10)$$

Typically, the Henyey-Greenstein (HG) phase function is used to describe the SPF of dispersive medium such as water and atmosphere with the expression as:

$$\tilde{\beta}(\theta) = \frac{1 - g^2}{4\pi(1 + g^2 - 2g \cos \theta)^{3/2}}, \quad (11)$$

where g is the average cosine of θ in all scattering directions and almost equals 0.924 for all water types. At first and before any interactions, the photon passes a certain distance called the step size (Δs):

$$\Delta s = -\frac{\log \xi^S}{c(\lambda)}, \quad (12)$$

where ξ^S is a uniform random variable between $[0, 1]$. After that, the new coordinates of the photon's position are updated according to:

$$\dot{x} = x + \mu_x \Delta s, \dot{y} = y + \mu_y \Delta s, \dot{z} = z + \mu_z \Delta s, \quad (13)$$

The interaction of the photon with the scattering point leads to energy loss and deviation of the photon from the transmission direction. Therefore, the photon's energy level (weight) is updated by:

$$W_{Post} = \left(1 - \frac{b}{c}\right) W_{Pre}, \quad (14)$$

where W_{Post} and W_{Pre} represent the weights after and before the interaction, respectively. The new azimuth (ϕ) and elevation (θ) angles also need to be calculated due to changing the photon transmission direction after the scattering point, as follows:

$$R = 2\pi \int_0^{\hat{\theta}} \tilde{\beta}(\theta) \sin \theta d\theta, \quad (15)$$

$$\phi = 2\pi R, \quad (16)$$

where R is a uniform random variable between $[0, 1]$.

Then, from (15), $\hat{\theta}$ can be obtained as:

$$\cos \hat{\theta} = \frac{1}{2g} \left[1 + g^2 - \left(\frac{1 - g^2}{1 - g + 2gR} \right)^2 \right]. \quad (17)$$

Finally, the new transmission direction vector can be calculated as [26]:

$$\begin{aligned} \mu'_x &= -\mu_y \sin \hat{\theta} \cos \phi + \mu_x (\cos \hat{\theta} + \sin \hat{\theta} \sin \phi), \\ \mu'_y &= -\mu_x \sin \hat{\theta} \cos \phi + \mu_y (\cos \hat{\theta} + \sin \hat{\theta} \sin \phi), \\ \mu'_z &= -(\mu_x^2 + \mu_y^2) \sin \hat{\theta} \sin \phi / \mu_z + \mu_z \cos \hat{\theta}, \end{aligned} \quad (18)$$

At the Rx side, a photon can be detected when its position and the arrival angle are within the Rx's aperture and Field of View (FOV) and its weight is higher than the threshold level. The process of photon scattering continues until the photon is received at the PD or disappeared by losing all its weight. The threshold weight at the PD is assumed 10^6 in this study.

C. Receiver

At the Rx side, the received signal is detected by an optical Rx composed of a single PD and a trans-impedance amplifier (TIA). The regenerated electrical received signal is given by:

$$y(t) = s(t) * h_c(t) + n(t), \quad (19)$$

where $n(t)$ is the additive white Gaussian noise with the power $P_n = N_0 B_{Rx}$, N_0 is the noise power. Note, spectral density, and B_{Rx} is the bandwidth of the Rx. $n(t)$ is mostly dominated by the ambient light induced shot noise. Then, the received signal is amplified and converted to a parallel signal via the S/P block. Following CP removal, an N -point FFT block is employed to transform the TD signal $y[n]$ to the Frequency Domain (FD) signal $Y[k]$. The upper half of the signal is removed due to the use of Hermitian symmetry at the Tx side and then the least square (LS) method is utilized to calculate the complex-valued channel frequency response (CFR) based on pilot symbols, as follows:

$$H_p(m) = \frac{Y_p(m)}{X_p(m)}, \quad m = 1, \dots, N_p \quad (20)$$

where m is the number of pilot sub-carrier, N_p is the number of pilots in one OFDM symbol, $Y_p(m)$ is the received pilot symbol that are extracted from every eight subcarriers of the received OFDM signal, and $X_p(m)$ is the transmitted pilot symbol. Then, linear interpolation is performed to compute the CFR function on the remaining subcarriers as:

$$\hat{H}[(m-1)L + \ell] = H_p(m) + \frac{1}{L} [H_p(m+1) - H_p(m)], \quad \ell = 1, \dots, L \quad (21)$$

where $L = 8$ is the distance in subcarriers between two consecutive pilots.

A one-tap frequency-domain equalizer (FDE) with the minimum mean square error (MMSE) method is used to compensate for the channel deficiencies. The MMSE coefficients are calculated as follows:

$$C_k^{MMSE} = \frac{\hat{H}_k^*}{\hat{H}_k \hat{H}_k^* + 1/\gamma}, \quad (22)$$

where k is the number of subcarriers, and γ is the signal to noise ratio (SNR). So, the decision variable is calculated as:

$$\hat{S}_k = C_k^{MMSE} Y[k], \quad (23)$$

Finally, \hat{S}_k is converted to a serial data stream using a parallel-serial (P/S) block prior to QAM demodulation [28].

D. RCF Method

Fig. 2 illustrates the schematic block diagram of the



Fig. 2: Block diagram of the RCF method.

The so-called Repetitive Clipping and Filtering (RCF) is only applied at the Tx end of the OFDM system, however it influences the performance of the Rx signal at the receive end. The RCF algorithm proceeds in four steps [29], [30]:

Step 1: The amplitude of L -times oversampled time domain signal $x[m]$ is clipped while its phase remains unchanged:

$$x_c[m] = \begin{cases} A & |x[m]| < A \\ x[m] & |x[m]| \geq A \end{cases} \quad (24)$$

where A is the threshold clipping level that equals to:

$$A = CR \times \sigma \quad (25)$$

where CR and σ are the clipping ratio and root mean squared value (RMS) of $x[m]$, respectively.

Step 2: The clipping is followed by the FD filtering to reduce OOB distortion caused by clipping. Hence, the clipped TD signal $x_c[m]$ is passed through a filter consisting of FFT and IFFT operators. First, $x_c[m]$ is converted back into the discrete FD as $X_c[k]$ using an FFT. Then, the OOB components of $X_c[k]$ are set to zero while the in-band components are left unchanged. The resultant signal is as follows:

$$\hat{X}_c[k] = [X_c[0], \dots, X_c[\frac{N}{2} - 1], 0, \dots, 0, X_c[NL - \frac{N}{2} + 1], \dots, X_c[NL - 1]] \quad (26)$$

Step 3: Then, $\hat{X}_c[k]$ is converted into the TD using the LN -point IFFT block.

Step 4: The filtering technique can lead to peak regrowth. So, repeat K times step 1 to step 3, where K is a positive integer usually chosen between one and four until the amplitude of the OFDM signal is set to a specified threshold level [20].

Result and Discussion

In this section, we present computer simulation results for the proposed system, which are obtained using MATLAB and real-world parameters. All simulation parameters are given in Table 1.

RCF method. To avoid large PAPR, the clipping technique as a hard limiter is applied to the amplitude of the complex values of the IFFT output. Following, the filtering technique is designed to alleviate or cancel Out of Band (OOB) distortion dependent on the oversampling value however, however it cannot correct in-band distortion.

Table 1: System parameters

Symbol	Parameter	Value
-	Modulation type	QAM
M	Modulation order	4
N	(I)FFT size	256, 1024
-	Pilot type	Comb
N_p	Number of pilots	30 ($N = 256$), 114 ($N = 1024$)
N_a	Number of active subcarriers	224 ($N = 256$), 910 ($N = 1024$)
R_b	Bit rate (Mbps)	6
N_{CP}	Cyclic prefix length	$N/4$
f_{LED}	LED cut-off frequency (MHz)	3
θ	LED divergence angle	60°
FOV	PD field of view	60°
W_s	Signal bandwidth (MHz)	3
f_s	Sampling frequency (MHz)	12
Δx		1 m
L	Clipping ratio	4
CR		0.4 : 0.2 : 4

In Fig. 3, the CCDF function is used to evaluate the PAPR performance of 4-QAM DCO OFDM UOWC system with RCF method for a range of Clipping Ratio (CR) values. The clipping and filtering level, depth, and the number of subcarriers is one, 5 m, and 1024, respectively. The CCDF of PAPR is defined as CCDF (PAPR) = P (PAPR, respectively $> PAPR_0$) where $PAPR_0$ is the target PAPR value. It is clearly shown that utilizing the RCF method leads to the PAPR reduction up to ~4 dB, at CCDF of 10^{-2} and CR = 4 compared with the original signal. In addition, it can be seen that the CCDF of PAPR can be remarkably reduced by decreasing CR. The best PAPR reduction is achieved for the lowest CR = 0.4 which is 6.9 and 7.35 dB at CCDF = 10^{-2} and CCDF = 10^{-3} , respectively, compared with the original signal. In fact,

the PAPR performance of the proposed system becomes worse as the CR value increases due to a decrease in the

A value which is equal to $CR \times \sigma$.

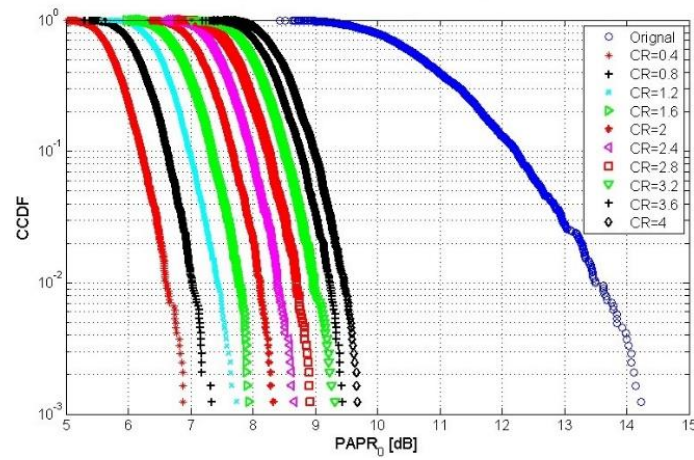


Fig. 3: CCDF performance of the proposed system for one time clipping and filtering and various CR values.

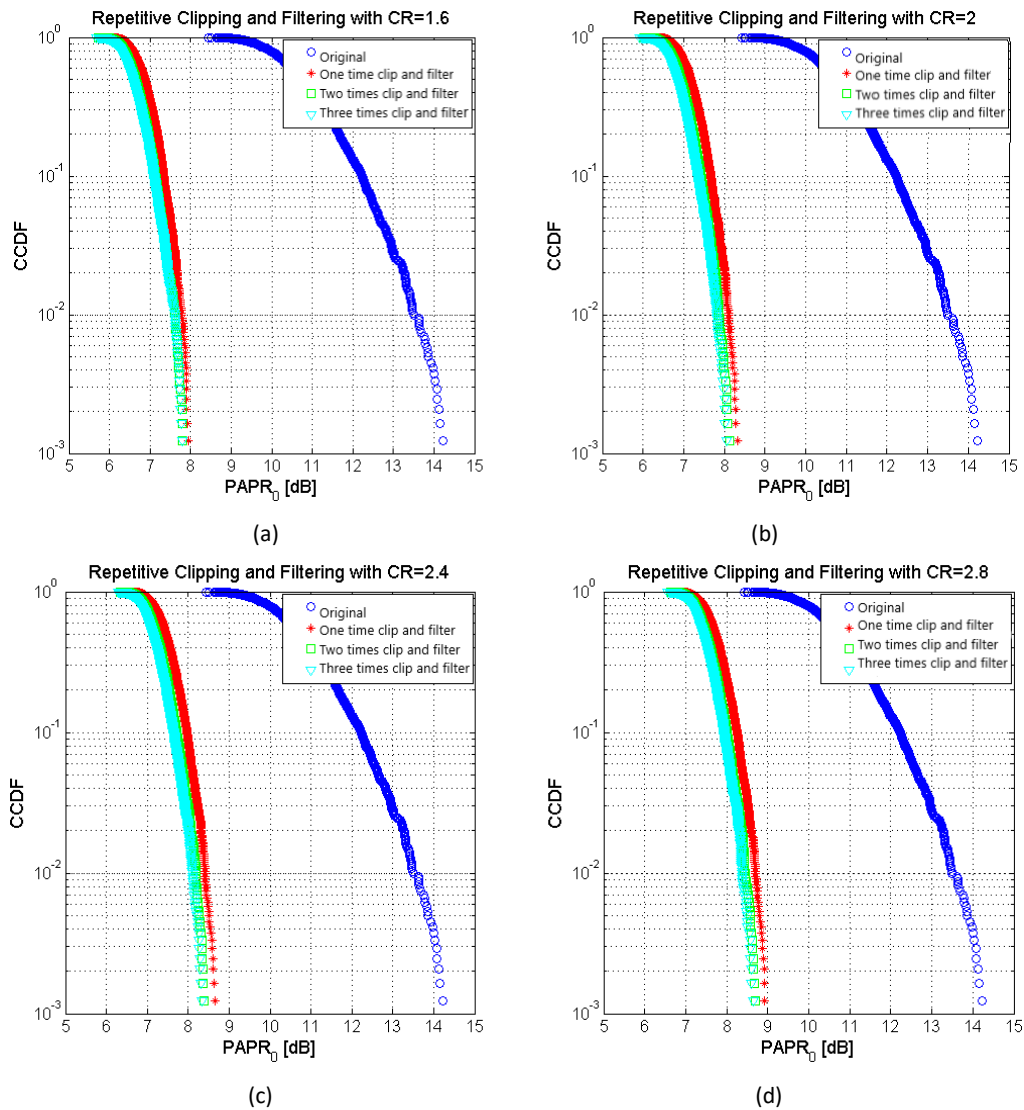


Fig. 4: The CCDF performance of the RCF: (a) CR=1.6, (b) CR=2, (c) CR=2.4 and (d) CR=2.8.

The CCDF performance is also examined when the clipping and filtering levels are 2 and 3 times at the CR of 1.6, 2, 2.4, and 2.8, as shown in Fig. 4. For example, for three times clipping and filtering scenario, the results demonstrate that at CCDF of 10^{-2} , the PAPR gains for the CR = 1.6, 2, 2.4, 2.8 are equal to 7.60, 7.85, 8.19, and 8.41 dB, respectively.

The system BER and EVM performances are also evaluated for different CRs in a 4-QAM DCO OFDM system using RCF method.

As can be seen in Fig. 5, by increasing the gain of CRs both BER and EVM performances of the system are improved.

For instance, CRs of 4, 3.2, 2.8 and 2.4 reach the target BER about 10^{-4} at E_b/N_0 of 13, 14, 16 and 22, respectively.

Next, the performance of the DCO-OFDM UOWC system with and without the RCF method is assessed while two different link lengths of 1 m and 5 m are considered.

As can be seen clearly in Fig. 6 (a), employing RCF method with CR = 3.6 results in improved BER performance up to 10 dB compared with the system without RCF method at target BER = 10^{-4} . In addition, by increasing the link length to 5 m, the proposed method with CR of 3.6 reached the target BER of 10^{-4} at E_b/N_0 of 16 dB.

Finally, we examined the effect of the RCF method on

the error floor.

According to Fig. 7, utilizing the RCF method with higher CR at both depths of 1 m and 10 m leads to error floor at low BER values.

In summary, the CCDF of PAPR can be remarkably reduced by decreasing CR. Clearly, utilizing the RCF method leads to the PAPR reduction up to ~4 dB. In addition, the CCDF performance is increased by the number of clipping and filtering iterations, however the improvement is not significant for more than once as evidenced by Fig. 4.

To conclude, clipping and filtering techniques eliminate the out-of-band radiation by clipping the time-domain signal to a predefined level and subsequently filtering.

The relatively small in-band distortion is combatted using low-order signal constellation, coding, and/or clipping noise cancellation techniques, where we have not considered all these techniques in this research. To suppress peak regrowth due to filtering, RCF techniques can be used.

Their convergence rate decreases significantly after the first few iterations. Also, the increased number of iterations leads to increased computational complexity, especially when the number of subcarriers is very large. The convergence rate can be improved by setting the clipping threshold to a level slightly lower than the required level.

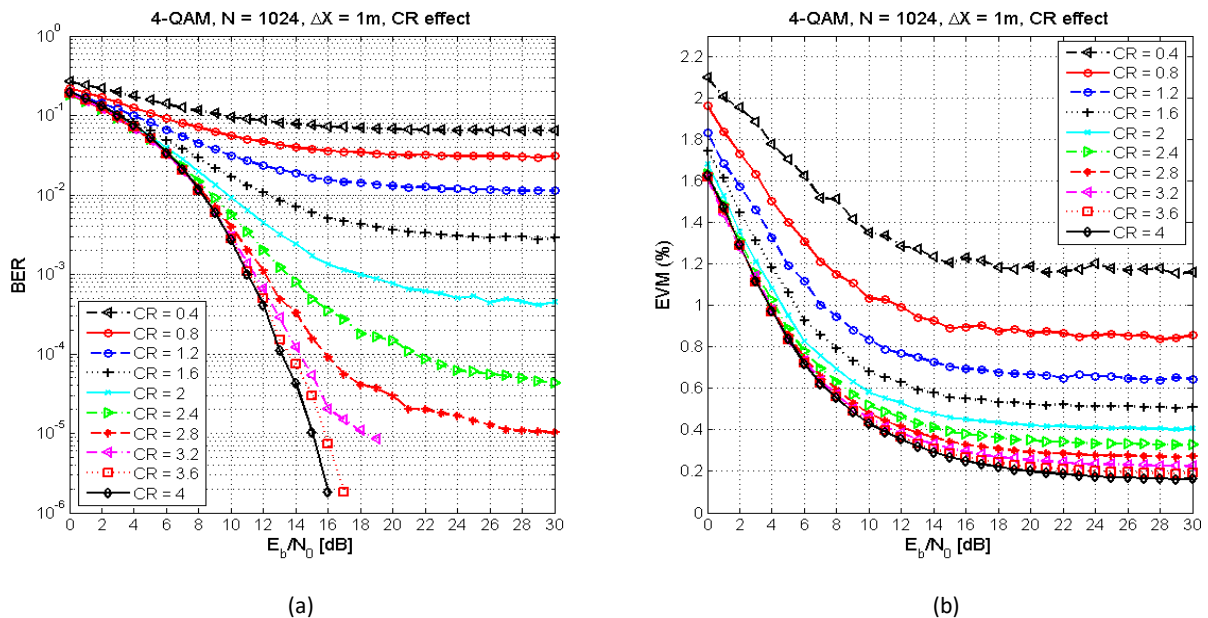


Fig. 5: BER and EVM vs. E_b/N_0 : (a) BER performance and (b) EVM performance.

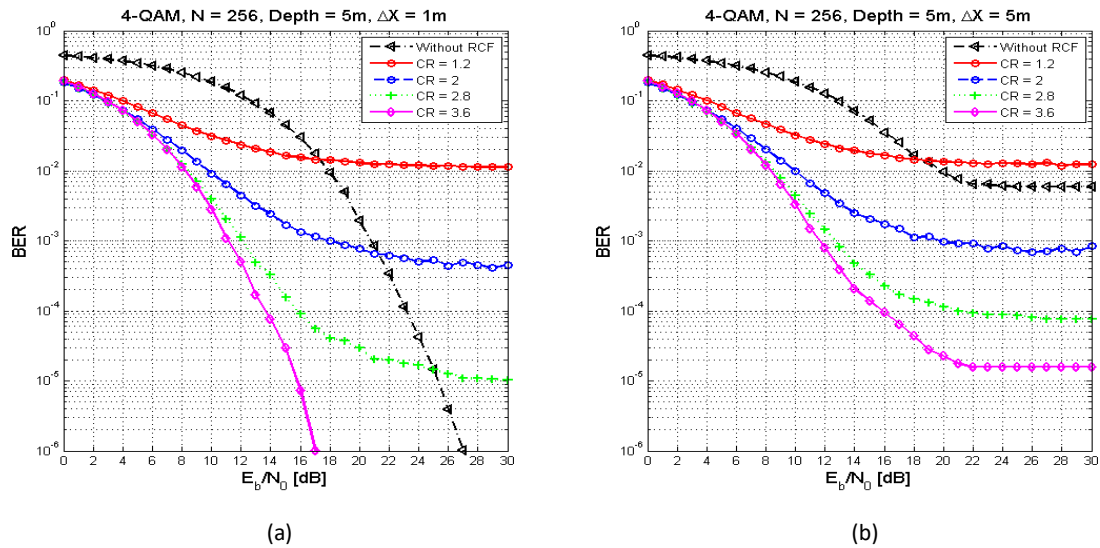


Fig. 6: BER Performance of the DCO OFDM UWOC system with and without RCF Method: (a) $\Delta X = 1$ m and (b) $\Delta X = 5$ m.

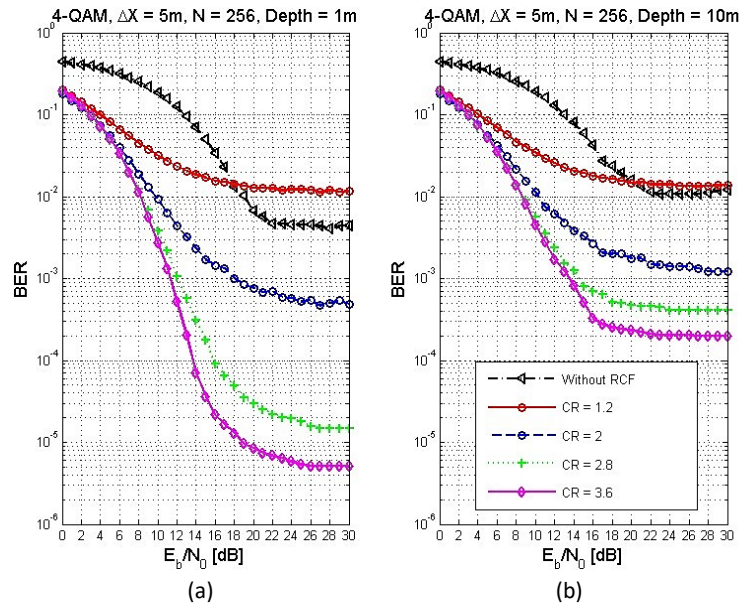


Fig. 7: The effect of the RCF method on the error floor: (a) depth=1m, (b) depth=10m.

Conclusion

In this article the performance of the DCO-OFDM-based UWOC system was evaluated. The MCML method with the HG model of the SPF is employed for channel modeling. The results showed that the system performance is limited by increasing the link length, the number of subcarriers, and depth. Also, the RCF method was successfully utilized to reduce the PAPR of the DCO-OFDM UWOC system and improve BER performance so that the reduced PAPR for CR=4 as the extreme case for CCDF values of 10^{-2} and 10^{-3} are 9.51 dB and 9.72 dB, respectively. It was also shown that by decreasing the CR, the PAPR is reduced at the cost of the BER and EVM degradation. To reduce PAPR even more, other techniques such as multiple signal representation, coding, and Discrete Fourier Transform (DFT) precoding with RCF may be used, which are of our future subject for research.

Author Contributions

All the Authors contributed to all part of preparing and writing of this paper.

Acknowledgment

The author would like to thank the editor and reviewers for their helpful comments.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

UWC Underwater Wireless Communications

ROV	Remotely Operated Vehicles
RF	Radio Frequency
ACO-OFDM	Asymmetrically Clipped Optical OFDM
DCO-OFDM	DC-biased Optical OFDM
UOWC	Underwater Optical Wireless Communications
ISI	Inter Symbol Interference
MIMO	Multiple Input- Multiple Output
OOK	On-Off Keying
OFDM	Orthogonal Frequency Division Multiplexing
PAPR	Peak-to-Average Power Ratio
VLC	Visible light Communication
RCF	Repeated Clipping and Filtering
IM/DD	Intensity Modulation/Direct Detection
MC	Monte Carlo
FOV	Field of View
TIA	Trans-Impedance Amplifier

References

- [1] S. Hessien, S. C. Tokgoz, N. Anous, A. Boyacı, M. Abdallah, K. A. Qaraqe, "Experimental evaluation of ofdm-based underwater visible light communication system," *IEEE Photonics J.*, 10(5), 2018.
- [2] G. S. Spagnolo, L. Cozzella, F. Leccese, "Underwater optical wireless communications: Overview," *Sensors*, 20(8): 1-14, 2020.
- [3] M. F. Ali, D. N. K. Jayakody, Y. Li, "Recent trends in underwater visible light communication (UVLC) systems," *IEEE Access*, 10: 22169 – 22225, 2016.
- [4] I. N'doye, D. Zhang, M. S. Alouini, T. M. Laleg-Kirati, "Establishing and maintaining a reliable optical wireless communication in underwater environment," *IEEE Access*, 9: 62519 – 62531, 2021.
- [5] H. Kaushal, G. Kaddoum, "Underwater optical wireless communication," *IEEE Access*, 4: 1518 – 1547, 2016.
- [6] Z. Zeng, S. Fu, H. Zhang, Y. Dong, J. Cheng, "A survey of underwater optical wireless communications," *IEEE Commun. Surv. Tutorials*, 19(1): 204-238, 2017.
- [7] S. Zhu, X. Chen, X. Liu, G. Zhang, P. Tian, "Recent progress in and perspectives of underwater wireless optical communication," *Prog. Quantum Electron.*, 73: 100274, 2020.
- [8] K. Mamatha, K. B. N. S. K. Chaitanya, S. Kumar, A. Arockia Basil Raj, "Underwater wireless optical communication - A review," presented at the Int. Conf. Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021.
- [9] N. Saeed, A. Celik, Tareq Y. Al-Naffouri, M. S. Alouini, "Underwater optical wireless communications, networking, and localization: A survey," *Ad Hoc Networks*, 94, 2019.
- [10] J. Xu, A. Lin, X. Yu, Y. Song, M. Kong, F. Qu, J. Han, W. Jia, N. Deng, "Underwater laser communication using an OFDM-modulated 520-nm laser diode," *IEEE Photonics Technol. Lett.*, 28(20): 2133-2136, 2016.
- [11] J. Xu, M. Kong, A. Lin, Y. Song, X. Yu, F. Qu, J. Han, N. Deng, "OFDM-based broadband underwater wireless optical communication system using a compact blue LED," *Opt. Commun.*, 369: 100-105, 2016.
- [12] R. Kraemer, J. S. Tavares, F. Pereira, H. M. Salgado, L. M. Pessoa, "5.36 Gbit/s OFDM optical wireless communication link over the underwater channel," presented at the 12th Int. Symposium on Communication Systems, Networks and Digital Signal Processing (CSNDSP), Porto, Portugal, 2020.
- [13] N. J. Jihad, S. M. Abdul Satar, "Performance study of ACO-OFDM and DCO OFDM in optical camera communication system," presented at 2nd Al-Noor International Conference for Science and Technology (NICST), Baku, Azerbaijan, 2020.
- [14] B. Noursabbaghi, G. Baghersalimi, O. Mohammadian, "Performance evaluation of an OFDM-based underwater wireless optical communication depth by considering depth-dependent variations in attenuation," presented at 2nd West Asian Colloquium on Optical Wireless Communications (WACOWC), Tehran, Iran, 2019.
- [15] A. Nayak, A. Goen, "A review on PAPR reduction techniques in OFDM system," *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.*, 5 (4): 2767-2772, 2019.
- [16] M. Munjure Mowla, M. Yeakub Ali, R. A. Aoni, "Performance comparison of two clipping based filtering methods for PAPR reduction in OFDM signal," *J. Mobile Network Commun. Telematics (IJMNCT)*, 4(1): 23-34, 2014.
- [17] F. B. Offiong, S. Sinanović, W. O. Popoola, "On PAPR reduction in pilot-assisted optical OFDM communication systems," *IEEE Access*, 5: 8916-8929, 2017.
- [18] J. Bai, Y. Li, Y. Yi, W. Cheng, H. Du, "PAPR reduction based on tone reservation scheme for DCO-OFDM indoor visible light communications," *Opt. Express*, 25(20): 24630-24638, 2017.
- [19] J. Wang, Y. Xu, X. Ling, R. Zhang, Z. Ding, C. Zhao, "PAPR analysis for OFDM visible light communication," *Opt. Express*, 24(24): 27457-27474, 2018.
- [20] J. Bai, C. Cao, Y. Yang, F. Zhao, X. Xin, A. H. Soliman, J. Gong, "Peak-to-average power ratio reduction for DCO-OFDM underwater optical wireless communication system based on an interleaving technique," *Opt. Eng.*, 57(8), 2018.
- [21] J. Armstrong, "Peak-to-average power reduction for OFDM by repeated clipping and frequency domain filtering," *Electron. Lett.*, 38(5): 246-247, 2002.
- [22] J. Bai, Y. Li, W. Cheng, Y. Yang, Z. Dua, Ya. Wang, "PAPR reduction for IM/DD-OFDM signals in underwater wireless optical," presented at the 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 2018.
- [23] T. Essalih, M. A. Khalighi, S. Hranilovic, H. Akhouayri, "Optical OFDM for SiPM-Based underwater optical wireless communication links," *Sensors*, 20(21), 2020.
- [24] M. Nassiri, G. Baghersalimi, Z. Ghassemloo, "Optical OFDM based on the fractional Fourier transform for an indoor VLC system," *OSA*, 60(9): 2664-2671, 2021.
- [25] C. Gabriel, M. A. Khalighi, S. Bourennane, "Monte Carlo-based channel characterization for underwater optical communication systems," *J. Opt. Commun. Networking*, 5(1): 1-12, 2013.
- [26] D. Chen, C. Li, Z. Xu, "Performance evaluation of OOK and system based on APD receiver," presented at the 16th Int. Conf. on Optical Communications and Networks (ICOON): 1-3, 2017.
- [27] S. K. Sahu, P. Shanmugam, "A theoretical study on the impact of particle scattering on the channel characteristics of underwater optical communication system," *Opt. Commun.*, 40(8): 3-14, 2018.
- [28] J. Alkhasraji, C. Tsimenidis, "Coded OFDM over short range underwater optical wireless channels using LED," presented at the Int. Conf. of Oceans, Aberdeen, UK, 2017.

- [29] T. Sri Sudha, G. Sasibhushana Rao, "Clipping based PMPR reduction techniques for LTE-OFDM systems," in Proc. Int. Con. on Intelligent Data Communication Technologies and Internet of Things (ICICI): 1023-1031, 2018.
- [30] Z. S. Hadi, B. M. Omran, "Peak-to-Average power reduction using repeated frequency domain filtering and clipping in OFDM," Int. J. Comput. Appl., 122(11): 6-10, 2015.

Biographies



Behzad Noursabbaghi received his B.Sc. degree from Chabahar Maritime University and M.Sc. degree from the University of Gilan, Iran, in 2016 and 2019, respectively, all in Electrical Engineering. His research interests are focused on Visible Light Communications and Underwater Wireless Optical Communications.

- Email: behzadnoursabbaghi@yahoo.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Gholamreza Baghersalimi received his B.Sc. degree from the University of Tehran, Iran, M.Sc. degree from Tarbiat Modares University, Iran, and Ph.D. degree from the University of Leeds, UK, all in Electrical Engineering. Now, he is an Associate Professor in the Department of Electrical Engineering, the University of Guilan. His research interests are in the area of fiber-optic

communication systems, optical wireless communications especially visible light communications.

- Email: bsalimi@guilan.ac.ir
- ORCID: [0000-0003-1305-1109](https://orcid.org/0000-0003-1305-1109)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://guilan.ac.ir/~bsalimi>



Atiyeh Pouralizadeh received her B.Sc. and M.Sc degrees from the University of Guilan, Iran, in 2017 and 2021, respectively, all in Electrical Engineering. Her research interests are focused on Data and Signal processing, Optical Wireless Communications (OWC), Visible Light Communications (VLC), Machine Learning (ML), Deep Learning (DL) techniques and Neural

Networks (NNs).

- Email: atiyehpouralizadeh@gmail.com
- ORCID: [0000-0002-7974-2271](https://orcid.org/0000-0002-7974-2271)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Ozra Mohammadian Chakhansar received her B.Sc. degree from Mohaghegh Ardabili University and M.Sc. degree from the University of Guilan, Iran, in 2015 and 2019, respectively, all in Electrical Engineering. Her research interests are focused on non-orthogonal multiple-access techniques, random access protocols, orthogonal frequency division multiple-access.

- Email: ozramohammadian@gmail.com
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Copyrights

©2023 The author(s). This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, as long as the original authors and source are cited. No permission is required from the authors or the publishers.



How to cite this paper:

B. Noursabbaghi, G. Baghersalimi, A. Pouralizadeh, O. Mohammadian, "PAPR reduction in OFDM UOWC system employing Repetitive Clipping and Filtering (RCF) method," J. Electr. Comput. Eng. Innovations, 11(2): 301-310, 2023.

DOI: [10.22061/jecei.2022.9061.569](https://doi.org/10.22061/jecei.2022.9061.569)

URL: https://jecei.sru.ac.ir/article_1818.html





Research paper

Centrality and Latent Semantic Feature Random Walk (CSRW) in Large Network Embedding

M. Taherparvar¹, F. Ahmadi Abkenari^{1,2,*}, P. Bayat¹

¹Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran.

²Faculty of Computer Engineering and Information Technology, Payam Noor University, Tehran, Iran.

Article Info

Article History:

Received: 03 October 2022
Reviewed: 10 November 2022
Revised: 30 December 2022
Accepted: 21 January 2023

Keywords:

BTM topic modelling
Centrality criteria
Deep learning
Network embedding
Social network analysis

*Corresponding Author's Email
Address:
Fateme.Abkenari@pnu.ac.ir

Abstract

Background and Objectives: Embedding social networks has attracted researchers' attention so far. The aim of network embedding is to learn a low-dimensional representation of each network vertex while maintaining the structure and characteristics of the network. Most of these existing network embedding methods focus on only preserving the structure of networks, but they mostly ignore the semantic and centrality-based information. Moreover, the vertices selection has been done blindly (greedy) in the existing methods.

Methods: In this paper, a comprehensive algorithm entitled CSRW stands for centrality, and a semantic-based random walk is proposed for the network embedding process based on the main criteria of the centrality concept as well as the semantic impact of the textual information of each vertex and considering the impact of neighboring nodes. In CSRW, textual analysis based on the BTM topic modelling approach is investigated and the final display is performed using the Skip-Gram model in the network.

Results: The conducted experiments have shown the robustness of the proposed method of this paper in comparison to other existing classical approaches such as DeepWalk, CARE, CONE, COANE, and DCB in terms of vertex classification, and link prediction. And in the criterion of link prediction in a Subgraph with 5000 members, an accuracy of 0.91 has been reached for the criterion of closeness centrality and is better than other methods.

Conclusion: The CSRW algorithm is scalable and has achieved higher accuracy on larger datasets.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Life without social media is unimaginable in today's world. The media in different forms play a prominent role in our vitality. In fact, media is one of the influential pillars of information constitution and attainment. In the meantime, social networks have become the foundation of borderless communication while their remarkable impacts on people's lives and continuance are immeasurable. One source of big data with its three mainstream characteristics of volume, velocity, and

variety originates from social networks. Moreover, in the shadow of social behaviors, these data may include human collaborations, interactions, and communications that are categorized as highly nonlinear and complex problems.

A graph as a structure of vertices and edges is an adequate tool for representing high-dimensional data such as networks of human collaborations, wireless sensors, research paper citations, etc. The scale of complex networks ranges from hundreds to billions of vertices, making the efficient network analysis process

challenging. A very suitable solution for this problem is the concept of network embedding, in which each node is mapped to a low-dimensional vector while the global and local characteristics of the network will be maintained [1], [2].

Also, hidden information related to vertices should be extracted based on each network's characteristics. This information could indicate the specifications that highly affect the correct analysis of the network. For example, the neighborhood information and semantics of vertices (in different orders) could help to properly advance the random walk in the network embedding process. Therefore, the vector representations of nodes could be employed in tasks such as network analysis and classification, node clustering, and link prediction [3]-[5]. Considering the high potential of network embedding, there are two main challenges in this domain. First, the scale of real-world networks is big data. Hence, the learning task may take months or fail. Second, network data are often complex and high-dimensional, which makes it very challenging to design a suitable model with the aim of preserving the network structure. The network embedding process consists of two steps: First, sampling the network data as a corpus and then embedding it using the Word2vec approach [6]. Meanwhile, DeepWalk is a pioneering method among network embedding approaches [1].

In networks that have been constructed based on real-world scenarios, most vertices have a low degree with only a few nodes of a high degree. So, criteria such as the degree or different centrality approaches of each node alone are not optimal indicators for selecting a vertex in the random walk process.

Recently, many efforts have been devoted to the development of network embedding algorithms. Early research mainly emphasized reducing the dimensions of the network based on the feature extraction process. However, the high cost of calculating the adjacency matrix of large-scale networks is a major challenge in these approaches. Recently, inspired by the success of Word2Vec, interesting research endeavors have been conducted in network embedding frameworks that result in DeepWalk [1], Node2vec [2], and LINE [7] approaches. These classical methods have shown promising performance in many machine learning applications.

The DeepWalk method creates random walks for each vertex and uses them as background information to learn the representations of the vertices [1]. Node2Vec extends DeepWalk by utilizing two predefined parameters to control the random walk method, which is a trade-off between breadth-first and depth-first search traversal approaches [2]. DeepWalk and Node2Vec face the problem of insufficient sampling in dense networks, so some local patterns are not reflected in these

perspectives. In addition, some research focus on the examination and extraction of vertices' features, such as text [8] or labels [10].

Our proposed approach of centrality and semantic-aware random walk (CSRW) in this paper for network embedding, employs one criterion among the set of degree, degree centrality, load centrality, closeness centrality, and eigenvector centrality. After loading the corresponding value of each node, the random walk generation process will begin in the following manner: The first node is randomly selected to start the process then the next node is selected from the neighbours of the previous vertex in such a way that the average selected criterion of that node and the nodes of its next hops is more optimal than the rest of the neighboring nodes (the number of hops is set from one to four).

As depicted in Fig. 1, if we consider the number of orders to be equal to 4, node No. 3 has a higher selection priority than nodes No. 2 and No. 4 in terms of the average degree criterion.

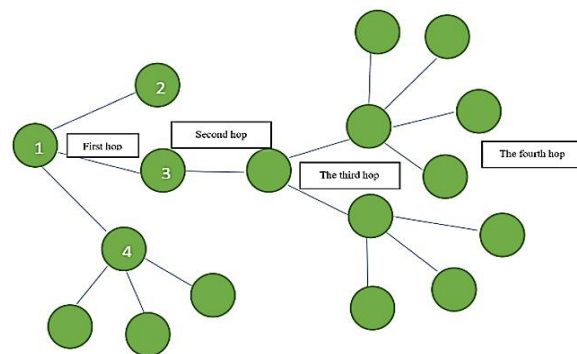


Fig. 1: The process of selecting nodes in the random walk process.

Therefore, a random walk is generated for all nodes in the network graph. Finally, the generated walk sequences are sent to the Skip-Gram model in the form of text strings to produce display vectors. Simultaneous with the random walk generation process, the textual content of the adjacent vertices is collected in a single document per node in the sequence of random walks. Then, the pattern in this document is extracted using a topic modeling approach in a word pairs fashion. Finally, the output of the context analysis constituent is linked with the display vector obtained from the Skip-Gram model.

The contributions and innovations of this paper are as follows:

- In this paper, a centrality and semantic-based random walk process entitled CSRW is proposed for the network embedding process with the aim of overcoming the problem of blind selection of vertices.
- The novelty of the proposed CSRW approach is that it focused on two dimensions, different criteria, and the semantic information and connection of vertices.

- The proposed approach is scalable. The increase and decrease in the size of the input network will not affect the effectiveness of the CSRW approach.
- Regarding the fact that networks from real-world scenarios have low degree vertices as a majority and high degree vertices as a minority, the greedy approaches for node selection cannot reflect the local and global properties of the network. To overcome this problem, the CSRW approach is proposed for the vertex selection process considering the effect of the evaluation-based criteria scores of neighboring nodes. So, another innovation of the algorithm is the better reflection of the local and global characteristics of the nodes in the network.

The method presented in this paper has been evaluated in several real-world networks such as reference networks.

The experimental results have proven that the proposed algorithm outperforms other community and text-based approaches in the fields of vertex classification, and link prediction.

The rest of this paper is organized as follows. Section 2 briefly presents previous research and algorithms related to the random walk strategies in network and topic modeling approaches. In section 3, the proposed algorithm of this paper is discussed with formal expressions along with mappings and explanations. Section 4 contains the results of the conducted experiments for verification of the proposed methods' effectiveness. Moreover, the parameter sensitivity of the proposed approach is analyzed. Finally, section 5 summarizes the discussion on the proposed method's framework, conclusions, and future work.

Related Work

Feature learning has been widely used in different filed as computer vision [10], [11] and natural language processing [12], [13]. With the development of the Internet, large amount of data are produced by complex networks. Unstructured data is a challenging issue in network embedding, and many methods are proposed to learn features of network.

In recent years, deep learning methods have been employed as an alternative to feature vector-based learning. These methods have used deep learning to learn representation vectors. They generate random walks with the help of different network search strategies and provide the input as contextual information to the Skip-Gram model [1].

Perozzi et al. (2014) confirmed the similarity between vertices in the random walk sequence in terms of the words in their contexts.

They proposed a DeepWalk model that uses Skip-Gram architecture to extract feature vectors from a sequence of

random walks [1]. DeepWalk was the first method that used the Skip-Gram model to generate feature vectors. Although the DeepWalk method has shown good performance in vertex classification, because of not considering the neighborhood information of higher orders, it could not maintain the global structure of the network.

Tang et al. (2015) proposed the LINE algorithm in which they employ first and second-order neighborhoods together with preserving local information to learn node representations.

In the LINE method, two independent functions are defined for the first and second-order neighborhoods. The LINE method and the DeepWalk are unable to learn the representation vector for boundary nodes in the network [7].

Grover et al. (2016) proposed Node2Vec random walks based on strategies such as depth-first and breadth-first traversal approaches. This algorithm considers only the second-order neighborhoods and cannot reach the nodes whose distance from the starting node of the random walk is more than two. Therefore, like the DeepWalk method, it cannot maintain the global structure of the network [2].

Chen et al. (2019) presented the lateral information network embedding, which defined a semantic neighborhood to model the shape of each node, then applied random walks to explore this neighborhood [14]. Wang et al. (2016) proposed a deep model with a semi-supervised architecture entitled SDNE, which maps data to a nonlinear hidden space and can simultaneously optimize first-order and second-order neighborhoods [15].

Community detection in the network is one of the common methods in network embedding. Li et al. (2019) presented a network embedding method based on evolutionary algorithms that can maintain the neighborhood and communities of vertices in the network by optimizing a multi-objective function [16]. Chen et al. (2016) proposed a method with valuable group information for large-scale networks by considering the internal structures of groups and the information between them [17]. Kikha et al. (2018) presented an algorithm called CARE, which uses the Louvain community detection method to detect the communities of nodes in the network and construct a sequence of random walks. The CARE algorithm employs the Louvain method to discover communities [18]. Wang et al. (2020) proposed CANE algorithm, which describes the embedding of community-aware network through adversarial training. The CANE method minimizes the community assignment error with the aim of improving network embedding [19].

Criteria associated with vertices in the network have

many usages in network embedding. Shi et al. (2019) proposed a network propagation embedding method with the aim of overcoming limitations such as the tendency to select high-degree nodes [20].

The drawback of their research is neglecting the global structure of very complex networks in the random walk. Chen et al. (2019) presented a generalizable model that uses both edge and node centrality information to learn low-dimensional vector representations that can maintain different vertex centrality information [21]. Zhao et al. (2019) proposed an integrated framework for social and behavioral recommendations with network embedding and introduced a joint network embedding approach as a pre-training step for hidden user representations [22].

Li et al. (2019) represented an unsupervised network embedding model for encoding edge relationship information, thus feature representation of vertices can be further captured [23].

The aim of topic models is to utilize observed text in order to infer hidden topic distribution. Some researchers have used topic models regarding authors' collaboration networks to infer the research community [24]. Mai et al. (2008) have presented a general solution of text mining with a network-based structure entitled NetPLSA for topic optimization in the network [25].

Wu et al. (2019) proposed a multi-task dual attention LSTM model to learn network representations for specific applications [26]. The model can capture structure, content and label information, then adjust vertex representations according to the downstream task. Yuan et al. (2019) proposed an algorithm called COANE, which uses the LDA topic modeling approach to detect the community of vertices in the network and construct a sequence of random walks [27].

In our other paper (2022), the DCB algorithm is proposed as a network attribute embedding framework that includes the contextual information of vertices in the network embedding process. In this research, the topic modeling based on word pairs has been utilized to investigate the relationship and semantic analysis of nodes [28]. Chen et al. (2022) proposed semantic feature-aware embedding via optimized random walk and paragraph2vec. By using textual semantics instead of contextual semantics, this method has been able to achieve higher accuracy in complex networks than the deep walking method [29].

In the recent works, random walk with the aim of embedding the network usually uses a single criterion, for example, random selection, centrality criterion, semantic analysis of texts, communities, etc. In the methods presented in this article to improve the quality of network embedding, two parameters of centrality criteria and semantic features are used.

In this paper, a random walk generation based on centrality and semantic information (CSRW) is proposed for network embedding. This method consists of two sections.

The first part includes the selection of nodes based on the criteria of centrality and the semantic influence of nodes in the network embedding. As mentioned before, the greedy selection of the highest degree or the largest centrality criterion for each node cannot reflect the properties of the network locally and globally. To solve this problem, a new approach is presented in this paper in the selection of vertices.

In the random walk generation process, a node is selected with the optimal value of the average evaluation criterion of itself and its neighboring nodes in the next order (in one to four hops).

In the next section, we will describe the proposed algorithm of this research:

Methodology

In this section, we will present a detailed description of the proposed CSRW algorithm and a brief overview of the existing node evaluation criteria.

In this paper, an algorithm called CSRW is proposed that generates a random walk based on the centrality and semantic information in network embedding, which is based on the average of the degree criteria or different centralities for the neighboring nodes of each vertex (in different orders) and semantic analysis of feature vectors of network nodes.

Among the important features of social networks that are able to maintain the local and global structure, we consider the average information of different degree or centrality criteria for the neighboring nodes to a vertex (in different hops) and the semantic analysis of textual features specific to each node as the dominant ones.

Suppose $G = (V, E, T)$ is a feature graph where V is a set of vertices; $E \subseteq V \times V$ edges represent the relationships between the vertices and T represents the text content of the vertices; In particular, the textual information of each vertex $v \in V$ is related to the word sequence $T_v = (W_1, W_2, \dots, W_n)$ where $n_v = |T_v|$.

Network embedding tries to create a matrix of features with low dimensions called $\phi \in \mathbb{R}^{|V| \times d}$. $d \ll |V|$ defines the dimensions of the hidden representation space d is less than $|V|$.

The Skip-Gram model has been used to obtain the best mapping performance of the Φ function.

A. Nodes Evaluation Criteria

Most of the classical methods for network embedding are focused on maintaining the network structure and generally do not pay proper attention to utilizing and collecting information related to the criteria and centrality nuggets of different vertices. Metrics with the

ability to measure the importance of each node separately are practical in many applications. Of course, a greedy approach based on different criteria in selecting nodes is not able to reflect the characteristics of the network properly. In this paper, the main focus is on how to use criteria such as degree and different centrality-based metrics.

So in this domain, we employ various criteria such as degree criteria, degree centrality, closeness centrality [30], eigenvector centrality [31], and load centrality [32]. Our proposed CSRW model in this paper is able to maintain all mentioned centrality-based measures in its structure. In this paper, representation learning aims to preserve network structure, centrality information, and semantic connection between nodes. To this aim, we proposed a general model that employs various centrality criteria.

We briefly explain each criterion in continue:

I) Degree of node criterion

Simply, the number of connections of a node to other vertices in the graph is called the degree of that node. The centrality criterion means how important a node is in a social network. In the graph discussion, there are different types of centrality-based metrics, which can be used to identify impactful nodes in the network.

II) Degree centrality criterion

In a network graph, degree centrality is measured by the total number of direct edges with other nodes according to the basic formula in (1) [30]:

$$C_d(N_i) = \sum_{j=1}^n X_{ij} (i \neq j) \quad (1)$$

X_{ij} is a link between node i and j . with the increase in the size of the networks, in order to reduce the impact of the network size on this centrality criterion, formula (1) was standardized as formula (2) [30]:

$$C'_d(N_i) = \frac{\sum_{j=1}^n X_{ij}}{(n-1)(n-2)} (i \neq j) \quad (2)$$

$\sum_{j=1}^n X_{ij}$ represents the number of direct edges connected to node N , and n is equal to the total number of nodes in the network graph. Based on Fig. 2, the number of direct edges connected to node A is equal to 2, so, the value of C_d is equal to 2, and after standardization, the value of C'_d will be equal to 0.167 [30].

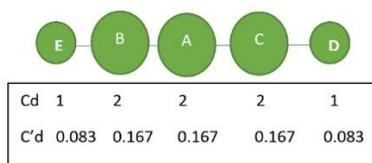


Fig. 2: An example graph representing the degree centrality metric [30].

III) Closeness centrality criterion

In a graph, the closeness centrality of a vertex is the average length of the shortest path between the node and all other vertices in the network graph. Therefore, the more central a node is, the closer it is to other vertices.

Equation (3) is the basic equation of closeness centrality, which is equal to the total number of steps from node N to all nodes of the network [30].

$$C_c(N_i) = \frac{1}{[\sum_{j=1}^n d(N_i, N_j)]} (i \neq j) \quad (3)$$

According to Fig. 2, node A is located next to nodes B and C. Therefore, the distance of node A to these two nodes is equal to one order, and its distance to nodes E and D are equal to two orders. Hence, the closeness centrality value for node A is equal to $C_c(N_i) = \frac{1}{1+1+2+2} \approx 0.167$, $C'_c \approx 0.67$. The results for the rest of the nodes are shown in to Fig. 3 [30]:



Fig. 3: An example graph representing the closeness centrality metric [30].

IV) Eigenvector centrality criterion

The eigenvector is the largest eigenvalue of an adjacency matrix, which can be a good measure of network centrality metric calculation. Unlike the degree-based metrics, which weight each edge equally, the eigenvector weights connections based on their centrality. The centrality of the eigenvector can be calculated as the weighted sum of not only direct edges but also indirect connections of any length.

Formula (4) illustrates the eigenvector centrality metric. It describes the centrality of the eigenvector x in two ways, as a matrix equation, and as a summation. The centrality of a node is proportional to the sum of the centralities of the nodes connected to it. λ is the largest eigenvalue of A and n is the number of vertices [31]:

$$A_x = \lambda_x, \lambda_{x_i} = \sum_{j=1}^n a_{ij} x_{ij}, i = 1, \dots, n \quad (4)$$

V) Load centrality criterion

The measure of load centrality for each node is the fraction of the shortest paths that pass through that vertex. This criterion is obtained based on formula (5):

$$LC(v) = \sum_{s,d \in V} \theta_{s,d}(v) \quad (5)$$

The load centrality criterion employs an algorithm to calculate paths with the minimum weight between pairs of nodes (s, d). The variable $\theta_{s,d}$ is the value sent from node s to node d . It is assumed that this value is always

transmitted to the next node with the lowest value. $\theta_{s,d}(v)$ is the total amount of value sent from node v . It is normally assumed that $s \neq d$ and $d \neq v$ [32].

B. Centrality and Semantic-Based Random Walk for Network Embedding

In this paper, the random walk algorithm based on the

centrality and semantic information entitled CSRW is proposed for the network embedding process.

The diagram of the CSRW template is illustrated in Fig. 4. And the flowchart of the random walk is illustrated in Fig. 5.

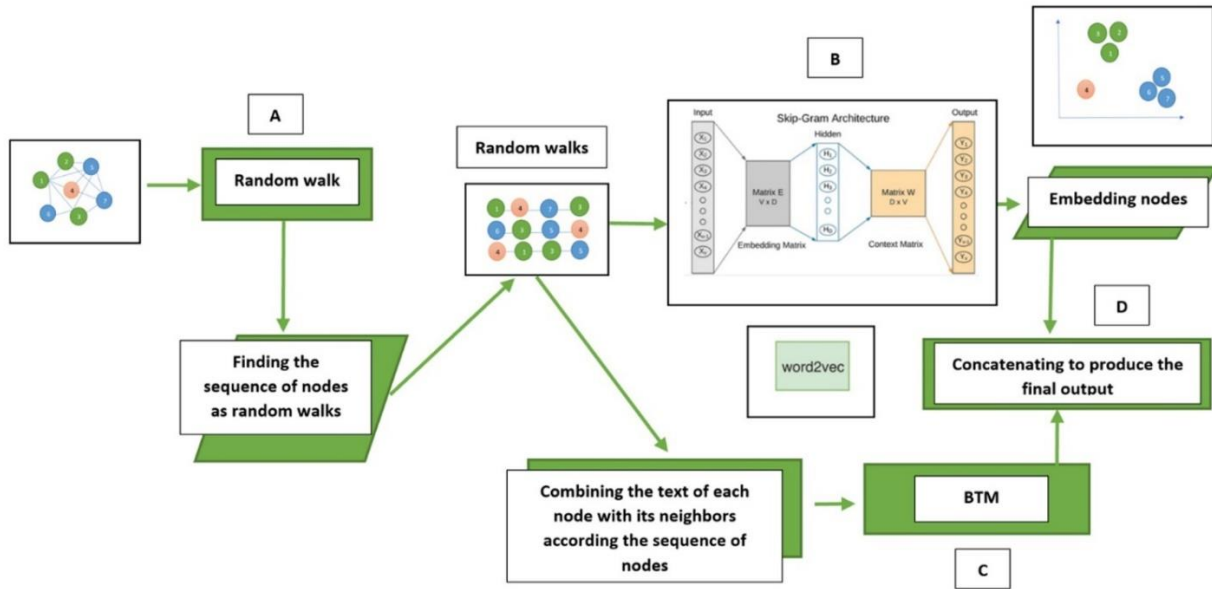


Fig. 4: Diagram of the proposed method of this paper (CSRW).

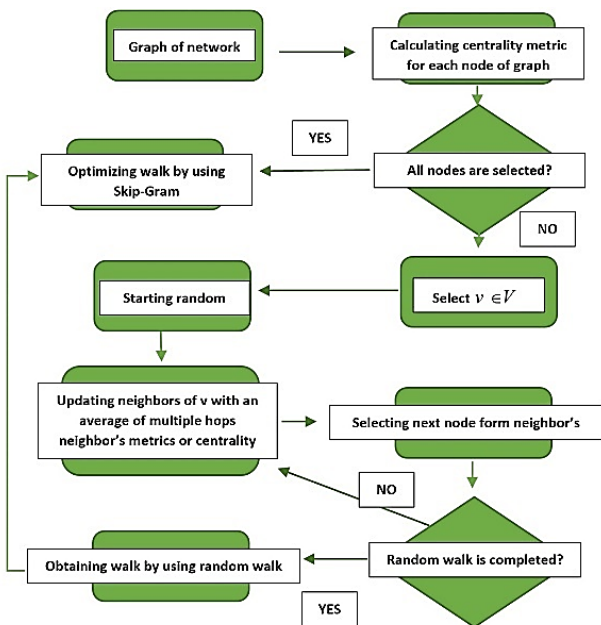


Fig. 5: Random walk flowchart process in CSRW method.

The diagram of CSRW method is illustrated in Fig. 4. This diagram composes four main parts:

- Part A: The aim of the random walk part is to find a sequence of nodes in such a way that the local and

global characteristics of the nodes in the network are preserved. The random walk used in the CSRW method is shown in the flowchart in Fig. 5.

- Part B: A sequence of vertices $s = (v_1, v_2, \dots, v_{|s|})$ obtained by a random walk through the network is considered a word sequence and each vertex in the sequence is considered a word. In the next step, the CSRW can obtain embedding nodes using the Skip-Gram model, which aims to maximize the average log of observing a vertex:

$$\max_{\theta} \frac{1}{|s|} \sum_{i=1}^{|s|} \log P_r(\{v_{i-w}, \dots, v_{i+1}, \dots, v_{i+w}\} | v_i) \quad (6)$$

- Part C: The BTM topic model is used for the contextual analysis of the collected documents from neighbors. In this part, the text content of the neighboring nodes is combined and analyzed. Latent semantic features are extracted in this section.
- Part D: In this part the embedding nodes and the semantic features are concatenating to produce final output.

In this section, the centrality and semantic-based random walk algorithm for network embedding (CSRW) are presented. The main purpose of this algorithm is to create representation vectors for network vertices. In

CSRW, the evaluation of nodes as the next selected vertex in the random walk sequence is based on the average of criteria such as degree, closeness, eigenvector and load centrality-based metrics. The evaluation is in such a way that each node is verified by the average value of its criterion and its neighboring nodes in the next order (one to four hops).

Among the evaluated neighboring nodes, the vertex that has a more optimal value with a higher priority for selection is chosen as the next vertex in the random walk process by employing the random selection of the Roulette Wheel genetic algorithm. The proposed method in this paper solves the problem of blind selection of nodes.

In this scenario, a node is selected that has a higher right to be chosen than other neighboring vertices in the future random walk sequence. Algorithm 1 depicts the steps of CSRW.

Algorithm 1: Framework of CSRW

Input: graph: $G(V, E, T)$; Window size: w ;
 representation dimension: d ;
 walks per vertex: γ ; walk length: l .

Output: matrix of network representations: $\phi \in \mathbb{R}^{|V| \times d}$

- 1: $B = \text{Benchmark measurement}(G)$
- 2: Sample ϕ from $u^{|V| \times d}$
- 3: **for** $i = 0$ to γ **do**
- 4: $\vartheta = \text{Shuffle}(V)$
- 5: **for** each vertex $v \in \vartheta$ **do**
- 6: $W_v = \text{Random Walk}(G, v, B, l)$
- 7: $D_v = \text{ContextAggregation}(G, v, B)$
- 8: $\text{SkipGram}(\phi, S_v, w)$
- 9: **end for**
- 10: $Pr_t = \text{BTM}_t(D)$
- 11: **end for**
- 12: $\phi = \phi \oplus Pr_t$
- 13: **return** ϕ

Fig. 6: CSRW algorithm.

As shown in Fig. 6, in line 1, the desired criterion is measured for all network nodes. In line 2, Before learning the representation vectors for the network nodes, the U matrix is randomly generated to produce the nodes' representation vectors in the next section. The algorithm is now able to learn the final representation vectors in lines 3 to 11.

Before repeating the grid nodes, in line 4, at the start of each pass, the vertices are shuffled to prevent the node visiting order in the ϕ . The core task of the network embedding is done in line 6, where a random walk is generated for the selected vertex. Line 7 shows the semantic analysis section of the textual information of the

nodes, which is explained in algorithm 3. In line 10, we aggregate text feature of a vertex to that's of its neighbors as D and input them into the text-based BTM model. Finally, the generated paths and the results of the node's semantic analysis are used to update the node's presentation in line 12.

Fig. 7 illustrates the *Random Walk* algorithm for the CSRW algorithm. A random walk starting at node v is denoted by W_v . A random walk sequence for node v can be represented by random variables such as $W_v^1, W_v^2, \dots, W_v^k$.

To create a customized random walk starting from node v , first, all the neighbors of that node are extracted. Then a random variable r between 0 and 1 is created. α is random variable to select from neighbors.

If random variable is less than α then the criteria related to the neighboring nodes of the current node are checked.

This section in the random walk can have four types: In each type, the average criterion of the nodes of that order neighbors is examined for example in the second type, the average criterion of neighbor nodes up to the second order is examined.

Finally, a node is randomly selected from the final list obtained for the neighbors of the current vertex using the genetic algorithm of Roulette Wheel method.

Based on the algorithm illustrated in Fig. 7, random walks are generated independently. Hence, the current algorithm can be parallelized to speed up the embedding process.

In addition, if some new nodes are added or removed from the network, the random walk is calculated only for the new vertices.

The contextual analysis of the nodes in Fig. 6 are in performed in lines 7 and 10. In line 7, $D_v = \text{ContextAggregation}(G, v)$ in which node v and network graph G are used in the text aggregation section to collect text documents related to node v and its first-order neighbors.

The text aggregation section is shown in Fig. 8 entitled Context Aggregation Algorithm. In Fig. 6, line 10, $Pr_t = \text{BTM}(D)$ of the BTM topic model is used for the contextual analysis of the collected documents D_v .

The text type of nodes in social networks is short. In the Context Aggregation algorithm, the text content of the neighboring nodes is combined and analyzed. In another paper (2021), it has been shown that the BTM model has a better performance in relation to short texts than other topic models [33].

Finally, in Fig. 6, line 12, the $\phi = \phi \oplus Pr_t$ vectors obtained from the random walk analysis and the contextual analysis of the nodes are combined to produce the final ϕ vectors. Fig. 8 shows the process of summarizing the text. To reduce the deviation between

the posterior population distribution and the actual population distribution, it should be considered that the length of a document should not be less than the total number of documents [34].

$$v.\log|v| \ll \text{length}(D_v) \quad (7)$$

Line 5 of Fig. 8 shows how to assign probabilities to neighboring vertices.

The selection of the next vertices is based on the selection strategy of the Roulette Wheel genetic algorithm.

Algorithm 2: Random Walk

Input: graph: $G(V, E, T)$; Source node of RW: v_i ;

Benchmark measurement of graph: B ;

walk length: l ; Random variable to select from

neighbours: α .

Output: A path with max length l : **Random Walk**

```

1: initialize RW with  $v_i$ 
2: While length(path) <  $l$ 
3:   if the current node has neighbours
4:     if ((random(0,1) =  $r$ ) <  $\alpha$ )
5:       For a node in current_neighbours:
6:         For nodes in hop_1's node:
7:           list= Benchmark measurement list of
              nodes
8:         For nodes in hop_2's node:
9:           list= Benchmark measurement list of
              nodes
10:        For nodes in hop_3's node:
11:          list= Benchmark measurement list
              of nodes
12:        For nodes in hop_4's node:
13:          list= Benchmark measurement
              List of nodes
14:        Calculate the average of the list and add it
              into score list.
15:        next node for RW= Random from members
              Sorted score list
17:   else:
        Select another  $v_j$  at random from members
         $v_j$ 's current_neighbours
17: else:
        backtrack in the path and select the last
        node which has neighbors that are in the
        path
18: end While.
```

Fig. 7: Random walk for CSRW algorithm.

Algorithm 3: Context Aggregation

Input: graph: $G(V, E, T)$; Source node of RW v_i ;

Benchmark measurement of graph: B ;

Output: the contextual text information: D_v

```

1: Initialize  $D_v$  with  $T_v$ 
2: While length( $D_v$ ) <  $\gamma.\log|v|$  do
3:   if current vertex has neighbors, then
4:     for each neighbor vertex  $u$  of  $v$  do
5:       list= Benchmark measurement list of nodes
6:     end for
7:     select a vertex  $u$  list based on Roulette Wheel
8:      $D_v = D_v \oplus T_u$ 
9:   else
10:     $D_v = D_v \oplus T_v$ 
11:   end if
12: end while
13: return  $D_v$ 
```

Fig. 8: The algorithm combines the semantic of each vertex with its neighbors.

Result and Discussion

In this section, methods, experimental data sets, and parameter settings will be described. Then, the proposed algorithm of this paper is evaluated in two supervised learning tasks such as vertex classification and link prediction, and will be compared with other existing approaches.

As described in the literature review section, DeepWalk is an advanced network embedding algorithm that uses natural language processing for network embedding. CARE is an algorithm for community-aware network embedding and obtains community-related information using the Louvain method, and finally, random walks are converted into low-dimensional representation vectors using the Skip-Gram model. Also, CONE and COANE are algorithms related to network embedding that obtain community information with topic models and convert the generated random walk into low-dimensional representation vectors by using the Skip-Gram method.

DCB algorithm is an embedding of network attributes that can include the information and content of vertices' text in the network embedding process. This paper used the BTM topic model to examine the relationship and semantic analysis of nodes. In this section, we will compare our new method of CSRW with the classic methods of DeepWalk, CARE, CONE, COANE, and DCB on the following datasets of Cora and DBLP.

1) Data set

The Cora dataset contains 2708 machine learning articles from 7 classes and 5429 edges among the articles.

Each vertex represents an article and the citation relationships between documents form a typical complex network [35].

DBLP V12 contains 4 million articles and 45 million edges between them, and the date of this dataset is 09/04/2020. In this paper, two subgraphs with the number of 2000 and 5000 nodes are used for implementation from DBLP dataset [36].

Also, the content of the title of each article is used as the feature information. In Cora data set, the titles extracted from the main dataset had missing values, and in this regard, the link of the articles was used as a replacement. The characteristics of the dataset are shown in Table 1.

Table 1: The dataset used for the experiments

	Data set	Nodes	Edges	Labels
(1)	Cora	2708	5429	7
(2)	DBLP_2000	2000	4013	4
(3)	DBLP_5000	5000	11587	4

The display vector dimension is set to $d = 128$ for all datasets above. For DeepWalk, the number of walks is set to 20, the walk length is set to 20, and the window size w is set to 10.

In order to provide a fair comparison, the parameter settings used for DCB, CARE, CONE, COANE, and CSRW correspond to the values used for the DeepWalk. In all the above cases, the value of the variable k , the number of topics, is considered equal to 14.

II) Comparison based on link prediction metric

Link prediction is a task to estimate the probability of links between nodes in a graph. It is a supervised and semi-supervised learning task. The model is trained using a subset of link that have truth labels. For predicting link existence, the truth may just be whether the edge exists in the original data, rather than a separate label. Link prediction can also be done as a downstream task from node representation learning, by combining node embedding vectors for the source and target nodes of the edge and training a supervised or semi-supervised classifier against the result.

A standard evaluation criterion, the area under the curve (AUC), is adopted here, which indicates the probability that potentially connected vertices are more similar than unrelated ones.

The CSRW algorithm has been implemented in four different versions from one to four hops and each of them is implemented based on the criteria of degree, degree centrality, closeness centrality, eigenvector centrality and load centrality. In the implementation of the first type,

the next node is randomly selected among the neighboring valued nodes. In the second type, the neighboring nodes are valued in such a way that the average value of each node is calculated with its second-order neighboring nodes.

In the third type, the averaging process continues until the neighbors of the third order, and in the fourth type, it continues until the fourth order. So, the CSRW algorithm has four execution types for each measurement criterion.

In this section, regarding Fig. 9 to Fig. 14, it should be noted that the bar graphs related to closeness centrality, degree centrality, degree, eigenvector centrality, and load centrality are calculated based on average values obtained from the implementation of four types of CSRW. For example, the degree bar graph is the average of the values for the execution of CSRW_Degree_Hop1, to CSRW_Degree_Hop4.

Regarding the datasets of DBLP_2000, DBLP_5000, and Cora, two different implementations have been performed with the aim of showing the high importance of context analysis in network embedding with no semantic analysis and impact of the text and by considering the semantic analysis of the nodes' context.

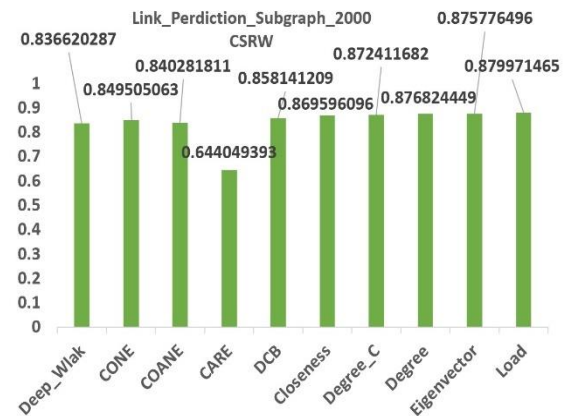


Fig. 9: ACU scores on link prediction criterion for DBLP_2000 for CSRW regardless of context.

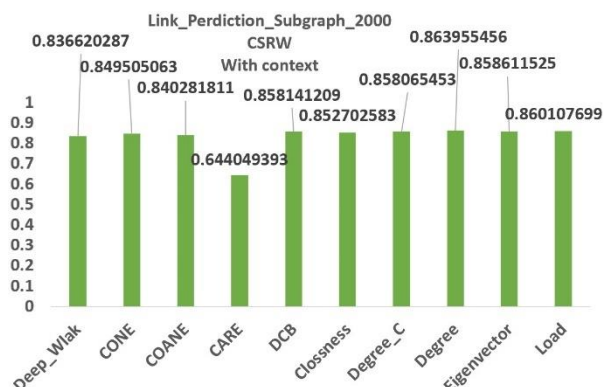


Fig. 10: ACU scores on link prediction criterion for DBLP_2000 for CSRW with semantic impact consideration.

Fig. 9 and Fig. 10 show the values obtained from the implementation of the CSRW algorithm without and with considering the semantic information in the DBLP_2000 dataset regarding the link prediction criterion. It can be seen from Fig. 9 and Fig. 10 that the last five bars which belong to our proposed method achieve the highest scores in the link prediction metric.

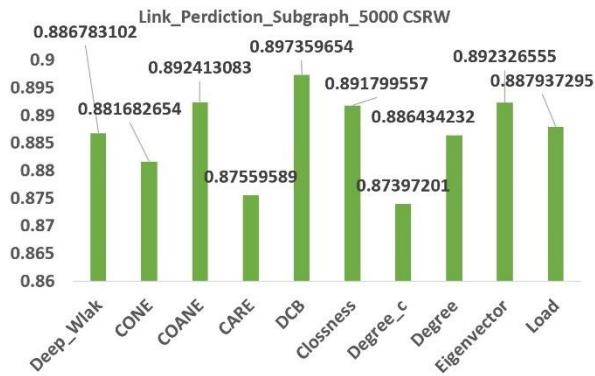


Fig. 11: ACU scores on link prediction criterion for DBLP_5000 for CSRW regardless of semantic.

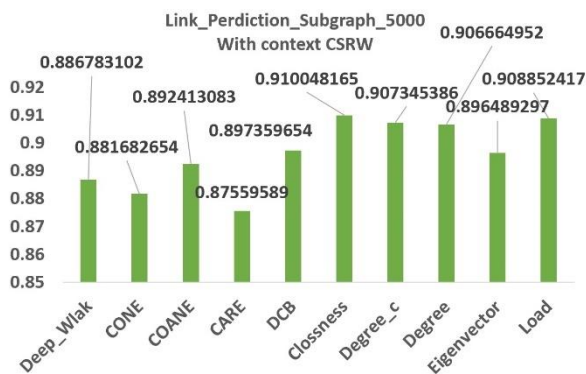


Fig. 12: ACU scores on link prediction criterion for DBLP_5000 for CSRW considering the semantic impact.

Based on Fig. 11 and Fig. 12 of CSRW implementation without and with context consideration in the DBLP_5000 dataset, it is clear that the semantic impact of the nodes in the accuracy obtained from the network embedding is very important.

As shown in Fig. 12, regarding the implementation of the CSRW algorithm, considering the context analysis in the DBLP_5000 dataset, the highest accuracy level of 0.91 has been obtained with CSRW_Closness centrality. Also, the methods of degree, degree centrality, and load centrality have obtained accuracies above 0.9. That is, the average performance of CSRW_Closness centrality in the first to fourth order is equal to 0.91 and has performed better than other classical methods. In

Fig. 16, the accuracies of different CSRW runs for different metrics have been illustrated.

Based on Fig. 9 to Fig. 12 of the CSRW implementation without and with context consideration in the DBLP_2000 dataset and DBLP_5000 dataset, it is clear that the semantic impact of the nodes and Scalability of the dataset in the accuracy obtained from the network embedding is very important. the accuracy obtained has improved with the increase in the size of the data set.

Based on Fig. 12, considering the context analysis in the DBLP_5000 dataset, the highest accuracy has been obtained with different centrality criteria. so, the obtained accuracy will be better, when both centrality criteria and context analysis are used.

By examining these results, it can be seen that the CSRW algorithm is scalable considering the semantic analysis and has achieved the highest accuracy compared to other classical methods.

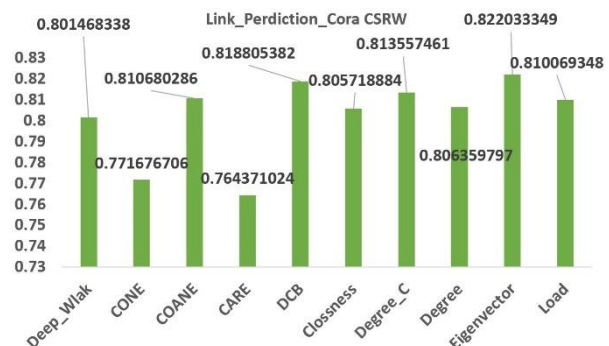


Fig. 13: ACU scores on the link prediction criterion for the Cora dataset for CSRW without considering the semantic.

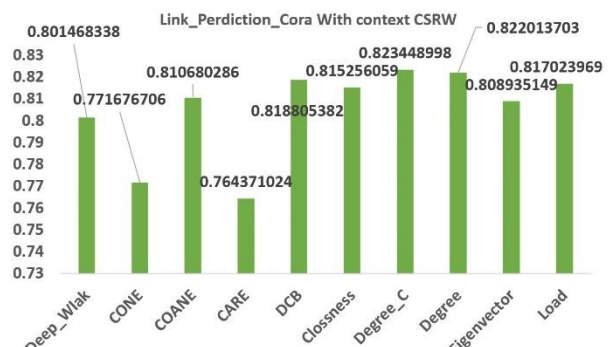


Fig. 14: ACU scores on the link prediction criterion for the Cora dataset for CSRW with semantic consideration.

In the Cora dataset based on Fig. 13, it can be observed that all the new methods, except the eigenvector centrality method, have not achieved high accuracy. But according to Fig. 14, it can be seen that under employing the semantic analysis of nodes in CSRW, all the new methods have reached high accuracy in comparison to the classical methods. Parts 1, 2, 3, 4 and 5 of Fig. 15 shows that the execution of the four types of CSRW algorithms considering the semantics for all five criteria. For example, part 1 shows the accuracy of the

implementation of the CSRW_Clossness algorithm for one to four hops.

This comparison has been made with the aim of showing that increasing the hops to some extent increases the accuracy of the implementation. Increase the number of hops in the neighbor valuation process increases the number of loops in the CSRW algorithm and finally the time complexity increases. To increase the number of hops, you should pay attention to the following points:

- Memory and CPU power of the system to run the algorithm.
- The size of the network or dataset.

Finally, in order to achieve ideal accuracy, a trade-off must be made between two computational complexity and algorithm execution times. So, increasing the number of orders or hops in the CSRW algorithm should be continued until the improvement is achieved.

According to part 3 of Fig. 15, in the degree criterion, the accuracy of the algorithm has been improved by increasing the number of hops. And in parts 1 and 5 of Fig. 15 in closeness and load centrality the accuracy in lower

hops has worked better. So, it is a better criterion to reach higher accuracy in lower hops with less complexity and time.

Fig. 16 illustrates the fact that with a high-dimensional network, i.e., a DBLP_5000, the CSRW embedding methods have achieved higher accuracy in lower orders in comparison to Fig. 15.

III) Comparison based on Vertex Classification

Vertex classification is used to evaluate the quality of the obtained representations, where L2-regularized logistic regression is employed as a supervised classifier. In the experiments, the training size of input datasets is increased from 10% to 90%. Precision, recall, Micro-F1, and Macro-F1 measures are applied to evaluate performance of different algorithms.

The experiments are repeated for 10 times and the average classification accuracy with different training ratio on the Cora dataset is shown in Fig. 17, Fig. 18, Fig. 19 and Fig. 20. The reason for choosing the Cora dataset at this stage is the number of seven tags that the articles are classified based on them. CSRW performs significantly better than other methods.

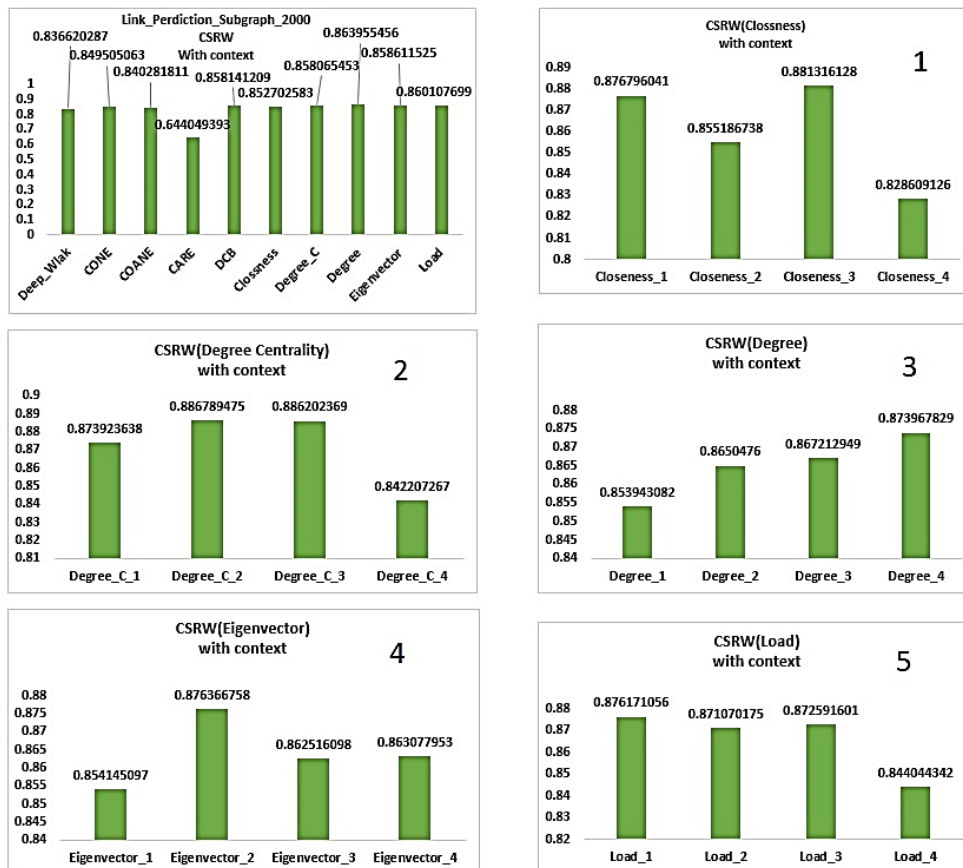


Fig. 15. Implementation of different variants of CSRW with semantic consideration.

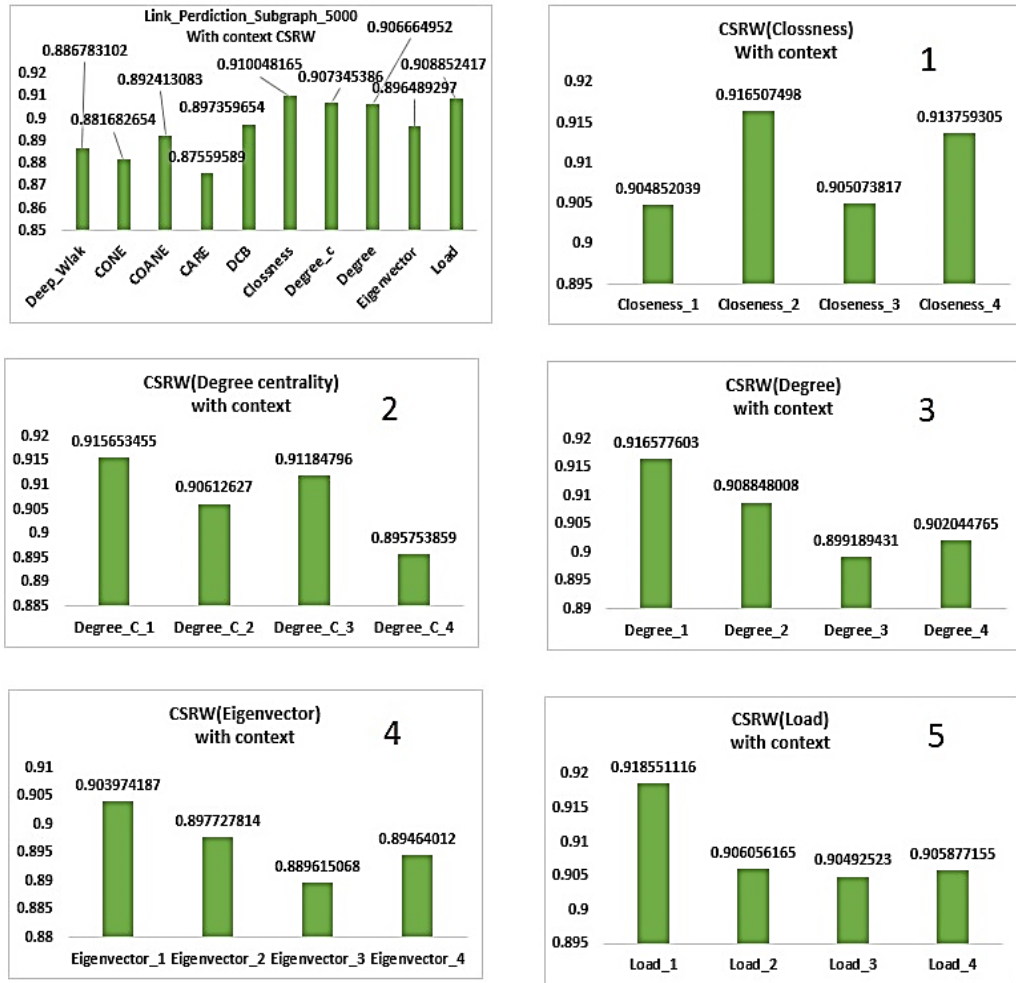


Fig. 16 Different runs of CSRW with semantic consideration in a DBLP_5000.

	precision (%) of vertex classification on subset of Cora								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Wlak	0.620267	0.735636	0.751928	0.776447	0.779706	0.796268	0.78179	0.810533	0.81112
CONE	0.54328	0.614857	0.672727	0.666667	0.691194	0.710181	0.714856	0.770066	0.745304
COANE	0.740948	0.743416	0.740367	0.757877	0.741338	0.769592	0.774685	0.759104	0.750292
CARE	0.401248	0.60314	0.622189	0.637459	0.661083	0.67522	0.67428	0.673484	0.670309
DCB	0.674852	0.743308	0.775674	0.782438	0.804739	0.803491	0.805691	0.827026	0.855566
Clossness	0.676629	0.725933	0.757182	0.778333	0.787549	0.782727	0.807588	0.81508	0.867522
Degree_c	69%	74%	76%	77%	79%	81%	81%	82%	86%
Degree	0.664572	0.739418	0.756709	0.771052	0.795269	0.795596	0.815553	0.815658	0.864983
Eigenvector	0.623988	0.724598	0.742497	0.764716	0.790447	0.79852	0.802363	0.80871	0.868247
Load	0.657042	0.72445	0.755357	0.781467	0.790543	0.790375	0.79933	0.813574	0.857068

Fig. 17: Precision score on Cora dataset.

	Recall (%) of vertex classification on subset of Cora								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Wlak	0.664069	0.731426	0.754219	0.776	0.781388	0.79428	0.776138	0.802583	0.808118
CONE	0.719032	0.736041	0.738924	0.746462	0.739291	0.75738	0.762608	0.741697	0.723247
COANE	0.460623	0.603599	0.616561	0.627077	0.655835	0.668819	0.664207	0.662362	0.667897
CARE	0.546349	0.611906	0.661392	0.662769	0.686115	0.711255	0.703567	0.769373	0.730627
DCB	0.678835	0.740655	0.770042	0.782154	0.804284	0.802583	0.801968	0.824723	0.856089
Clossness	0.663249	0.723812	0.756593	0.777538	0.788774	0.782288	0.806888	0.814114	0.865793
Degree_c	0.681911	0.740309	0.76068	0.771231	0.788589	0.804889	0.808426	0.823801	0.860821
Degree	0.667248	0.738579	0.755142	0.770615	0.794867	0.795203	0.813346	0.813653	0.862103
Eigenvector	0.645406	0.719774	0.741825	0.764769	0.788405	0.797279	0.799815	0.809041	0.866328
Load	0.723004	0.755142	0.768987	0.789513	0.789513	0.787362	0.797355	0.810886	0.850138

Fig. 18: Recall score on the Cora dataset.

	Micro-F1 (%) of vertex classification on subset of Cora								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Wlak	0.664069	0.731426	0.754219	0.776	0.781388	0.79428	0.776138	0.804428	0.808118
CONE	0.719032	0.736041	0.738924	0.746462	0.739291	0.75738	0.762608	0.741697	0.723247
COANE	0.460623	0.603599	0.616561	0.627077	0.655835	0.668819	0.664207	0.662362	0.667897
CARE	0.514552	0.581208	0.634697	0.643794	0.663549	0.684687	0.69999	0.772793	0.714999
DCB	0.678835	0.740655	0.770042	0.782154	0.804284	0.802583	0.801968	0.824723	0.856089
Clossness	0.663249	0.723812	0.756593	0.777538	0.788774	0.782288	0.806888	0.814114	0.863293
Degree_c	0.681911	0.740309	0.76068	0.771231	0.788589	0.804889	0.808426	0.823801	0.860821
Degree	0.681911	0.740309	0.76068	0.771231	0.788589	0.804889	0.808426	0.823801	0.860821
Eigenvector	0.645406	0.719774	0.741825	0.764769	0.788405	0.797279	0.799815	0.809041	0.856328
Load	0.659249	0.723004	0.755142	0.779692	0.789513	0.787362	0.797355	0.810886	0.850138

Fig. 19: Micro-F1 score on the Cora dataset.

	Macro-F1 (%) of vertex classification on subset of Cora								
	10%	20%	30%	40%	50%	60%	70%	80%	90%
Deep_Wlak	0.583304	0.71307	0.742073	0.756243	0.763717	0.777506	0.764953	0.796252	0.798318
CONE	0.701254	0.711586	0.716973	0.720391	0.703544	0.733319	0.738022	0.720704	0.707439
COANE	0.359383	0.581234	0.559166	0.612954	0.643715	0.659506	0.656507	0.657173	0.638063
CARE	0.514552	0.581208	0.634697	0.643794	0.663549	0.684687	0.69999	0.772793	0.714999
DCB	0.651602	0.71335	0.748199	0.76854	0.790261	0.794474	0.79461	0.816848	0.839823
Clossness	0.637908	0.70201	0.735927	0.756183	0.77449	0.763297	0.79173	0.801792	0.844276
Degree_c	0.652644	0.722914	0.746479	0.755803	0.773975	0.790789	0.804402	0.807173	0.843487
Degree	0.628174	0.717102	0.736098	0.754681	0.78126	0.778483	0.800028	0.801043	0.84867
Eigenvector	0.568314	0.698901	0.723646	0.748658	0.773177	0.782594	0.786372	0.788884	0.855871
Load	0.613795	0.697015	0.734937	0.755955	0.774846	0.774268	0.781337	0.794344	0.853862

Fig. 20: Macro-F1 score on the Cora dataset.

IV) Parameter Sensitivity

The effect of the number of communities or topics (k) is shown in Fig. 21.

The k parameter varies from 9 to 24 and it shows the link prediction values for the CSRW method for different

criteria of closeness centrality, degree centrality, degree centrality, eigenvector centrality, and load centrality for a DBLP_5000.

Here, load centrality has provided a better result than other criteria.

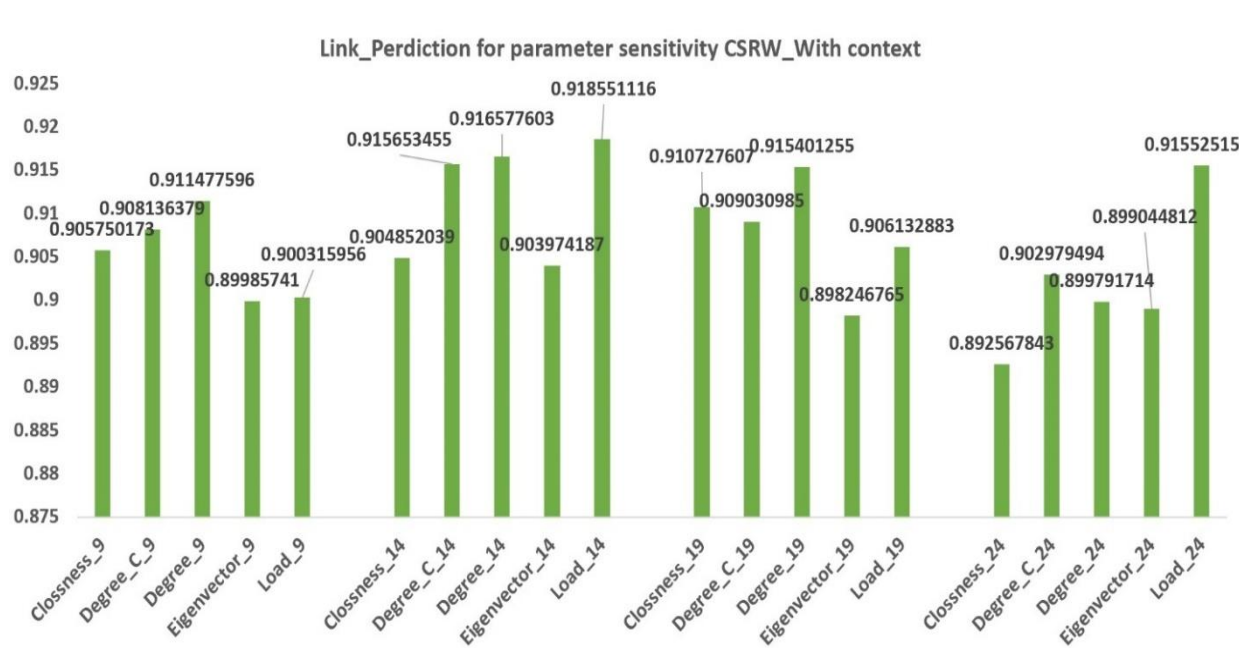


Fig. 21: The effect of the number of communities on the link prediction criterion for the CSRW algorithm.

To determine the best value of k , the averaging process has been performed for five methods of proximity centrality, degree, degree centrality, eigenvector centrality and load centrality in CSRW under setting different values for k .

According to Fig. 22, the best average value for k is 14. In this paper, the value of k is considered equal to 14 in all the implementations.

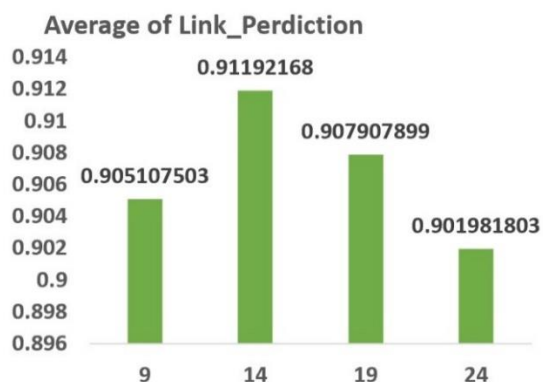


Fig. 22: The average values of the link prediction values of different criteria for the number of communities in the CSRW method.

Conclusion and Future Work

In this paper, the random walk algorithm based on the centrality and semantic consideration named CSRW is presented for the network embedding process. This method consists of two parts: random walk and semantic analysis.

In the random walk section, various criteria such as degree, degree centrality, closeness centrality, eigenvector centrality, and load centrality have been employed.

The non-greedy usage of these criteria is the specific innovation of this paper. Each node is measured based on the average of its criteria and the higher-order vertices. In the semantic analysis section, the textual information of each node is aggregated with the neighboring vertices of the next order and is analyzed by the BTM topic model. Finally, the final vector is obtained by combining the vectors obtained from the random walk and contextual analysis.

The proposed approach of this paper overcomes the shortcomings of classical research and has reached a high efficiency in the network embedding phase.

The experimental results on real-world concept-based networks show the effectiveness and robustness of CSRW compared to five basic methods DeepWalk, CARE, CONE, COANE, and DCB. The CSRW method is focused on two dimensions, different criteria, and the semantic connection of nodes.

In future works, we plan to expand this work to obtain the most influential vertices and maximize their influence throughout the network.

The methods presented in this article focus on one type of node. However, real-world networks are usually composed of different types of vertices, relationships, and explicit information.

Therefore, in the continuation of this research, the proposed method can be extended on heterogeneous

networks. And also, in the following, it is possible to optimize the use of network embedding to obtain influential nodes and maximize the impact in social networks. In the future work, we plan to expand this work in terms of obtaining Influential nodes and optimizing the network.

Author Contributions

M. Taherparvar: Programmer, Validation, Conceptualization, Visualization, Investigation, collected the dataset, Writing-Reviewing and Editing, Writing - Original draft preparation. F. Ahmadi Abkenari: Supervision, Project administration, Conceptualization, Methodology, Visualization, Investigation, Writing-Reviewing and Editing, Programmer, Writing - Original draft preparation. P. Bayat: Writing-Reviewing and Editing. All authors discussed the results.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

CSRW	Centrality, and a Semantic-based Random Walk
BTM	Biterm Topic Models for Short Text

References

- [1] B. Perozzi, R. Al-Rfou, S. Skiena, "Deepwalk: Online learning of social representations," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 701–710, 2014.
- [2] A. Grover, J. Leskovec, "Node2vec: Scalable feature learning for networks," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 855–864, 2016.
- [3] S. Bhagat, G. Cormode, S. Muthukrishnan, "Node classification in social networks," Social Network Data Analytics (Springer) Ed. Charu Aggarwal, 2011.
- [4] A. Faroughi, R. Javidan, "CANF: Clustering and anomaly detection method using nearest and farthest neighbor," Future Gener. Comput. Syst., 89: 166–177, 2018.
- [5] S. Aslan, M. Kaya, "Topic recommendation for authors as a link prediction problem," Future Gener. Comput. Syst., 89: 249–264, 2018.
- [6] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv:1301.3781, 2013.
- [7] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, "Line: Large-scale information network embedding," in Proc. International Conference on World Wide Web: 1067–1077, 2015.
- [8] H. Gao, H. Huang, "Deep attributed network embedding," in Proc. International Joint Conference on Artificial Intelligence: 3364–3370, 2018.
- [9] X. Huang, J. Li, X. Hu, "Label informed attributed network embedding," in Proc. ACM International Conference on Web Search and Data Mining: 731–739, 2017.
- [10] J. Butepage, M. J. Black, D. Kragic, H. Kjellstrom, "Deep representation learning for human motion prediction and classification," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition: 1591–1599, 2017.
- [11] X. Du, J. J. Y. Wang, "Support image set machine: Jointly learning representation and classifier for image set classification," Knowledge-Based Syst., 78: 51–58, 2015.
- [12] J. Li, J. Li, X. Fu, M. A. Masud, J. Z. Huang, "Learning distributed word representation with multi-contextual mixed embedding," Knowledge-Based Syst., 106: 220–230, 2016.
- [13] M. Janner, K. Narasimhan, R. Barzilay, "Representation learning for grounded spatial reasoning," Trans. Assoc. Computer. Ling., 6: 49–61, 2018.
- [14] Z. Chen, T. Cai, C. Chen, Z. Zheng, G. Ling, "SINE: Side information network embedding," in Proc. International Conference on Database Systems for Advanced Applications: 692–708, 2019.
- [15] D. Wang, P. Cui, W. Zhu, "Structural deep network embedding," in Proc. International Conference on Knowledge Discovery and Data Mining: 1225–1234, 2016.
- [16] M. Li, J. Liu, P. Wu, X. Teng, "Evolutionary network embedding preserving both local proximity and community structure," IEEE Trans. Evol. Comput., 24(3): 523–535, 2019.
- [17] J. Chen, Q. Zhang, X. Huang, "Incorporate group information to enhance network embedding," in Proc. International Conference on Information and Knowledge Management: 1901–1904, 2016.
- [18] M. M. Keikha, M. Rahgozar, M. Asadpour, "Community aware random walk for network embedding," Knowl. Based Syst., 47–54, 2018.
- [19] J. Wang, J. Cao, W. Li, S. Wang, "CANE: community-aware network embedding via adversarial training," Knowl. Inf. Syst., 63: 411–438, 2020.
- [20] Y. Shi, M. Lei, H. Yang, L. Niu, "Diffusion network embedding," Pattern Recognit., 88: 518–531, 2019.
- [21] H. Chen, H. Yin, T. Chen, Q.V.H. Nguyen, W.-C. Peng, X. Li, "Exploiting centrality information with graph convolutions for network representation learning," in Proc. IEEE International Conference on Data Engineering: 590–601, 2019.
- [22] W. Zhao, H. Ma, Z. Li, X. Ao, N. Li, "SBRNE: An improved unified framework for social and behavior recommendations with network embedding," in Proc. International Conference on Database Systems for Advanced Applications: 555–571, 2019.
- [23] Q. Li, J. Zhong, Q. Li, Z. Cao, C. Wang, "Enhancing network embedding with implicit clustering," in Proc. International Conference on Database Systems for Advanced Applications: 452–467, 2019.
- [24] M. RosenZvi, T. Griffiths, M. Steyvers, P. Smyth, "The author-topic model for authors and documents," in Proc. Uncertainty in Artificial Intelligence: 487–494, 2004.
- [25] Q. Mei, D. Cai, D. Zhang, C. Zhai, "Topic modeling with network regularization," in Proc. International Conference on World Wide Web: 101–110, 2008.
- [26] L. Wu, D. Wang, S. Feng, Y. Zhang, G. Yu, MDAL: "Multi-task Dual Attention LSTM Model for Semi-supervised Network Embedding," in Proc. International Conference on Database Systems for Advanced Applications: 468–483, 2019.
- [27] Y. Gao, M. Gong, Y. Xie, H. Zhong, "Community-oriented attributed network embedding," Knowledge-Based Systems, 193: 105418, 2019.
- [28] M. Taherparvar, F. Ahmadi Abkenari, P. Bayat, "Attribute network embedding based on maintaining the structure and semantic

features of the graphs,” in Proc: International Conference on The New Horizons in The Electrical Engineering, Computer and Mechanical, 2022.

- [29] L. Chen, Y. Li, X. Deng, Z. Liu, M. Lv, T. He, “Semantic-aware network embedding via optimized random walk and paragraph2vec,” *J. Comput. Sci.*, 63: 101825, 2022.
- [30] J. Zhang, Yu. Luo, “Degree centrality, betweenness centrality, and closeness centrality in social network,” in proc. International Conference on Modelling, Simulation and Applied Mathematics, 2017.
- [31] P. Bonacich, “Some unique properties of eigenvector centrality,” *Social Network*, 29(4): 555-564, 2007.
- [32] L. Maccari, L. Ghio, A. Guerrieri, A. Montresor, R. Lo Cigno, “On the distributed computation of load centrality and its application to DV routing,” in proc. IEEE Conference on Computer Communications, 2018.
- [33] M. Taherparvar, F. Ahmadi Abkenari, P. Bayat, “Conformance evaluation of topic modeling approaches on web-based short text dynamic graph databases,” in proc. International Conference on Web Research, 2021.
- [34] J. Tang, Z. Meng, X. Nguyen, Q. Mei, M. Zhang, “Understanding the limiting factors of topic modeling via posterior contraction analysis,” in Proc. International Conference on Machine Learning: 190–198, 2014.
- [35] A. K. McCallum, K. Nigam, J. Rennie, K. Seymore, “Automating the construction of internet portals with machine learning,” *Information Retrieval*: 127–163, 2000.
- [36] <https://www.aminer.org/citation>, last access June 17, 2023.

Biographies



Mohadeseh Taherparvar received B.Sc. in software engineering from Azad University of Lahijan, Iran, in 2008, MSc software engineering from Azad University of Qazvin, Iran, in 2013 and she is PHD candidate in Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran. Her research interests are Data Mining, Deep Learning, and Optimization Algorithms. she also

has experience in Python.

- Email: mtaherparvar@phd.iaurasht.ac.ir
- ORCID: 0000-0002-7822-5088
- Web of Science Researcher ID:NA
- Scopus Author ID:NA
- Homepage: NA



Fatemeh Ahmadi-Abkenari received the M.Sc. degree in information technology from the Polytechnique (Amirkabir) University of Tehran, Iran, in 2007, and the Ph.D. degree in computer engineering from UTM, Malaysia, in 2012. She is currently an Assistant Professor with the Faculty of Computer Engineering and Information Technology, Payam-Noor University, Rasht Branch, Iran. Her main research interests include

machine learning, data mining, text mining, sentiment and opinion mining, artificial neural networks, deep learning, and natural language processing.

- Email: Fateme.Abkenari@pnu.ac.ir
- ORCID: 0000-0001-5175-6826
- Web of Science Researcher ID:NA
- Scopus Author ID:NA
- Homepage: NA



Pyman Bayat received the M.Sc. degree from Islamic Azad University, Arak Branch, Iran, and the Ph.D. degree in computer engineering from UCSI University, Malaysia. He is currently an Assistant Professor with the Faculty of Computer Engineering, Islamic Azad University, Rasht Branch, Iran. His main research interests include distributed systems, image processing, and data mining.

- Email: bayat@iaurasht.ac.ir
- ORCID: 0000-0003-2291-1369
- Web of Science Researcher ID:NA
- Scopus Author ID:NA
- Homepage: NA

How to cite this paper:

M. Taherparvar, F. Ahmadi Abkenari, P. Bayat, “Centrality and Latent Semantic Feature Random Walk (CSRW) in large network embedding,” *J. Electr. Comput. Eng. Innovations*, 11(2): 311-326, 2023.

DOI: 10.22061/jecei.2023.9279.600

URL: https://jecei.sru.ac.ir/article_1834.html





Research paper

Mutual Coupling Reduction in MIMO Microstrip Antenna by Designing a Novel EBG with a Genetic Algorithm

R. Shirmohamadi Suiny¹, M. Bod^{2,*}, G. Dadashzadeh¹

¹Department of Electrical Engineering, Shahed University, Tehran, Iran.

²Communications Engineering Department, Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:

Received 26 October 2022

Reviewed 05 December 2022

Revised 31 December 2022

Accepted 21 January 2023

Keywords:

Multiple-input–Multiple-output (MIMO) antenna

Mutual coupling reduction

Decoupling

Electromagnetic Band-Gap (EBG)

Genetic Algorithm (GA)

Microstrip array

*Corresponding Author's Email Address:

Mohammadbod@sru.ac.ir

Abstract

Background and Objectives: Multi-input multi-output (MIMO) antennas have been of interest in wireless communications in recent years. In these systems, many antennas are placed next to each other. The most important issue in the design of MIMO antennas is mutual coupling. Many methods have been proposed to reduce the mutual coupling of MIMO antennas. Many of these methods require an additional substrate on top or bottom of the antenna. In the reduction of mutual couplings electromagnetic band-gap (EBG) structures are preferred because they are coplanar with the antenna and can be compactly designed. In this paper, to reduce mutual coupling in MIMO antennas, a novel compact EBG structure based on the genetic algorithm optimization is proposed.

Methods: The method proposed in this paper to design an optimal EBG structure is to use a genetic algorithm (GA). In this method, an EBG unit cell is designed by a binary code, and then the 7×2 EBG structure of the unit cell is placed between two antenna elements with $\lambda/2$ distance. The optimization algorithm tries to find the best unit cell to reduce the mutual coupling between two elements. After 70 generations in the genetic algorithm, the GA determines a compact structure of EBG elements which reduces mutual coupling significantly.

Results: Two-element patch antennas with and without the proposed EBG structure are fabricated and the mutual couplings between array elements are measured at 5.68GHz in both cases. It is shown that the proposed compact EBG structure reduced the isolation of the two antennas by 27 dB. This decrease in mutual coupling is much higher than in the previous papers. The proposed EBG has little effect on other antenna radiation parameters such as S11 and radiation patterns.

Conclusion: In general, in this paper, a compact and coplanar EBG structure is proposed to significantly reduce the mutual coupling in MIMO antennas. The method presented in this paper can be used for other MIMO antenna configurations at other frequencies and the proposed method will create a completely optimal structure to reduce mutual coupling.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

One of the important methods in increasing the capacity of the telecommunication system in recent years has

been the Multiple Input Output (MIMO) technology. Multiple-input-multiple-output (MIMO) technology not only provides high data transfer rates but also increases

communication reliability and provides multiplexing gain. For these reasons, MIMO technology was used in the fourth generation of wireless communication and is being widely used in the 5G and beyond in the form of massive MIMOs [1]-[3].

In MIMO technology, antennas are designed with small interelement distances for wide-angle beamforming purposes. Due to the small space and a large number of elements, strong mutual coupling effects occur between antenna elements. Mutual coupling degrades the radiation performance [4], as well as the available throughput [5]. Therefore, in recent years, extensive research has been done on the methods of reducing mutual coupling in MIMO antennas in universities and industry [6]-[28].

In general, the methods of reducing mutual coupling can be divided into two categories of non-coplanar [6]-[10] and coplanar [11]-[25] methods. In coplanar methods, the antenna element and the decoupling structure are both located on the same layer. While in non-coplanar methods, the decoupling structure is placed higher or lower than the antenna element.

Among the non-coplanar methods, we can refer to the methods such as: using the near-field resonator [6], the array-antenna decoupling surface [7], the metasurface-based decoupling method [8], T-shaped decoupling network [9], and using of a transmission-line for decoupling [10]. Although these methods have had some success in reducing the mutual coupling of antennas, however, all these works lead to a high antenna profile and need an additional substrate layer. Therefore the non-coplanar methods are unattractive especially when low-profile performance is desired.

The coplanar methods do not require an additional substrate and provide a more compact antenna structure. Among the coplanar methods, we can refer to the methods such as: using different types of metamaterials [11]-[13], defect ground structures (DGS) [15], [16], and Electromagnetic band-gap (EBG) structures [17]-[26]. Among these methods, the advantage of EBGs is their simple construction compared to metamaterials, which is an important issue and shows the superiority of these structures.

EBG structures also increase the gain and reduce the back lobe in the antenna [19]. While the DGS for example reduces the antenna gain and increases the back radiations. In addition, EBG can help to reduce the dimensions of the designed antenna [21].

In massive MIMO applications where elements with small distances are used, EBG structures should also be designed with optimal and small dimensions. If the EBG gets too close to the antenna, antenna matching is affected negatively due to reactive coupling between the array elements and EBG [22]. Therefore, the

miniaturization of EBG has been one of the issues of interest to researchers [23]-[25].

A compact EBG structure for low-profile applications is presented in [24]. However, the reported structure required a distance between the two antennas of more than 1λ and provided only a 6 dB improvement in isolation. In [25], the miniaturization of the EBG structure is done with the help of a fractal pattern and its combination with DGS. The final structure has provided a 16 dB improvement in isolation. But using the DGS structure degrades many advantages of EBG structures and increases the back radiation.

The EBG structures presented so far either have a specific design formulation or have introduced parameters that are changed by the designer to obtain the desired response [17]-[26]. All these methods require extensive trial and error simulation which is time-consuming in design. Also, all the proposed structures are presented for a specific antenna structure at a specific frequency.

In this paper, a novel EBG structure is generated by the genetic algorithm (GA) to reduce mutual coupling between array elements. In the proposed method a unit cell of the EBG structure is generated by the GA algorithm. This unit cell is then repeated with equal distances in length and width and the final EBG structure is created. The method proposed in this paper to generate an EBG structure with the genetic algorithm can be implemented in any desired antenna structure at any frequency.

Also, the electromagnetic simulations required to obtain the appropriate EBG have been performed automatically by linking MATLAB with the electromagnetic simulator.

To demonstrate the proposed method two-element patch antennas with and without EBG are fabricated and measured. In the measurement results, a 27dB improvement in the array isolation can be seen, while the distance between the two elements of the antenna remains 0.5λ .

This paper is organized as follows. Section II describes how GA is employed to design the EBG structure. Section III presents simulation results and reports measurements on the constructed array including EBG. Finally, the conclusions of this study are reviewed in Section IV.

Design Procedure

As it is known, the EBG structures are periodic and consist of a large number of unit cells. In this paper, a unit cell of EBG is first designed by the genetic algorithm. Then, the unit cell is repeated along X and Y directions in a specific number and distances so that a novel EBG structure is created. The method for designing the novel EBG structure has already been used to design a wideband monopole antenna [27], [28], and microstrip filters [29], [30].

The reference antenna is a microstrip array including two patch elements designed at the 5.68 GHz frequency. This antenna is etched on an FR4 substrate with a 1.6 mm height and $\epsilon_r=4.4$ and loss tangent of 0.03 as shown in Fig. 1. The distance between its elements is $\lambda/2$ and the periodic EBG structure is placed between the antenna elements. Other designed parameters of the microstrip array are: $L = 11.7 \text{ mm}$, $W = 10.9 \text{ mm}$, $d = 14.9 \text{ mm}$, $L_a = 58.19 \text{ mm}$, $W_a = 40.33 \text{ mm}$, $S = 26.6 \text{ mm}$, $h = 1.6 \text{ mm}$.

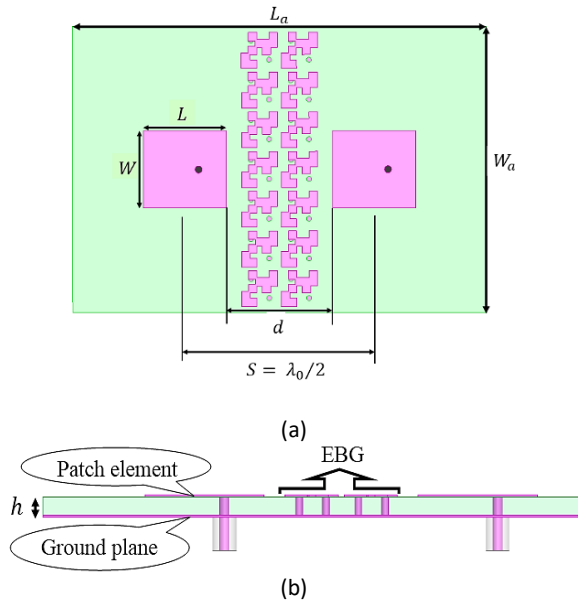


Fig. 1: Geometry of the reference patch antenna array and the EBG structure. The distance between elements is $\lambda_0/2$. (a) Top view, (b) Side view.

To design the novel EBG with the GA, a unit cell of the EBG structure is decoded as a binary chromosome. This chromosome is mapped to a 5x5 square patch as shown in Fig. 2. As it can be seen, the GA chromosome has five genes with a five-bit binary sequence, therefore a 25-bit binary chromosome is used to define the EBG unit cell. These binary bits determine the presence or absence of 1 mm square patches. If the corresponding bit includes 1, the square is filled with a metal patch; when the corresponding bit is 0, the square is left free. These square patches also overlap so that when metal squares are connected only at one point, their electrical interconnection is assured.

Square patches employed in the EBG cell have 1mm edges and overlapping is considered to be 0.2mm; thus, each square has 1.2mm edges. In each unit cell, two connections to the ground (vias) are also considered. These vias are orthogonal to the structure of the planar section of the unit cell. The location of the vias is in the middle of each EBG cell and remains unchanged in the optimization process. As described in [22], the vias create inductive impedance in the EBG structure and block the surface wave propagation.

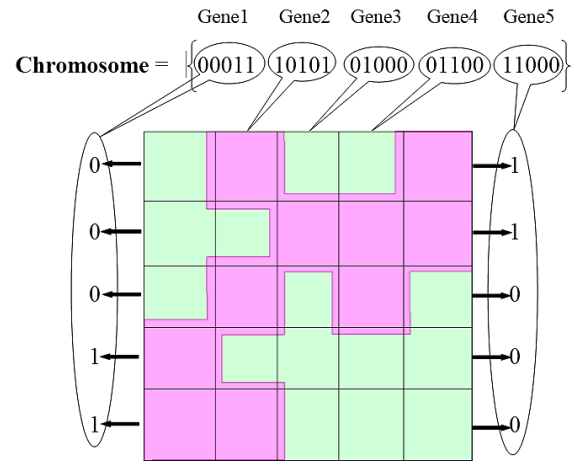


Fig. 2: GA chromosome implementation to the unit cell of the EBG.

Fig. 3 shows the complete dimensions of the proposed unit cell. The two vias with 0.72mm diameters are shown as a circle in this figure. The design EBG includes two columns and seven rows of the GA unit cells. The period of repeating unit cells is 5.7mm and therefore the gap between each unit cell is $g=0.5\text{mm}$.

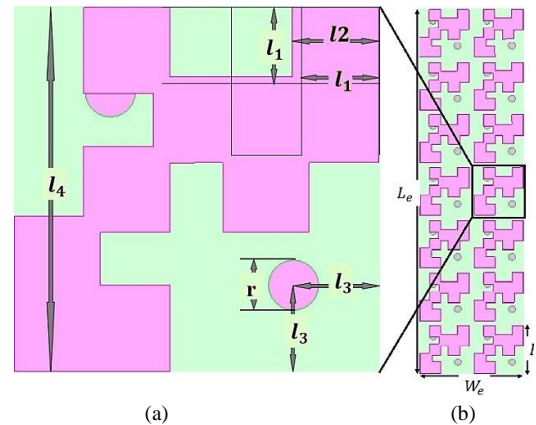


Fig. 3: The Proposed EBG structure using GA. (a) Top view of the EBG unit cell. (b) Top view of the EBG. The parameters are: $l_1 = 1\text{mm}$, $l_2 = 1.2\text{mm}$, $l_3 = 4\text{mm}$, $l_4 = 5.2\text{mm}$, $r = 0.72\text{mm}$, $W_e = 1$, $L_e = 39.76\text{mm}$, $P = 5.7\text{mm}$.

GA is started by generating a forty randomly initial population. In each generation selection, cross-over, and mutation functions are used to create next-generation populations. The rates of selection, cross-over, and mutation functions are considered as 0.2, 0.7, and 0.1 respectively.

GA is employed using MATLAB software and the chromosome generated by GA is first transformed to a unit cell in HFSS software and then converted to the EBG structure by repeating along X and Y directions. The EBG evaluation in the HFSS is done after the implementation of the 7x2 EBG structure. It should be mentioned that each GA chromosome changes the EBG structures and the other parameters of the array antenna remain

unchanged. The S-parameter obtained by HFSS simulation is exported to MATLAB and evaluated by the cost function. A lower value assigned to an EBG by cost function leads to a higher probability of using that EBG in subsequent generations.

The GA cost function has $|S_{11}|$ and $|S_{21}|$ parameters in the desired bandwidth and is evaluated as follows:

$$\text{Cost1} = \sum_{i=1}^N w_i (|S_{12}(f_i)|) \quad (1)$$

$$\text{Cost2} = \sum_{i=1}^N w_i (|S_{11}(f_i)| + |S_{12}(f_i)|) \quad (2)$$

in which N , is the number of sampling frequencies considered for the reduction of the mutual coupling, w_i represents the weighting value at the i -th sampling frequency, and f_i is the i -th sampling frequency.

In the decision-making process, the cost function is first considered to reduce mutual coupling in the intent frequency band by (1) and the matching definition is neglected. When the mutual coupling is reduced, a member of the generation is selected such that the matching condition and mutual coupling reduction are satisfied as defined in (2). In this step, weight coefficients for both conditions are considered to be the same. Weighting coefficients in (1) and (2) are chosen for better convergence of the genetic algorithm. For example, if the value of $|S_{12}|$ is better than -50 dB, the coefficient of that frequency becomes zero. Also, larger weighting coefficients are used at frequencies close to antenna resonance, where the importance of reducing mutual coupling is higher.

Table 1: Summary of the steps of designing the Proposed EBG structure

Step	Procedure
01	GA parameters and criteria are set in Matlab.
02	Matlab creates random 25-bit binary chromosomes.
03	The generated chromosomes are decoded as pixel EBG unit cells.
04	The generated unit cells send to HFSS by a Matlab link.
05	In HFSS the unit cells are repeated in the 7×2 array and create different GA-EBG structures.
06	HFSS simulates the GA-EBG structures between a predefined array of antennas.
07	The S11 and S12 results send back to Matlab.
08	GA evaluates the simulated GA-EBG structures with equations (1) and (2). The next generation of GA chromosomes is created from the previous ones by mutation, selection, and cross over functions.
09	
10	If there is no convergence, the algorithm continues from the third step.

Table 1 shows a summary of the design steps of the proposed GA EBG structure. The desired isolation between array elements is considered more than 45 dB and after 70 generations this isolation is obtained and the algorithm is terminated. The final microstrip array antenna with the obtained GA-EBG is shown in Fig. 1. The advantage of employing EBG using GA is that the designer can monitor all design steps accurately and obtain the desired response by changing parameters.

Results and Discussion

After designing GA-EBG, the simulation results of two microstrip elements with and without EBG are compared in Fig. 4. As can be seen, both antennas have resonance around 5.7 GHz frequency. The relative bandwidth of the antenna without GA-EBG is 4% from 5.54 to 5.78 GHz, while the antenna with GA-EBG has a bandwidth of about 3.5% from 5.6 to 5.8 GHz. In this figure, the designed EBG structure shifts the antenna resonance slightly. This issue can be solved by changing the dimensions of the microstrip antenna in the presence of the EBG structure.

The $|S_{21}|$ result of the microstrip arrays is also shown in Fig. 4. As can be seen, $|S_{21}|$ is -21dB if the array does not include EBG and is -50 dB if it includes EBG; therefore, the proposed EBG improves the mutual coupling by about 29dB in the simulation results.

The proposed two-element antennas with and without EBG are fabricated and their characteristics are measured as shown in Fig. 5. The printed circuit board technology is used for manufacturing these antennas.

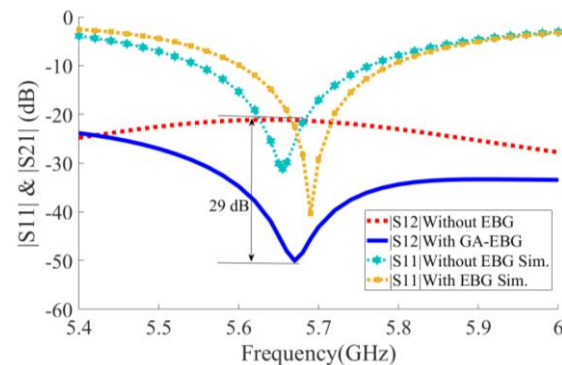


Fig. 4: Comparison of simulated scattering parameters for an E-plane coupled antenna pair (with GA-EBG and without EBG).

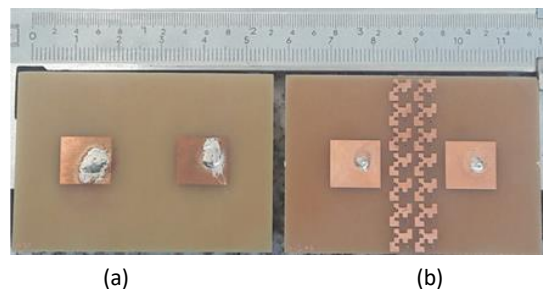


Fig. 5: The fabricated prototype of the two-element patch antennas. (a) Without EBG, and (b) with the proposed GA-EBG.

The S-parameters of these microstrip arrays are measured using the vector network analyzer 3413E of Agilent and the results are shown in Fig. 6 (a) and (b). As can be seen in these figures, the measurement results are in very good agreement with the simulation results. In the measurement results at the frequency of 5.68GHz, $|S_{21}|$ is -21dB and by adding GA-EBG, it becomes -48 dB; thus, by adding EBG, the mutual coupling is improved by 27 dB. By comparing the measurement and simulation results in Fig. 4 and 6, it can be seen that the mutual coupling values in the measurement are about 2 dB higher than the simulation values. This issue could be due to the non-ideal nature of the antennas in the manufacturing conditions.

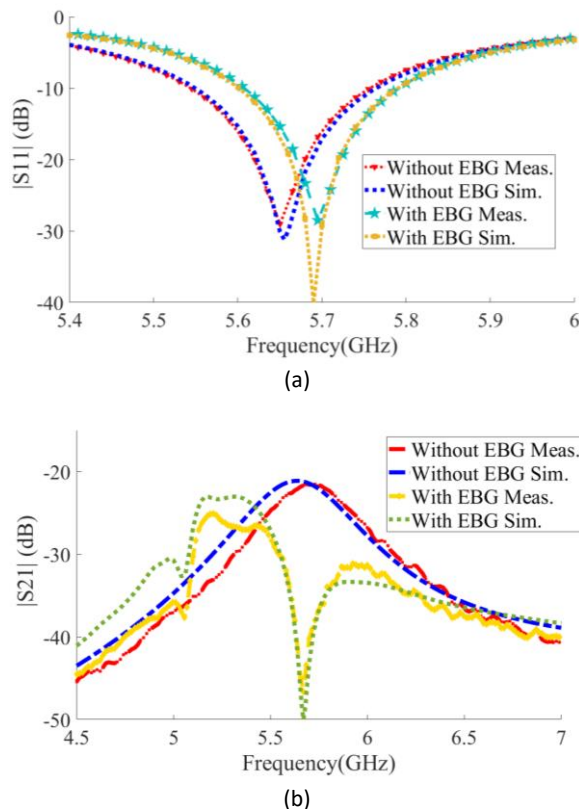


Fig. 6: Comparing the measurement and simulation results of the two-element antenna array with and without the presence of the proposed EBG. (a) S11 result, (b) S21 result.

Fig. 7 shows the amplitude of electric field distribution on the antennas, both in the presence and absence of EBG structure. In this figure, only one patch antenna (i.e. patch 1) is excited to observe the effects of isolation and mutual coupling on the side antenna (i.e. patch 2). As can be seen, in the case where there is no EBG, the stimulated field has reached the second antenna from the first antenna. However, when the EBG is added to the structure it does not allow surface waves to propagate and there are very few fields around the second antenna. Therefore, in this condition, the mutual coupling is greatly reduced. To study the effects of adding EBG on the radiation of the

antenna, the co and cross-polar radiation pattern of the antenna with and without EBG is measured and shown in Fig. 8. As can be the copolar radiation pattern of the antenna with the proposed EBG is almost match with the antenna without EBG. In this figure, it is also clear that, unlike previous works, the back radiation of the antenna has not changed much with the presence of EBG. However, the structure with the presence of EBG has slightly damaged the cross-polarization due to the creation of asymmetric currents. It should be noted that a slight increase in cross-polarization has been seen in most structures based on EBG.

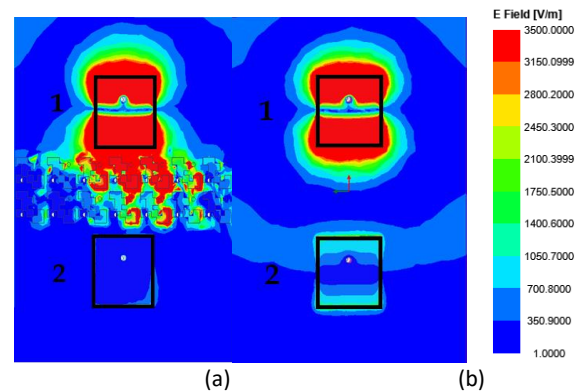


Fig. 7: The amplitude of electric field distribution on the antennas, (a) Without EBG, and (b) With GA-EBG, at 5.7 GHz frequency.

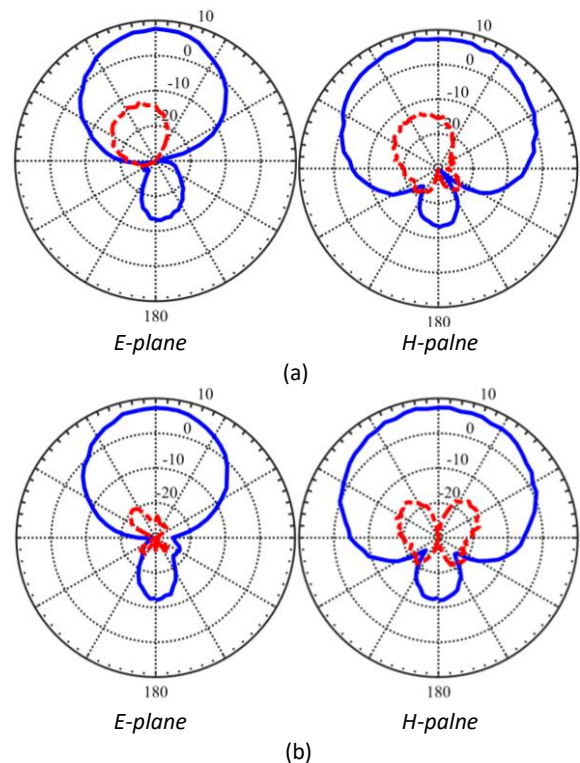


Fig. 8: Measured E- and H-plane radiation patterns of the proposed antenna with and without proposed EBG at 5.68GHz (a) the antenna with EBG, and (b) the antenna without GA-EBG. The solid line represents co-polarization, and the dashed line represents cross-polarization.

Finally, to compare the EBG designed in a paper with the previous paper, Table 2 is given. In this table, the structure presented in this paper is compared with six similar examples. As can be seen, the structure of this paper has made a significant improvement in the amount of mutual coupling, while the distance between two elements of the array remains $\lambda/2$.

In general, considering all the results presented in this paper, it can be concluded that a novel, compact, low-cost EBG structure has been presented to reduce the mutual coupling between MIMO antennas. The method presented in this paper can be used for other MIMO antenna configurations at other frequencies.

Table 2: Comparison of characteristics of various EBG Types with the proposed EBG structure

EBG Design Type	Freq. (GHz)	Mutual Coupling Reduction	Space Between Elements	Substrate (ϵ_r , height (mm))
SRR [11]	3.6	19dB	$0.5\lambda_0$	(4.4, 1)
CSRR [13]	5	10dB	$0.5\lambda_0$	(3.4, 1.27)
EBG+DGS [19]	5.8	22.3dB	$0.4\lambda_0$	(4.6, 1.6)
Uniplanar EBG [24]	5.6	13.5dB	$0.5\lambda_0$	(10.2, 2.54)
Fractal EBG [25]	5	16dB	$0.5\lambda_0$	(2.65, 1)
This Work	5.7	27dB	$0.5\lambda_0$	(4.4, 1.6)

Conclusion

In this paper, a novel compact EBG structure is proposed to reduce the mutual couplings between array elements. The proposed EBG is designed based on 7×2 unit cells of EBG created by the genetic algorithm. Two-element patch antennas with and without the proposed EBG are fabricated and the mutual couplings between array elements are measured in both cases. In the measurement and simulation results, the proposed EBG structure has improved the mutual coupling by more than 27 dB. In comparison to the previous paper, the proposed EBG has made a significant improvement in the amount of mutual coupling reduction, while the distance between two elements of the array and the height of the substrate remain as low as possible.

Author Contributions

R. shirmohamadi, M. Bod and G. Dadashzadeh developed the proposed antenna idea and performed the analytic simulations and measurements. M. Bod has

written the manuscript. R.shirmohamadi and G. dadashzadeh edited/reviewed the paper.

Acknowledgment

The authors would like to thank the anonymous reviewers and the editors of JECEI for their valuable comments and suggestions for improving quality of the paper.

Conflict of Interest

The author declares that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

Abbreviations

DGS	Defect ground structure
EBG	electromagnetic band-gap
GA	genetic algorithm
HFSS	High Frequency Simulation Software
f_i	the i th sampling frequency
MIMO	Multiple-input-multiple-output
SRR	Split ring resonator
CSRR	Complementary split ring resonator
w_i	the weighting value at the i th sample
λ_0	corresponding wavelength

References

- [1] T. L. Marzetta, "Massive MIMO: An introduction," Bell Labs Tech. J., 20: 11–22, 2015.
- [2] A. O. Martinez, J. Ø. Nielsen, E. De Carvalho, P. Popovski, "An experimental study of massive MIMO properties in 5G scenarios," IEEE Trans. Antennas Propag., 66(12): 7206–7215, 2018.
- [3] M. A. Jensen, J. W. Wallace, "A review of antennas and propagation for MIMO wireless communications," IEEE Trans. Antennas Propag., 52(11): 2810–2824, 2004.
- [4] L. Savy, M. Lesturgie, "Coupling effects in MIMO phased array," presented at the IEEE Radar Conf., Philadelphia, PA, USA, 2016.
- [5] K. H. Chen, J. F. Kiang, "Effect of mutual coupling on the channel capacity of MIMO systems," IEEE Trans. Veh. Technol., 65(1): 398–403, 2016.
- [6] M. Li, B. G. Zhong, S. W. Cheung, "Isolation enhancement for MIMO patch antennas using near-field resonators as coupling-mode transducers," IEEE Trans. Antennas Propag., 67(2): 755–764, 2019.
- [7] K.-L. Wu, C. Wei, X. Mei, and Z.-Y. Zhang, "Array-antenna decoupling surface," IEEE Trans. Antennas Propag., 65(12): 6728–6738, Dec. 2017.
- [8] F. Liu, J. Guo, L. Zhao, G.-L. Huang, Y. Li, Y. Yin, "Dual-band metasurface-based decoupling method for two closely packed dual-band antennas," IEEE Trans. Antennas Propag., 68(1): 552–557, 2020.

- [9] X. J. Zou, G. M. Wang, Y. W. Wang, H. P. Li, "An efficient decoupling network between feeding points for multielement linear arrays", *IEEE Trans. Antennas Propag.*, 67(5): 3101-3108, 2019.
- [10] Y.-M. Zhang, S. Zhang, J.-L. Li and G. F. Pedersen, "A transmission-line-based decoupling method for MIMO antenna arrays," *IEEE Trans. Antennas Propag.*, 67(5): 3117-3131, 2019.
- [11] Z. Qamar, L. Riaz, M. Chongcheawchamnan, S. A. Khan, M. F. Shafique, "Slot combined complementary split ring resonators for mutual coupling suppression in microstrip phased arrays," *IET Microw., Antennas Propag.*, 8(15): 1261-1267, 2014.
- [12] N. Supreeyattikul, N. Teerasuttakorn, "Improved isolation of a dual-band MIMO Antenna using modified S-SRRs for millimeter-wave applications," in *Proc. 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*: 388-391, 2020.
- [13] M. M. Bait-Suwailam, O. F. Siddiqui, O. M. Ramahi, "Mutual coupling reduction between microstrip patch antennas using slotted-complementary split-ring resonators," *IEEE Antennas Wireless Propag. Lett.*, 9: 876-878, 2010.
- [14] Y. Li, H. Yang, H. Cheng, J. Wu, Y. Yang, L. Hua, Y. Wang, "Isolation enhancement in dual-band MIMO antenna by using metamaterial and slot structures for WLAN applications," *J. Phys. D: Appl. Phys.*, 55(32), 2022.
- [15] Z. Niu, H. Zhang, Q. Chen, T. Zhong, "Isolation enhancement for 1×3 closely spaced E-Plane patch antenna array using defect ground structure and Metal-Vias," *IEEE Access*, 7: 119375-119383, 2019.
- [16] H. Xing, X. Wang, Z. Gao, X. An, H. X. Zheng, M. Wang, E. Li, "Efficient isolation of an MIMO antenna using defected ground structure," *Electronics*, 9(8): 1265, 2020.
- [17] K. S. Parvathi, S. R. Gupta, "Novel dual-band EBG structure to reduce mutual coupling of air gap based MIMO antenna for 5G application," *AEU Int. J. Electron. Commun.*, 138: 153902, 2021.
- [18] X. Tan, W. Wang, Y. Wu, Y. Liu, A. A. Kishk, "Enhancing isolation in Dual-Band Meander-Line multiple antenna by employing split EBG structure," *IEEE Trans. Antennas Propag.*, 67(4): 2769-2774, 2019.
- [19] Y. F. Cheng, X. Ding, W. Shao, B. Z. Wang, "Reduction of mutual coupling between patch antennas using a polarization-conversion isolator," *IEEE Antennas Wireless Propag. Lett.*, 16: 1257-1260, 2016.
- [20] F. Yang, Y. Rahmat-Samii, "Mutual coupling reduction of microstrip antennas using electromagnetic band-gap structure," in *Proc. IEEE AP-S Dig.*, 2: 478-481, 2001.
- [21] R. Baggen, M. Martinez-Vazquez, J. Leiss, S. Holzwarth, L. S. Dioli, P. de Maagt, "Low profile Galileo antenna using EBG technology," *IEEE Trans. Antennas Propag.*, 56(3): 667-674, 2008.
- [22] Z. Iluz, R. Shavit, R. Bauer, "Microstrip antenna phased array with electromagnetic bandgap substrate," *IEEE Trans. Antennas Propag.*, 52(6): 1446-1453, 2004.
- [23] M. Coulombe, S. Farzaneh, C. Caloz, "Compact Elongated Mushroom (EM)-EBG Structure for enhancement of patch antenna array performances," *IEEE Trans. Antennas Propag.*, 58(4): 1076-1086, 2010.
- [24] S. D. Assimonis, T. V. Yioultis, C. S. Antonopoulos, "Design and optimization of uniplanar EBG structures for low profile antenna applications and mutual coupling reduction," *IEEE Trans. Antennas Propag.*, 60(10): 4944-4949, 2012.
- [25] X. Yang, Y. Liu, Y. Xu, S. Gong, "Isolation enhancement in patch antenna array with fractal UC-EBG structure and cross slot," *IEEE Antennas Wireless Propag. Lett.*, 16: 2175-2178, 2017.
- [26] M. J. Al-Hasan, T. A. Denidni, A. R. Sebak, "Millimeter-wave compact EBG structure for mutual coupling reduction applications," *IEEE Trans. Antennas Propag.*, 63(2): 823-828, 2015.
- [27] A. J. Kerkhoff, R. L. Rogers, H. Ling, "Design and analysis of planar monopole antennas using a genetic algorithm approach," *IEEE Trans. Antennas Propag.*, 52 (10): 2709-2718, 2004.
- [28] M. John, M. J. Ammann, "Wideband printed monopole design using a genetic algorithm," *IEEE Antennas Wireless Propag. Lett.*, 6: 447-449, 2007.
- [29] A. Mallahzadeh, M. Bod, "Method for Designing Low-pass Filters with a Sharp Cut-off," *IET Microw. Antennas Propag.*, 8(1): 10-15, 2014.
- [30] M. Bod, A. R. Mallahzadeh "Band-pass filter design using modified CSRR-DGS," *Int. J. RF Microwave Comput. Aided Eng.*, 24(5): 544-548, 2014.

Biographies



Reza Shirmohamadi Suiny received a B.S. degree from Shahid Rajaei University, Iran, in electric engineering in 2009 and an M.S. degree from Shahed University, Iran, in communication engineering in 2017. His research interests include microstrip antenna designs, electromagnetic band-gap structures, numerical methods in electromagnetics, and RF/microwave circuit and system design.

- Email: r.shirmohamadi@shahed.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Mohammad Bod was born in Tehran, Iran, in 1986. He received a Ph.D. degree in electrical engineering from the Amirkabir University of Technology University, Tehran, in 2018. From 2019 to 2021, he was a Post-Doctoral Researcher at Amirkabir University with a fellowship awarded by the Iran National Science Foundation (INSF). He is currently an Assistant Professor at the Electrical Engineering Department, Shahid Rajaei Teacher Training University, Tehran. He has authored or co-authored 20 journal papers and two Persian books. His research interests include phased array radar, antenna and passive microwave component design, and numerical methods in electromagnetic.

- Email: mohammadbod@sru.ac.ir
- ORCID: [0000-0003-2687-1368](https://orcid.org/0000-0003-2687-1368)
- Web of Science Researcher ID: AAH-5551-2019
- Scopus Author ID: 54918504900
- Homepage: <https://www.sru.ac.ir/mohammadbod/>



Gholamreza Dadashzadeh was born in Urmia, Iran, in 1964. He received a B.Sc. degree in communication engineering from Shiraz University, Shiraz, Iran, in 1992 and an M.Sc. and Ph.D. degrees in communication engineering from Tarbiat Modares University (TMU), Tehran, Iran, in 1996 and 2002, respectively. From 1998 to 2003, he was the Head Researcher of the Smart Antenna for Mobile Communication Systems (SAMCS) and WLAN 802.11 Project with the Radio Communications Group, Iran Telecomm Research Center (ITRC), Tehran, where he was the Dean of the Communications Technology Institute (CTI), from 2004 to 2008. He is currently a full Professor with the Department of Electrical Engineering, Shahed University, Tehran. He has authored or co-authored more than 130 papers in referred journals and international conferences in the area of antenna design and smart antennas. Dr. Dadashzadeh is a member of the Iranian Association of Electrical and Electronics Engineers of Iran

(IAEEE). He received the First Degree of National Researcher from the Ministry of ICT of Iran in 2007.

- Email: gdadashzadeh@shahed.ac.ir
- ORCID: 0000-0002-8479-6784
- Web of Science Researcher ID: AAH-5551-2019
- Scopus Author ID: 14527009100
- Homepage:
<http://research.shahed.ac.ir/WSR/WebPages/Teacher/TEn.aspx?TID=233>

How to cite this paper:

R. Shirmohamadi Suiny, M. Bod, G. Dadashzadeh, "Mutual coupling reduction in MIMO microstrip antenna by designing a novel EBG with a genetic algorithm," J. Electr. Comput. Eng. Innovations, 11(2): 327-334, 2023.

DOI: [10.22061/jecei.2023.9375.615](https://doi.org/10.22061/jecei.2023.9375.615)

URL: https://jecei.sru.ac.ir/article_1835.html





Review paper

Brand New Categories of Cryptographic Hash Functions: A Survey

B. Sefid-Dashti¹, J. Salimi Sartakhti^{1,*}, H. Daghigh²

¹Department of Computer Engineering, University of Kashan, Kashan, Iran.

²Faculty of Mathematical Science, University of Kashan, Kashan, Iran.

Article Info

Article History:

Received 28 November 2022
Reviewed 28 December 2022
Revised 13 January 2023
Accepted 23 March 2023

Keywords:

Optical hash function
Memory-hard function
Bandwidth-hard function
Physical unclonable function
Quantum hash function
Application-specific hash function

Abstract

Background and Objectives: Cryptographic hash functions are the linchpins of mobile services, blockchains, and many other technologies. Designing cryptographic hash functions has been approached by research communities from the physics, mathematics, computer science, and electrical engineering fields. The emergence of new hash functions, new hash constructions, and new requirements for application-specific hash functions, such as the ones of mobile services, have encouraged us to make a comparison of different hash functions and propose a new classification.

Methods: Over 100 papers were surveyed and reviewed in detail. The research conducted in this paper has included four sections; article selection, detailed review of selected articles, data collection, and evaluation of results. Data were collected as new hash function properties, new hash function constructions, new hash function categories, and existing hash function attacks which are used to evaluate the results.

Results: This paper surveys seven categories of hash functions including block cipher-based functions, algebraic-based functions, custom-designed functions, Memory-hard Functions (MHFs), Physical Unclonable Functions (PUFs), quantum hash functions and optical hash functions. To the best of our knowledge, the last four mentioned categories have not been sufficiently addressed in most existing surveys. Furthermore, this paper overviews hash-related adversaries and six hash construction variants. In addition, we employed the mentioned adversaries as evaluation criteria to illustrate how different categories of hash functions withstand the mentioned adversaries. Finally, the surveyed hash function categories were evaluated against mobile service requirements.

Conclusion: In addition to new classification, our findings suggest using PUFs with polynomial-time error correction or possibly bitwise equivalents of algebraic structures that belongs to post-quantum cryptography as candidates to assist mobile service interaction requirements.

*Corresponding Author's Email
Address: salimi@kashanu.ac.ir

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

A cryptographic hash function is an integral part of a variety of applications such as digital signatures [1], authentication by static passwords, authentication by One-Time Passwords (OTP) [2], [3], data integrity [4], holographic encryption [5], Elliptic Curve Integrated

Encryption Scheme (ECIES) [6], Merkle tree [7], WS-Security [8], [9], data anonymization [10], Blockchain [11]-[13], cryptocurrencies [14], [15], video similarity search [16], and hash chain based strong password authentication [17], to name a few. Nine uses of cryptographic hash functions have been reviewed by

Alkandari et al. [18] in detail.

Hash functions were introduced in the late 1970s [19]. Ever since, hash function has received great interest, which has led to the construction of a wide variety of hash functions as well as attacks that attempt to invert or forge hash values. For example, the five-year NIST SHA-3 competition which culminated in the selection of a hardware-effective algorithm, *Keccak* [20], [21], as the winner in 2012 demonstrated numerous hash functions and a security analysis of them. The Password Hashing Competition that ran from 2013 to 2015 culminated in the selection of an MHF algorithm, *Argon2* [22], as a winner. To address constrained environments such as the continuously growing Internet of Things (IoT), NIST Lightweight competition was initiated in 2018 and is due to end in the year 2022. In March 2021, 10 lightweight cypher submissions were selected for the final round of the competition, and 3 out of the 10 finalists provide lightweight hashing functionalities. *SM3* hash function [23] was introduced as a new Chinese standard in 2010. *Streebog* hash function [24] was introduced as a new Russian standard in 2013. In 2015, *Kupyna* hash function [25] was introduced as a new standard in Ukraine.

Importantly, designing variants of hash functions has been undertaken by research communities from the fields of physics, mathematics, computer science, and electrical engineering, and this has led to the introduction of new categories of hash functions. Moreover, attacks based on physics, mathematics, computer science, and electrical engineering have been developed to compromise the security of a wide variety of hash dependent applications.

This research surveys seven categories of hash functions, namely block cipher-based functions, algebraic-based functions, custom-designed hash functions, MHFs, PUFs, quantum hash functions, and optical hash functions. To the best of our knowledge, the last four mentioned categories have not been sufficiently addressed in most existing surveys [18], [19], [26]–[29].

On the other hand, with the proliferation of smartphones and tablets, mobile devices are introduced as a new computational platform for enterprise applications and other software systems, and are considered as major participants in IoT and related constrained environments (e.g. smart homes, smart cities). Although mobile devices are strong enough to consume and/or provide some services, they suffer from computational and communicational constraints on their resources and generally experience intermittent connectivity. A clear understanding of what exactly is needed from an application-specific hash function is an urgent requirement. Hence, we evaluate the surveyed categories against mobile software requirements. The remainder of this paper is structured as follows:

I. Due to significant developments in the literature and

the use of cryptographic hash functions, today, new cryptographic hash functions impose more properties than traditional cryptographic hash functions. These supplementary properties are discussed in Section II.

II. SHA-3 competition (2007–2012) and Password Hashing Competition (2013–2015) fostered the design and analysis of processor-centric and memory-centric cryptographic hash functions, respectively. In turn, such events led to the introduction of new iterative and noniterative hash function constructions, six of which are overviewed in Section III. This section also reviews two hash function combiners.

III. New categories of hash functions including PUFs, quantum hash functions, and optical hash functions, MHFs along with the Bandwidth-hard Functions (BHF) subcategory, and some attacks affecting each category are presented in Sections IV and V. Section IV briefs on what affecting attacks entail. Investigated hash functions and the proposed seven-category classification are presented in the Section V. The attacks presented in Section IV are used as evaluation criteria in Section V.

IV. Mobile services suffer from computational and communicational problems. Hence, lightweight but not less secure cryptographic hash functions which secure interactions of resource constrained devices is an urgent need. Requirements which influence mobile services to choose some variants of cryptographic hash functions are presented in Section VI. In addition, Section VI discusses how each hash function category fits the mobile service requirements and why this research suggests PUFs with some enhancements and possibly bitwise equivalents of algebraic structures for mobile service consumption.

V. Section VII discusses the selection of appropriate hash functions for four application scenarios.

VI. Finally, in Section VIII the paper is summed up and conclusions are provided.

Definition

A hash function maps an input message of arbitrary length to a fixed length output which is called “hash,” “hash value,” or “message digest.” A hash function with n -bit output length is called an n -bit hash function. A good hash function produces random and uniform outputs. An output sequence resulted from applying a hash function in succession is called a “has chain” [2].

In addition to message, some hash functions may either accept a salt or a secret key. The former is called a salted hash function. Salts are randomly generated for each input message and are used for password hashing. The latter is usually used to build message authentication codes (MACs) and is called a keyed hash function. In other words, it serves as a checksum. In contrast to salts, keys are secrets and are not supposed to vary for different

messages.

Depending on the application, a hash function h may need to support some or all of the following properties:

- I. It maps arbitrary length input x to $h(x)$ efficiently. An efficient implementation may be achieved in software or hardware or both.
- II. One-way property or pre-image resistant property: For any given y in the image of h , it is not computationally feasible to find a message x such that $y = h(x)$.
- III. Second pre-image resistant property: For any given message x , it is not computationally feasible to find a message x' such that $x \neq x'$ and $h(x) = h(x')$.
- IV. Collision resistant property: It is not computationally feasible to find a pair x and x' such that $x \neq x'$ and $h(x) = h(x')$.
- V. Second collision resistant property: An attacker should not be able to use a given collision $h(x_1) = h(x'_1)$ to find another collision $h(x_2) = h(x'_2)$.
- VI. Hiding property: Given $h(r||x)$ so that r is chosen from a high min-entropy probability distribution and $||$ denotes concatenation of values, it is not computationally feasible to find x [30]. This property is a variant of one-way property and originates from blockchain terminology.
- VII. Puzzle friendliness property: Given r and $h(r||x)$ so that r comes from a spread-out set and h is an n -bit hash function, it is computationally infeasible to find x in time significantly less than 2^n [30]. Bitcoin mining is a race to solve such a computational puzzle.
- VIII. Chosen-Target-Forced-Prefix (CTFP) preimage resistance property: Committing a hash value h , without knowing the prefix of the message that will be hashed should be difficult [31], [32].
- IX. Chosen-Target-Forced-Midfix (CTFM) preimage resistance property: Committing a hash value h , without knowing any part of the message that will be hashed should be difficult [33].
- X. Application-Specific Integrated Circuit (ASIC) resistance property: It should not be easy to compute on ASIC machines [34].
- XI. Robustness property, aka robust video hashing property: For a given pair x and x' such that $x \neq x'$, $h(x) = h(x')$ as long as x and x' represent the same video content s , even though they represent it in different manners. In plain English, a hash function h used for video hashing should be robust against content-preserving changes such as encoding and blurring [16], [35].

The first four properties are mentioned in many references, but the rest are more or less new. *Property I* emphasizes that a hash function may be used by resource-constrained devices or to provide a fingerprint for a possibly very large file. An example of this property

is a parameter provided by SHA-3 hash function to trade-off security and performance [20], [21]. As another example, some hash functions such as MD-6 provide parallel implementation to speed up hashing a long message on multicore processors [36].

A hash function that supports *Properties I* and *II* is called a *one-way hash function* [1], [37]. A *cryptographic hash function* is a one-way hash function that provides second pre-image and collision resistant properties.

Since the introduction of cryptographic hash functions in the late 1970s, lots of hash functions have emerged that support pre-image resistance and second pre-image resistance properties; providing collision resistance, however, is more challenging. Fortunately, while few of hash function applications, such as digital signature, rely on collision resistance, for others providing pre-image resistance and second pre-image resistance properties is sufficient [2], [19]. Incidentally, there are collision-free hash functions as well, such as the lattice-based hash function proposed by Goldreich et al. [38].

Regarding the special ways that hash functions are employed in blockchain, hiding and puzzle friendliness properties are defined. *Properties VI* and *VII* harden bitcoin mining by reducing its surface of vulnerability, but as bitcoin lacks *Property X*, there are ASIC machines which speed up mining with reduced cost per bitcoin mined. *Properties VIII* and *IX* are preventive criteria to resist against herding attack (Section IV-A-4). Finally, *Property XI* focuses video hashing design on semantic content changes [16], [35] extracted from segmented video structural elements such as video shots [39]. Illustrated with UML class diagram, Fig. 1 depicts how hash functions, one-way hash functions, and cryptographic hash functions are subsequently extended (denoted by UML Generalization relationship) by adding pre-image property and both second pre-image and collision resistant properties, respectively. Fig. 1 further shows how application-specific hash functions, such as blockchain specific and video hash functions, enrich the required properties to satisfy application-specific requirements.

One may wonder whether a practical hash function without one-way property exists. Murmur hash [40] is an example of a hash function which is not designed for one-wayness. Non-cryptographic hash functions (NCHFs) [41] provide fast lookup capability. This paper concentrates on cryptographic hash functions, referred to hereafter as hash functions.

Constructions and Combiners

Designing a hash function entails making important decision on how to mix input message bits all together. While a large number of hash functions exists, they all have been designed based on a handful of constructions.

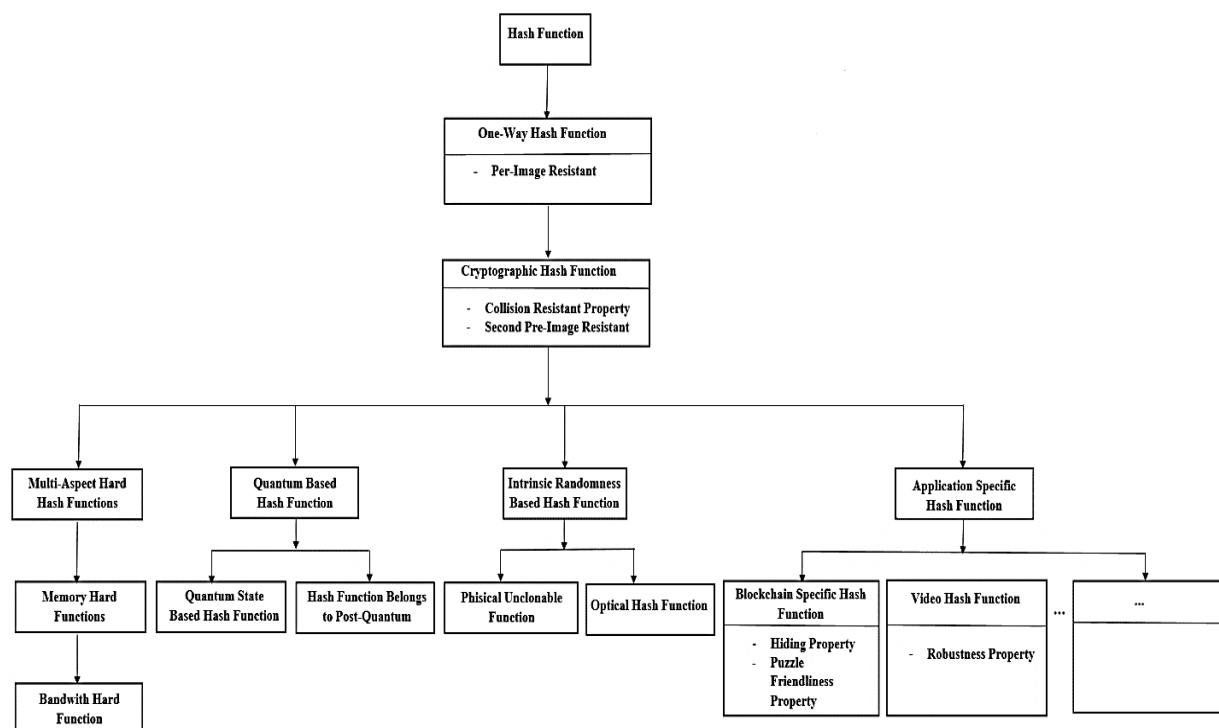


Fig. 1: Hierarchy of hash functions.

Hash function constructions are important in combining all bits of arbitrary-length messages in a way that holds properties such as collision resistance. These constructions split an arbitrary-length message into equal-sized blocks and iterate through the blocks to combine block bits all together. Some constructions combine block bits themselves, while others apply a compression function on each block and combine the results.

A compression function is a one-way function which takes a fixed length block of message along with a chaining variable as input, mixes the bits of input with each other, and returns a shorter, fixed-length output.

The way that a hash function construction combines the results of its underlying compression function is called *domain extension* [42]. For some *domain extensions*, if the underlying compression function has a security property such as collision resistance, that *domain extension* can produce hash functions that retain that property. For example, it is proven that hash functions based on the Merkle-Damgård construction (Section III-B-1) which use a fixed initial value along with an appropriate padding are collision resistant as long as their corresponding compression function is collision resistant.

Moreover, some other *domain extensions* such as the Zipper hash construction [43] (Section III-B-3) produce hash functions which hold properties such as collision resistance regardless of their underlying compression function.

This section reviews four iterative and two noniterative hash function constructions. The former includes most common constructions such as Merkle-Damgård and Sponge, while the latter includes tree style and graph-based hash function constructions. Other rarely used hash function constructions, such as Widepipe and the Hash Iterated Framework, are not included in this survey. Both of the omitted constructions aim to solve internal collision problems. The former uses output transformation, while the latter uses a salt and a counter to achieve this goal [44].

Finally, this section reviews two *hash function combinators* (simply *combiner* henceforth). A *combiner* combines the output of two hash functions or the output of the compression functions of two hash functions [45]. As an example, bitcoin uses double SHA-256 (i.e. $\text{SHA-256}(\text{SHA-256}(\text{message}))$) and a combination of RIPEMD-160 and SHA-256 (i.e. $\text{RIPEMD-160}(\text{SHA-256}(\text{message}))$) that are examples of combining hash functions in a sequential order. As another example, a combination of MD-5 and SHA-1 was used by SSL/TLS [46]. Concatenation combinators and XOR combinators are also used [45]. *Merkle tree* and *Zipper hash* combinators are reviewed in this section. The former combines the outputs of a hash function in tree style, while the latter combines the outputs of two different compression functions in reverse order.

A. Noniterative Constructions

This section first reviews Merkle tree and then

discusses tree- and graph-based constructions. These constructions map arbitrary-length input to tree leaves or graph walks and process the resulting tree or graph.

1. **Merkle Tree:** Merkle tree [7] is a combiner and uses a binary tree structure to allow the integrity of large data sets to be verified quickly. One of its recent applications is bitcoin. Fig. 2 depicts an example of a Merkle tree [14]. The tree's leaves are data blocks we want to hash. The hash of each leaf node is stored in its immediate parent node. Then, the hash of each pair of nodes is concatenated and hashed together, until there is one root hash known as the Merkle root [14]. Data integrity of a block is verified by checking hashes from that block to the root node (Fig. 3 [14], [30]). A tree consisting of n nodes requires verifying about $\log n$ items [30], including verifying hash of that data block and its sibling-node (if it exists), and then proceeds upward until it reaches the top.

2. **MD-6 Tree style construction:** MD-6 [36] uses a 4-ary tree structure to achieve parallelism along with alternative sequential mode. As a source of parallelism, each round of its compression function uses 16

parallelizable loops. Moreover, it parallelizes a quaternary Merkle tree-like structure with a height adjustment parameter (L). Regarding L , there are three modes of operation:

- $L = 64$ as the default and means fully tree-based mode.
- $L = 0$ means sequential mode and uses a Merkle-Damgård construction.
- Specifying a number greater than 0 and less than 64 means hybrid mode. First using L level tree, and then sequential mode.

Fig. 4 shows an example of an MD-6 tree [36]. MD-6 uses a 4-to-1 compression function at each internal node of the tree. Tree leaves store blocks of data to be hashed, and internal nodes store the results of applying compression function on the concatenated data of four child nodes. The compression function at the root node is flagged to return truncated result as a MD-6 hash value.

MD-6 was submitted to the SHA-3 competition, but due to an error found in its security proof against differential attacks [19], it did not proceed to the second round of that competition.

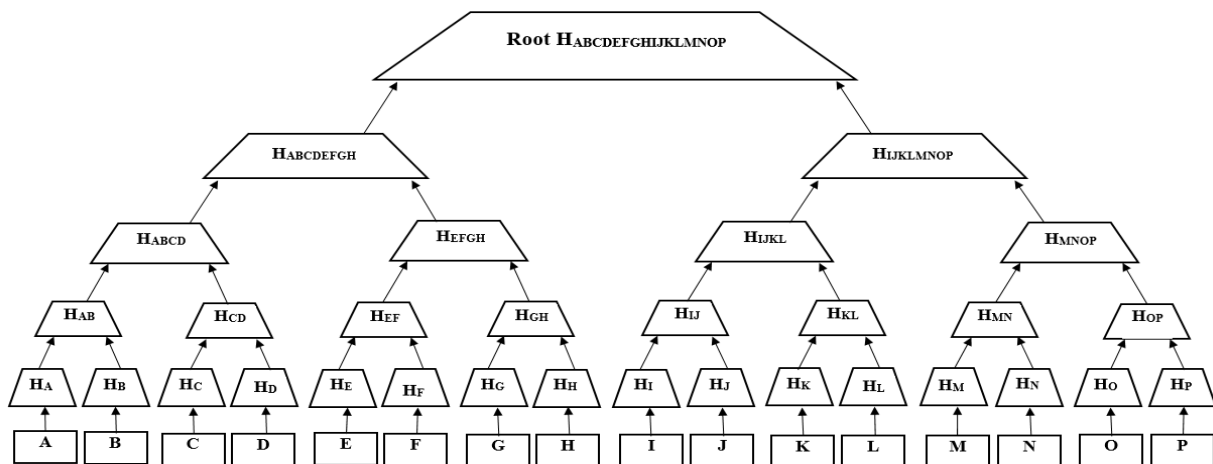


Fig. 2: An example of a Merkle tree construction [14].

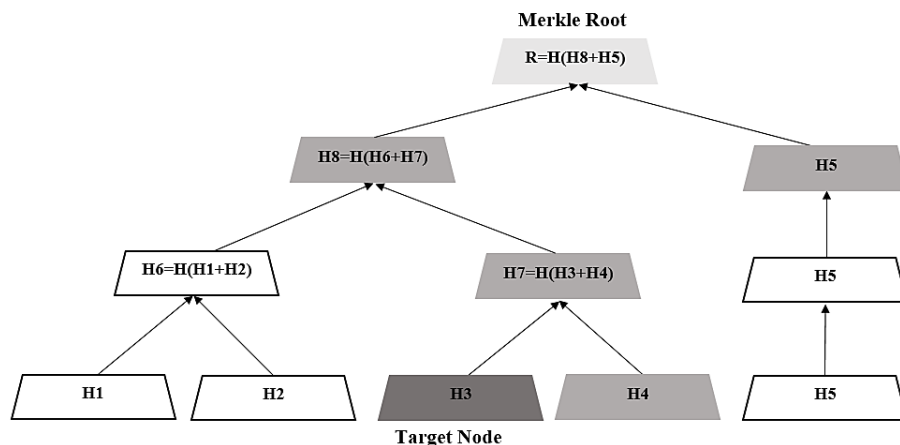


Fig. 3: Verifying hashes from a block to the root node [14], [30].

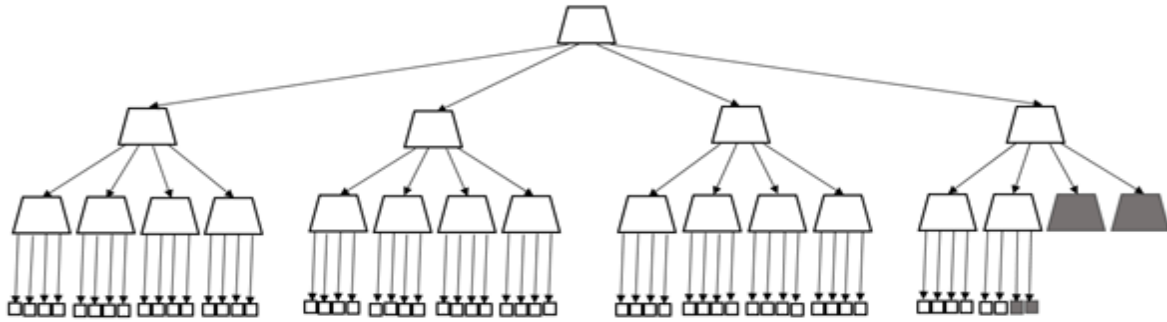


Fig. 4: An example of an MD-6 tree construction [36].

3. Graph based constructions: There are a number of hash functions defined based on Cayley graphs which are expanders too. A Cayley graph is one that encodes a group based on its generator set. An expander graph is a sparse but highly connected graph, so that each small set of vertices has many neighbors. Cayley graphs which map non-Abelian finite groups and are expanders were used to design hash functions. An example is the elliptic curves-based graph hash function defined by Charles et al. [47]. Regarding the hardness of finding cycles in an expander graph, this graph hash function used the input message to walk around an expander graph and defined collision-resistance as equivalent to finding a cycle in such a highly connected graph.

In addition, the preimage resistance of some graph hash functions depends on the hardness of the Factorization problem in non-Abelian groups [48].

B. Iterative Constructions

These constructions iterate through an arbitrary-length input to compute bitwise operations such as XOR

on fixed-length blocks of that input. Each iteration mixes an input block with either an initial value or the output of its previous iteration. The input message will be padded if its length is not an integer multiple of the block size. Hash functions based on such constructions are known as iterated cryptographic hash functions [49].

1. Merkle-Damgård construction: Merkle-Damgård construction [50], [51] was used by known hash functions MD-5, SHA-1, and SHA-2. It allows the construction of collision-resistant hash functions from collision-resistant compression functions when fixed initial values are used and the length of the input message is appended to it [19]. The same, however, is not true about pre-image resistance and second pre-image resistance properties [52]. Fig. 5 represents this construction [28]; M_i labelled boxes represent message blocks, F labelled trapezoids represent compression functions, solid lines represent dataflows, and other symbols intuitively represent initial value and output digest. This notation is common in cryptography literature with some exceptions that are considered irrelevant.

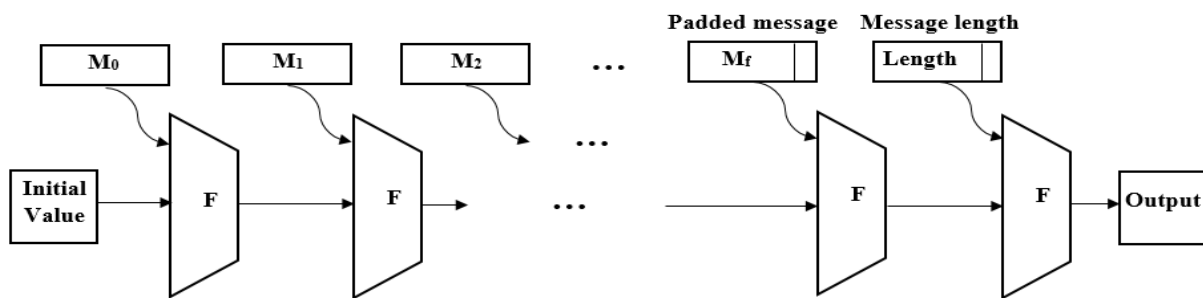


Fig. 5: Merkle-Damgård construction [28].

2. Shoup construction: The Shoup construction aims to achieve pre-image resistance and is depicted in Fig. 6 [53], [54]. It is similar to the Merkle-Damgård construction along with some mask bits that are XORed with the results of the compression function at each iteration [53], [54]. Bitwise XOR operations are represented by the \oplus symbol.

3) Zipper hash construction: Zipper hash combines the results of two different compression functions in reverse order. Hence, it is a hash function combiner and a hash function construction as well. Regarding the second collision-resistant property, this construction aims to prevent the use of a successful attack on a compression function to attack a hash function which applies it. It

employs two compression functions, f_0 and f_1 , which process input blocks in the reverse order [43].

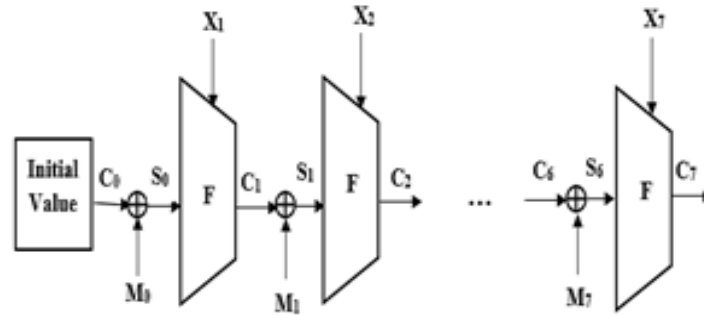


Fig. 6: Shoup construction [53], [54].

It is depicted in Fig. 7 [43] and includes the follow steps:

Step 1: Pad input message M so that the length of padded message $P(M)$ is a multiple of block size (i.e. input size of compression functions f_0 and f_1). Say blocks M_1, M_2, \dots, M_t .

Step 2: Compute h_1 as $f_0(M_1, \text{Initial Value})$, and h_2, \dots, h_t as $h_i = f_0(M_i, h_{i-1})$.

Step 3: Compute h'_1 as $f_1(M_t, h_t)$, and h'_2, \dots, h'_t as $h'_i = f_1(M_{t-i+1}, h'_{i-1})$.

Step 4: Compute output transformation function $g(h'_t)$ as hash value of input message.

The output transformation function is represented by a g labelled trapezoid.

A second pre-image attack on *Zipper hash* was introduced [42], although the time complexity of this attack was not much better than the time complexity of the brute force attack (i.e. $O(2^n)$). In addition, *Herding attack* (Section IV.A.4) was extended to attack the *Zipper hash* and other hash function constructions which process each message block more than once [55].

4: Sponge construction: The Sponge construction [20], [21] is used by the Keccak hash function which won the SHA-3 competition. This construction takes the padding algorithm as input and adds zero initiated bits which are called capacity (c) to the processing bits of each iteration which are called bit-rate (r). The ratio of capacity bits to bit-rate determines the balance between security and performance [21].

Fig. 8(a) shows how an input message is padded and processed by appending capacity bits to each block in each iteration [21]. Accomplishing such iteration through all blocks is called the *absorbing phase* which processes $b = r + c$ bits at each iteration. In addition, the Sponge construction allows users to customize the output size. If the length of required output (l) is not greater than b , then the first l bit of b is returned as output; if $l > b$, however, then the *squeezing phase* begins, so that the first r bit of the output of all *squeezing* iterations are concatenated and returned as output. Fig. 8(b) shows the squeezing phase [21].

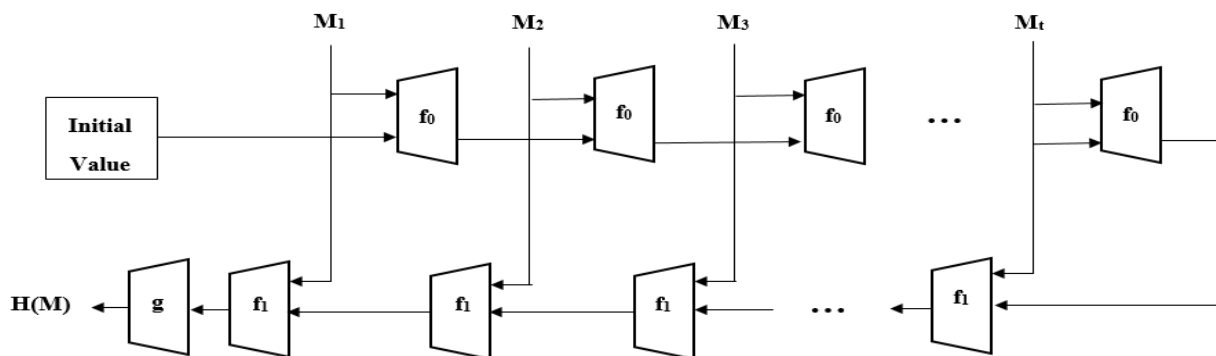


Fig. 7: Zipper hash construction [43].

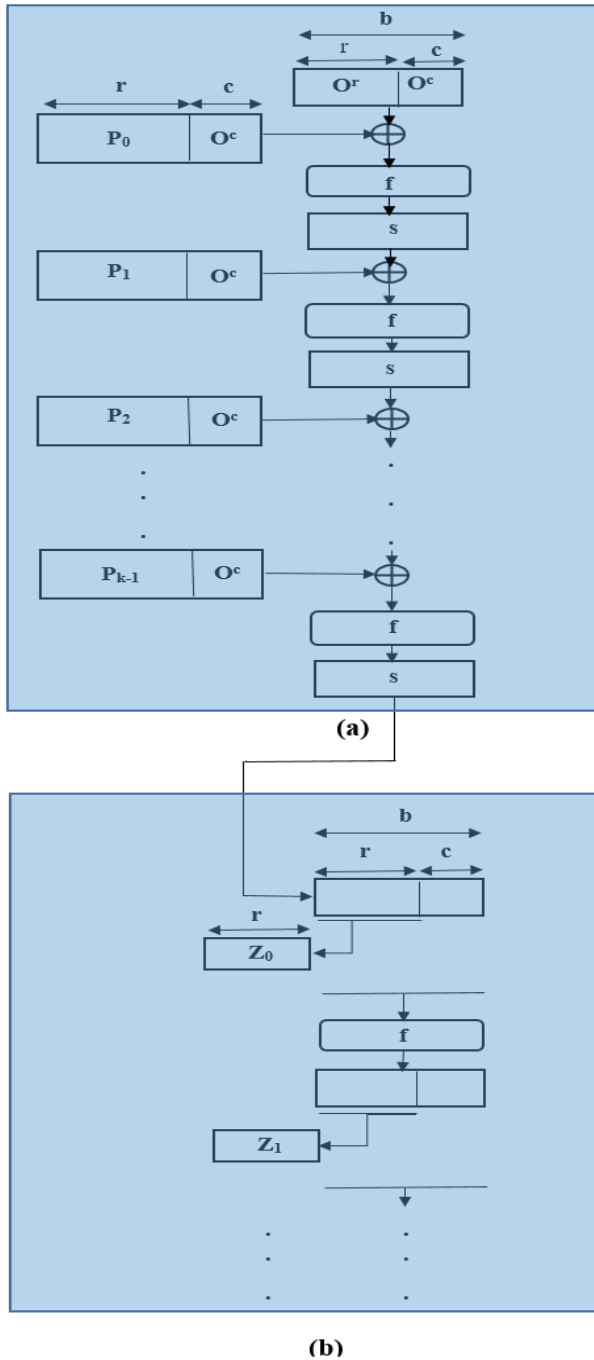


Fig. 8: Sponge construction, (a) input padding (b) squeezing output [21].

Attacks and Adversaries

How a hash function resists different attacks is the most important criterion for gaining wide acceptance. Loosely speaking, there are four categories of approaches to make an attack on a hash function: generic attacks, cryptanalysis attacks, quantum adversaries, and implementation specific adversaries. This section briefly describes these categories, and Table 1 depicts the target, method, and complexity of each attack category. The parameter n used in the last column of Table 1 denotes the length of input message which will be hashed.

A. Generic Attacks

Generic attacks are slow, but they apply to all hash functions, regardless of their algorithms and corresponding implementations. Thus, these attacks define a lower band for the output length of secure hash functions [56]. These attacks call a hash function or its compression function a number of times and seek relationships between the results. As a generic attack uses the black box model, it may cause exponential time complexity in the form of $\theta(2^{(n-k)/a})$, where n is output length of hash function, a indicates the possible order reduction by statistical methods (e.g., birthday attack and herding attack), and k is the order reduction achieved in the cost of $\theta(2^k)$ space (herding attack and rainbow tables). See Table 2.

1: Brute force attacks: A brute force attack on an n -bit hash function evaluates that function on $\theta(2^n)$ distinct input values to find (second) pre-images; considering multiple targets, say $\theta(2^t)$ targets, the cost can be reduced to $\theta(2^{n-t})$, while this degradation can be answered by parameterization of the hash function [19]. Furthermore, in some cases such as password hashing, rainbow tables, which are cached tables of precomputed hash values, may accelerate these attacks and trade increased space usage with decreased time. But random salting [57] and automatic padding [58] prevent such lookup table creations.

A brute force attack shows the worst case to find a pre-image or second pre-image on an n -bit hash function. It determines a lower bound for the output length of hash function to resist pre-image and second pre-image attacks (Similarly, birthday attack defines a lower bound for the output length of hash function to resist against collision attacks). For example, 224 bits is the lower bound used by SHA-2 and SHA-3 hash functions.

2: Birthday attacks: These algorithms find a collision based on the so-called birthday paradox in the cost of $\theta(2^{n/2})$ with a probability greater than $\frac{1}{2}$.

Seemingly unintuitive, the birthday paradox states that 23 people are sufficient to have a shared birthday occurrence with $\frac{1}{2}$ probability, i.e. the probability of finding a shared birthday (i.e. collision) for t people whose birthdays are independently distributed among the $n = 365$ days of a non-leap year is $\theta(t^2/n)$ if $t < n^{1/2}$ and is a constant value otherwise [56]; the exact value is computed by the possibility that each investigated person does not share their birthday with previously investigated persons and subtracting that product value from 1 [1]; this probability is denoted in (1).

$$p = 1 - \prod_{i=0}^{t-1} (365 - i) / 365. \quad (1)$$

This attack stores $O(2^{n/2})$ values, and it may be possible to trade off required time against memory as described by Katz and Lindell [56].

Table 1: Categories of Attacks on Hash Functions; Targets and Methods. n , a , k , b , c , d and e present output length of hash function, possible order reduction by statistical methods, order reduction achieved in the cost of $\theta(2^k)$ space, polynomial time constant value, polynomial time constant value, polynomial time constant value, sub-exponential time constant value, and , polynomial time constant value respectively

Category	Target		Method Elements	Time Complexity
Generic attacks	The output of hash function (hash value) or the output of compression function		Statistical methods and probability theory	$\theta(2^{(n-k)/a})$; where k and a are constant values
Cryptanalysis	Steps of algorithm		Detecting no random behavior in parts of a hash algorithm	From $\theta(n^b)$ to $\theta(2^n)$, where b is a constant value
Quantum adversaries	Steps of algorithm		Quantum solution for classically non-polynomial steps of algorithm, such as Integer Factorization and Discrete Logarithm.	From $\theta(n^c)$ to $\theta(2^{n/d})$, where c is a constant value
Implementation specific attacks	Physical security attacks	Dependency of Time and power consumption to executed operations and processed data. Electromagnetic fields which are emitted by processors.	Time measurements to verify the correlation between a partial key value and the expected running time, power traces, and also measuring near- and far- field of processors	$\theta(n^e)$, where e is a constant value – few time complexities.
	Software implementation attacks	Steps of algorithm implemented in a vulnerable programming language or in a vulnerable manner.	For example, buffer overflow for algorithms implemented in C language without boundary checking	

To counter this attack, one may use Universal One-Way Hash Functions (UOWHF), which are a class of hash functions that are indexed by a parameter (key) and select function instance based on selected challenge input [2].

3: Meet-in-the-middle attacks: These attacks apply to multiple encryption schemes such as double DES and find matches between encrypted values of one scheme and decrypted values of another scheme [59]. Derived from encryption, these attacks were applied for finding pre-images of reduced variants of common hash functions such as MD4 [60], [61], MD-5 [61], SHA-1 [62], [63] and SHA-2 [60], [64]. For example, Aoki et al. [64] divided the steps of the compression function and used a pre-image of the compression function to gain a pre-image of the hash function. As another example, Knellwolf and Khovratovich [62] employed the meet-in-the-middle technique along with differential cryptanalysis (differential cryptanalysis is discussed in section IV-B-3) to attack SHA-1.

4: Herding attack: A herding attack, aka the Nostradamus attack, finds (second) pre-images on a hash function by searching collisions among precomputed

compression functions.

It uses the birthday paradox to find the mentioned collisions and constitutes a diamond-shaped network of these collisions to determine a hash value that can be declared as a commitment to some predictions about the future.

At a point in the future, a second pre-image of that value which includes some happened events will be published as evidence to support that assertion [65], [66]. This attack finds a suffix that can be appended to a message, so that the concatenated message results in a hash value which is equal to the hash value claimed by attacker.

Mennink [32] improved the flexibility of the attack by adjusting trade-off between the speed of attack and the length of the (second) pre-image.

The herding attack was designed to target hash functions based on the Merkle–Damgård construction. Moreover, Andreeva et al. [55] showed the success of herding attacks on four other hash function constructions, namely concatenated, zipper, hash-twice, and tree hash constructions.

B. Cryptanalysis

Cryptanalysis exploits logical weaknesses in a hash algorithm to invert or forge hash values [67].

These attacks are generally more efficient than generic attacks, but their applicability is limited to either a specific hash function or a specific implementation of a hash function.

This section overviews four attacks in this category: length extension attack, algebraic cryptanalysis, differential cryptanalysis, and rebound attack.

The first exploits the lack of output transformations, and the second breaks codes by solving equivalent equations. The others detect primitives that hold properties leading to a non-random behavior through a number of rounds. Some cryptanalysis attacks operate in polynomial time (e.g., length extension attack and differential cryptanalysis), while others operate in exponential time and space complexity (e.g., rebound attack). See Table 2.

1: Length Extension Attack: Some hash function constructions, such as the known Merkle–Damgård one, process subsequent blocks, mix the results subsequently, and provide the internal state of the processed blocks as a hash value.

Exposure of the internal state makes the hash function vulnerable to length extension attack.

Message authentication is an example of an application which is susceptible to length extension attack. Applications may authenticate messages by prepending a secret value to the message and computing the hash of the concatenated message at both sides (i.e. sender and receiver) [67].

Such applications are susceptible to length extension attack if they use a vulnerable hash algorithm and the attacker has access to the message and its hash value, and they know or guess the length of the secret, although they do not know the secret itself.

This attack is implemented by initiating the hash algorithm with a given internal state, which is the hash value of a secret prepended to a message, and appending attacker data as subsequent blocks by subsequently feeding the algorithm.

Next, the attacker will submit the computed hash value along with a concatenated message that involves the original message, padding of the original algorithm, and attacker data to the receiver.

Output transformation is a solution to resist length extension attack [19] that is employed by hash functions such as Modular Arithmetic Secure Hash (MASH) [2] and MD-6 [36].

2: Algebraic Cryptanalysis: Algebraic cryptanalysis is a method for attacking hash functions by solving polynomial systems of equations [68]. Some hash functions are reduced to instances of a satisfiability

problem [69]. Such encoding of cryptographic algorithms and the subsequent reasoning is called *logical cryptanalysis* [70]. There are many examples of this type of attack to find second pre-images on round reduced variants of MD-4 [71], MD-5 [71] and SHA-1 [71], pre-images on a round reduced variant of MD-4 [72], and Keccak [73], [74].

3: Differential Cryptanalysis: Differential cryptanalysis seeks the relation between input differences and corresponding output differences. It is quite common to see exclusive OR (XOR) as the difference operator. In addition, operators such as modular subtraction have been used to successfully attack MD-5 [75] and SHA-1 [76] hash functions.

C. Quantum Adversaries

This section discusses quantum adversaries. Companies such as IBM, Google, D-Wave, and Microsoft have developed quantum computers using various types of qubits.

D-Wave practical quantum devices have attracted research interest [77]. While up to eight trapped-ion qubits, about ten nuclear magnetic resonance qubits, and about ten optic qubits were considered as the maximum number of qubits in 2010 [78], in 2017, D-Wave announced and shipped its new commercial quantum computer equipped with 2000 qubits [79] (D-Wave uses Adiabatic quantum computation instead of gate-based quantum computation).

In 2019, D-Wave announced a new 5000 qubit device too.

Moreover, Microsoft announced that the company is going to offer a full-fledged topological quantum computing system which includes hardware, software, and programming languages, so that a free preview of the programming language which supports simulation of up to 30 logical qubits on personal computers (or up to 40 logical qubits on Azure) would be released by the end of the 2017 [80]. Microsoft Quantum Development Kit including Q# programming language is a released part of this stack.

Moreover, programming languages and software development kits (SDKs) such as Google qsim [81], IBM Qiskit [82], D-Wave Ocean [83], Scaffold [84], Quipper [85], and Microsoft LIQUi> [86] facilitate the transition from high-level quantum algorithms to low-level gate representation, different architectures, error correction, and so on.

The emergence of these commercial quantum computers (D-Wave and in future Microsoft) connoted the existence of both opportunity of quantum cryptography schemes and threat of quantum adversaries.

Tackling the latter is referred to as post-quantum cryptography.

Table 2: Categories of Hash Functions – Analysis and applicability of attacks

Category	Algorithm	Advantages	Disadvantages	Applicable attacks
Hash functions based on a block cipher	A block cypher algorithm strengthened with one-wayness.	Reuse of existing block cypher effort and benefit from compact implementation due to the cypher reuse	Each function of these category maps an invertible block cypher algorithm to a noninvertible hash function. The second vulnerability map is between 64- and 128-bit block length of block cyphers and 224- to 1024-bit length of hash values required for securing against generic attacks. Furthermore, any vulnerability on the underlying block cypher may lead to a vulnerability on the associated hash function.	There are successful known side channel attacks for block cyphers such as DES and AES. These attacks are candidates for attacking the corresponding hash functions.
Hash functions based on algebraic structures	Reduction of pre-image resistance, second pre-image resistance, and collision search to computationally hard problems such as Integer Factorization and square root.	Provable security	Computation of these hash functions includes operations such as modular multiplication that are time-consuming and impose dependency between processed data and consumed time and power. Hence, many of the hash functions in this category are slow and are prone to side channel attacks.	Firstly, some computationally hard problems such as Integer Factorization which belong to BQP computational complexity class have known attacks for specific values of parameters. Secondly, due to the use of data dependent operations such as modular multiplication, these hash functions are prone to side channel attacks.
Custom designed hash functions	Designed from scratch and usually use bitwise operations to cause confusion and diffusion.	Usually have better performance than algebraic-based hash functions	The lack of provable security is commonplace in these hash functions.	The use of bitwise operations (e.g., XOR, AND, and circular shifts) improves their performance and reduces their vulnerability against side channel attacks. In different categories, however, successful attacks on these hash functions are reported.
Physical unclonable functions	Instead of explicit algorithms, they benefit from intrinsic randomness resulted from manufacturing variations.	The absence of any explicit algorithms makes cryptanalysis and quantum adversaries inapplicable. In addition, PUFs are tamper resistant. Hence, active physical attacks are inapplicable, too.	The use of PUFs requires skill in hardware description languages such as VHDL. In essence, PUFs map fixed length inputs to fixed length outputs, while arbitrary length input is desired. Moreover, they require error correction enhancements.	Some variants of side channel attacks (i.e. power analysis and timing attacks) enhanced by machine learning are used to read the output of these functions.
Quantum hash functions	Problems that are zero-knowledge against general quantum attacks are implemented by either a classical or a quantum algorithm.	Security is proved by reduction to computationally hard problems which belong to NP-BQP. That is, they belong to NP computational complexity class but not BQP.	Problems such as Approximate Closest Lattice Vector require considerable computational resources even for the computation of a hash value.	Being new, there are no reported attacks on these hash functions, and there is no evident proof of their strength against cryptanalysis.
Memory hard functions	Algorithms with huge amounts of memory usage and memory considerations including input-independent memory addressing, input-dependent memory addressing, and planned number of passes over the memory	Increasing the cost of ASIC-based attacks in terms of memory usage and energy consumption	As they intentionally lack the efficient input-to-output mapping property (Property I in Section 2), many resource constrained devices cannot afford to use these functions.	Cryptanalysis attacks which employ time-memory trade-off may target MHFs with input-independent memory addressing. Furthermore, side channel attacks may target MHFs with input-dependent memory addressing.
Optical hash functions	Use “confusion” and “diffusion” of modulated light instead of computation of a compression function	Low processing amounts due to taking advantage of natural randomness instead of computation	Complexity of setting up an optoelectronic system in a noisy environment	Being new, there are no reported attacks on these hash functions, and although there is no explicit algorithm, there is no evident proof of their strength against cryptanalysis.

In brief, quantum computing upsides include:

- I. Significant speedup: There are quantum algorithms for some computationally hard problems such as Factoring and Ground State Estimation that are exponentially faster than the best classical algorithms for those problems [87]. Such problems belong to the Bounded-error Quantum Polynomial (BQP) computational complexity class which can be solved efficiently on a quantum computer with a bounded probability of error [88].

Their disadvantages include:

- I. Error correction: Resisting communication channel noise errors such as bit-flip errors and phase errors, and tolerating computational faults such as faulty logic gates are necessary and are achieved through error correction techniques such as employing redundant qubits [88].
- II. Scalability problems: Existence of noise and entanglement phenomena cause scalability problems [89].

Shor [90] introduced polynomial time algorithms for Factorization and Discrete Logarithms on quantum computers. Grover's quantum searching algorithm [91], [92] can find a 256-bit AES key in about 2^{128} quantum operations [93] and is used to find hash pre-images [94]. Furthermore, there are quantum attacks to find hash collisions [95].

In contrast to problems such as Factorization and Discrete Logarithms which have polynomial time quantum algorithms [90], post-quantum cryptography [96] tends to introduce problems that cannot be solved by quantum computers in polynomial time. Watrous [97] proved that problems such as Graph Isomorphism and Graph 3-coloring are zero-knowledge against general quantum attacks. Kashefi and Kerenidis [98] defined several quantum one-way functions such as Graph Non-Isomorphism, Approximate Closest Lattice Vector, and Group Non-Membership and generalize their results for any hard instance of Circuit Quantum Sampling problem as a candidate quantum one-way function.

D. Physical Security: Side Channel Attacks

Classical cryptanalysis views steps of algorithms as transformation of inputs to outputs. Conversely, physical security views specific characteristics imposed by an implementation of those steps which are running on a specific processor in a specific environment. Physical attacks may or may not depackage the chip; such situations are called invasive or noninvasive attacks, respectively. In addition, physical attacks may or may not try to tamper with the proper functioning of the device and are called active or passive attacks, respectively [2]. Side-channel attacks, or environmental attacks, exploit dependency of information such as running time, power consumption, and electromagnetic emissions of operated

data and performing instructions to (statistically) learn about an algorithm's internal state [2], [99] or expose the device's secrets. The SHA-3 finalists were evaluated against three variants of side channel attack: timing attack, power analysis, and electromagnetic analysis. The evaluation declared the sufficient security margin of all finalists and found collisions on the round reduced variant of Keccak [99]. Cryptographic algorithms prevent such attacks by avoiding the use of data-dependent or power-dependent operations such as multiplications, data-dependent rotations, and table lookups.

PUFs (See Section V-D) are tamper resistant variants of hash functions, but there are polynomial time side channel attacks on PUFs [100] that enable the attacker to read the generated output value.

In addition to physical security, there are adversaries which consider an implementation of a security primitive from the viewpoint of software and programming language flaws. The buffer overflow found on the C language implementation of MD-6 is an instance of such software implementation attacks [19].

Hash Function Categories

This section describes cryptographic hash functions in seven categories and analyses the strengths and vulnerabilities of each category (See Table 2). The proposed seven-category classification includes hash functions based on a block cipher, hash functions based on algebraic structures, custom-designed hash functions, PUFs, quantum hash functions, MHFs, and optical hash functions. To the best of our knowledge, the last four mentioned categories have not been sufficiently addressed in most existing surveys [18], [19], [26]-[29].

A. Hash Functions Based on Block Ciphers

Developed mostly based on DES and AES, these hash functions reuse underlying block ciphers to achieve a compact implementation. The main challenges of these hash functions lie in designing a noninvertible construction based on an invertible block cipher. The SHA-3 finalist BLAKE [101] and Russian standard hash Streebog [24] are two known hash functions of this category.

B. Hash Functions Based on Algebraic Structures

Most hash functions in this category use computationally hard problems such as Factorization, Discrete Logarithm, Knapsack, Lattice Problems, and Elliptic Curves and prove their security by reduction [102]. Some of these hash functions, though, allow the insertion of trapdoors to construct collisions by the person who chooses the design parameters [2]. The functions based on modular arithmetic suffer from being slow. There are many attacks for specific instances of hard problems, such as RSA [103]. As an example, collision resistancy of Very Smooth Hash (VHA) [104] is reduced to find nontrivial

modular square roots, but this function is not pre-image resistant [105]. Modular Arithmetic Secure Hash (MASH) was published as an International Organization for Standardization (ISO) standard on December 1998 and was reviewed and re-confirmed as current version of standard in 2022 [106]. It has strong output transformation but its security is not supported by a mathematical proof. Finite field is used to define some hash functions [107]. A recent survey on hash functions based on computational problems defined on lattices was provided by Mishra et al [108]. Furthermore, hash functions based on Cellular Automata [109] are newly introduced members of this category.

Finally, another important family of hash functions comprises chaos-based hash functions. A chaotic system behaves in an unpredictable but deterministic manner and is highly sensitive to initial conditions, so a very small change in its initial state may have a large effect on its later state. A chaotic map is a mathematical function which states such a chaotic behavior in one- or multi-dimensions. As an example, Teh et al. [110] presented a compression function based on a one-dimensional chaotic map and used Merkle–Damgård construction to process arbitrary-length messages.

C. Custom-Designed Hash Functions

Known cryptographic hash functions including MD-2, MD-4, SHA-1, SHA-2, and SHA-3 (Keccak) are instances of this category. These algorithms are designed independent from other security primitives. Although these hash functions do not provide provable security and their security depends on confusion and diffusion, the use of bitwise operations such as XOR, AND, and circular shifts leads to low processing time and partial security against side channel attacks, even though there are some reports of such attacks [111].

D. Physical Unclonable Functions (PUFs)

PUFs are hardware based security primitives and provide challenge response behavior based on manufacturing variations that occur on a small scale. Their intrinsic unpredictability stems from random elements (e.g., various gate delay) in their manufacturing process [112], [2]. Depending on the usage, this challenge response behavior may be provided in an invertible or non-invertible manner [113]. An individual PUF device, however, cannot be practically cloned or copied, even with access to the exact manufacturing process that produced it in the first place. This intrinsic randomness reduces computational costs, thus making PUFs a candidate for the security of resource-constrained devices such as embedded systems [114], and IoT [113].

There are two notable PUF types: Weak PUFs and Strong PUFs; the former accepts one or a few challenges and is employed as a secret key for device specific

encryption, while the latter accepts, possibly, an exponential number of challenges and is considered as a physical hash function [115]. SRAM PUFs and their variants are the most popular implementation of Weak PUFs and Arbiter PUFs, and their variants are the most popular implementation of electrically Strong PUFs. Weak PUFs suffer from cloning and invasive attacks (e.g., Helfmeier et al. [116] created a physical clone of a SRAM PUF using Focused Ion). Cloning and invasive attacks are hardly applicable on Strong PUFs. The most common attacks on Strong PUFs are modeling attacks [117], side channel attacks [118], and the combination of both [100], [119].

To conclude, PUFs benefit from the following advantages:

- I. Instead of storing a hash value or a secret key on the device that includes both security consideration and additional device memory cost, the PUF response is derived when needed [115].
- II. Most types of PUFs are tamper-resistant [115], but there are some side channel attacks enhanced by machine learning [100].

and suffer from the following disadvantages:

- I. PUFs are prone to error and need to employ an error correction mechanism. Depending on PUF type, error correction may be executed on a PUF holding device or on a communication server [115].
- II. In contrast to non-physical approaches, PUFs are prone to aging [115].

In essence, PUFs are maps between fixed length inputs and fixed length outputs, while arbitrary length input is desired. Therefore, PUFs are widely used for authentication and rarely used for integrity checks (a common application of hash functions).

Finally, PUFs based on nanotechnology are the recently reported trend of PUF design [120].

E. Quantum Hash Functions

There are two sub-categories of quantum hash functions, i.e. hard problems which belong to postquantum cryptography and hash functions based on quantum state. The former was described in Section 4.3, and the latter is discussed in the current section. In addition to the mentioned subcategories, there are quantum hash functions which operate on classical inputs and produce classical outputs [121].

Ziatdinov [122] and Yang et al. [121] attributed the first state-based quantum hash function to Buhrman et al. [123], who introduced the notion of quantum fingerprinting. Abelayev and Vasiliev [124], [125] introduced quantum hash functions that map input data to quantum states so that the functions have pre-image resistance (sampling property), second pre-image resistance, and collision resistance properties. Abelayev et al. [126] discussed the reverse relation between the pre-

image resistance and collision resistance properties of quantum hash functions and introduced a construction to build balanced quantum hash functions.

F. Memory-Hard Functions

There are cases such as cryptocurrency mining and password hashing in which a hash function without an efficient input-to-output mapping property (*Property I* in Section 2) is desired. In contrast to the design goals of distributed electronic payment systems such as Bitcoin, multicore CPUs, GPUs, and dedicated ASIC modules are used to accelerate cryptocurrency mining at a low cost. This consolidates the computing power of the network. Some ASIC miners are roughly 200,000 times faster and 40,000 times more energy efficient than a modern multicore CPU [127]. Dictionary attacks on hashed password databases are further examples of such parallel computation.

The ASIC resistance property (*Property X* in Section II) aims to reduce attackers' massively parallel advantage. To this end, MHFs [128] and BHF [127] were introduced to increase the hardware capital cost and energy consumption, respectively. Percival [128] put forward the MHFs idea that with an increase in the size of a hash derivation circuit, the number of possible circuits on a given area of silicon will decrease. Furthermore, he introduced the *scrypt* hash function [128], [129] as the first instance of MHF.

Input-independent memory addressing, input-dependent memory addressing, and number of passes over the memory are major considerations in designing an MHF. For example, *Argon2* hash function [22] includes the following three variants:

- I. *Argon2d*: It uses data-dependent memory access and targets the design of cryptocurrency Proof-of-Work (PoW).
- II. *Argon2i*: It uses data-independent memory access to resist side channel attacks and includes more passes over the memory in comparison with *Argon2d*. *Argon2i* aims to secure password hashing.
- III. *Argon2id*: It is not a part of *Argon2* hash function proposal [22] and use a sequential composition of data-dependent and data-independent memory accesses. First half pass uses data-independent memory access and the second half uses data-dependent memory access.

As a last example of MHFs, Zamanov et al. [34] evaluated the memory demand of Equihash and Ethash algorithms. The former increases PoW memory usage based on the birthday problem, while the latter fills a huge amount of memory and searches within it.

Although MHFs incur additional capital costs, ASICs require far less energy than CPUs. To this end, BHFs define a large number of planned memory accesses to avoid the energy saving of ASIC hash engines [127].

G. Optical Hash Functions

Because of physical properties of light such as velocity and its parallel nature, light-based computing is promising and has been shown to outperform electronic computing in some cases [130]. Optical hash functions are photoelectric systems which encode blocks into images known as the "information plane" [131] and replace computations of a compression function with "confusion" and "diffusion" of modulated light [132]. Amplitude-only spatial light modulator, phase-only spatial light modulator, charge coupled devices along with lenses [131], half mirrors [131], and/or scattering media [132] are the basic constituents of such systems. As an example, Wen-Qi et al. [132] proposed an optical hash function which is based on scattering media and provides the avalanche effect and collision resistance. As another example, He and Peng [131] proposed two optical hash functions based on phase-truncated Fourier transform and interference phenomena (i.e. two beam interference). Last but not least on our list of examples, as noise inherent in free space setup can affect the security and performance of beam interference and phase truncation-based hash functions, Kumar et al. [133] proposed an optical hash function based on superposition.

Mobile Service Requirements

Mobile devices can consume some services and also provide some other services, but they have several constraints on their resources which may jeopardize the Quality of Service (QoS). On the other hand, as mobile devices roam between environments, they are exposed to more attacks than stationary computers. Hence, lightweight but not less secure cryptographic hash functions which secure interactions of resource-constrained devices are urgently needed. Mobile service requirements are as follows:

- I. Roaming may cause inaccessibility of some resources and accessibility to some others. To aid service continuity, hash functions are used to identify identical alternative resources and mutual authentication of the mobile device and remote servers [134].
- II. Most mobile devices have low processing power in comparison with desktop computers.
- III. Most mobile devices have small memory size in comparison with desktop computers.
- IV. Limited battery capacity makes energy consumption an important consideration for mobile devices. Not only does WS-Security hash computation required by service invocation consume energy, but also the battery usage of hash computation is important to avoid power analysis side channel attacks [135].
- V. Mobile device bandwidth is limited by the network interfaces of that device and by the network being

used. This limit mediated mobile WS-Security solutions usage [136], [137].

VI. From time to time, mobile devices undergo connection intermittence caused not only by roaming, but also by things such as other wireless devices, microwave ovens, and other devices with poorly shielded cabling.

VII. Some mobile devices have multiple network interfaces such as Wi-Fi, Bluetooth, NFC, and GPRS (in addition to LoRaWAN and ZigBee for IoT). To benefit from multi-homed architectures, authentication and integrity achieved by hash functions are urgent needs for mobile service communications [138].

Hence, low processor usage, thrifty memory usage, and limited battery usage are urgent needs of application-specific hash functions for mobile services. In addition, due to connection intermittence and bandwidth limitation, mobile security-related computations such as hash computation can hardly be delegated to servers that are available through wireless connections. For simplicity, the application-specific hash function for mobile services will be referred to hereafter as mobile hash functions.

Such mobile hash functions need to cope with the mentioned limitations, and it is desirable that they benefit from multi-homed architectures. Table 3 shows the appropriateness of each hash function category for satisfying mobile service requirements. As Table 3 outlines, optical hash functions and state-based quantum hash functions are not applicable for mobile devices. Algebraic-based functions benefit from provable security but have high computational costs. Bitwise equivalent of algebraic structures that belongs to post-quantum cryptography seemed like a good idea, but we could not find such algebraic-based hash functions in practice. PUFs have very low computational costs and communicate just challenge-responses. In addition, PUFs are available for IoT nodes [139]. Hence, we suggest PUFs with polynomial-time error correction for mobile service hashing.

Application Scenarios

All applications do not have the same requirements for security and performance. There are a number of application scenarios for cryptographic hash functions. Four scenarios and their corresponding analysis to select appropriate cryptographic hash functions are presented in Table 4. The first scenario benefits from the parallel processing capability of hash functions such as MD-6. The second scenario uses the intrinsic randomness of PUFs to lighten hash computation load for resource constrained sensor nodes. The third shows the usage of hash chains for process authentication. Finally, the last scenario shows the need for output transformation in the lack of encryption.

Conclusion

Massive usage, significant competitions such as the SHA-3 competition, the Password Hashing competition and the NIST lightweight competition, and nationwide hash standards [20], [21], [23]-[25] have led to the introduction of new hash functions and new hash function constructions. To the best of our knowledge, recent research and competitions make the following futuristic trends possible: Resource constrained devices are used in IoT solutions such as smart farming and smart cities. Security plays a crucial role in the success such systems so that employing hash functions need to be both resource efficient and side-channel resistant [141]. Hence, lightweight hash functions received great attention in recent years so that IoT specific hash functions emerged and NIST lightweight competition is ongoing since 2018 [142]-[144]. In contrast to the lightweight design of these hash functions, it is important that a hash function cannot be computed too fast on massively parallel computers and quantum computers. Hence, evaluation of hash functions on quantum computers is a recent measure to avoid brute force attacks [145].

Table 3: Appropriateness of each Hash Function category for satisfying mobile service requirements

Row	Hash Function Category		Mobile Service Hash Consideration				
			Processing	Memory	Battery	Security	Applicability
1	Hash functions based on a block cipher		High				
2	Hash functions based on algebraic structures		High			Proven	
3	Custom-designed hash functions		Low				
4	Physical unclonable functions		Very low	None or very low (depending on PUF type)			
5	Quantum hash functions	Quantum states	No reported work (have not found yet)				Not applicable
		Post-quantum cryptography	High				
6	Memory-hard functions			High			
7	Optical hash functions		No reported work (have not found yet)				Not applicable

Table 4: Application scenarios – selecting appropriate Hash Function

Row	Scenario Name	Scenario	Analysis
1	A file server on a multiprocessor host	A multiprocessor file server stores some large multimedia files. This server needs to provide the hash value of each file as a checksum. Users can download files along with corresponding checksums. To ensure a file has not been tampered with after the checksum was created, user computes the hash of the downloaded file and compares it with the checksum.	Computing hash for large files connotes the need for fast computation. It may be obtained by using a fast hash function such as BLAKE [101], [140] (BLAKE 2 or 3) or a multiprocessing support hash function such as MD-6 [36]. The multiprocessor server indicates the latter function as choice.
2	Message authentication in a sensor network	A sensor network sends monitored data to a server. A hash function is used for message authentication. Each sensor node has limited memory and limited processing speed. More importantly, each sensor node operates with limited battery energy and will die as its energy is consumed.	Resource constraints of sensor nodes and the reverse relationship between energy consumption and node lifetime suggest the use of intrinsic properties of sensors instead of running a hash algorithm on these nodes. Hence, PUFs [139] are appropriate for this scenario.
3	One-time passwords	In a geographically distributed organization, it is required that two processes hosted on different servers authenticate and communicate with each other. There is no deployed authentication (or encryption) facilities such as Primary Key Infrastructure (KPI).	This scenario may benefit from one-time passwords that are a hash chain made by consecutive computation of hash values and using the hash values in descending order (using last value first). Any hash function that supports the one-way property is appropriate for this scenario, so that an eavesdropper cannot use an observed password to compute the next valid password.
4	Authentication and integrity without encryption	A key is shared between sender and receiver. To send a message, the sender hashes that message prepended by the shared key. Then the message along with the hash value is transmitted to the receiver. Having the shared key, the receiver will hash the received message prepended by the shared key and compares it with the received hash value.	This scenario is prone to length extension attack (Section 4.2.1). It allows the attacker to forge messages with the same prefix. Hence, both authentication and integrity will be lost. Section 4.2.1 pointed out that exposure of the internal state of the hash function causes this vulnerability. Hence, hash functions benefitting from output transformation such as SHA-3 (Keccak) and MASH (section 5.2) are appropriate for this scenario.

As mentioned, PUFs based on nanotechnology are the recently reported trend of PUF design [120]. Last but not least, optical computing has a long history to trace back and was introduced 60-year ago [146], but optical hash functions were introduced in recent years are among the futuristic trend of hash functions. In addition, application-specific properties have been defined for applications such as cryptocurrency and video hashing. In this article, we discussed 11 properties of hash functions (Section 2), overviewed the concepts of compression function and domain extension, and outlined four iterative and three noniterative hash function constructions and combiners (Section 3). The current research also investigated those hash functions and proposed a seven-category classification (Section 5). To the best of our knowledge, four out of seven categories have not been sufficiently addressed in most existing surveys [18], [19], [26]–[29]. In addition, this article discussed some attacks affecting each category (Table 2) and summarized what effective attacks entail (Section 4).

Furthermore, considering the prevalence of mobile devices, this paper discussed mobile service requirements on hash functions (Section 6), outlined how each hash function category fits these requirements (Table 3), and suggested (strong) PUFs with polynomial-time error correction for mobile service hashing. In addition, the bitwise equivalent of algebraic structures that belong to post-quantum cryptography seemed like a good idea, but we could not find such algebraic-based hash functions in practice. Finally, to clarify the usage, four application scenarios and their corresponding analysis to select appropriate cryptographic hash functions were presented (Table 4). The authors aim to extend this work by extracting patterns which fulfill the 11 properties discussed in second section. This extension, along with the other mentioned benefits, can assist design, choice, and analysis of hash functions.

Author Contributions

Second and third authors supervised this research by sketching roadmap, and evaluating the results at each

step. First author searched in authentic journals and research repositories to gather all relevant papers, and read the selected papers in details. In addition, he made a comparison of investigated hash functions.

All authors discussed and analyzed the results and cooperatively summed up the work.

Acknowledgment

The authors gratefully thank the anonymous reviewers and the editor of JECEI.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>MHF</i>	Memory-hard Functions
<i>BHF</i>	Bandwidth-hard Functions
<i>PUF</i>	Physical Unclonable Function
<i>SHA</i>	Secure Hash Algorithm
<i>UOWHF</i>	Universal One-Way Hash Functions
<i>WS-Security</i>	Web Services Security

References

- [1] J. Hoffstein, J. Pipher, J. H. Silverman, *An introduction to mathematical cryptography*. New York: Springer, 2008.
- [2] H. C. A. van Tilborg, S. Jajodia, Eds., *Encyclopedia of cryptography and security*. Springer Science+Business Media, 2011.
- [3] J. Keller and S. Wendzel, "Reversible and plausibly deniable covert channels in one-time passwords based on hash chains," *Appl. Sci.*, 11(2): 731, 2021.
- [4] W. Stallings, *Cryptography and network security: principles and practice*, sixth ed. Pearson Education, 2014.
- [5] C. Wang, S. J. Li, D. Wang, Q. H. Wang, "P-28: A method of holographic encryption based on hash function," *Dig. Tech. Pap.*, 47(1): 1228–1230, 2016.
- [6] L. C. Washington, *Elliptic curves: number theory and cryptography*, second ed. Boca Raton, FL: Chapman & Hall/Crc, 2008.
- [7] R. C. Merkle, "A certified digital signature," in *Proc. Conf. on the Theory and Application of Cryptology*: 218–238, 1989.
- [8] J. Rosenberg, D. L. Remy, *Securing web services with WS-security: demystifying WS-security, WS-policy, SAML, XML signature, and XML encryption*. Indianapolis, Ind.: Sams, 2004.
- [9] A. Nadalin, C. Kaler, R. Monzillo, P. Hallam-Baker, Eds., *Web services security: SOAP message security 1.1*. OASIS, 2006. Last accessed: Jan. 7, 2023.
- [10] L. Demir, A. Kumar, M. Cunche, C. Lauradoux, "The pitfalls of hashing for privacy," *IEEE Commun. Surv. Tutor.*, 20(1): 551–565, 2018.
- [11] M. Wang, M. Duan, J. Zhu, "Research on the security criteria of hash functions in the blockchain," in *Proc. the 2nd ACM Workshop on Blockchains, Cryptocurrencies, and Contracts*: 47–55, 2018.
- [12] S. Abed, R. Jaffal, B. J. Mohd, M. Al-Shayegi, "An analysis and evaluation of lightweight hash functions for blockchain-based IoT devices," *Cluster Comput.*, 24(4): 3065–3084, 2021.
- [13] A. Kuznetsov, I. Oleshko, V. Tymchenko, K. Lisitsky, M. Rodinko, A. Kolhatin, "Performance analysis of cryptographic hash functions suitable for use in blockchain," *Int. j. comput. netw. inf. secur.*, 13(2): 1–15, 2021.
- [14] A. M. Antonopoulos, *Mastering bitcoin: Programming the open blockchain*, 2nd ed. O'Reilly Media, 2017.
- [15] J. Garay, A. Kiayias, N. Leonardos, "The bitcoin backbone protocol: Analysis and applications," in *Proc. Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques*: 281–310, 2015.
- [16] G. Wu, J. Han, Y. Guo, L. Liu, G. Ding, Q. Ni, L. Shao, "Unsupervised deep video hashing via balanced code for large-scale video retrieval," *IEEE Trans. Image Process.*, 28(4): 1993–2007, 2019.
- [17] M. S. Jan, M. Afzal, "Hash chain based strong password authentication scheme," in *Proc. 13th Int. Bhurban Conf. on Applied Sciences and Technology (IBCAST)*: 355–360, 2016.
- [18] A. A. Alkandari, I. F. Al-Shaikhli, M. A. Alahmad, "Cryptographic hash function: A high level view," in *Proc. 2013 Int. Conf. on Informatics and Creative Multimedia*: 128–134, 2013.
- [19] B. Preneel, "The First 30 Years of Cryptographic Hash Functions and the NIST SHA-3 Competition," in *Proc. Cryptographers' track at the RSA Conf.*: 1–14, 2010.
- [20] G. Bertoni, J. Daemen, M. Peeters, G. Van Assche, "Keccak," in *Proc. 32nd Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques*: 313–314, 2013.
- [21] W. Stallings, "Inside SHA-3," *IEEE Potentials*, 32(6): 26–31, 2013.
- [22] A. Biryukov, D. Dinu, D. Khovratovich, "Argon2: new generation of memory-hard functions for password hashing and other applications," in *Proc. 2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbruecken, Germany: 292–302, 2016.
- [23] S. Shen, X. Lee, R. Tse, W. Wong, Y. Yang, "The SM3 cryptographic hash function," draft-sca-cfrg-sm3-02, 2018.
- [24] V. Dolmatov, A. Degtyarev, "GOST R 34.11-2012: hash function," *RFC 6986*, 2013.
- [25] R. Oliynykov, I. Gorbenko, O. Kazymyrov, V. Ruzhentsev, O. Kuznetsov, Y. Gorbenko, A. Boiko, O. Dyrda, V. Dolgov, A. Pushkaryov, "A new standard of Ukraine: The Kupyna hash function," *Cryptology ePrint Archive*, DSTU 7564: 2014, 2015.
- [26] S. Bakhtiari, R. Safavi-Naini, J. Pieprzyk, "Cryptographic hash functions: A survey," *Department of Computer Science, University of Wollongong, Technical Report 95-09*, Jul. 1995.
- [27] J. Delvaux, R. Peeters, D. Gu, I. Verbauwhede, "A survey on lightweight entity authentication with strong PUFs," *ACM Comput. Surv.*, 48(2): 1–42, 2015.
- [28] I. Mironov, "Hash functions: Theory, attacks, and applications," *Microsoft Research, Silicon Valley Campus*, 1–22, Nov. 2005.
- [29] R. Purohit, U. Mishra, A. Bansal, "A survey on recent cryptographic hash function designs," *Int. J. Emerging Trends & Technology in Computer Science (IJETTCS)*, 2(1): 2278–6856, 2013.
- [30] A. Narayanan, J. Bonneau, E. W. Felten, A. Miller, S. Goldfeder, *Bitcoin and cryptocurrency technologies: A comprehensive introduction*. Princeton, NJ: Princeton University Press, 2016.
- [31] M. Rjaško, "On chosen target forced prefix preimage resistance," *Tatra Mt. Math. Publ.*, 47(1): 115–135, 2010.
- [32] B. Mennink, "Increasing the flexibility of the herding attack," *Inf. Process. Lett.*, 112(3): 98–105, 2012.
- [33] E. Andreeva, B. Mennink, "Provable chosen-target-forced-midfix preimage resistance," in *Int. Workshop on Selected Areas in Cryptography*: 37–54, 2011.
- [34] A. R. Zamanov, V. A. Erokhin, P. S. Fedotov, "ASIC-resistant hash functions," in *Proc. 2018 IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*: 394–396, 2018.
- [35] H. Chen, Y. Wo, G. Han, "Multi-granularity geometrically robust video hashing for tampering detection," *Multimed. Tools Appl.*, 77(5): 5303–5321, 2017.
- [36] R. L. Rivest, B. Agre, D. V. Bailey, C. Crutchfield, Y. Dodis, K. E. Fleming, A. Khan, J. Krishnamurthy, Y. Lin, L. Reyzin, E. Shen, J. Sukha, D. Sutherland, E. Tromer, Y. L. Yin, "The MD6 hash function—a proposal to NIST for SHA-3," *Submission to NIST*, 2(3), 2008.
- [37] R. Reischuk, M. Hinkelmann, "One-way functions - mind the trap - escape only for the initiated," in *Proc. Algorithms Unplugged*, Berlin, Heidelberg: Springer Berlin Heidelberg, 131–139, 2011.

- [38] O. Goldreich, S. Goldwasser, S. Halevi, "Collision-free hashing from lattice problems," in Proc. Studies in Complexity and Cryptography. Miscellanea on the Interplay between Randomness and Computation, Berlin, Heidelberg: Springer Berlin Heidelberg, 30-39, 2011.
- [39] W. Hu, N. Xie, L. Li, X. Zeng, S. Maybank, "A survey on visual content-based video indexing and retrieval," IEEE Trans. Syst. Man Cybern. C Appl. Rev., 41(6): 797-819, 2011.
- [40] A. Appleby, Murmurhash 3.0, 2016. Last accessed: Aug. 25, 2022.
- [41] C. Estébanez, Y. Saez, G. Recio, P. Isasi, "Performance of the most common non-cryptographic hash functions," Softw. Pract. Exp., 44(6): 681-698, 2014.
- [42] S. Chen and C. Jin, "A second preimage attack on zipper hash," Secur. Commun. Netw., 8(16): 2860-2866, 2015.
- [43] M. Liskov, "Constructing an ideal hash function from weak ideal compression functions," in Proc. 13th Int. Workshop on Selected Areas in Cryptography: 358-375, 2006.
- [44] B. Denton, R. Adhami, "Modern hash function construction," in Proc. the Int. Conf. on Security and Management (SAM): 479-483, 2011.
- [45] Z. Bao, I. Dinur, J. Guo, G. Leurent, L. Wang, "Generic attacks on hash combiners," J. Cryptology, 33(3): 742-823, 2019.
- [46] M. Fischlin, A. Lehmann, D. Wagner, "Hash function combiners in TLS and SSL," in Proc. Cryptographers' Track at the RSA Conf.: 268-283, 2010.
- [47] D. X. Charles, K. E. Lauter, E. Z. Goren, "Cryptographic hash functions from expander graphs," J. Cryptology, 22(1): 93-113, 2009.
- [48] C. Petit, J. J. Quisquater, "Cryptographic hash functions and expander graphs: The end of the story?," in Proc. The New Codebreakers, Berlin, Heidelberg: Springer Berlin Heidelberg: 304-311, 2016.
- [49] B. A. Forouzan, Cryptography & network security. Maidenhead, England: McGraw Hill Higher Education, 2007.
- [50] R. C. Merkle, "One way hash functions and DES," in Proc. Conf. on the Theory and Application of Cryptology: 428-446, 1989.
- [51] I. B. Damgård, "A design principle for hash functions," in Conf. on the Theory and Application of Cryptology: 416-427, 1989.
- [52] E. Andreeva, G. Neven, B. Preneel, T. Shrimpton, "Seven-property-preserving iterated hashing: ROX," in Proc. 13th Int. Conf. on the Theory and Application of Cryptology and Information Security: 130-146, 2007.
- [53] V. Shoup, "A composition theorem for universal one-way hash functions," in Int. Conf. on the Theory and Applications of Cryptographic Techniques: 445-452, 2000.
- [54] I. Mironov, "Hash functions: From merkle-damgård to shoup," in Proc. Int. Conf. on the Theory and Applications of Cryptographic Techniques: 166-181, 2001.
- [55] E. Andreeva, C. Bouillaguet, O. Dunkelman, J. Kelsey, "Herdin, second preimage and trojan message attacks beyond Merkle-Damgård," in Proc. 16th Int. Workshop on Selected Areas in Cryptography: 393-414, 2009.
- [56] J. Katz, Y. Lindell, Introduction to modern cryptography, 2nd ed. Philadelphia, PA: Chapman & Hall/CRC, 2014.
- [57] K. Malvoni, J. Knezovic, "Are your passwords safe: Energy-Efficient Bcrypt Cracking with Low-Cost Parallel Hardware," in Proc. 8th USENIX Workshop on Offensive Technologies (WOOT 14): 1-7, 2014.
- [58] H. J. Mun, S. Hong, J. Shin, "A novel secure and efficient hash function with extra padding against rainbow table attacks," Cluster Computing, 21(1): 1161-1173, 2017.
- [59] E. Conrad, S. Misenar, J. Feldman, Cissp Study Guide, 2nd ed. Waltham, MA, USA: Syngress Publishing, 2012.
- [60] J. Guo, S. Ling, C. Rechberger, H. Wang, "Advanced meet-in-the-middle preimage attacks: First results on full Tiger, and improved results on MD4 and SHA-2," in 16th Int. Conf. on the Theory and Application of Cryptology and Information Security: 56-75, 2010.
- [61] K. Aoki, Y. Sasaki, "Preimage attacks on one-block MD4, 63-step MD5 and more," in Proc. 15th Annu. Int. workshop on selected areas in cryptography: 103-119, 2008.
- [62] S. Knellwolf, D. Khovratovich, "New preimage attacks against reduced SHA-1," in Proc. 32nd Annu. Cryptology Conf.: 367-383, 2012.
- [63] K. Aoki, Y. Sasaki, "Meet-in-the-middle preimage attacks against reduced SHA-0 and SHA-1," in Proc. 29th Annu. Int. Cryptology Conf.: 70-89, 2009.
- [64] K. Aoki, J. Guo, K. Matusiewicz, Y. Sasaki, L. Wang, "Preimages for step-reduced SHA-2," in Proc. 15th Int. Conf. on the Theory and Application of Cryptology and Information Security: 578-597, 2009.
- [65] J. Kelsey, T. Kohno, "Herdin hash functions and the Nostradamus attack," in Proc. Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques: 183-200, 2006.
- [66] M. Stamp, R. M. Low, Applied cryptanalysis: breaking ciphers in the real world. Hoboken, N.J.: Wiley-Interscience, 2007.
- [67] W. Stallings, Network security essentials: Applications and standards, 4th ed. Prentice Hall, 2010.
- [68] G. V. Bard, Algebraic Cryptanalysis. Springer, 2009.
- [69] D. Jovanović, P. Janičić, "Logical analysis of hash functions," in 5th Int. Workshop on Frontiers of Combining Systems: 200-215, 2005.
- [70] F. Massacci, L. Marraro, "Logical cryptanalysis as a SAT problem," J. Automated Reasoning, 24(1): 165-203, 2000.
- [71] F. Legendre, G. Dequen, M. Krajecki, "Encoding hash functions as a sat problem," in Proc. 2012 IEEE 24th Int. Conf. on Tools with Artificial Intelligence, 1: 916-921, 2012.
- [72] D. De, A. Kumarasubramanian, R. Venkatesan, "Inversion attacks on secure hash functions using SAT solvers," in 10th Int. Conf. on Theory and Applications of Satisfiability Testing: 377-382, 2007.
- [73] P. Morawiecki, M. Srebrny, "A SAT-based preimage analysis of reduced Keccak hash functions," Inf. Process. Lett., 113(10-11): 392-397, 2013.
- [74] E. Homsirikamol, P. Morawiecki, M. Rogawski, M. Srebrny, "Security margin evaluation of SHA-3 contest finalists through SAT-based attacks," in Proc. 11th IFIP Int. Conf. on Computer Information Systems and Industrial Management: 56-67, 2012.
- [75] X. Wang, H. Yu, "How to break MD5 and other hash functions," in Proc. 24th Annu. Int. Conf. on the Theory and Applications of Cryptographic Techniques: 19-35, 2005.
- [76] X. Wang, Y. L. Yin, H. Yu, "Finding collisions in the full SHA-1," in Proc. 25th Annu. Int. Cryptology Conf.: 17-36, 2005.
- [77] W. Vinci, T. Albash, A. Mishra, P. A. Warburton, D. A. Lidar, "Distinguishing classical and quantum models for the d-wave device," Cornell University Library, 2014.
- [78] E. Knill, "Quantum computing," Nature, 463(7280): 441-443, 2010.
- [79] "D-Wave announces first order for 2000Q quantum computer," ID Quantique, 24-Feb-2017. Last accessed: Jan. 10, 2023. Available at: D-Wave Announces D-Wave 2000Q Quantum Computer and First System Order — D-Wave Government (dwavefederal.com).
- [80] "With new Microsoft breakthroughs, general purpose quantum computing moves closer to reality," Stories, Sep. 25, 2017. Last accessed: Jan. 10, 2023.
- [81] "Quantum simulator," Google Quantum AI. Last accessed: Jan. 10, 2023.
- [82] "qiskit.org," Qiskit.org. Last accessed: Jan. 10, 2023.
- [83] "D-wave ocean software documentation — ocean documentation 5.3.0 documentation," Dwavesys.com. Last accessed: Jan. 10, 2023.
- [84] A. J. Abhari et al., "Scaffold: Quantum programming language," Princeton univ NJ dept of computer science, Rep. TR-934-12, 2012. Last accessed: Jan. 10, 2023.
- [85] A. S. Green, P. L. Lumsdaine, N. J. Ross, P. Selinger, B. Valiron, "Quipper: a scalable quantum programming language," in Proc. the 34th ACM SIGPLAN Conf. on Programming language design and implementation: 333-342, 2013.

- [86] "Language-Integrated Quantum Operations: LIQUi>," Microsoft Research. Last accessed: Jan. 10, 2023.
- [87] S. Patil, A. JavadiAbhari, C. F. Chiang, J. Heckey, M. Martonosi, F. T. Chong, "Characterizing the performance effect of trials and rotations in applications that use Quantum Phase Estimation," in Proc. 2014 IEEE Int. Symposium on Workload Characterization (IISWC): 181-190, 2014.
- [88] M. A. Nielsen, I. L. Chuang, Quantum Computation and Quantum Information: 10Th Anniversary Edition. Cambridge, England: Cambridge University Press, 2010.
- [89] S. Imre, "Quantum computing and communications – Introduction and challenges," Comput. Electr. Eng., 40(1): 134-141, 2014.
- [90] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in Proc. 35th Annu. Symposium on Foundations of Computer Science: 124-134, 1994.
- [91] L. K. Grover, "A fast quantum mechanical algorithm for database search," in Proc. the twenty-eighth Annu. ACM symposium on Theory of computing: 212-219, 1996.
- [92] L. K. Grover, "Quantum mechanics helps in searching for a needle in a haystack," Phys. Rev. Lett., 79(2): 325-328, 1997.
- [93] D. J. Bernstein, "Grover vs. mceliece," in Third Int. Workshop on Post-Quantum Cryptography: 73-80, 2010.
- [94] P. Wang, S. Tian, Z. Sun, N. Xie, "Quantum algorithms for hash preimage attacks," Quantum Eng., 2(2): 2020.
- [95] X. Dong, S. Sun, D. Shi, F. Gao, X. Wang, L. Hu, "Quantum collision attacks on AES-like hashing with low quantum random access memories," in Proc. 26th Int. Conf. on the Theory and Application of Cryptology and Information Security: 727-757, 2020.
- [96] D. J. Bernstein, "Introduction to post-quantum cryptography," in Post-Quantum Cryptography, Berlin, Heidelberg: Springer Berlin Heidelberg, 1-14, 2009.
- [97] J. Watrous, "Zero-Knowledge against Quantum Attacks," SIAM j. comput., 39(1): 25-58, 2009.
- [98] E. Kashefi, I. Kerenidis, "Statistical Zero Knowledge and quantum one-way functions," Theor. Comput. Sci., 378(1): 101-116, 2007.
- [99] S. J. Chang et al., "Third-round report of the SHA-3 cryptographic hash algorithm competition," National Institute of Standards and Technology, Gaithersburg, MD, Rep. 7896, Nov. 2012.
- [100] U. Rührmair et al., "Efficient Power and Timing Side Channels for Physical Unclonable Functions," in Proc. 16th Int. Workshop on Cryptographic Hardware and Embedded Systems, 476-492, 2014.
- [101] J. P. Aumasson, W. Meier, R. C. W. Phan, L. Henzen, The hash function BLAKE. Springer-Verlag Berlin Heidelberg, 2014.
- [102] A. Bauer, E. Jaulmes, E. Prouff, J.-R. Reinhard, J. Wild, "Horizontal collision correlation attack on elliptic curves: - Extended Version -," Cryptogr. Commun., 7(1): 91-119, 2015.
- [103] S. Y. Yan, Cryptanalytic attacks on RSA. Springer, 2008.
- [104] S. Contini, A. K. Lenstra, R. Steinfeld, "VSH, an efficient and provable collision-resistant hash function," in Proc. 25th Int. Conf. on the Theory and Applications of Cryptographic Techniques: 165-182, 2006.
- [105] M. J. O. Saarinen, "Security of VSH in the real world," in Proc. 7th Int. Conf. on Cryptology in India: 95-103, 2006.
- [106] ISO/IEC 10118-4:1998 Information technology — Security techniques — Hash-functions — Part 4: Hash-functions using modular arithmetic, ISO/IEC 10118-4:1998, Dec. 1998.
- [107] S. Kölbl, E. Tischhauser, P. Derbez, A. Bogdanov, "Troika: a ternary cryptographic hash function," Des. Codes Cryptogr., 88(1): 91-117, 2020.
- [108] N. Mishra, S. H. Islam, S. Zeadally, "A comprehensive review on collision-resistant hash functions on lattices," J. Inf. Secur. Appl., 58: 102782, 2021.
- [109] V. Manuceau, "About a fast cryptographic hash function using cellular automata ruled by far-off neighbours," Int. j. eng. trends technol., 69(2): 39-41, 2021.
- [110] J. S. Teh, K. Tan, M. Alawida, "A chaos-based keyed hash function based on fixed point representation," Cluster Comput., 22(2): 649-660, 2018.
- [111] M. Zohner, M. Kasper, M. Stöttinger, S. A. Huss, "Side channel analysis of the SHA-3 finalists," in Proc. 2012 Design, Automation & Test in Europe Conf. & Exhibition (DATE): 1012-1017, 2012.
- [112] C. Herder, M. D. Yu, F. Koushanfar, S. Devadas, "Physical unclonable functions and applications: A tutorial," Proc. IEEE. Electr. Electron. Eng., 102(8): 1126-1141, 2014.
- [113] T. F. Lee, W. Y. Chen, "Lightweight fog computing-based authentication protocols using physically unclonable functions for internet of medical things," J. Inf. Secur. Appl., 59: 102817, 2021.
- [114] A. P. Fournaris, N. Sklavos, "Secure embedded system hardware design – A flexible security and trust enhanced approach," Comput. Electr. Eng., 40(1): 121-133, 2014.
- [115] U. Rührmair, D. E. Holcomb, "PUFs at a glance," in 2014 Design, Automation & Test in Europe Conf. & Exhibition (DATE): 1-6, 2014.
- [116] C. Helfmeier, C. Boit, D. Nedospasov, J. P. Seifert, "Cloning physically unclonable functions," in 2013 IEEE Int. Symposium on Hardware-Oriented Security and Trust (HOST): 1-6, 2013.
- [117] U. Rührmair, J. Sölter, "PUF modeling attacks: An Introduction and overview," in Proc. 2014 Design, Automation & Test in Europe Conf. & Exhibition (DATE): 1-6, 2014.
- [118] J. Delvaux, I. Verbauwhede, "Side channel modeling attacks on 65nm arbiter PUFs exploiting CMOS device noise," in Proc. 2013 IEEE Int. Symposium on Hardware-Oriented Security and Trust (HOST): 137-142, 2013.
- [119] A. Mahmoud, U. Rührmair, M. Majzoobi, F. Koushanfar, "Combined modeling and side channel attacks on strong PUFs," Cryptology ePrint Archive, 2013.
- [120] Y. Gao, S. F. Al-Sarawi, D. Abbott, "Physical unclonable functions," Nat. Electron., 3(2): 81-91, 2020.
- [121] Y. G. Yang, J. R. Dong, Y. L. Yang, Y. H. Zhou, W. M. Shi, "Usefulness of decoherence in quantum-walk-based hash function," Int. J. Theor. Phys., 60(3): 1025-1037, 2021.
- [122] M. Ziatdinov, "Quantum Hashing. Group approach," Lobachevskii J. Math., 37(2): 222-226, 2016.
- [123] H. Buhrman, R. Cleve, J. Watrous, R. de Wolf, "Quantum fingerprinting," Phys. Rev. Lett., 87(16), 2001.
- [124] F. Abelayev, A. Vasiliev, "Quantum hashing," Cornell University Library, arXiv:1310.4922 [quant-ph], 2013.
- [125] F. Abelayev, M. Abelayev, "Quantum hashing via e-Universal hashing constructions and freivalds' fingerprinting schemas," in Proc. 16th Int. Workshop on Descriptive Complexity of Formal Systems: 42-52, 2014.
- [126] F. Abelayev, M. Abelayev, A. Vasiliev, "On the balanced quantum hashing," J. Phys. Conf. Ser., 681(1): 012019, 2016.
- [127] L. Ren, S. Devadas, "Bandwidth hard functions for ASIC resistance," in Proc. Theory of Cryptography Conf.: 466-492, 2017.
- [128] C. Percival, Stronger key derivation via sequential memory-hard functions. 1-16, 2009. Last accessed: Jan. 10, 2023.
- [129] C. Percival, S. Josefsson, The scrypt password-based key derivation function. Internet Engineering Task Force (IETF), No. rfc7914, Aug. 2016.
- [130] X. Li, Z. Shao, M. Zhu, J. Yang, Fundamentals of optical computing technology: forward the next generation supercomputer. Singapore: Springer, 2018.
- [131] W. He, X. Peng, "Optical one-way hash function," In: Advanced Secure Optical Image Processing for Communications, Al Falou, A. ed. IOP Publishing. 2018.
- [132] W. Q. He, J. Y. Chen, L. B. Zhang, D. J. Lu, M.-H. Liao, X. Peng, "Optical Hash function based on multiple scattering media," Acta Physica Sinica, 70(5): 054203, 2021.
- [133] A. Kumar, A. Fatima, N. K. Nishchal, "An optical Hash function construction based on equal modulus decomposition for authentication verification," Opt. Commun., 428: 7-14, 2018.
- [134] L. D. Tsobdjou, S. Pierre, A. Quintero, "A new mutual authentication and key agreement protocol for mobile client—server environment," IEEE trans. netw. serv. manag., 18(2): 1275-1286, 2021.

- [135] P. Kocher, J. Jaffe, B. Jun, P. Rohatgi, "Introduction to differential power analysis," *J. Cryptogr. Eng.*, 1(1): 5-27, 2011.
- [136] S. N. Srirama, M. Jarke, W. Prinz, "Mobile web services mediation framework," in *Proc. the 2nd workshop on Middleware for service oriented computing: held at the ACM/IFIP/USENIX Int. Middleware Conf.*: 6-11, 2007.
- [137] M. Asif, S. Majumdar, R. Dragnea, "Partitioning the WS execution environment for hosting mobile web services," in *Proc. 2008 IEEE Int. Conf. on Services Computing*, vol. 2: 315-322, 2008.
- [138] J. Li, W. Zhang, V. Dabra, K. K. R. Choo, S. Kumari, D. Hogrefe, "AEP-PPA: An anonymous, efficient and provably-secure privacy-preserving authentication protocol for mobile services in smart cities," *J. Netw. Comput. Appl.*, 134: 52-61, 2019.
- [139] J. Kong, F. Koushanfar, "Processor-based strong physical unclonable functions with aging-based response tuning," *IEEE Trans. Emerg. Top. Comput.*, 2(1): 16-29, 2014.
- [140] J. P. Aumasson, S. Neves, Z. Wilcox-O'Hearn, C. Winnerlein, "BLAKE2: simpler, smaller, fast as MD5," in *Proc. 11th Int. Conf. on Applied Cryptography and Network Security*: 119-135, 2013.
- [141] R. Chakraborty, A. Ghosh, V. E. Balas, A. A. Elgar, *Blockchain: Principles and Applications in IoT*. Boca Raton: Chapman and Hall/CRC, 2022.
- [142] C. Dobraunig, M. Eichlseder, F. Mendel, M. Schl  ffer, "Ascon v1.2: Lightweight Authenticated Encryption and Hashing," *J. Cryptol.*, 34(3), 2021.
- [143] P. Podimatas, K. Limn  tis, "Evaluating the Performance of Lightweight Ciphers in Constrained Environments-The Case of Saturnin," *Signals*, 3(1): 86-94, 2022.
- [144] S. Blanc, A. Lahmadi, K. Le Gouguec, M. Minier, L. Sleem, "Benchmarking of lightweight cryptographic algorithms for wireless IoT networks," *Wirel. netw.*, 28(8): 3453-3476, 2022.
- [145] W. K. Lee, K. Jang, G. Song, H. Kim, S. O. Hwang, H. Seo, "Efficient Implementation of lightweight hash functions on GPU and quantum computers for IoT applications," *IEEE Access*, 10: 59661-59674, 2022.
- [146] J. E. Midwinter, *Photonics in Switching*, 1st ed. Academic Press, 1993.

Biographies



Behrouz Sefid-Dashti received his B.S. and M.S. degrees in computer engineering from the Iran Information Technology Development and Islamic Azad University (Tehran North Branch), respectively. He has over 18+ years of experience in the field of software development, and is currently a Ph.D. candidate of software engineering at the University of Kashan. His career and experience include software analysis, design and architecture, and blockchain development. His research interests include blockchain, software architecture, and cryptography.

Behrouz Sefid-Dashti received his B.S. and M.S. degrees in computer engineering from the Iran Information Technology Development and Islamic Azad University (Tehran North Branch), respectively. He has over 18+ years of experience in the field of software development, and is currently a Ph.D. candidate of software engineering at the University of Kashan. His career and experience include software analysis, design and

- Email: b.sefiddashti@grad.kashanu.ac.ir
- ORCID: [0000-0002-3767-4303](https://orcid.org/0000-0002-3767-4303)
- Web of Science Researcher ID: HJP-5663-2023
- Scopus Author ID: 56123693700
- Homepage: NA



Javad Salimi Sartakhti is an assistant professor of artificial intelligence in the department of computer engineering at the University of Kashan, Iran. He obtained his B.Sc. degree in computer engineering from the University of Kashan and his M.Sc. degree in software engineering from the Tarbiat Modares University, Tehran, Iran, in 2008 and 2013, respectively. In January 2017, he obtained his Ph.D. degree in artificial intelligence at the Isfahan University of Technology. He ranked first among students of computer engineering in all three degrees. His main research interests are mechanism design and game theory, blockchain, machine learning, and Deep learning.

- Email: salimi@kashanu.ac.ir
- ORCID: [0000-0003-1183-1232](https://orcid.org/0000-0003-1183-1232)
- Web of Science Researcher ID: HJY-2812-2023
- Scopus Author ID: 51864592100
- Homepage: <https://faculty.kashanu.ac.ir/salimi/en>



Hassan Daghigh is an associate professor of mathematics at the University of Kashan, Iran. He received his M.Sc. in mathematics (Commutative Algebra) at the University of Tarbiat Modarres, Iran. He got his Ph.D. degree in mathematics (elliptic curves) at McGill university, Canada in 1998, under the direction of Henri Darmon. His research activities are now focused on number theory (elliptic curves, algebraic number theory) and applications in cryptography in particular lattice and isogeny based cryptography.

- Email: hassan@kashanu.ac.ir
- ORCID: [0000-0002-4242-769X](https://orcid.org/0000-0002-4242-769X)
- Web of Science Researcher ID: HJY-1325-2023
- Scopus Author ID: NA
- Homepage: <https://faculty.kashanu.ac.ir/daghigh/en>

How to cite this paper:

B. Sefid-Dashti, H. Daghigh, J. S. Sartakhti, "Brand new categories of cryptographic hash functions: A survey," *J. Electr. Comput. Eng. Innovations*, 11(2): 335-354, 2023.

DOI: [10.22061/jecei.2023.9271.598](https://doi.org/10.22061/jecei.2023.9271.598)

URL: https://jecei.sru.ac.ir/article_1840.html





Research paper

Modeling Transport in Graphene-Metal Contact and Verifying Transfer Length Method Characterization

B. Khosravi Rad¹, M. khaje², A. Eslami Majd^{2,*}

¹Optoelectronics and Nanophotonics Research Group, Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.

²Research electronic center, Malek Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:

Received 25 November 2022
Reviewed 10 January 2023
Revised 31 January 2023
Accepted 09 March 2023

Keywords:

Graphene-metal contact
Contact resistance
Transfer length method
Circuit modeling
Effective channel width

*Corresponding Author's Email
Address: a_eslamimajd@mut-es.ac.ir

Abstract

Background and Objectives: One of the common methods for measuring the contact resistance of graphene sheets is the transfer length or transmission line method (TLM). Apart from the contact resistance, TLM gives the resistance of the channel sheet and the effective transfer length of the measured samples. Furthermore, the implementation of TLM is simple. To analyze this method, one can use circuit modeling (CM).

Methods: An important parameter of TLM is the contact resistance between the metal electrode and the graphene channel. To compare this parameter with other measures, it is normalized by multiplying it by the channel width. In this research, for TLM analysis, all the components of the structure including electrodes, graphene channel, and metal-graphene contact are modeled in a circuit.

Results: PSpice and MATLAB are integrated for TLM circuit modeling. The metal electrodes and the graphene channel are modeled based on the values of the resistances measured in the laboratory using the van der Pauw method and the resistances reported in the article in ohms per square. Moreover, the metal-graphene contact resistance is considered based on the values reported in the literature in ohms-micrometers.

Conclusion: The modeling results show that, in addition to the effective transfer length, the effective transfer width can be defined on a contact, according to the dimensions of the structure. Therefore, the channel width is a vague characteristic of the TLM measurement, which plays a very important role in measuring contact resistance. Furthermore, the contact resistance and the resistance of the channel sheet are independent of each other and of the distance between the contacts. If defects in the graphene channel are randomly distributed along the channel between the contacts, they do not have a significant impact on the contact resistance, while they increase the resistance of the graphene sheet provided that they do not disrupt the channel. Indeed, for a 10% defect (or 90% coverage along the channel), the resistance of the sheet increases by 16%. In addition, by using this modeling, parameters such as the distribution of the contact current, the sources of errors, and their influence in determining the contact resistance and resistance of the channel sheet are investigated.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Contact resistance is one of the main limiting factors in the performance of short-channel graphene-based

transistors [1]. The performance of an RF transistor is characterized by two parameters, the cut-off frequency (f_T) and the maximum oscillation frequency (f_{max}), which respectively represent the frequencies at which the current and power gain are unity and modulated by the gate; the cutoff frequency is obtained in (1), [2], [3], [4].

$$f_T = \frac{g_m}{2\pi C_g} \quad (1)$$

where C_g is the gate capacitance, and g_m is the conductivity. As a geometric parameter of field-effect transistors, the cutoff frequency has an inverse relationship with the square of the gate length and is almost independent of the width of the transistor. [2]. Nevertheless, better performance is observed in graphene transistors with much longer gate lengths, compared to Si and elements in groups 3 and 5 of the periodic table. Moreover, higher values of cutoff frequency have been reported for graphene transistors. Another important parameter in the RF transistors is f_{max} , which is indicated in (2) [5].

$$f_{max} = \frac{f_T}{\sqrt{(g_D(R_G + R_{SD}) + 2\pi f_T R_G C_G)}} \quad (2)$$

where g_D is the channel conductivity, R_G is the gate resistance, and R_{SD} is the drain-source resistance. Some studies have demonstrated that there is a significant difference between the values of f_T and f_{max} values in graphene RF transistors. To increase f_T , the carrier mobility should be increased [5], and three factors should be investigated and optimized: 1- the defects that depend on the growth process in graphene, 2- the impurities that cause scattering at the interface between the gate dielectric and graphene channel, and 3- contact resistance [3].

The frequency performance of graphene is limited by two factors, that is, the mobility of carrier in the graphene channel and the contact resistance of graphene-metal. The mobility of carrier in devices with a large channel length is dominant relation to the contact resistance. However, when the length of the channel is reduced, the role of the contact resistance in improving the speed of the device becomes critical and dominant [6]. Therefore, to improve the performance of graphene devices at high frequencies, the contact resistance in short-channel devices should be reduced as much as possible because the increase in contact resistance reduces the cutoff frequency [7], [8], [9]. A very important point is to characterize the metal-graphene contact resistance that can be made in several ways one of which is the TLM measurement [10], [11], [12], [13], [14]. In this method, important parameters, such as transmission length, contact resistance, and channel sheet resistance, are obtained [15], [16], [17], [18]. Then, each of the

mentioned parameters is varied under the influence of different factors, such as contamination of the metal-graphene interface, defects or the presence of pollution particles in the graphene channel [19], and metal-graphene geometry, especially at the interface [20].

To better understand the transmission line method, TLM circuit modeling can be used. In this method, choosing geometric and physical values in appropriate ranges for metal-graphene contact helps to determine the limitations of the measurement method and provide correction coefficients for the measured values. In 2014, Vincenzi Giancarlo et al. investigated the contact current crowding or current transfer length (L_t) in circuit diagrams [21]. furthermore, in 2015, Zhang Peng et al. investigated the contact current crowding and its relationship with specific contact resistance. These authors demonstrated that the transfer length increases with the increase in the specific contact resistance [22]. In 2017, González-Díaz et al. also used a sample with a sheet resistance of 1 kΩ/□ (1kilo ohm/square) in a circuit model and evaluated the sources of error caused by finite size contacts in the van der Pauw measurement method [23]. The structure of TLM includes metal electrodes and a channel between them. The distances between the electrodes, or, the length of the channel, increase in an ascending manner. In TLM, by measuring the resistance between adjacent electrodes and plotting a graph of the obtained values, according to the distances between the contacts, the importance of contact resistance, transfer length, and sheet resistance are obtained. In practice, the TLM measurement method is used for graphene, as shown in Fig. 1. The measurement accuracy increases by increasing the number of contacts and changing the dimensions.

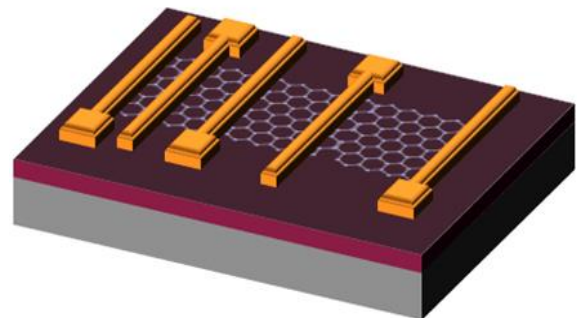


Fig. 1: Schematic of TLM measurement method.

The simple and one-dimensional circuit model of metal-graphene contact has been discussed in Refs. [24], [25], [26], [27] are shown in Fig. 2. As it is evident, these articles have only referred to the current distribution under the contact or the effective transfer length of the electrode.

In the present study, for the first time, the circuit model of the metal-graphene contact, in other words, the TLM method is simulated in three dimensions. Using the mentioned model, sources of error, current distribution

under the contact, and an estimate of the effective transfer length and its relationship with the dimensions of the structure are discussed and analyzed. Also, the accuracy or inaccuracy of the model is analyzed. In addition to the mentioned cases, for the first time, the effective width of the channel is also checked, representing the part of the width of the graphene channel where the maximum amount of current enters it. The current entering the graphene channel from the metal electrode, instead of uniformly covering the entire channel, the part of the channel closer to the current source contributes more to the current flow. In other words, the electric current chooses the shortest path to pass.

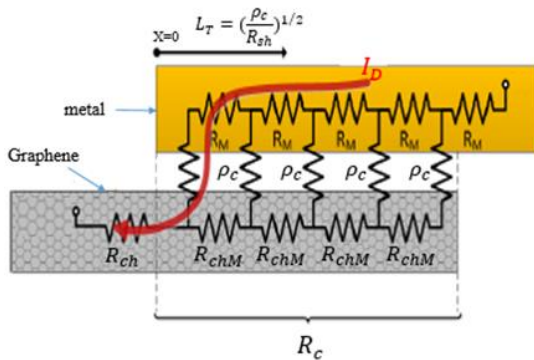


Fig. 2: One-dimensional circuit model of metal-graphene contact.

Results and Discussion

The transfer length method includes metal electrodes, the substrate, the channel between the electrodes, and the contact resistance between them. A network of electrical resistances models the graphene sheet and metal electrodes in a circuit form. To model the graphene channel and metal contacts, a square unit cell is considered whose dimensions are the same as those of the graphene channel and metal electrodes. Then, each unit cell of metal contacts and graphene channel is modeled independent of the dimensions of the unit cell, using four resistors with values of $2R_{shm}$ and $2R_{shg}$, respectively (see Fig. 3). In this model, the resistances of the channel and metal sheet (R_{shm} and R_{shg}) are expressed in ohm/square.

Table 1: TLM modeling parameters

Measurement parameters	Unit	Values
Contact length (L_c)	μm	1
Channel width (w)	μm	5
Contact distances (d)	μm	1, 2, 3, 4, 5, 6
Channel sheet resistance (R_{shg})	Ω/\square	500
Metal sheet resistance (R_{shm})	Ω/\square	5
Contact resistance (R_c)	$\Omega.\mu\text{m}$	200

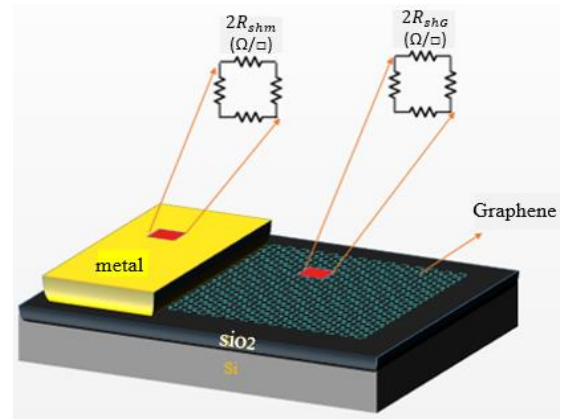


Fig. 3: Circuit schematic of metal electrode and graphene channel.

After determining the dimensions of the channel and metal electrodes, they are modeled using PSpice according to Fig. 4. This figure shows Ni metal electrodes with a sheet resistance of $5 \Omega/\square$ corresponding to 20 nm of the metal, which is measured using the van der Pauw method. Also, the sheet resistance of monolayer graphene, with a thickness of 0.34nm, is considered $500 \Omega/\square$. This value is based on the values obtained from the van der Pauw measurement method carried out on the CVD graphene transferred on a SiO_2/Si substrate in the laboratory. In addition, the contact resistance at the metal-graphene interface is modeled with a value of $200 \Omega.\mu\text{m}$ according to the literature (e.g., see Refs. [18], [28], [29], [30], [31] that report values in the range of 100-2500 $\Omega.\mu\text{m}$). In this case, the normalized contact resistance values per micron are used. As shown in Fig. 4, the dimensions of metal contacts and graphene channels are small, so the expansion of the mentioned structure using PSpice is very time-consuming and challenging. Then, for more flexibility and the ability to change the system of TLM, the netlist related to the resistances of metal electrodes, graphene channel, and contact resistances, along with the contacts between them, was produced by MATLAB. The generated netlist was executed in PSpice, and its results were checked.

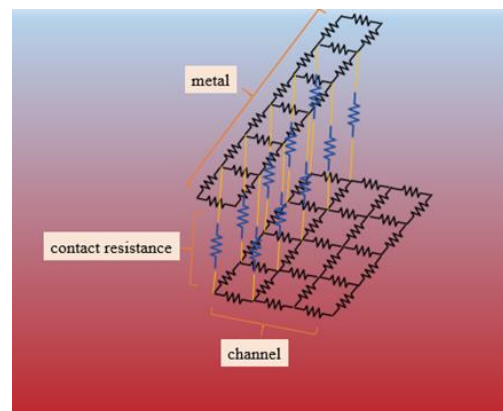


Fig. 4: Circuit modeling of the metal electrode, graphene channel, and contact resistance.

Finally, the TLM diagram was plotted, as shown in Fig. 5. This figure's results confirm the model's correctness.

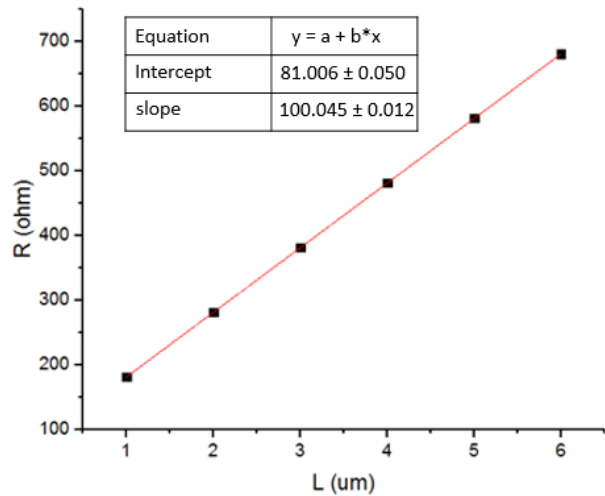


Fig. 5: TLM measurement diagram for graphene in circuit modeling mode.

Another parameter evaluated in this study is the transfer length of the metal electrode. To extract this parameter and show the current distribution under the contact, the dimensions in Table 1, with a 2 μm increase in the electrode length, a 10 μm increase in the channel width, and a channel length of 1 μm , were considered. Then, the results of the current distribution under the contact were obtained, as shown in Fig. 6. The horizontal and vertical axes represent the length of the metal electrode and the width of the channel, respectively. Here, the current distribution under the contact and the effective length of the metal electrode are involved in the transfer.

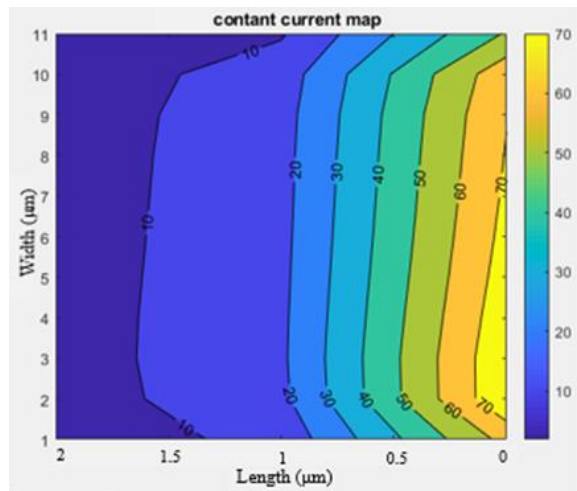


Fig. 6: Current distribution from the metal electrode to the graphene channel. The maximum current is transferred along less than 1 μm of the metal electrode.

A. Effect of Channel Width on the Accuracy of Measuring Contact Resistance and Channel Sheet Resistance

After examining the transfer length parameter, the

effect of channel width (w) on the measurement of contact resistance and sheet resistance was investigated for a fixed channel length. In this case, the modeling based on the TLM measurement method was used for different widths of the graphene sheet, and the results are shown in Table 2. These results indicated that from one point on, if the entire width of the channel is applied to normalize contact resistance, the measurement error increases, and contact resistance becomes more significant than the desired value in the modeling. It was found that by increasing the, instead of the entire width of the channel, only a part of it participates in the current transfer. In other words, a part of the channel width significantly contributes to the current transfer. The reason for this phenomenon can be the resistance of the metal sheet on the graphene channel, which depends on the type of metal and the sheet thickness. To solve this problem, the graphene channel's width and, consequently the electrode's dimensions should be reduced as much as possible so that the passing current has a shorter path to flow. Also, the thickness of the metal sheet should be such that its resistance is low. suppose the resistance of the metal electrode and channel width is small, instead of taking a shorter path for entering the channel. In that case, the current can flow throughout the metal electrode and then enter the channel uniformly. In this case, the effective width of the channel is equal to the width of the channel.

Table 2: Effective widths obtained for different channel widths

Channel width (w , μm)	effective channel width (w_{ef} , μm)	Model contact resistance ($R_c \cdot w$, $\Omega \cdot \mu\text{m}$)	Measured contact resistance ($R_c \cdot w$, $\Omega \cdot \mu\text{m}$)
5	4.94	200	202.5
6	5.6	200	214.35
7	5.98	200	234.15
8	6.3	200	254
9	6.52	200	276.3
10	6.7	200	299

To show the current distribution across the width of the channel, the dimensions of the contacts and the channel were considered according to Table 3. The results in terms of different widths (w) are shown in Fig. 7. From this figure, for a width greater than 20 μm , the uniformity of the current distribution across the width of the channel starts to change. Also, this issue was investigated, for different values of the metal sheet resistance, and it was shown that the values of metal sheet resistance, which depend on the thickness of the metal sheet, affect the mentioned results. Therefore, if the structure dimensions are not carefully considered, an error can occur in the

normalized values of the metal-graphene contact resistance.

Table 3: Parameters for plotting the effective width of the channel

Measurement parameters	Unit	Values
L_c	μm	2
w	μm	20, 30, 40, 50
d	μm	10
R_{shG}	Ω/\square	500
R_{shm}	Ω/\square	5
R_c	$\Omega.\mu\text{m}$	200

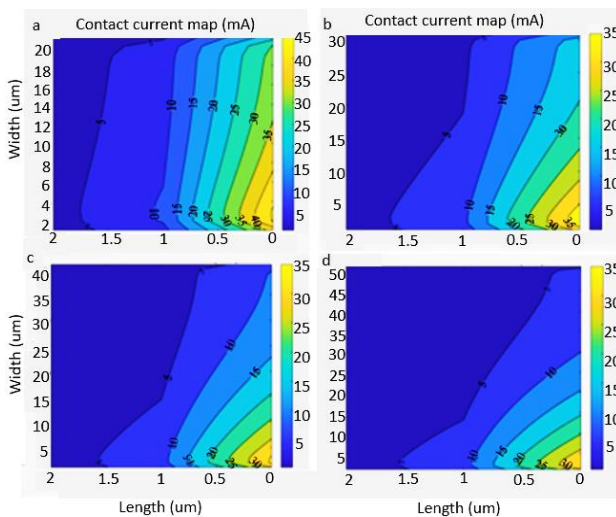


Fig. 7: Current distribution from the metal to the graphene channel. (a) The channel width is 20 μm . (b) The channel width is 30 μm . (c) The channel width is 40 μm . (d) The channel width is 50 μm .

B. Effect of Distances Between Contacts and Channel Defects on Measuring Channel Sheet and Contact Resistances

The TLM measurement method was modeled to investigate the effect of the distances between electrodes or the length of the channel in the measurement of contact resistance. The results were plotted for different distances between the contacts. In this case, the channel is uniform and homogeneous. According to Fig. 8, the contact and channel sheet resistance values are the same for different channel dimensions.

However, when 10% of defects are considered for the graphene channel, the TLM measurement diagram (Fig. 9) shows that the contact resistance decreases by a few percent compared to the defect-free state, which is not very noticeable, while the channel sheet resistance increases by 16%.

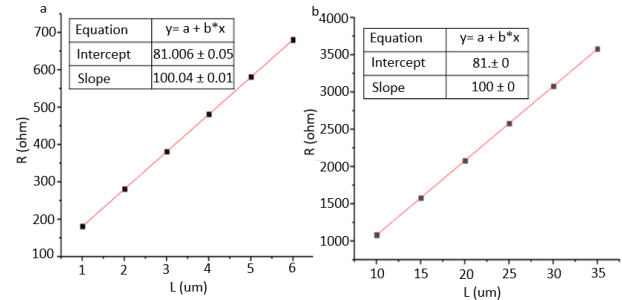


Fig. 8: TLM measurement diagram for distances of (a) 1, 2, 3, 4, 5, 6 and (b) 10, 15, 20, 25, 30, and 35 microns between electrodes.

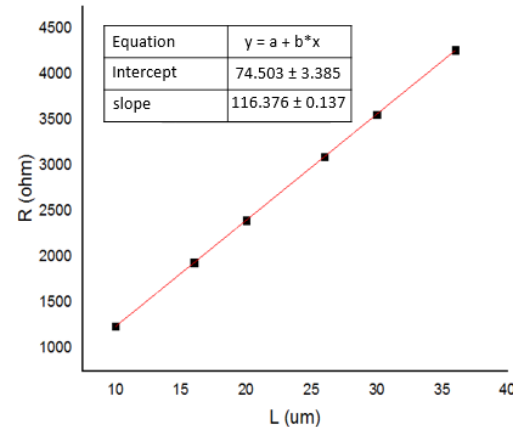


Fig. 9: TLM measurement diagram for 10% defects in the channel and inter-contact distances of 10, 16, 20, 26, 30, and 36 μm ; channel sheet resistance has increased by 16%.

When the graphene channel defect is 20%, the contact resistance is almost unchanged, but the channel sheet resistance increases by 32% (Fig. 10). Therefore, it can be concluded that if the defect is spread throughout the graphene sheet or the area between the electrodes (provided that it does not interrupt the channel), then the contact resistance remains almost constant while the channel sheet resistance increases.

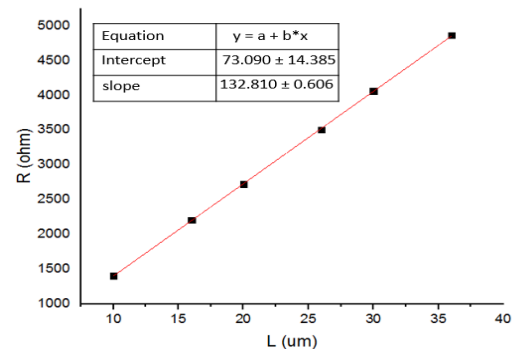


Fig. 10: TLM measurement diagram for 20% defects in channel and inter-contact distances of 10, 16, 20, 26, 30, and 36 μm ; channel sheet resistance has increased by 32%.

C. Effect of Error in Measuring Distances Between Electrodes on Contact Resistance and Channel Sheet Resistance

Other parameters affecting the TLM measurement

results are relative errors in measuring the distances between contacts and their resistances. Therefore, the TLM diagram was plotted by assuming an error of 5% in the measurement of distances between the contacts for contact distances of 10, 15, 20, 25, 30, and 35 micrometers. According to Fig. 11, the error in measuring of distances between the contacts does not affect the value of the contact resistance, but it changes the channel sheet resistance.

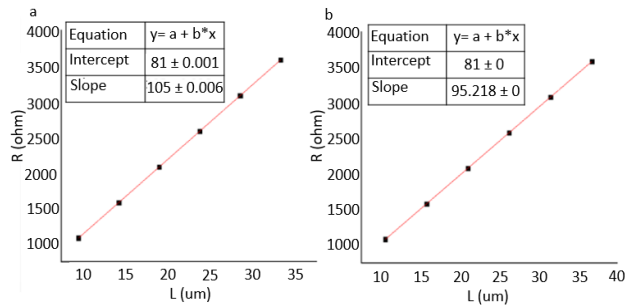


Fig. 11: TLM diagram for 5% error in measuring the distances between contacts for distances of (a) 9.5, 14.25, 19, 23.75, 28.5, 33.25 and (b) 10/5, 15/75, 21, 26/25, 31/5, 36/75 microns.

D. Effect of Error in Measuring the Resistance Between Contacts on Metal-Graphene Contact Resistance and Graphene Sheet Resistance

In this case, assuming an error of 5% in measuring the resistance between the electrodes, the contact and channel sheet resistance values show a 5% error (see Fig. 12). If the error rate is added to the measured resistance values between the electrodes, these parameters increase by 5%. Otherwise, the same amount of error is obtained again for contact and channel sheet resistance, with the difference that their values are reduced this time.

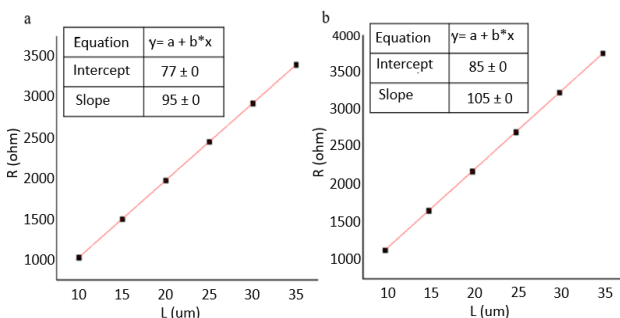


Fig. 12: TLM measurement diagram for 5% error in measuring the resistance between metal electrodes. The error percentage has (a) been summed with and (b) subtracted from the actual value of the measured resistances.

E. Proposing TLM Structure and Providing Channel Width Correction Coefficients

It was considered that the normalized resistance of the contact depends on the dimensions of the TLM structure and the channel width. Therefore, a structure with

specific dimensions and appropriate correction coefficients was presented. In this structure, the metal sheet resistance was considered $3 \Omega/\square$, obtained using the van der Pauw measurement method for a thickness of 30 nm of Ni metal sheet. Other characteristics of the structure are shown in Table 4.

Table 4: Modeling parameters of the proposed TLM structure

Measurement Parameters	Unit	Values
L_C	μm	2
w	μm	20, 30, 40, 50
d	μm	10
R_{shG}	Ω/\square	500
R_{shm}	Ω/\square	5
R_C	$\Omega.\mu\text{m}$	200

By comparing the values obtained from the measurement diagram with the values from modeling, the correction coefficients are determined according to Table 5. These coefficients are obtained by dividing the modeling contact resistance by the measured contact resistance.

Table 5: Correction coefficients for measured contact resistance

Modeling contact resistance ($\Omega.\mu\text{m}$)	Modeling channel width (μm)	Measured contact resistance ($\Omega.\mu\text{m}$)	Correction coefficients
100	10	166.5	0.601
200	10	264.2	0.760
300	10	363.5	0.825
400	10	466	0.859
500	10	576.5	0.867

Conclusion

In this article, for the first time, circuit modeling of the TLM method has been done, and by using this modeling, parameters such as current distribution under the contact, effective transfer length, error sources, and their influence in determining the contact resistance and channel sheet resistance have been investigated. The modeling showed that the maximum current is transferred in a length of fewer than $1 \mu\text{m}$ of the metal electrode. Most importantly, it has been shown for the first time that in addition to the effective transfer length, the effective channel width, which plays a vital role in determining the normalized contact resistance and channel sheet resistance, is an ambiguous aspect of the TLM measurement method that has not been mentioned

so far. It was found that by increasing the width of the channel, only a part of it participates in the current transfer. In other words, a part of the channel width significantly contributes to the current transfer. The reason for this phenomenon can be the resistance of the metal sheet on the graphene channel, which depends on the type of metal and the sheet thickness. In addition, it was shown that the measurement of the contact and channel sheet resistance are independent of each other and the distances between electrodes. Furthermore, the contact resistance is almost constant for the defects in the graphene channel, but the channel sheet resistance increases. The effect of the error in measuring the distances/resistances between the contacts on the contact resistance and channel sheet resistance was investigated. The error in the measurement of spaces between the contacts does not affect the value of the contact resistance, but it changes the channel sheet resistance. However, the error in measuring the resistance between the electrodes causes the values of the contact resistance and the channel sheet resistance to change by the same amount of error.

Author Contributions

All the authors participated in the conceptualization and implementation, and B. Khosravi Rad wrote the manuscript.

Acknowledgment

The authors would like to acknowledge the Faculty of Electrical & Computer Engineering, Malek Ashtar University of Technology, and the Microelectronic laboratory, for their support and contribution to this study.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>TLM</i>	Transfer Length Method
<i>CM</i>	Circuit Modeling
<i>ETL</i>	Effective Transfer Length
<i>CR</i>	Contact Resistance
<i>CSR</i>	Channel Sheet Resistance
<i>ETW</i>	Effective Transfer width
<i>MSR</i>	Metal Sheet Resistance

References

- [1] L. Anzi et al., "Ultra-low contact resistance in graphene devices at the Dirac point," *2D Mater.*, 5(2): 025014, 2018.
- [2] Y. M. Lin et al., "Development of graphene FETs for high frequency electronics," in *Proc. 2009 IEEE International electron devices Meeting (IEDM)*: 1-4, 2009.
- [3] K. Kim, J. Y. Choi, T. Kim, S. H. Cho, H. J. Chung, "A role for graphene in silicon-based semiconductor devices," *Nature*, 479(7373): 338-344, 2011.
- [4] J. Zheng et al., "Sub-10 nm gate length graphene transistors: operating at terahertz frequencies with current saturation," *Sci. Rep.*, 3(1): 1-9, 2013.
- [5] W. Zhu et al., "Graphene radio frequency devices on flexible substrate," *Appl. Phys. Lett.*, 102(23): 233102, 2013.
- [6] Y. Wu et al., "High-frequency, scaled graphene transistors on diamond-like carbon," *Nature*, 472(7341): 74-78, 2011.
- [7] M. R. Islam, M. A. Haque, M. Fahim-Al-Fattah, M. N. K. Alam, M. R. Islam, "Dynamic performance of graphene field effect transistor with contact resistance," in *Proc. 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*: 21-25, 2016.
- [8] Y. Wu et al., "RF performance of short channel graphene field-effect transistor," in *Proc. 2010 International Electron Devices Meeting*: 9.6.1-9.6.3., 2010.
- [9] A. Mehrfar, A. Eslami Majd, "Enhancement of the photoresponse in the platinum silicide photodetector by a graphene layer," *J. Electr. Comput. Eng. Innovations (JECEI)*, 10(2): 363-370, 2022.
- [10] D. K. Schroder, *Semiconductor material and device characterization*. John Wiley & Sons, 2015.
- [11] V. Passi et al., "Contact resistance Study of "edge-contacted" metal-graphene interfaces," in *Proc. 2016 46th European Solid-State Device Research Conference (ESSDERC)*: 236-239, 2016.
- [12] B. K. Bharadwaj, D. Nath, R. Pratap, S. Raghavan, "Making consistent contacts to graphene: effect of architecture and growth induced defects," *Nanotechnology*, 27(20): 205705, 2016.
- [13] A. Quellmalz et al., "Influence of humidity on contact resistance in graphene devices," *ACS applied materials & interfaces*, 10(48): 41738-41746, 2018.
- [14] V. Passi et al., "Ultralow specific contact resistivity in metal-graphene junctions via contact engineering," *Adv. Mate. Interfaces*, 6(1): 1801285, 2019.
- [15] J. Anteroinen, W. Kim, K. Stadius, J. Riikonen, H. Lipsanen, J. Rynanen, "Extraction of graphene-titanium contact resistances using transfer length measurement and a curve-fit method," *Int. J. Mate. Metall. Eng.*, 6(8): 807-810, 2012.
- [16] J. Moon et al., "Ultra-low resistance ohmic contacts in graphene field effect transistors," *Appl. Phys. Lett.*, 100(20): 203512, 2012.
- [17] F. Giubileo, A. Di Bartolomeo, "The role of contact resistance in graphene field-effect devices," *Prog. Surf. Sci.*, 92(3): 143-175, 2017.
- [18] A. Gahoi, S. Wagner, A. Bablich, S. Kataria, V. Passi, M. C. Lemme, "Contact resistance study of various metal electrodes with CVD graphene," *Solid-State Electron.*, 125: 234-239, 2016.
- [19] S. Burzhuev, "Decreasing Graphene Contact Resistance by Increasing Edge Contact Length," *University of Waterloo*, 2016.
- [20] M. Houssa, A. Dimoulas, A. Molle, *2D materials for nanoelectronics*. CRC Press, 2016.
- [21] G. Vincenzi, "Graphene: FET and Metal Contact Modeling. Graphène: modélisation du FET et du contact métallique," *Université Paul Sabatier-Toulouse III*, 2014.

- [22] P. Zhang, Y. Lau, R. Gilgenbach, "Analysis of current crowding in thin film contacts from exact field solution," *J. Phys. D: Appl. Phys.*, 48(47): 475501, 2015.
- [23] G. González-Díaz et al., "A robust method to determine the contact resistance using the van der Pauw set up," *Measurement*, 98: 151-158, 2017.
- [24] F. Liu, W. T. Navaraj, N. Yogeswaran, D. H. Gregory, R. Dahiya, "van der Waals contact engineering of graphene field-effect transistors for large-area flexible electronics," *ACS nano*, 13(3): 3257-3268, 2019.
- [25] K. A. Jenkins, "Graphene in high-frequency electronics: This two-dimensional form of carbon has properties not seen in any other substance," *Am. Sci.*, 100(5): 388-398, 2012.
- [26] F. Urban, G. Lupina, A. Grillo, N. Martucciello, A. Di Bartolomeo, "Contact resistance and mobility in back-gate graphene transistors," *Nano Express*, 1(1): 010001, 2020.
- [27] P. Zhang, Y. Lau, "An exact field solution of contact resistance and comparison with the transmission line model," *Appl. Phys. Lett.*, 104(20): 204102, 2014.
- [28] W. S. Leong, H. Gong, J. T. Thong, "Low-contact-resistance graphene devices with nickel-etched-graphene contacts," *ACS nano*, 8(1): 994-1001, 2014.
- [29] K. Nagashio, T. Nishimura, K. Kita, A. Toriumi, "Contact resistivity and current flow path at metal/graphene contact," *Appl. Phys. Lett.*, 97(14): 143514, 2010.
- [30] A. Venugopal, L. Colombo, E. Vogel, "Contact resistance in few and multilayer graphene devices," *Appl. Phys. Lett.*, 96(1): 013512, 2010.
- [31] S. M. Popescu, A. J. Barlow, S. Ramadan, S. Ganti, B. Ghosh, J. Hedley, "Electroless nickel deposition: an alternative for graphene contacting," *ACS Appl. Mater. Interfaces*, 8(45): 31359-31367, 2016.

Biographies



Babak Khosravi Rad was born in Khorramabad in Iran, 1993. He received the B.Sc. degree in Electrical Engineering from Lorestan University of Technology, Lorestan, Iran. Also, he received the M.Sc. degree in Electrical Engineering from Malek-Ashtar University of Technology and He is a Ph.D. student in electronics from Tarbiat Modares University, Tehran, Iran. His research interests include GFET, detector, MEMS, and semiconductor field effect devices.

- Email: khosravirad_babak@modares.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Mehdi Khaje received the M.Sc. and Ph.D. degree in Electrical Engineering from Malek-Ashtar and Orumiyeh University of Technology, respectively. His research interests include superconducting amplifying and detector devices, MEMS, and semiconductor field effect devices.

- Email: m.khaje@mut.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Abdollah Eslami Majd was born in Hamadan, Iran, on March 23, 1976. He received the B.E. degree in applied physics from Bu-Ali Sina University, Hamadan, Iran in 1998. He received M.E. degree in atomic and molecular physics from Amir Kabir University of Technology Tehran Polytechnic, Tehran, Iran in 2001. He received the Ph.D. Degree in photonics from Laser and Plasma Institute of Shahid Beheshti University, Tehran, Iran in 2011. Since joining electrical engineering and electronic department of Malek Ashtar University of Technology in 2012, he has engaged in research and development of starry light in the satellite camera, laser induced breakdown spectroscopy (LIBS) and hemispherical resonator gyroscope (HRG). He is co-author of more than 30 publications. Dr. Eslami is a member of Optics and Photonics Society of Iran and Physics Society of Iran.

- Email: a_eslamimajd@mut-es.ac.ir
- ORCID: 0000-0002-7538-3160
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

B. Khosravi Rad, M. Khaje, A. Eslami Majd, "modeling transport in graphene-metal contact and verifying transfer length method characterization," *J. Electr. Comput. Eng. Innovations*, 11(2): 335-362, 2023.

DOI: [10.22061/jecel.2023.9266.597](https://doi.org/10.22061/jecel.2023.9266.597)

URL: https://jecel.sru.ac.ir/article_1841.html





Research paper

Actor Double Critic Architecture for Dialogue System

Y. Saffari, J. S. Sartakhti *

Department of Electrical and computer engineering, University of Kashan, Kashan, Iran.

Article Info

Article History:

Received 26 November 2022

Reviewed 28 December 2022

Revised 08 January 2023

Accepted 01 March 2023

Keywords:

Dialogue system

Actor-critic

Double DQN

Task-based

*Corresponding Author's Email

Address: salimi@kashanu.ac.ir

Abstract

Background and Objectives: Most of the recent dialogue policy learning methods are based on reinforcement learning (RL). However, the basic RL algorithms like deep Q-network, have drawbacks in environments with large state and action spaces such as dialogue systems. Most of the policy-based methods are slow, cause of the estimating of the action value using the computation of the sum of the discounted rewards for each action. In value-based RL methods, function approximation errors lead to overestimation in value estimation and finally suboptimal policies. There are works that try to resolve the mentioned problems using combining RL methods, but most of them were applied in the game environments, or they just focused on combining DQN variants. This paper for the first time presents a new method that combines actor-critic and double DQN named Double Actor-Critic (DAC), in the dialogue system, which significantly improves the stability, speed, and performance of dialogue policy learning.

Methods: In the actor critic to overcome the slow learning of normal DQN, the critic unit approximates the value function and evaluates the quality of the policy used by the actor, which means that the actor can learn the policy faster. Moreover, to overcome the overestimation issue of DQN, double DQN is employed. Finally, to have a smoother update, a heuristic loss is introduced that chooses the minimum loss of actor-critic and double DQN.

Results: Experiments in a movie ticket booking task show that the proposed method has more stable learning without drop after overestimation and can reach the threshold of learning in fewer episodes of learning.

Conclusion: Unlike previous works that mostly focused on just proposing a combination of DQN variants, this study combines DQN variants with actor-critic to benefit from both policy-based and value-based RL methods and overcome two main issues of both of them, slow learning and overestimation. Experimental results show that the proposed method can make a more accurate conversation with a user as a dialogue policy learner.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Task-based dialogue systems aim to interact with users to achieve their goals, such as movie ticket booking [1], restaurant reservation [2], and booking flights [3]. Dialogue systems, based on their structure of training, are divided into two categories: end-to-end and pipeline. The end-to-end ones directly map the user's action in a natural language way to the agent's response using a

sequence-to-sequence model with supervised learning [4]. Pipeline methods separate the system into some interdependent units, mainly natural language understanding (NLU), natural language generation (NLG), and dialogue management unit [5]. The dialogue management unit has the dialogue state tracker (DST) and the agent's policy, mostly represented by a neural network.

The goal of a dialogue manager is to learn the dialogue policy. The policy decides which action should be chosen based on the input state. In other words, the agent is trained in the dialogue management unit to learn the dialogue policy, so that it can interact directly with the knowledge base to create the appropriate action for each input from the user or user simulator [4].

Supervised learning and RL are used to train a task-based dialogue system [6]. Based on the task, the input space can be very large and make the DST misunderstand the input from the user, cause of errors in the NLU unit. Therefore, to overcome this issue and insufficient available annotated data, Reinforcement learning (RL) can be used to learn the policy automatically using interacting with a user simulator **Error! Reference source not found.** Then learning the dialogue policy, which can be viewed as a Markov decision process (MDP) [8], [9], [10], is often formulated as an RL problem [11].

Two main categories of model-free RL algorithms are policy-based and value-based learning algorithms [12]. Deep Q-networks (DQN) is one of the most successful value-based RL algorithms [13]. As DQN has a maximization of overestimated action values, it tends to prefer overestimated to underestimated values. This overestimation makes DQN learn unrealistically high action values [14]. Double DQN, is one of the variants of DQN that solves the issue of overestimation in DQN by decoupling the action selecting function and Q value estimator [15].

Dialogue policy learning for tasks with large state-action space, such as booking a movie ticket, needs to take a lot of dialogue turns to be able to explore in such a large space, leading to a long trajectory and finally delayed and sparse reward signals. To deal with reward sparsity, it's possible to employ intrinsic reward after each action to guide the exploration. Another class of RL algorithms takes advantage of value-based and policy-based RL algorithms methods, called actor-critic architecture. The critic is used to approximate the value and the actor tried to learn the policy, while the critic helps to update the actor by determining how much the action selected by the agent is good [16].

Many of the sequential decision-making problems, such as game playing, use DQN and its extension of it to learn the policy [17], [18]. DQN has been used in the context of dialog policy learning too [4], [19] but there is less effort in studying the extensions of DQN, especially in combination with other categories of RL algorithms such as actor-critic which is a combination of value and policy-based algorithms.

This study presents a new model for dialogue policy learning in a task-based dialogue system that combines the double DQN and actor-critic approaches to take advantage of both.

According to our information, this study proposes the first method that combines double DQN with the actor-critic method in a dialogue system environment.

The main contributions of this paper are as follows:

- The whole structure of the proposed method is actor-critic. The critic used to employ intrinsic reward after each action to guide the exploration and makes the learning process more stable in high dimensional state-action space.
- To overcome the overestimation in DQN, double DQN is employed in the actor unit.
- To have a smoother and faster update in the agent, the minimum of two errors was calculated based on the critic, and double DQN was obtained for optimizing the actor.

The proposed model is evaluated on the movie ticket booking dataset. This dataset was collected via Amazon Mechanical Turk, with annotations provided by domain experts [4], [20]. The evaluation results show that the proposed model operates well in the dialogue system tasks with high dimensional state-action space and the training process is more efficient than the basic algorithms, double DQN, and actor-critic.

The rest of the paper is organized as follows. First, briefly, researches that have investigated DQN methods in the dialogue system or have presented a new architecture of the combination of methods are reviewed. Then is provided some necessary background knowledge of the double DQN and actor-critic methods. The next section involves the main contribution of this study, which discusses the presented method which combines double DQN with the actor-critic network, and presents the details of the method and components of the dialogue system. In the results and discussion section, the experimental results of the proposed method on the movie ticket booking dataset are presented. Finally, the conclusion section, concludes the paper and makes suggestions for future research.

Related Work

There are many DRL-based approaches that are configured for dialogue management, but most of the complex DRL-based models that combine the variants of DQN or other types of RL algorithms, have been investigated in a game environment.

Many researchers had studied deep reinforcement learning (DRL) algorithms (a combination of RL and deep learning) to improve their training performance. However, the use of RL faces several problems such as the requirement for a very large number of train data, the instability of learning, slow learning, and convergence problems caused by the overestimation.

The first DRL method is DQN which for the first time proposed to play Atari games by Google Deep Mind [17],

[13]. Using a convolutional neural network (CNN), DQN extracts features from a screen of the game and then uses Q-learning [21] to learn how to play the game. A lot of research has been conducted on versions of DQN, to improve it.

Normal policy-based methods are slow as they have to estimate the value of each action through multiple episodes, and then normalize the sum of the future discounted rewards for each action. Actor-critic is a combination of value and policy-based methods. The critic gives directions to the actor, which means that the actor can learn the policy faster. Actor-critic needs less computation too because the policy is explicitly stored, and actor-critic methods can learn the optimal probabilities of selecting various actions. There is a lot of research that proposed models based on the actor-critic in all environments such as asynchronous advantage actor-critic (A3C) [22], A2C [23], actor-critic using Kronecker-factored trust region (ACKTR) [24], and soft actor-critic.

The actor-critic method has achieved superior performance on sequential decision-making problems [16], [19], [25]. Recently, some actor-critic algorithms are applied for dialogue policy learning, such as A2C [22], eNAC [26], and ACER [27]. ACER is the most efficient off-policy actor-critic method that, unlike basic actor-critic methods, it uses experience replay and some other methods to reduce the bias and variance of estimators. Su et al. employed the actor-critic model in dialogue policy optimization and showed that it can make convergence faster and more stable than other RL methods such as DQN [14].

In some of the complex stochastic environments, DQN leads to a suboptimal policy because of the overestimation of action values. This overestimation happens because of the noise on the approximations due to the generalization [14]. Hasselt et al. [15] presented a double-Q estimator for value-based RL methods to decrease the overestimated Q values. This leads to a more stable learning process and improves performance. While the double-Q estimator is the most popular method to solve the overestimation problem, there are other methods to solve this problem such as dropout techniques on DQN [28], cross DQN algorithm [29], and decreasing learning rate [30].

In the following, some popular research that tried to improve DQN in dialogue systems has been described. Firstly, studies that investigated DQN variants with different setups in a dialogue system environment to find the most suitable model have been described and concluded that DDQN could not outperform other variants such as dueling DQN. Then the works that investigated the actor-critic method as a combination of policy and value-based methods to overcome the

problems of DQN have been described and concluded that actor-critic can help to increase the speed of learning in the dialogue system. Also, some of the comparisons between DQN variants and actor critic showed that actor critic can outperform all of them including DDQN this can be because of the large state action space in the dialogue system and the huge help of critic to overcome the slow exploration and learning. Moreover, results of related works show that double DQN is ineffective in an actor-critic because due to the slow-changing policy in an actor-critic, the current and target value estimates remain too similar to avoid maximization bias. Reference [31] developed a novel variant of double Q-learning which limits possible overestimation. And their results demonstrate that mitigating overestimation can greatly improve the performance of modern algorithms.

References [32] and [33] investigated extensions of DQN such as double DQN, distributional DQN, and dueling DQN, individually and in combination, in the task-based dialogue system to find the most suitable model in a dialogue system. Finally, they concluded that choosing a specific algorithm for a specific task is not the right thing to do, and it depends on the state-action space, type of task, and the dataset. However, they have chosen the dueling network as the best choice over other methods or combinations. It seems that the network structure of basic DQN cannot model the Q-function perfectly while dueling DQN using an extra value function performs better in a dialogue system.

Fatemi used deep policy networks which are trained with an advantage actor-critic method for statistically optimized dialogue systems. The training process is done in a two-step approach: supervised learning and batch-based learning. The main benefit of their method, which paves the way for developing trainable end-to-end dialogue systems, was a combination of supervised and RL. They showed that the RL method based on an actor-critic architecture can exploit a small amount of data and can be used to improve the convergence speed of RL in dialogue systems. They compared the results of DQN, double DQN, advanced actor-critic (A2C), and SARSA algorithms on the dialogue state tracking challenge 2 (DSTC2) [34] dataset for the restaurant domain. According to experimental results, A2C had the fastest and best performance in convergence [35].

To reach to a faster and more stable learning, Gao proposed an adversarial A2C model which could perform well in the function of the dialogue system for ordering movie tickets. The proposed model trains a discriminator through an expert data file and online experiences. Then the trained discriminator is used as an additional critic to guide policy learning [23]. Also, Yen-Chen Wu proposed an actor-double-critic model to improve the performance stability of the DRL in a voice-based dialogue system for

restaurant ordering [25].

As mentioned in this section, most of the improvements in DQN and the combination of it with actor-critic methods are employed in game environments. Based on our information, there is no work that tries a combination of actor critic and double DQN in dialogue system environment. As the dialogue system environment in this study has a discrete space and usually dialogue systems in the real world suffer from slow convergence, two most popular extensions, double DQN to overcome the overestimation and actor-critic to faster convergence have been chosen to analyze and combine them to employ in this study.

Background

In this section firstly, the structure of whole dialogue systems described and after formulation the dialogue policy learning process as a Markov decision process, the two RL method that have been used in proposed new model in this study, are described.

An end-to-end dialogue system has a user or a user simulator with a user goal. A user goal represents the user's goal of the conversation at a particular task. After choosing a goal, the user's utterance passed through the NLU unit that the output of this unit is a lower-level representation of a natural language sentence, called a semantic frame. The DST takes the user's action and the history of the current conversation to build a state representation as input for the agent to learn the policy. The policy of the agent unit chooses an action using interaction with a database to fill the information slots. The agent's output, which is the action in the form of a semantic frame, is sent to the NLG component, and it converts the action to a natural language format for the user [23].

In a discrete space of dialogue, at each time step t , the current state $s_t \in S$ of the environment is sent to the agent. The agent responds by selecting an appropriate action $a_t \in A$. The user gets this action and, based on the affection of it in the environment, gives to the agent a signal named as reward $r_{t+1} \in R$ and new state $s_{t+1} \in S$. Formalization of this cycle as a Markov decision process (MDP) is in the form of $\langle S, A, R_{t+1}, S_{t+1}, \gamma \rangle$ [36].

A. Double DQN

Deep Q-network is the beginning of the development of RL into more complex decision-making problems. It tries to teach a network to predict the $Q(s, a)$ value of action a by receiving a state s . Using of target network trick, the loss to update the parameters of the agent at DQN with Q' as target network, formulated as (1), where γ is discount factor which $\gamma \in [0, 1]$, and α is learning rate [15]:

$$L = \left(R_{t+1} + \gamma \max_{a'} Q'(s_{t+1}, a') - Q(s_t, a_t) \right) \quad (1)$$

To solve the problem of the overestimation bias due to the maximization step in the conventional DQN, can decouple action selection and evaluation in DQN and rewrites the loss as (2) to update double DQN (DDQN):

$$L = (R_{t+1} + \gamma Q'(s_{t+1}, a) - Q(s_t, a_t)) \quad (2)$$

$$a = \operatorname{argmax}_{a'} Q(s_{t+1}, a') \quad (3)$$

B. Actor-Critic

Policy optimization methods are divided into two main categories: policy-based methods such as policy gradient and value-based such as Q-learning. But both of these major methods have drawbacks, which, combining two methods, can be complementary. Fig. 1 shows the general architecture of an actor-critic, combining policy-based and value-based approaches. The actor unit is used to generate the action.

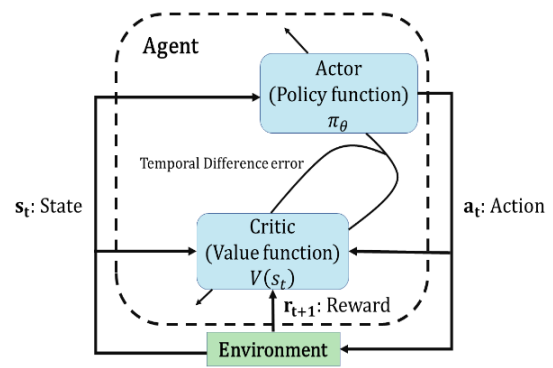


Fig. 1: Actor-Critic architecture.

The critic unit is supposed to approximate the value function and evaluate the quality of the policy used by the agent [37]. The evaluation method is the temporal difference (TD) error in (4), where V is the value function implemented by the critic [16]:

$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (4)$$

The critic uses the TD error to update itself (value function's parameters: w) in the gradient direction:

$$w \leftarrow w + \alpha \delta_t \nabla_w V(s) \quad (5)$$

The actor too updates itself (the policy parameters: θ) using the evaluation from the critic:

$$\theta \leftarrow \theta + \alpha \delta_t \nabla_\theta \ln \pi(s, a) \quad (6)$$

Methodology

In this study as illustrated in Fig. 2, a novel hybrid architecture, combining actor-critic and DDQN in a dialogue system, has been proposed. The overall neural network architecture for policy learner is an actor-critic that for overcoming the overestimation of Q values, DDQN is used to calculate the unbiased Q values.

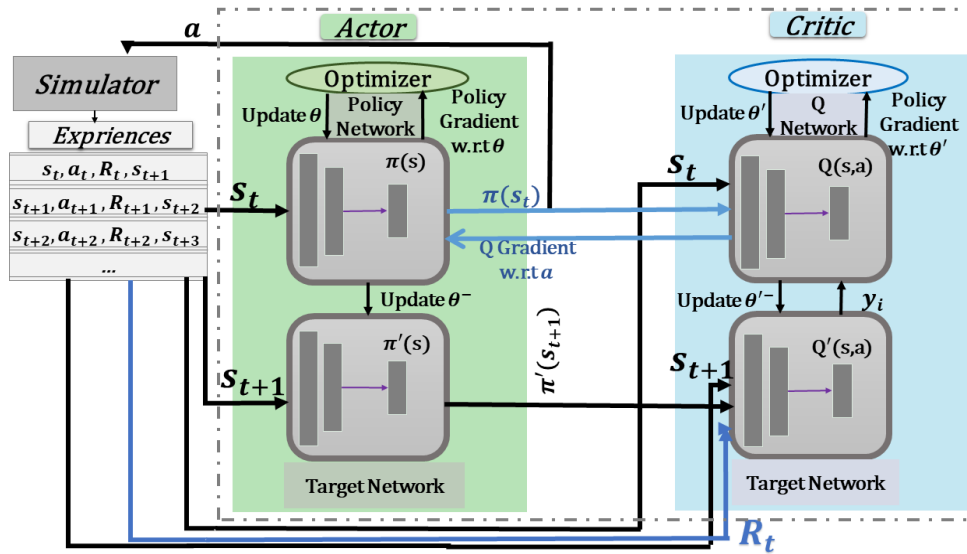


Fig. 2: Architecture of Double Actor-Critic: DAC.

The structure of the proposed policy learner model has been explained in detail in the proposed method subsection in further.

A. Dialogue System

In following firstly, the structure of the dialogue system framework and the elements of RL in this study like state and action, will be explained and then the proposed method will be explained in detail.

The dialogue framework used to apply the proposed agent is an end-to-end dialogue system that for simplicity, remains only the dialogue management component, and the NLU and NLG components have been removed. Fig. 3 shows the architecture of the dialogue system in this study.

To train end-to-end, a user simulator is needed that can automatically interact with the dialogue system. The simulator first generates a user goal while the agent does not know it and tries to accomplish that goal. Fig. 4 shows an example of the general format of a user goal. A user goal consists of two sections, inform slots and request slots. Inform slots are pairs of slot-value that express user restrictions and conditions. Requested slots are slots that the user wants to find some value for, during the conversation with the agent [4].

The inner function of the simulator is to produce action at each time step on a rule-based policy, based on the current state. The initial selected action by the user, must have at least one request slot and include the name of the movie as an information slot.

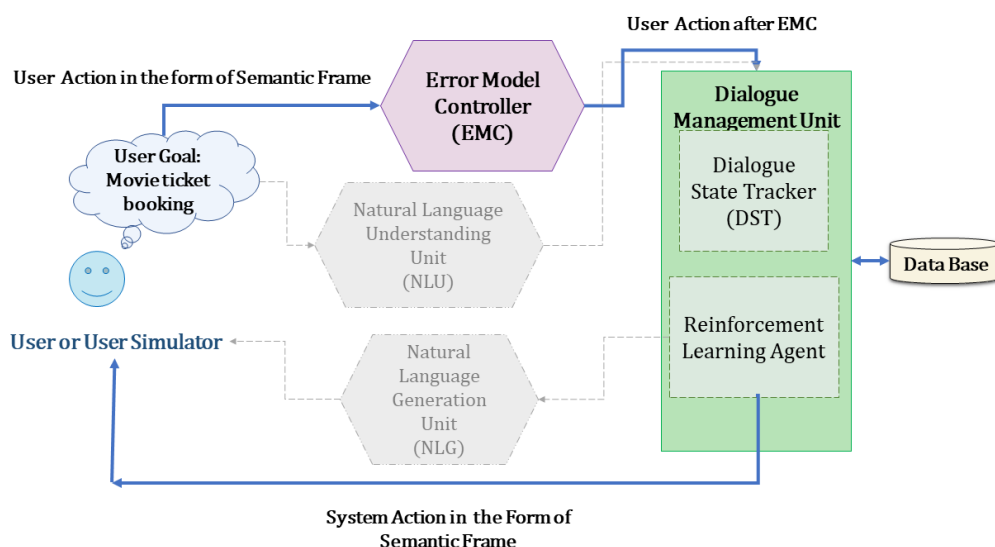


Fig. 3: Architecture of the used End-to-End dialogue system.

```
["request - slots": ["date": "UNK",
                    "starttime": "UNK"],
 "inform - slots": ["numberofpeople": "5",
                   "moviename": "avengers"]
```

Fig. 4: Example of user goal.

Briefly, a conversation cycle between the user simulator and the agent, as shown in Fig. 3, can be expressed as follows: 1) The user simulator produces an initial action according to certain rules. 2) The action of the user simulator, after applying noise in EMC, enters the DST in dialogue management. 3) Based on the information slots received from the simulator, the history and database are updated through interaction with the database. 4) A state is generated based on information retrieved from the database and history. 5) The generated state enters to the agent in the dialogue management and, based on the policy, action is generated. 6) The generated action, if its intent is informative, returns to the DST to quantify the information slots by interacting with the database and the history information in the DST. 7) The updated action enters to the user simulator. If the goal of the simulator will be completed and the desired ticket is found, the dialogue will close, otherwise, it will either end unsuccessfully or the dialogue will continue. To optimize this policy, in this research, dialogue management has been formulated as an RL learning problem. The state of this study is made up of useful information about the current state and conversation such as the last user action, last agent action, and round-num. The round-num is encoded to let the agent know if the episode was close to its max limit number of rounds that the agent might take a match-found action. A match-found action happens when the match ticket has been found. The steps taken to reach a match found action made a negative signal for the agent. As NLG and NLU components are ignored in the proposed system, then all the actions in this study, including the actions on the user side and agent side, are in the form of a semantic frame. Action in the form of a semantic frame including an intent. The intent represents the type of action and the rest of the action is split into inform slots which contain constraints and request slots which the sender does not know. In Fig. 5 Fig. 6 the format of action and state that are used in this work have been shown.

Finally, since the NLU component has been removed in this study, an error model controller (EMC) is created to simulate the error resulting from NLU. EMC boosted the dialogue system and make it more resistant to the noise caused by the error in NLU. The error used in this study is changing slot values with a probability.

```
["intent": "request",
 "inform - slots": ["city": "seattle"],
 "request - slots": ["date": "UNK", "starttime": "UNK"]]
```

Fig. 5: Example of Action in this study.

```
State = [user - action - rep,
        user - inform - slots - rep,
        user - request - slots - rep,
        agent - action - rep,
        agent - inform - slots - rep,
        agent - request - slots - rep,
        current - slots - rep,
        turn - rep,
        turn - onehot - rep,
        kb - binary - rep,
        kb - count - rep]
```

Fig. 6: Format of a state in this study.

B. Proposed Method

As mentioned before, dialogue policy learning can be formalized as a Markov decision process (MDP) in the form of $\langle S, A, R_{t+1}, S_{t+1}, \gamma \rangle$. The policy π is defined as a function $\pi: S \times A \rightarrow [0,1]$ that with probability $\pi(s, a)$ takes an action a in state s . The goal of RL is to find an optimal policy π^* , a policy that maximizes the value function which is the expected rewards of an episode in each state [36]. In this study, a hybrid deep reinforcement framework called Double Actor-Critic (DAC) was proposed, and ϵ -greedy search is employed during the training (please see Fig. 2). Using of a policy gradient method can suffer from a large variance. To overcome this issue, instead of increasing the size of batch size which causes a bad effect on sample efficiency, has been introduced a critic. The critic reduces the variance and, at the cost of bias, makes an improvement on the sample efficiency. In other words, reducing the cumulative reward using subtracting it with a baseline which here is the critic value, will make smaller gradients, and then smaller and more stable updates in policy learning will happen [14]. As the space of action-state of the studied task in this research is large, hence, it was decided to use the actor-critic architecture to learn policies better. To have more stability in training, batch learning is used too. The whole structure of DAC is actor-critic and the algorithm of DAC shows in Fig. 7. To approximate the true value function of the current policy π_θ , a critic $Q_w(s, a)$ was introduced. The objective function is (7):

$$J(\theta) = E Q_{\pi_\theta}(s, \pi_\theta(s)) \quad (6)$$

Using the deterministic policy gradient theorem, the gradient of the objective function is as (8):

$$\Delta_\theta J(\theta) = E [\Delta_\theta \pi_\theta(s) \Delta_a Q_{\pi_\theta(s,a)} | a = \pi_\theta(s)] \quad (7)$$

Then the Bellman operator, with expectation, is taken with respect to the next state s' , converted to (9) (line 5):

$$Q_\pi(s, a) = r(s, a) + \gamma E [Q(s', \pi(s')) | s, a] \quad (8)$$

And the TD error as final loss at this actor-critic network becomes (10) (line 7):

$$E [Q_w(s, a) - r(s, a) + \gamma E [Q'_w(s', \pi'_\theta(s')) | s, a]] \quad (9)$$

Algorithm 2 DAC Architecture

```

1: Inputs:
   initial learning rates  $\alpha_0$  and  $\beta_0$ , target
   network replacement frequency  $C$ 
2: Initialize:
   network  $Q$  with weights  $(\theta, w)$  at random
   target  $Q'$  with weights  $(\theta', w') \leftarrow (\theta, w)$ 
3: for  $t=1$  to  $T$  do
4:   Sample  $M$  transitions  $(s_i, a_i, r_i, s_{i+1})$  uniformly
5:    $Y_{i(AC)} = r_i + \gamma Q_{w'}(s_i, \arg \max_{b \in A} Q_{\theta'}(s_{i+1}, b))$ 
6:    $Y_{i(DDQN)} = r_i + \gamma Q_{\theta'}(s_i, \arg \max_{b \in A} Q_{\theta}(s_{i+1}, b))$ 
7:    $\delta_w = \frac{1}{M} \sum_i \Delta_w(Y_{i(AC)} - Q_w(s_i, a_i))$ 
8:    $\delta_\theta = \frac{1}{M} \sum_i \Delta_\theta \pi_\theta(s_i) E[\Delta_w Q_w(s_i, a_i)]|_{a=\pi_\theta(s_i)}$ 
9:    $\delta_{DDQN} = \frac{1}{M} \sum_i \Delta_\theta(Y_{i(DDQN)} - Q_\theta(s_i, a_i))$ 
10: end for
11: update:
12:  $\delta_\theta = \min(\delta_{DDQN}, \delta_\theta)$ 
13:  $\theta \leftarrow \theta + \alpha \delta_\theta$ 
14:  $w \leftarrow w + \beta \delta_w$ 
15: Every  $C$  times, update target network:  $(\theta', w') \leftarrow (\theta, w)$ 
16: Output: policy parameters  $\theta$ 

```

Fig.7: The algorithm of DAC.

Hence, the objective function to optimize the policy in actor-critic is as (11) where (θ', w') referred to the parameters of actor and critic's target network (line 2) to stabilize learning in TD error and T is the number of steps in the episode (line 8).

$$\Delta_\theta J(\theta) = E[\sum_{i=0}^{T-1} \Delta_\theta \log \pi_\theta(s_i) Q_w(s, a)] \quad (10)$$

One of the problems with DQN in this task is that due to the maximization operator in predicting $q(s, a)$, the Q value of some actions is likely to be overestimated. This problem led the agent to constantly learn some non-optimal actions. In this task, in Q -learning, the value function is updated with a greedy target at step $t + 1$:

$$y = r + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \quad (11)$$

However, if the target is susceptible to error ϵ , then the maximum over the value along with its error will generally be greater than the true maximum [14]:

$$E_\epsilon[\max_{a_{t+1}} (Q(s_{t+1}, a_{t+1}) + \epsilon)] \geq \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \quad (12)$$

To solve this problem, two separate networks are used, one to select the appropriate action and the other to calculate the value of actions, and since these two separate models are trained, it is less likely to estimate the same actions. To implement the proposed architecture, there are two similar networks called online and target as actor part, that per each some fixed number of episodes, the weights of the online network are copied to the target. In the target Q value calculation, for updating network weights, the Q value is obtained based on the selected action of the online model but through the target model. While in DQN a maximum is used to calculate the Q of the next state, but here first the best action of the next state is selected and then the Q value of that action is extracted from the target network. Finally, to have smaller gradients, the $Q(s, a)$ calculated using the online network in the actor part, at the episode

$< e = s, a, r', s' >$, is subtracted by the (14) (line 6):

$$Y^{double} = r' + \gamma Q_{target}(s', A) \quad (13)$$

which A is the optimal action obtained by the online network and the result of this subtracting is assumed as a double loss (line 9). To have a smoother update in the agent, the minimum of two losses of actor-critic (10) and double was obtained for updating of actor, in each batch (line 12). The target network of DDQN updates periodically based on the weights of actor (line 15), and critic updates based on the target Q value that is calculated in the critic component. Updating of actor-critic done using stochastic gradient V with learning rates α and β respectively (lines 13,14), which adjusted are using ADAM [38].

Results and Discussion

This study considers a task-based dialogue system for helping users to book movie tickets. During the conversation, the dialogue system gathers information about the customer's desires and ultimately books the movie tickets. The environment then assesses a binary outcome (success or failure) at the end of the conversation, based on whether or not the movie is found in a limited time. The focus of this study is on the evaluating the DQN methods and the proposed novel model, DAC.

C. Dataset

The basic conversational data were collected via Amazon Mechanical Turk, with annotations provided by domain experts. The annotated dataset that has been used in this study is for movie ticket booking [4], [20] It consists of 2890 dialogue sessions, with approximately 7.5 turns per session on average. The annotated data includes 29 dialogue actions and 11 slots, of which most of the slots are inform slots, that users can use to constrain the search, and some are request slots, of which users can ask for values from the agent.

D. Analysis

For analyzing the performance and improvements in the presented new architecture, the dialogue system of the proposed method was trained with DDQN and the proposed model DAC. Then based on the success rate factor, these models were compared and the results of these experiments are shown in follow. All implementations have been performed in Google's Colab environment. The hyperparameter settings are given in Table 1.

Table 1: The parameters of model

Parameter	Value
discount factor	0.9
max memory size	100000
DQN hidden size	80
batch size	16
learning rate	1e-3
warm up memory	7000
number of episodes	15000
train frequency	100

E. Results

In this section, the actor-critic and DDQN and the combination of them implemented and the performance of them at the move ticket booking task, have been analyzed. The metric that has been used to measure the quality of the agent was the success rate. Per each transition, a reward based on the reward function and a success score is returned. A good policy should have a high success rate. The success rate (15) is known as the fraction of dialogues that are done successfully where the user goal matches the information the system acquired during the interaction:

$$\text{SuccessRate} = \frac{\text{Period Success Total}}{\text{fixed frequency of train}} \quad (14)$$

I) Double DQN

Previous works [17] have experimented with variants of DQN on Atari games. Wang [39] did research that experimented with variants of DQN in the same task and dataset in the dialogue system and reported that DDQN improves very little in most environments while performing better in restaurant and taxi domains. But it seems not to affect significantly in the movie domain. Based on the setup of learning at experiments, as shown in Fig. 8 of DQN and DDQN, the behavior of both likely is the same. However, it looks like in training for future episodes, the results get better. Then, it can be concluded that DQN does not suffer a lot from the problem of overestimation in Q value predicting in this task, hence, the improvement of DDQN here is limited. This can happen, because of some reason, an agent at this task could predict Q values near to actual Q values and non-optimal actions are not given a higher Q value than the optimal best action.

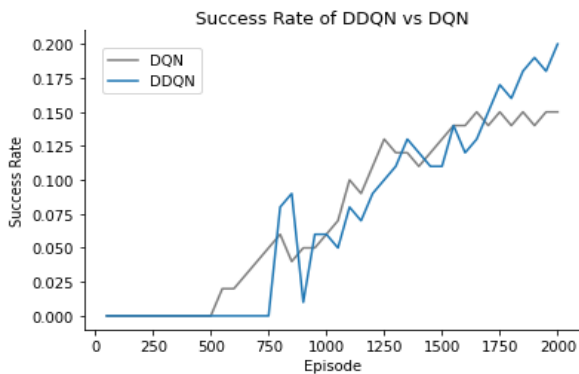


Fig. 8: Success rate of double DQN vs. DQN.

II) Actor Critic

Regular policy-gradient methods have slow convergence due to the large variances of the gradient estimates. The actor-critic methods attempt to reduce the variance using a critic network to estimate the value of the policy, which is then used to update the actor parameters [26]. As shown in Fig. 8 of DQN and actor-critic, the behavior of both likely is the same, but the actor-critic improved in stability success rate. It shows that the critic made actor decisions better and more accurately with more stability.

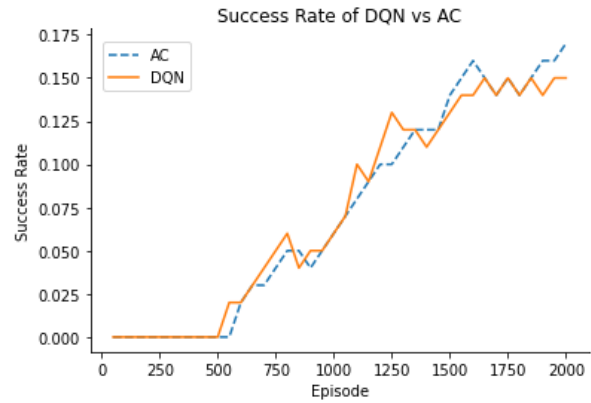


Fig. 9: Success rate of actor critic vs. DQN.

III) Double Actor Critic

In this paper, DDQN and actor-critic methods are combined to have the advantage of both. Based on the result of experiments on DDQN shows it doesn't have a significant effect to improve the DQN, but actor-critic made the agent more stable and faster. Hence, it is expected that the behavior of the proposed model, DAC, be affected by the actor-critic more and the behavior should be more stable. It is noteworthy that the proposed model reached the threshold of success rate faster too (please see Fig. 10). It can be concluded that the collaboration of the critic and target model of DDQN can result in better and solve some overestimation issues of DQN and actor-critic too. In Fig. 11, the behavior of all methods is shown.

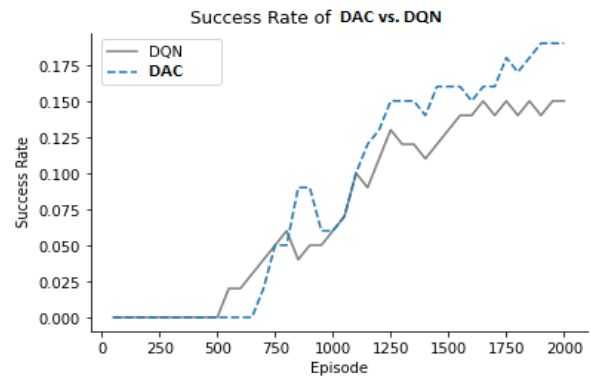


Fig. 10: Success rate of double actor critic vs. DQN.

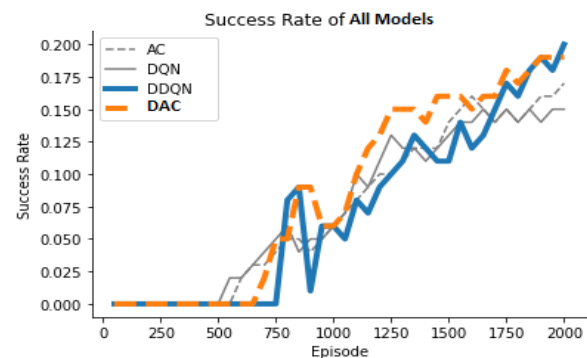


Fig. 11: Comparison of Actor-Critic, Deep Q Network (DQN), Double Deep Q Network (DDQN), and Double Actor-Critic (DAC) based on the success rate metric.

Conclusions

This study proposed a new architecture that combined two reinforcement methods, actor-critic and DDQN to have the advantage of both methods and examined it on a task-based dialogue system for booking the ticket movie task. DDQN has no significant effect on the results of DQN at this task and can conclude that this task does not suffer from notable overestimation that it can happen because of the task features. Actor-critic made more effect on the success rate and made the learning more stable and accurate. Finally, the investigated model, DAC, shows somehow comparable results. Actor-critic made it faster and more stable and can conclude that collaborating with the DDQN and actor-critic to make more accurate decisions, ends well as DDQN could solve new overestimation made by the actor-critic.

In a real application, the domain of dialogues is usually multi-domain and the diversity of actions is more and the training process is more complex. Then in this situation, learning is getting slower and more unstable. The biggest advantage of the proposed method in this study is a faster and more stable convergence in a complex domain.

In this study a user simulator with a rule-based policy has been used to interact with the agent in training. However, the best way is using a real user but it's not possible cause of the time and cost of it. This is obvious that the trained model with a user simulator is not excellent in the test step with the real user but the training with the simulator is a common and useful method. However, using a multi-agent interaction can be a good resolve that is out of the subject of this study.

The proposed idea of combining two methods and analysis from this research guide the future directions of applying more types of actor-critic methods to DQN variants for dialogue policy learning, although some methods may don't behave as well as they behave in Atari environments cause of different environmental features.

Author Contributions

Y.Saffari and JS.Sartakhti designed the experiments. Y.Saffari collected the data and carried out the data analysis. Y.Saffari and JS.Sartakhti interpreted the results and wrote the manuscript.

Acknowledgment

The author would like to thank the editor and reviewers for their helpful comments.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

rep	Representation
-----	----------------

NLU	Natural Language Understanding
NLG	Natural Language Generation
DST	Dialogue State Tracker

References

- [1] Z. C. Lipton, J. Gao, L. Li, X. Li, F. Ahmed, L. Deng, "Efficient exploration for dialog policy learning with deep {BBQ} networks & replay buffer spiking," arxiv preprint arxiv: 1608.05081, 2016.
- [2] T. H. Wen, D. Vandyke, N. Mrkšić, M. Gašić, L. M. Rojas-Barahona, P. H. Su, S. Ultes, S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics: 438-449, Valencia, Spain, 2017.
- [3] H. Cuayahuítl, S. Renals, O. Lemon, H. Shimodaira, "Hierarchical Dialogue optimization using semi-markov decision processes," in Proc. 8th Annual Conference of the International Speech Communication Association: Interspeech: 2693-2696, 2007.
- [4] X. Li, Y. N. Chen, L. Li, J. Gao, A. Celikyilmaz, "End-to-End task-completion neural dialogue systems," in Proc. Eighth International Joint Conference on Natural Language Processing, (1): 733-743, Taipei, Taiwan, 2017.
- [5] H. Sun, C. Zhao, S. Liu, H. Jiang, "A pipeline dialogue system scheme," in Proc. 2nd International Conference on Machine Learning and Computer Application: 1-5, Shenyang, China, 2021.
- [6] R. Fellows, H. Ihshaish, S. Battle, C. Haines, P. Mayhew, J. I. Deza, "Task-oriented dialogue systems: performance vs. quality-optima, a review," arxiv preprint. arxiv: 2112.11176, 2021.
- [7] M. I. Bahria, Z. Yan, "Supervised machine learning approaches: A survey," Int. J. Soft Comput., (5): 946-952, 2015.
- [8] R. Howard, Dynamic Programming and Markov Processes, The MIT Press, Cambridge, 1960.
- [9] S. Young, M. Gasić, B. Thomson, J. D. Williams, "Pomdp-based statistical spoken dialog systems: A review," Proc. IEEE, 101(5): 1160-1179, 2013.
- [10] J. D. Williams, S. Young, "Partially observable markov decision processes for spoken dialog systems," Comput. Speech Lang., 21(2): 393-422, 2007.
- [11] J. Williams, A. Raux, D. Ramachandran, A. Black, "The dialog state tracking challenge," in Proc. SIGDIAL: 404-413, 2013.
- [12] P. Swazinna, S. Udluft, D. Hein, T. Runkler, "Comparing model-free and model-based algorithms for offline reinforcement learning," IFAC-PapersOnLine, 55(15):19-26, 2022.
- [13] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, "Human-level control through deep reinforcement learning," Nature: 529-33, 26 Feb 2015.
- [14] S. Thrun, A. Schwartz, "Issues in using function approximation for reinforcement learning," in Proc. 4th Connectionist Models Summer School, 1993.
- [15] H. van Hasselt, A. Guez, D. Silver, "Deep reinforcement learning with double q-learning," arxiv preprint arxiv:1509.06461, 2015.
- [16] R. Chen, J. H. Goldberg, "Actor-critic reinforcement learning in the songbird," Curr. Opin. in Neurobiol., (65): 1-9, 2020.
- [17] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. A. Riedmiller, "Playing atari with deep reinforcement learning," DeepMind Technologies, 2013.
- [18] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. V. D. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, "Mastering the game of go with deep neural networks," Nature (529): 484, 2016.
- [19] B. Peng, X. Li, J. Gao, J. Liu, Y.-N. Chen, K.-F. Wong, "Adversarial advantage actor-critic model for task-completion dialogue policy learning," Int. Conf. IEEE, Acoustics, Speech and Signal Processing (ICASSP): 6149-6153, 2018.

- [20] X. Li, Z. C. Lipton, B. Dhingra, L. Li, J. Gao, Y. N. Chen, "A user simulator for task-completion dialogues," arxiv preprint arxiv:1612.05688, 2016.
- [21] C. J. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, (8): 279-292, 1992.
- [22] V. Mnih, A. Puigdomènech Badia, "Asynchronous methods for deep reinforcement learning," arxiv preprint arxiv:1602.01783v2, 2016.
- [23] J. Gao, M. Galley, L. Li, "Neural approaches to conversational ai, question answering, task-oriented dialogues and social chatbots," arxiv preprint arxiv:1809.08267, 2019.
- [24] Y. Wu, E. Mansimov, S. Liao, R. Grosse, "Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation," arxiv preprint arxiv:1708.05144, 2017.
- [25] Y. C. Wu, B. H. Tseng, M. Gas, "Actor-double-critic: incorporating model-based critic for task-oriented dialogue systems," *Findings of the association for computational linguistics: EMNLP*: 854–863, 2020.
- [26] J. Peters, S. Vijayakumar, S. Schaal, "Natural Actor-Critic," *ECML*: 280–291, 2005.
- [27] Z. Wang, V. Bapst, N. Hees, V. Mnih, R. Munos, K. Kavukcuoglu, N. D. Freitas, "Sample efficient actor-critic with experience replay," arxiv preprint arxiv:1611.01224, 2016.
- [28] M. Sabry, K. M. A. Amr, "On the reduction of variance and overestimation of deep q-learning," arxiv preprint arxiv:1910.05983v1, 2019.
- [29] X. Wang, A. Vinel, "Cross learning in deep q-networks," arxiv preprint arxiv:2009.13780v1, 2020.
- [30] Y. Chen, L. Schomaker, M. A. Wiering, "An Investigation Into the Effect of the Learning Rate on Overestimation Bias of Connectionist Q-learning," in *Proc. International Conference on Agents and Artificial Intelligence* 2021.
- [31] S. Fujimoto, H. van Hoof, D. Meger, "Addressing function approximation error in actor-critic methods," 2018.
- [32] Y. A. Wang, Y. N. Chen, "Dialogue environments are different from games: Investigating variants of deep q-networks for dialogue policy," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*: 1070-1076, 2019.
- [33] D. Vath, N. T. Vu, "To combine or not to combine? A rainbow deep reinforcement learning agent for dialog policy," *University of Stuttgart, Institute for Natural Language Processing (IMS)*, 2019.
- [34] M. Henderson, B. Thomson, J. D. William, "The second dialog state tracking challenge," in *Proc. 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014.
- [35] M. Fatemi, L. E. Asri, H. Schulz, J. He, K. Suleman, "Policy networks with two-stage training for dialogue systems," arxiv preprint arxiv: 1606.03152, 2016.
- [36] H. R. Chinaei, B. Chaib-draa, L. Lamontagne, "Learning observation models for dialogue POMDPs," in *Proc. Canadian Conference on Artificial Intelligence: Springer*(7310), 2012.
- [37] I. Grondman, L. Busoniu, G. A. D. Lopes, R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, 42(6): 1291–1307, 2012.
- [38] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," in *3rd Int.Conf. Learning Representations*, San Diego, 2015.
- [39] Z. Wang, T. Schaul, M. Hessel, H. V. Hasselt, M. Lanctot, F. De, "Dueling network architectures for deep reinforcement learning," *Computer Science, Machine Learning*, 2015.

Biographies



Yasaman saffari received her B.S. degree in Computer Engineering (Software) from Dr. Shariati Vocational and Technical Girls College in 2017 and her M.Sc. degree in Master of Arts in Computer Arts (Intelligent Simulators Design) from Faculty of Multimedia, Tabriz Islamic Art University in 2020. Her research interests include NLP, dialogue system, deep reinforcement learning.

- Email: y.saffari@grad.kashanu.ac.ir
- ORCID: [0000-0002-7178-0855](https://orcid.org/0000-0002-7178-0855)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Javad Salimi Sartakhti received his B.S. degree in Computer Engineering (Software) from University of Kashan in 2009, his M.Sc. degree in Computer Engineering (Software) from Tarbiat Modares University in 2012 and his Ph.D. degree in Computer Engineering (Software) from Isfahan University of Technology in 2016. His research interests include game theory & mechanism design, machine learning algorithms, deep learning and blockchain.

- Email: salimi@kashanu.ac.ir
- ORCID: [0000-0003-1183-1232](https://orcid.org/0000-0003-1183-1232)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: <https://faculty.kashanu.ac.ir/salimi/en>

How to cite this paper:

Y. Saffari, J. S. Sartakhti, "Actor double critic architecture for dialogue system," *J. Electr. Comput. Eng. Innovations*, 11(2): 363-372, 2023.

DOI: [10.22061/jecei.2023.9346.614](https://doi.org/10.22061/jecei.2023.9346.614)

URL: https://jecei.sru.ac.ir/article_1842.html





Research paper

Design of Miniaturized Microstrip Antenna with Semi-Fractal Structure For GPS/GLONASS/Galileo Applications

S. Komeyliani¹, M. Tayarani^{1,*}, S. H. Sedighy²

¹Department of Electrical Engineering, Iran University of Science and Technology, Tehran 1684613114, Iran.

²School of New Technologies, Iran University of Science and Technology, Tehran 16846-13114, Iran.

Article Info

Article History:

Received 08 November 2022

Reviewed 24 December 2022

Revised 08 February 2023

Accepted 01 March 2023

Keywords:

Miniaturized antenna

Low-profile antenna

Semi-fractal structure

Multiple feed configuration

Pure polarization

RHCP antenna

*Corresponding Author's Email Address:

m_tayarani@iust.ac.ir

Abstract

Background and Objectives: Microstrip patch antennas are widely used due to their advantages of compact size and easy fabrication compared to other types. However, they have low-performance parameters. As a result, several techniques are used to improve performance parameters in newly designed microstrip antennas. In this study, a novel miniaturized microstrip antenna with circular polarization (CP) is proposed for GNSS applications.

Methods: In the design process, the semi-fractal structure is used to reduce the antenna size. Circular polarization is generated using a three-feed configuration with 120° phase shift. The CP value is increased by use of perturbing slots and also removing the corners. The novel design of the feeding network and also considering the ground size same as the patch layer, keep the antenna size small. The co-axial probe is used in the feeding network and it is printed on Taconic RF-43 substrate with a low loss tangent of 0.0033. Numerical simulation is applied via CST commercial software to evaluate the antenna performance. The simulations are repeated in two other software, HFSS and FEKO, to validate the study.

Results: The proposed antenna has a compact size of 17.56 cm². The single-layer structure of the designed antenna leads to easy fabrication feature. The proposed antenna has a bandwidth of 55 MHz (1.558-1.614 GHz). It can operate at GPS L1 (1575 MHz), GLONASS G1 (1602 MHz), Galileo E1 (1589 MHz), and E2 (1561 MHz) bands. Results show a high front-to-back ratio (FBR) of 40 dB, RHCP gain of 3.45 dB, and pure CP with axial ratio (AR) beamwidth of 108°. Furthermore, the phase center variation (PCV) is less than 0.16 mm.

Conclusion: Key features of the proposed antenna are its novel fractal structure that leads to compact size, high front-to-back ratio, wide RHCP beamwidth with desirable bandwidth, and axial ratio beamwidth.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Microstrip antennas with circular polarization are widely used in satellite applications [1]–[8], such as global positioning systems (GPS) and global navigation satellite systems (GLONASS) to reduce multipath reflection effects. A circularly polarized radiated field is generally generated by the excitation of two orthogonal modes

with a 90° phase difference [9], [10]. In the single feed configuration, the feed should be located at a convenient position to excite orthogonal modes and produce CP [5], [11], [20], [21], [12]–[19].

Orthogonal modes are generated by perturbing antenna structure, such as creating narrow slits near the

edges of the patch [22], arc-shaped or orthogonality-located slots [14], [23], and also truncated corners of the patch [1]. Multiple feed patch antennas provide a larger CP purity and higher performance [1], but they generally have a larger size for feeding networks [24]. In most satellite applications, the size of the antenna is important [21], [25].

Fractal geometry is a low-cost method for miniaturizing the microstrip patch antenna which is used in several studies [26], [27]. Fractals are geometric shapes composed of multiple iterations of a single shape [6], which allows for a reduction in metallization and resonant frequency [28].

In this study, a novel design of a microstrip antenna with semi-fractal geometry is presented. The proposed antenna operates at GPS L1 (1575 MHz), GLONASS G1 (1602 MHz), Galileo E1 (1589 MHz), and E2 (1561 MHz) bands. The antenna is compact, low-profile, and planar. Details of the design process and simulation results are presented and discussed in the following sections.

Antenna Design

The proposed microstrip antenna consists of a patch layer, a Taconic RF-43 substrate, a ground plane, a feeding substrate, and a feeding network. To miniaturize the antenna with a low-cost method, a semi-fractal structure is employed. Then optimization is done to obtain the best geometry with the best performance.

The fractal structure, design process, and feeding techniques used in this study, are described as follows.

A. Fractal Structure

In the first step of producing the patch geometry, an equilateral triangle, as the first polygon, is selected. The main triangle size is obtained by optimization. The smallest triangle is selected which satisfies design constraints containing operating frequency and desirable bandwidth. The main triangle has a side length of 57 mm. In general, patch antenna has low bandwidth. It is proved that by removing the corners of the patch geometry, the bandwidth will be increased [29]. As a result, in the next step, the corners of the main triangle are removed.

Fig. 1 shows the current distribution on the surface of arced vertex triangle at $f=1.52$ GHz, the resonant frequency of the first step geometry. As is clear, more current is concentrated on the patch boundaries. So, by increasing the boundaries of the antenna geometry, the current will be increased and as a result, more radiation is achieved. For this reason, a semi-fractal structure is selected to develop the patch geometry.

The fractal part size at each step is chosen via optimization with the aim of the smallest patch satisfying the constraints, the same as the main triangle.

The fractal generation is started by removing the scaled, reversed shape triangle from the center of the patch as illustrated in Fig. 2-a. In the next step, again a

scaled, reversed triangle is truncated from the center of the patch as shown in Fig. 2-b. By repeating the previous steps with a scaled triangle, the geometry will be as Fig. 2-c. The final shape is obtained by removing two triangles from the external edges of the central triangle, which is illustrated in Fig. 2-d.

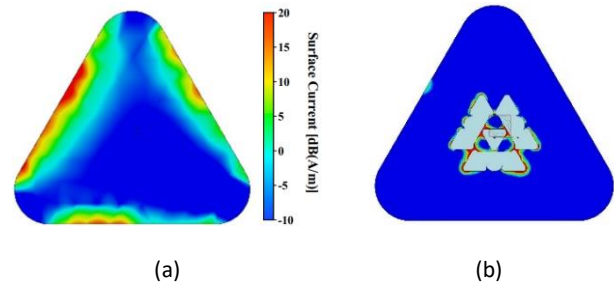


Fig. 1: Current distribution on the patch geometry (a) Before fractal generation, (b) After fractal generation.

The effects of each step geometry generation are evaluated in terms of return loss (S_{11}). As shown in Fig. 3, by each iteration, the resonant frequency is decreased, hence compression occurs. The compression ratio at the end of fractal structure generation is 12%.

As illustrated in Fig. 4, the edges of the designed patch structure, are removed by contraction of the reversed main triangle from the patch geometry. Converting the triangle shape to the hexagon, increases the antenna bandwidth, as shown in Fig. 5. In general, it can be concluded that tending the patch geometry to a circle shape will increase the antenna bandwidth.

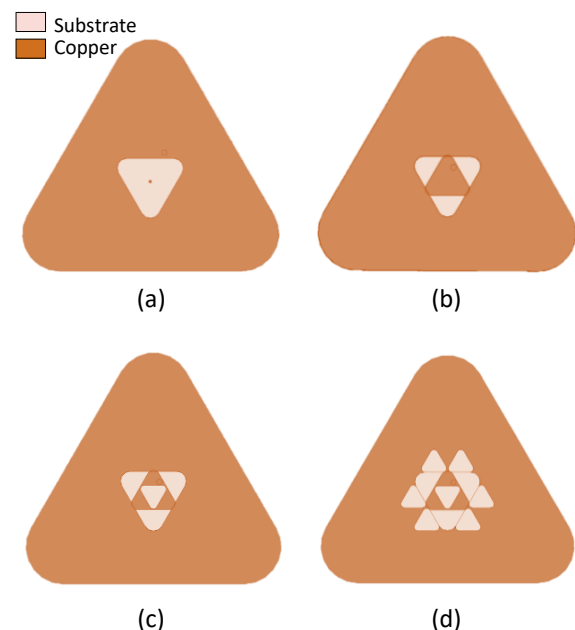


Fig. 2: The fractal structure generation steps from (a) to (d).

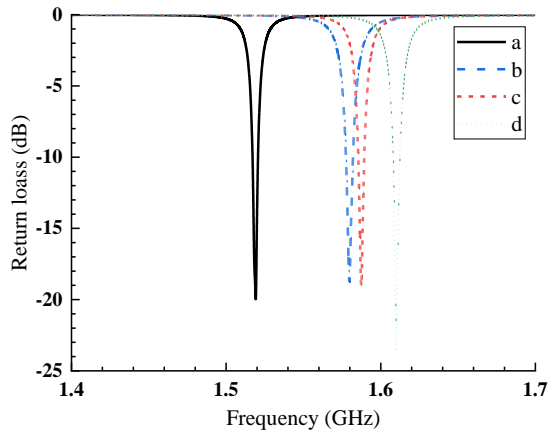


Fig. 3: Effects of fractal iterations.

B. Design Process

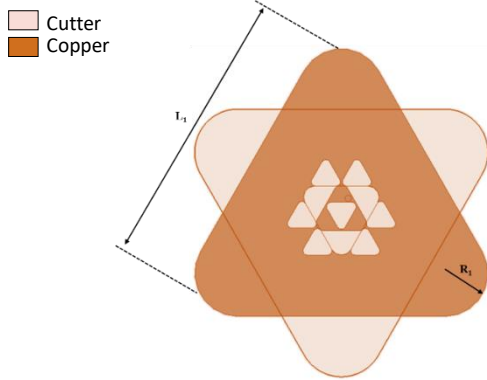


Fig. 4: Truncating edges of the designed antenna.

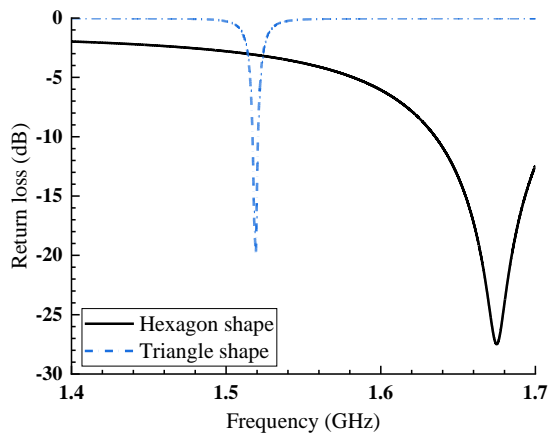


Fig. 5: Effect of converting the triangle to the hexagon.

The final geometry of the proposed microstrip antenna is shown in Fig. 6. The geometric parameters of the designed antenna and their optimum values are presented in Table 1.

As is clear, a hexagon slot and three circular slots are printed on the patch, which improve the antenna resonant frequency adjustment as illustrated in Fig. 7.

Table 1: The geometric parameters

Geometric parameters	Optimum value (mm)	Geometric parameters	Optimum value (mm)
L_1	39.55	R_1	14.14
L_2	26	R_2	1.24
L_3	22	R_3	1.8
L_4	10.2	R_f	3
L_5	6.79	t_1	0.16
L_6	3.99	t_2	0.2
L_7	3.38	ϕ_1	1.12
L_h	12.8	h	3

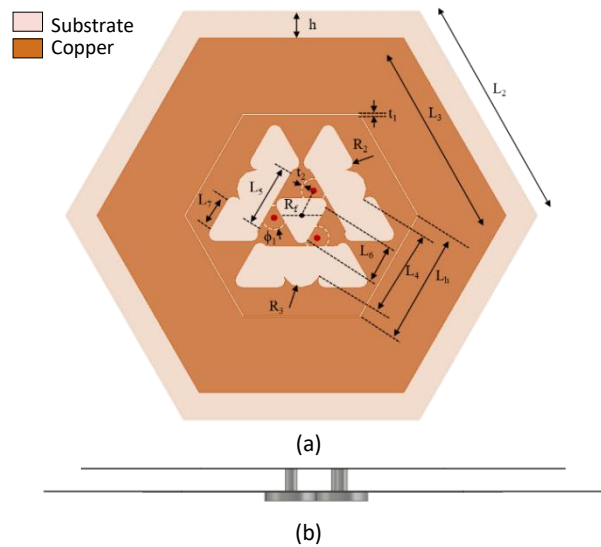


Fig. 6: The final geometry of the designed antenna; (a) Front view, (b) Side view.

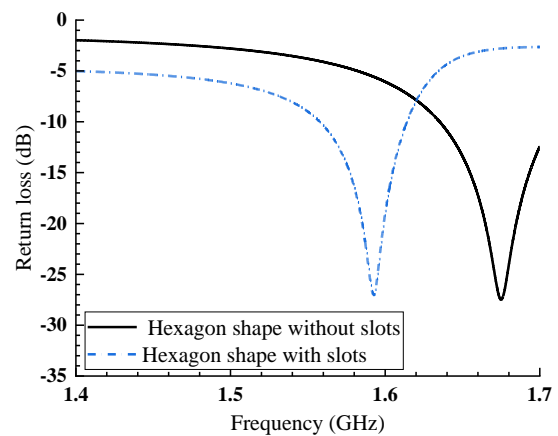


Fig. 7: The effect of adding slots.

The total compression ratio at the end of geometry design (hexagon shape with slots) is 30% with an increase of bandwidth. The hexagon patch layer is etched onto the upper side of the Taconic RF-43 substrate with relative permittivity of 4.3 and a loss tangent of 0.0033, and the feeding structure is printed on the other side. The substrate dimensions are 50mm×50mm×2mm, and the patch has a compact size of 40mm×40mm.

C. Parametric Study

The effects of key geometric parameters on the proposed antenna performance are analyzed and discussed. The parameters consist of the main triangle lateral length (L_1), the fractal triangle vertex arc radius (R_2), and the hexagon slot length (L_h). Sensitivity to the material has been checked by evaluating the variable relative permittivity (ϵ_r), between 4.1 and 4.5. The thickness of the substrate (t_{sub}) is also studied. Except for the studied parameter, other parameters have been constant.

The effect of the main triangle lateral length (L_1) on the proposed antenna bandwidth is depicted in Fig. 8.

As clearly shown, the optimum value of $L_1=39.55$ mm will result in operating at desirable frequency bands. The reduced L_1 does not satisfy below -10 dB return loss and increased L_1 shifts the operating range to lower frequencies.

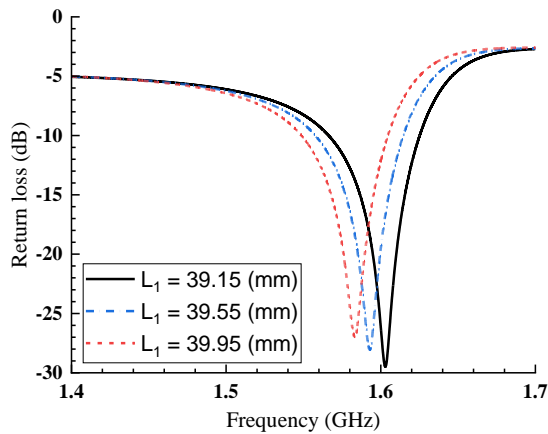


Fig. 8: The effect of the main triangle lateral length.

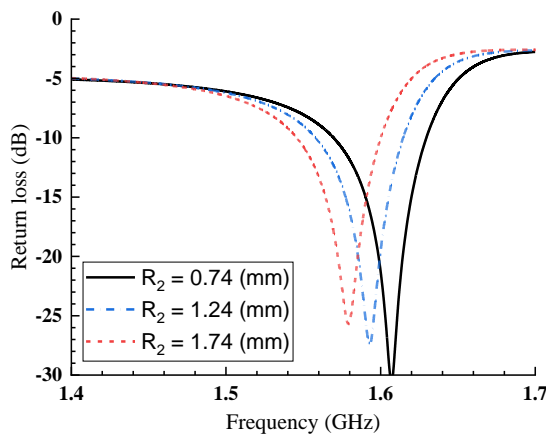


Fig. 9: The impact of the vertex arc radius.

Fig. 9 shows the impact of the vertex arc radius (R_2). The effect of R_2 parameter is the same as L_1 , and it has the optimum value of $R_2=1.24$ mm.

The hexagon slot length (L_h) effect on the performance of the proposed antenna is illustrated in Fig. 10. It is shown that by increasing L_h , the -10 dB return loss

bandwidth is improved and the operation bands will shift to high frequencies. The desirable performance is achieved at $L_h=12.8$ mm.

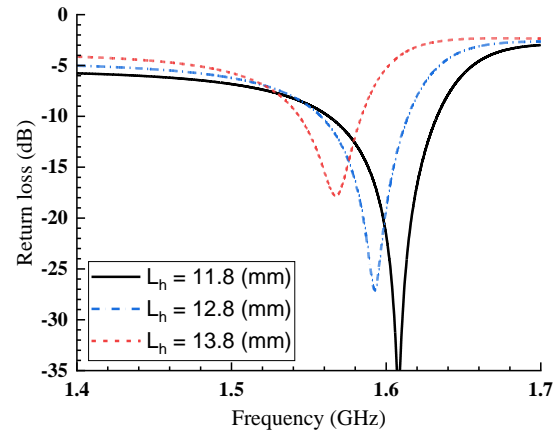


Fig. 10: The effect of hexagon slot length.

The substrate material sensitivity is studied by evaluating the effect of relative permittivity on return loss. Fig. 11 shows that an increase or decrease of ϵ_r , considerably changes the operating frequency range.

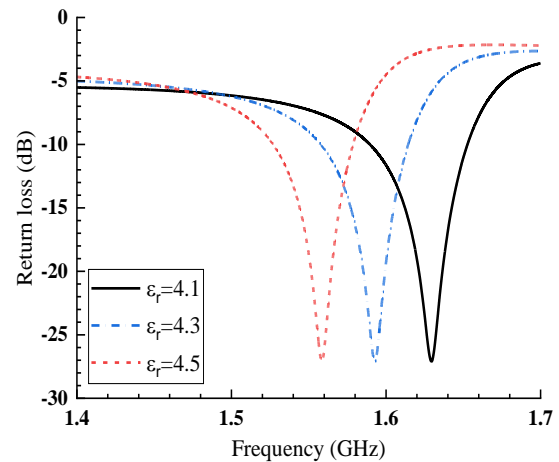


Fig. 11: The effect of substrate relative permittivity.

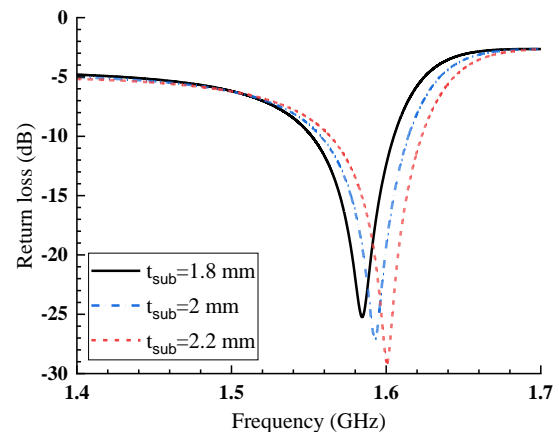


Fig. 12: The effect of substrate thickness.

Fig. 12 demonstrates the effect of substrate thickness. Decreasing t_{sub} shifts the resonant frequency to a lower value and an increase in it shifts to the higher value of resonant frequency. The optimum value of $t_{\text{sub}} = 2$ mm leads to operating at the desired frequency range.

D. Feeding Techniques

In this study, a three-feed configuration is designed as shown in Fig. 13. The feeding network is designed with a coaxial probe to be single layer and low-profile. Theoretical analysis is done to find the required temporal phase shift for CP. The electric field created by feeding port 1 on the Z-axis is called E_1 :

$$E_1 = \begin{bmatrix} e_{0x} \\ e_{0y} \\ e_{0z} \end{bmatrix} \quad (1)$$

The field created by the second feeding port is called E_2 . Considering the symmetric feeding configuration, the electric field resulting from E_2 is similar to E_1 , with the difference in spatial phase shift $\theta = \frac{2\pi}{3}$:

$$E_2 = e^{-j\varphi} \left[T\left(\frac{2\pi}{3}\right) \right] \begin{bmatrix} e_{0x} \\ e_{0y} \\ e_{0z} \end{bmatrix} \quad (2)$$

The T matrix represents the rotation of θ around the Z-axis. Also, the time delay between the excitation of the second port and the first port is considered as φ .

$$[T(\theta)] = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

Similarly, the electric field caused by the third feeding port is calculated as equation (4).

$$E_3 = e^{-2j\varphi} \left[T\left(\frac{4\pi}{3}\right) \right] \begin{bmatrix} e_{0x} \\ e_{0y} \\ e_{0z} \end{bmatrix} \quad (4)$$

Hence:

$$E_{\text{total}} = \left(1 + e^{-j\varphi} \left[T\left(\frac{2\pi}{3}\right) \right] + e^{-2j\varphi} \left[T\left(\frac{4\pi}{3}\right) \right] \right) \begin{bmatrix} e_{0x} \\ e_{0y} \\ e_{0z} \end{bmatrix} \quad (5)$$

E_{total} is simplified to equation (6).

$$E_{\text{total}} = \frac{1}{2} \begin{bmatrix} ue_{0x} - ve_{0y} \\ ve_{0x} + ue_{0y} \\ 2 \frac{1 - e^{-3j\varphi}}{1 - e^{-j\varphi}} e_{0z} \end{bmatrix} = \begin{bmatrix} e_x \\ e_y \\ e_z \end{bmatrix} \quad (6)$$

where u and v are defined as (7) and (8).

$$u = (1 + e^{j\alpha} + e^{2j\alpha}) + (1 + e^{j\beta} + e^{2j\beta}) \quad (7)$$

$$v = -j(1 + e^{j\alpha} + e^{2j\alpha}) + j(1 + e^{j\beta} + e^{2j\beta}) \quad (8)$$

Also, β is implied as (9).

$$\beta = -\varphi - \frac{2\pi}{3} \quad (9)$$

And α is defined as (10).

$$\alpha = -\varphi + \frac{2\pi}{3} \quad (10)$$

Therefore, the right-handed electric field is calculated from (11).

$$|R| = \left| \frac{e_x - je_y}{2} \right| = |(1 + e^{j\beta} + e^{2j\beta})| |(e_{0x} - je_{0y})| \quad (11)$$

Similarly, the left-handed electric field is obtained from (12).

$$|L| = \left| \frac{e_x + je_y}{2} \right| = |(1 + e^{j\alpha} + e^{2j\alpha})| |(e_{0x} + je_{0y})| \quad (12)$$

As shown in equation (11), $\beta=0$ leads to maximum R and thus pure right-handed circular polarization (RHCP). So according to equation (9) pure CP is achieved by setting the temporal and spatial phase shifts with the same value and negative sign. In this study the temporal phase shift for symmetric three-feed configuration is $-\frac{2\pi}{3}$. By setting $\varphi = -\frac{2\pi}{3}$ and $\alpha = \frac{4\pi}{3}$ the left-handed electric field shown in (12) will be zero.

The feeding network consists of transmission lines and a 3-way power divider. A Schematic of the feeding network elements is shown in Fig. 14. The length difference of the transmission lines, leads to the 120° phase difference between the ports. The power divider used in this study has a novel design. It consists of two lumped resistors of 75Ω for isolation and also some open-circuited lines for matching. The final design of the feeding network is simulated in ADS. The designed feeding network is fabricated on a Taconic RF-43 substrate with 0.508 mm thickness and a dielectric constant of 4.3.

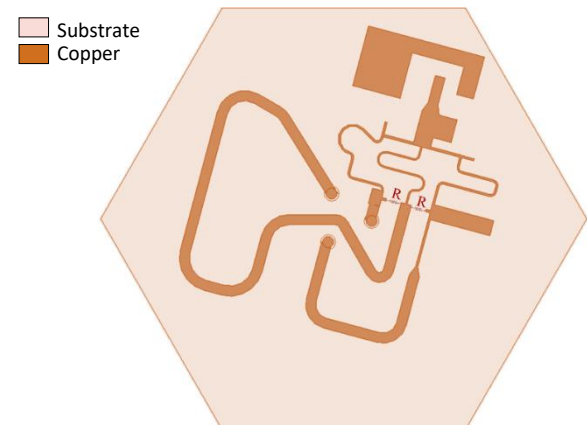


Fig. 13: The designed feeding network.

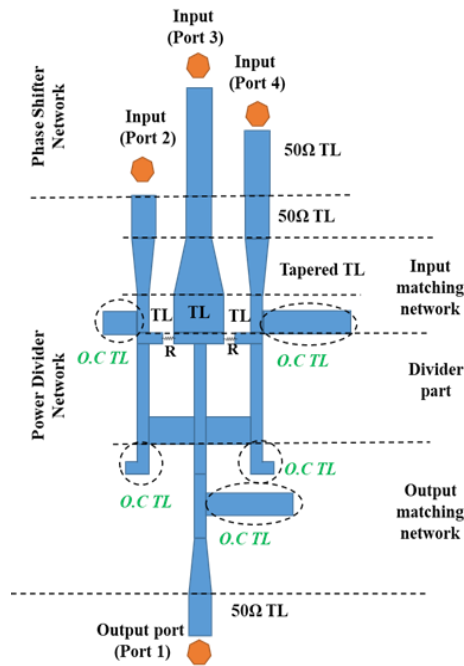


Fig. 14: Schematic of the feeding network elements.

Results and Discussion

The performance of the proposed antenna is evaluated via numerical simulation at CST Studio Suite software. To validate the results, the simulation is repeated in two other software, HFSS and FEKO. The results of the three simulations are depicted in terms of return loss, RHCP, and LHCP radiation pattern in Fig. 15 and Fig. 16, respectively. It is clear that all simulations are in good agreement and this validates the current study.

Fig. 15 shows the S_{11} versus frequency for the proposed antenna. As clearly shown, the simulated antenna is capable to operate at desirable frequency bands, (GPS L1, GLONASS G2, GALILEO E1, and E2), with a bandwidth of 56 MHz (1.558-1.614 GHz).

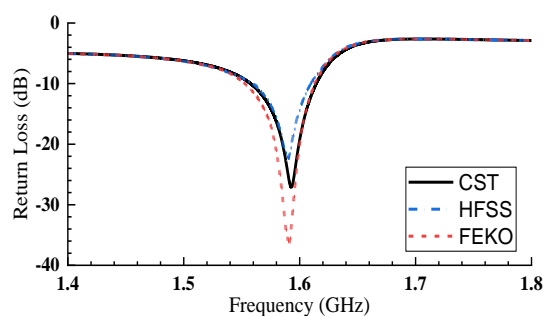


Fig. 15: Return loss.

The RHCP and LHCP radiation patterns of the designed antenna are illustrated in Fig. 16. Results show the RHCP beamwidth of 103° . It is also clear that the FBR of the proposed antenna is 40 dB. Furthermore, on the upper half plane, low LHCP gain is observed. The RHCP gain is 3.45 dB.

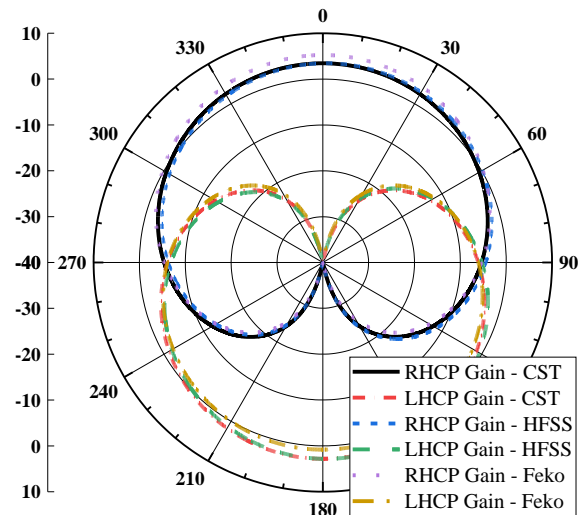


Fig. 16: RHCP and LHCP radiation pattern.

Fig. 17 demonstrates the AR beamwidth and the ratio of right-to-left radiation (R/L). It is clear that below 3 dB AR beamwidth is 108° . Also, in the range of 134° , R/L ratio has a value above 10,

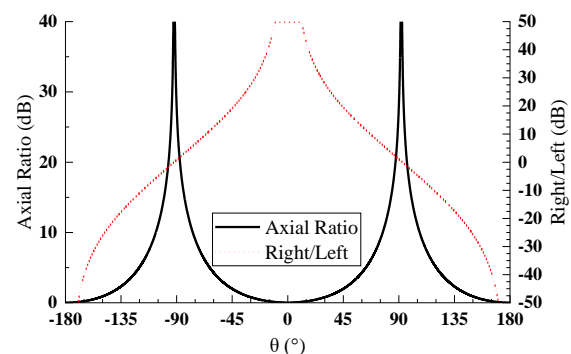
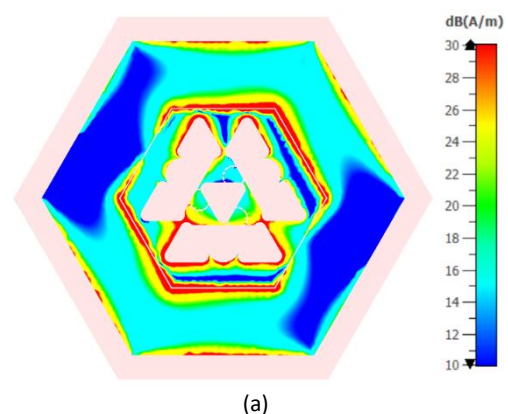


Fig. 17: Axial Ratio and R/L ratio versus theta at the resonant frequency.

The current distribution on the patch layer at the resonant frequency is shown in Fig. 18. Currents are depicted in three consecutive phases of 60° , 120° , and 180° . The current travels in a counter-clockwise direction with successive phase changes. This proves the circular polarization radiation of the proposed antenna.



(a)

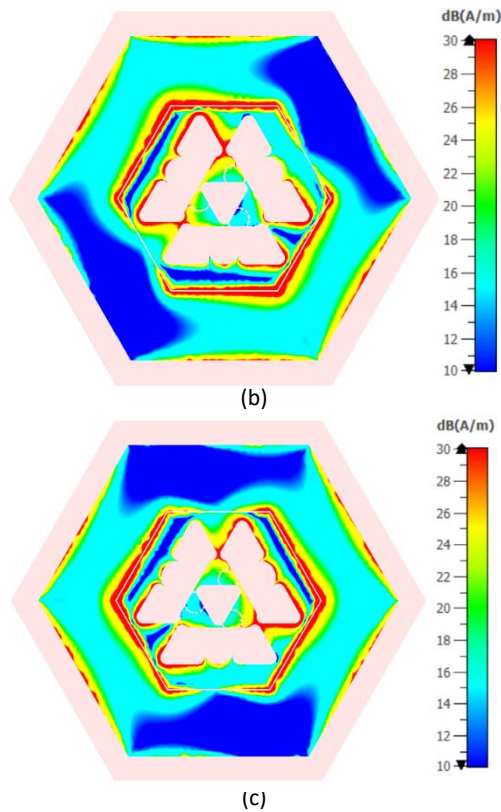


Fig. 18: Current distribution at different phase shifts; (a) 0° (b) 60° (c) 120°.

Phase stability is considered an important performance parameter for GPS antennas. To achieve phase stability, the antenna phase center variation (PCV) should tend to zero in an ideal case. In Fig. 19, coordinates of the phase center and also PCV for the proposed antenna are displayed towards frequency. As clearly depicted, at the whole antenna operation frequency (1.558-1.614 GHz), the PCV is less than 0.16mm.

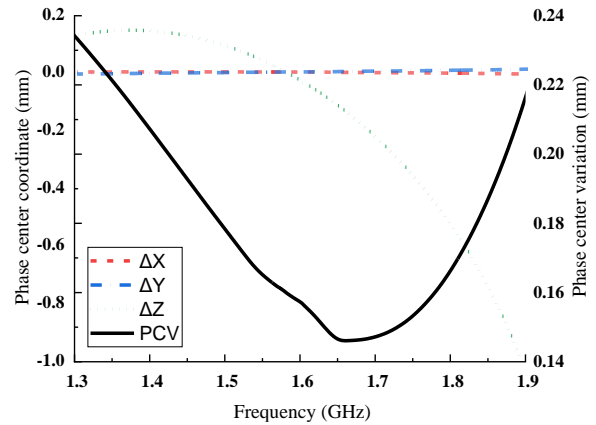


Fig. 19: Phase center coordinate and variation versus operating frequency.

Table 2: Comparison between the proposed antenna and previous studies.

Reference	Substrate	Resonant frequency (GHz)	Antenna area (mm ²)	Antenna height (mm)	FBR (dB)	RHCP beamwidth (deg)	Phase center variation (mm)	AR beamwidth (deg)
[1]	FR4	1.575	18496	24.2	25	85	1.2	134
[3]	FR4, Ceramic	1.575	900	4.8	20	83	NS*	153
[27]	organic magnetic substrate	1.575	3600	4	35	90	NS	NS
[30]	Rogers RT/Duroid 5880	1.575	19600	1.6	12	62	NS	NS
[31]	Taconic ORCER Cer – 10	1.575	10000	3.2	26.22	100	NS	100
[32]	FR4	1.575	3969	1.6	2.7	97	NS	101
[33]	RO3003C	1.575, 2.45	5320	3.1	NS	98	NS	121
Proposed antenna	Taconic RF-43	1.575	1756	2	40	103	0.365	108

* Not Stated

 : Best item

Conclusion

A novel miniaturized, low-cost, multiple-feed, circularly polarized microstrip antenna has been designed for GPS/GLONASS applications. The antenna can operate at GPS L1 (1575 MHz), GLONASS G1 (1602 MHz), Galileo E1 (1589 MHz), and E2 (1561 MHz) bands. For CP radiation, multiple feeds with the coupled technique are used. To miniaturize the antenna, semi-fractal geometry and also novel design of a feeding network are employed. Key features of the proposed antenna are its compact size, high FBR, wide RHCP beamwidth, desirable bandwidth, and AR beamwidth.

The designed antenna size is below 18 cm². Numerical simulation results show the proposed antenna has favorable AR of below 3 dB and back-ward radiation below -50 dB with FBR over 40 dB. Additionally, the proposed antenna has bandwidth of 56 MHz (1558-1614 MHz) and a wide RHCP beamwidth of 103°. Furthermore, the PCV is less than 0.16 mm.

Author Contributions

Each part of this paper was contributed by all three authors.

Acknowledgment

The authors would like to thank the reviewers and editor of the JECEI for their valuable comments.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

CP	Circular Polarization
RHCP	Right Hand Circular Polarization
LHCP	Left Hand Circular Polarization
AR	Axial Ratio
FBR	Front to Back Ratio
PCV	Phase Center Variation

References

- [1] F. Bilotti, C. Vegni, "Design of high-performing microstrip receiving gps antennas with multiple feeds," *IEEE antennas wirel. Propag. Lett.*, 9: 248–251, 2010.
- [2] Y. M. Cai, K. Li, Y. Yin, L. Zhao, "A multi-functional tri-mode patch antenna supporting dual-band gps and wireless communication system," *microw. Opt. Technol. Lett.*, 59: 2457–2462, 2017.
- [3] S. J. Jeong, "Compact circularly polarized antenna with a capacitive feed for gps/glonass applications," *etri j.*, 34: 767–770, 2012.
- [4] K. A. Yinusa, "A dual-band conformal antenna for gnss applications in small cylindrical structures," *IEEE antennas wirel. Propag. Lett.*, 17: 1056–1059, 2018.
- [5] Nasimuddin, Z. Ning Chen, X. Qing, "Dual-band circularly polarized s-shaped slotted patch antenna with a small frequency-ratio," *IEEE trans. Antennas propag.*, 58: 2112–2115, 2010.
- [6] M. A. S. M. Al-haddad, N. Jamel, A. N. Nordin, "Flexible antenna: a review of design, materials, fabrication, and applications," *J. Phys. Conf. Ser.*, 1878: 012068, 2021.
- [7] K. A. Yinusa, E. P. Marcos, S. Caizzone, "Robust satellite navigation by means of a spherical cap conformal antenna array," in *Proc. 2018 18th International Symposium on Antenna Technology and Applied Electromagnetics (ANTEM)*, 2018.
- [8] E. K. Kaivanto, M. Berg, E. Salonen, P. De maagt, "Wearable circularly polarized antenna for personal satellite communication and navigation," *IEEE trans. Antennas propag.*, 59(12): 4490–4496, 2011.
- [9] S. Komeylian, M. Tayarani, S. Hassan, "Design of a novel four-feed dual-band microstrip antenna with pure circular polarization and analysis of circular polarization parameters," *int. J. Nonlinear anal. Appl.*, 14, 2023.
- [10] S. Komeylian, M. Tayarani, S. Hassan, "A novel blind adaptive algorithm applied to new designed smart antenna array for interference suppression," *int. J. Nonlinear anal. Appl.*, 14, 2023.
- [11] K. Y. Lam, K. Luk, K. F. Lee, H. Wong, K. B. Ng, "Small circularly polarized u-slot wideband patch antenna," *IEEE antennas wirel. Propag. Lett.*, 10: 87–90, 2011.
- [12] Z. Ma, J. Chen, P. Chen, Y. F. Jiang, "Design of planar microstrip ultrawideband circularly polarized antenna loaded by annular-ring slot," *int. J. Antennas propag.*, 2021: 1–10, 2021.
- [13] X. Tang, H. Wong, Y. Long, Q. Xue, K. L. Lau, "Circularly polarized shorted patch antenna on high permittivity substrate with wideband," *IEEE trans. Antennas propag.*, 60: 1588–1592, 2012.
- [14] S. A. Rezaeieh, "Dual band dual sense circularly polarised monopole antenna for gps and wlan applications," *electron. Lett.*, 47: 1212, 2011.
- [15] X. sun, Z. Zhang, Z. Feng, "Dual-band circularly polarized stacked annular-ring patch antenna for gps application," *IEEE antennas wirel. Propag. Lett.*, 10: 49–52, 2011.
- [16] N. K. Suyan, F. Lal lohar, C. Dhote, Y. Solunke, "Design of circularly polarized irnss receiver antenna using characteristic mode analysis," *proc. Conect 2020 - 6th IEEE int. Conf. Electron. Comput. Commun. Technol.*, pp. 3–7, 2020.

- [17] J. Sze, W. Chen, "Axial-ratio-bandwidth enhancement of a microstrip-line-fed circularly polarized annular-ring slot antenna," *IEEE trans. Antennas propag.*, 59: 2450–2456, 2011.
- [18] C. Lin, F.-S. Zhang, Y.-C. Jiao, f. Zhang, x. Xue, "a three-fed microstrip antenna for wideband circular polarization," *IEEE antennas wirel. Propag. Lett.*, 9: 359–362, 2010.
- [19] Y. M. Cai, K. Li, Y. Z. Yin, X. Ren, "Dual-band circularly polarized antenna combining slot and microstrip modes for gps with his ground plane," *IEEE antennas wirel. Propag. Lett.*, 14: 1129–1132, 2015.
- [20] S. Gao, Q. Luo, F. Zhu, circularly polarized antennas. Chichester, uk: john wiley & sons, ltd, 2014.
- [21] S. Komeyliyan, S. Komeyliyan, F. H. Kashani, "Anisotropic uniaxial crystal as a substrate in spherical microstrip antenna with annular-circular patch and air gap layer," 2014 int. Work. Antenna technol. Small antennas, nov. Em struct. Mater. Appl. Iwat: 385–388, 2014.
- [22] S. M. Kim, K. S. Yoon, W. G. Yang, "Dual-band circular polarization square patch antenna for gps and dmb," *Microw. Opt. Technol. Lett.*, 49: 2925–2926, 2007.
- [23] G. Byun, H. Choo, S. Kim, "Design of a small arc-shaped antenna array with high isolation for applications of controlled reception pattern antennas," *IEEE trans. Antennas propag.*, 64: 1542–1546, 2016.
- [24] H. Chen, Y. Wang, Y. Lin, C. Lin, s. Pan, "Microstrip-fed circularly polarized square-ring patch antenna for gps applications," *IEEE trans. Antennas propag.*, 57: 1264–1267, 2009.
- [25] S. Komeyliyan, S. Komeyliyan, "Deploying an ofdm physical layer security with high rate data for 5g wireless networks," *IEEE canadian Conference on Electrical and Computer Engineering (CCECE)*, 2020: 1–7, 2020.
- [26] S. Shrestha, S. R. Lee, D. Y. Choi, "A new fractal-based miniaturized dual band patch antenna for rf energy harvesting," *int. J. Antennas propag.*, 2014: 1–9, 2014.
- [27] E. Wang, Q. Liu, "gps patch antenna loaded with fractal ebg structure using organic magnetic substrate," *prog. Electromagn. Res. Lett.*, 58: 23–28, 2016.
- [28] H. Malekpoor, M. Shahraiki, "Compact broadband microstrip triangular antennas fed by folded triangular patch for wireless applications," *adv. Electromagn.*, 10: 14–23, 2021.
- [29] M. Moore, Z. Iqbal, S. Lim, "A size-reduced, broadband, bidirectional, circularly polarized antenna for potential application in wlan, wimax, 4g, and 5g frequency bands," *prog. Electromagn. Res. C*, 114: 1–11, 2021.
- [30] A. R. Alajmi, M. A. Saed, "A pin-loaded microstrip patch antenna with the ability to suppress surface wave excitation," *prog. Electromagn. Res. C*, 62: 131–137, 2016.
- [31] K. K. So, H. Wong, K. M. Luk, C. H. Chan, "Miniaturized circularly polarized patch antenna with low back radiation for gps satellite communications," *IEEE trans. Antennas propag.*, 63: 5934–5938, 2015.
- [32] Z. Mar Phyto, T. May Nway, K. Kyu Kyu Win, H. Myo Tun, "Development of microstrip patch antenna design for gps in myanmar," *am. J. Electromagn. Appl.*, 8: 1, 2020.
- [33] Y. A. Sheikh, K. N. Paracha, S. Ahmad, A. R. Bhatti, A. D. Butt, S. K. A. Rahim, "Analysis of compact dual-band metamaterial-based patch

antenna design for wearable application," *arab. J. Sci. Eng.*, 47: 3509–3518, 2022.

Biographies



Saeed Komeyliyan was born in Tehran, Iran, in 1989. He received the B.Sc. degree from the IKIU, Qazvin, Iran, in 2011, the M.Sc. degree from Sharif University of Technology, Tehran, Iran in 2013, and he is a PhD student in the Faculty of Electrical Engineering, IUST, Tehran, Iran.

- Email: Komeyliyan.official@gmail.com
- ORCID: [0000-0001-6367-1985](https://orcid.org/0000-0001-6367-1985)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Majid Tayarani was born in Tehran, Iran, in 1962. He received the B.Sc. degree from the University of Science and Technology, Tehran, Iran, in 1988, the M.Sc. degree from Sharif University of Technology, Tehran, Iran in 1992, and the Ph.D. degree in communication and systems from the University of Electro-Communications, Tokyo, Japan, in 2001. From 1990 to 1992, he was a Researcher with the Iran Telecommunication Center, where he was involved with nonlinear microwave circuits. Since 1992, he has been a member of the faculty with the Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran, where he is currently an Associate Professor. His research interests are qualitative methods in engineering electromagnetic, electromagnetic compatibility (EMC) theory, computation and measurement techniques, microwave and millimeter-wave linear and nonlinear circuit design, microwave measurement techniques, and noise analysis in microwave signal sources.

- Email: m_tayarani@iust.ac.ir
- ORCID: [0000-0001-8605-0428](https://orcid.org/0000-0001-8605-0428)
- Web of Science Researcher ID:
- Scopus Author ID
- Home pgae: <http://ee.iust.ac.ir/content/13094/Dr.Tayarani>



Seyed Hassan Sedighy was born in Qaen, South Khorasan, Iran, in 1983. He received his B.Sc., M.Sc. and Ph.D. degrees all in Electrical Engineering from Iran University of Science and Technology (IUST), in 2006, 2008 and 2013, respectively. From December 2011 to July 2012, he was with the University of California, Irvine as a Visiting Scholar. He joined the School of New Technologies at IUST, as an Assistant

Professor in 2013.

- Email: sedighy@iust.ac.ir
- ORCID: [0000-0002-5813-5616](https://orcid.org/0000-0002-5813-5616)
- Web of Science Researcher ID:
- Scopus Author ID
- Home page: http://www.iust.ac.ir/page.php?slct_pg_id=13647&sid=94&slc_lang=en

How to cite this paper:

S. Komeylian, M. Tayarani, S. H. Sedighy, "Design of miniaturized microstrip antenna with semi-fractal structure for GPS/GLONASS/Galileo applications," J. Electr. Comput. Eng. Innovations, 11(2): 373-382, 2023.

DOI: [10.22061/jecei.2023.9352.609](https://doi.org/10.22061/jecei.2023.9352.609)

URL: https://jecei.sru.ac.ir/article_1843.html





Research paper

GSM based Water Salinity Monitoring System for Water Gate Management in Salt Farms

M. Enriquez*, A. Abella

College of Engineering, Occidental Mindoro State College, San Jose, Occidental Mindoro, Philippines.

Article Info

Article History:

Received 12 January 2023
Reviewed 05 February 2023
Revised 28 February 2023
Accepted 01 March 2023

Keywords:

Salt
Salinity
GSM module
Arduino uno
Solar sheets
DC motor

*Corresponding Author's Email
Address:
michelle_enriquez_d@dlsu.edu.ph

Abstract

Background and Objectives: Salt production is an ancient industry that still used primitive or traditional systems of evaporation. As technology continues to prosper in all aspects of life; the use of technology-based products is still a challenge in salt production. With the tedious activities and processes in salt farming; salt producers and salt farmers continue to look for alternatives to lessen the hard works. Salt farm activities initially started with the intrusion of saline water into the salt beds, but monitoring of the saline water is needed to ensure that only saline water can enter the salt farms to ensure the quantity and quality of salts.

Methods: This study aims to present a GSM-based water salinity monitoring system to lessen the frequent and manual monitoring of water salinity. The system is equipped with a solar panel, solar charger control, 12V battery, 12V relay, Arduino Uno, and GSM Module.

Results: The overall rating of 3.32 reflects that the developed system met the design functions; the materials are appropriate and the specifications meet the desired purpose; the system is efficient and consistent with its desired objectives of lessening the manual activities involved in the monitoring of water salinity. As the pH and conductivity sensors read the salinity value, it sends signals to the Arduino Uno; when the salinity level reads 34,000-35,000ppm a signal trigger the GSM Module to send a message to the gate valve. The performance efficiency of the system implied that the reaction of the Arduino Uno in triggering the GSM Module is in real-time as the salinity readings are received.

Conclusion: The real-time reaction of the Arduino Uno to send signals to the GSM Module proved the advantages of using the system and the automatic salinity readings can lessen the frequent and laborious activity in water salinity monitoring.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Salt is a mineral that is broadly spread in almost all continents [1]. As one of the ancient industries, the primitive or traditional method is still the emanating process used by most salt farmers, but the application of technology based methodologies in the salt production is becoming popular [2]. In the Philippines, the used of

ancient traditional and artisanal methods are commonly used relying greatly on solar evaporation.

The province of Occidental Mindoro, facing the abundant source of seawater contributed to 18% of the country's total salt production [3]. As the province is trying to reclaim its reputation in the salt industry, salt farmers are continuously seeking ways on increasing their harvest and ensure year-round production. The use of

technology-based products in salt production has always been one of the objectives of the Tamaraw Salt Producers Cooperative and salt farmers in the province, as studies and researches applied in the salt farms are limited. An attempt to mechanize the production process has been introduced in brine management by speeding up the crystallization process; and placing the Salt washer to purify and clean salts. However, salt farmers and managers desire technologies that could lessen manual labor and shorten processes involved in salt production.

Traditional salt production starts with the intrusion of saltwater in bed farms by manually opening a gate design to control the flow of saltwater. Water entering the salt farms is dependent on the required salinity before it allows to enter the salt beds and such activity requires constant monitoring, visual testing, and on-site visit as salinity changes quickly in time and is greatly influenced by the lateral water movement [4]. Human activities relating to the monitoring of water salinity have been described as tedious, extensive field works, and require a day-round observation.

At present, there are many technologies related to water level monitoring and the use of microcontrollers and sensors has been the favorite area of research. The application of sensors ranging from simple, web-based, and complex system architecture manifested the usefulness of microprocessors and sensors in water quality monitoring and agricultural use. A simple four electrodes conductivity sensor is used for the automatic logging of soil water salinity extracted from the wetting front of the irrigation cycle [5]. The wireless sensor modules: Zigbee communication was used in monitoring the precision of the agricultural system with temperature, soil humidity, light, and pH of the soil as the input parameters [6]. While the capability of the Zigbee wireless sensor network was explored in agriculture in monitoring environmental conditions like weather, soil moisture content, soil temperature, soil fertility, weed detection, water level, monitoring growth of the crop, precision agriculture, automated irrigation facility, and storage of agricultural products [7]. Zigbee wireless electronics was also used in developing remote water salinity based using a fiber-optic U-shaped sensor to strengthen the sensitivity and increase the range [8].

The use of microcontrollers and Internet-of-things provided the easiest means of water quality monitoring and showed applicability in assessing the water salinity of water bodies. A real-time water quality monitoring system using a wireless element capable of providing a warning to farmers was developed in fish farms and improving aquaculture [9]. A WIFI-microcontroller coupled with a parallel capacitive plate was developed and tested and found to increase the coverage range of the water salinity sensor [10]. A digital salinometer was

designed based on electrode sensor ATmega 328 microcontroller programmed using Arduino IDE with salinity value obtained by measuring the voltage of the difference in resistance [11]. An IoT-based real-time management system for monitoring the water in the fishponds was installed with the use of sensors and CCTV; and a food control system to manage the food intake and the water system in the pond [12]. A Modbus TCP/IP communication based on IoT integrated sensors in monitoring the water quality parameters like dissolved oxygen, pH, and water temperature [13]. Further, temperature sensors, pH, and salinity sensors were used in managing the water system of a shrimp farm using the Fuzzy logic processing method [14]. Also, a salinity detecting instrument using processor SOC MCU C8051F040 designed based on ARM7 microprocessors considered the effect of temperature on the conductivity of water was successfully installed to get the real-time data [15]. Further, an Arduino Uno chip coupled with GSM Module 6900a was developed for the monitoring of total dissolved solids as part of the water quality monitoring system [16]. Another microprocessor known as MSP432 connected to LoRA network read data such as water salinity, water level, and temperature was developed and used in the river salinity monitoring; the Raspberry Pi 3B+ was installed to allow the receiving and uploading of data to the cloud server connected to the Internet [17]. The capability of Arduino Uno coupled with Raspberry Pi 3B+ and LoRaWAN IoT protocol was used in monitoring the water quality parameters such as temperature, pH level, turbidity, salinity, and dissolved oxygen has an automatic correction to ensure the growth of the aquatic animal. The sensors used web applications to get information and monitor the parameters acceptable for fish growth [18].

Literature shows the capability of the microcontrollers like Zigbee and IoT –based technology in providing efficient monitoring of water in the fishponds and agricultural areas. Since the study area has a weak internet connection, the use of web-based applications is not feasible; this study aims to present a sensor-based water salinity monitoring system that can automatically open the gate to allow the saline water to enter the salt farms. The combination of the GSM module and Arduino Uno were used to detect the water salinity level in the study area. GSM module and Zigbee were used in monitoring the farm environment data such as soil moisture level, water level, temperature, and humidity by providing the farmer a text message [19]. GSM module coupled with a 16F877 microcontroller was used in establishing a link with the farmer, soil, and crop conditions by automatically sending text messages to the farm owners [20]. The use of the GSM module was also applied in monitoring the water quality parameters for a prawn farm by sending text messages on the status of pH,

temperature, and dissolved oxygen [21]. Further, the microprocessor called Arduino Mega 2560 was used as a command controller in monitoring the pH, water turbidity, and water temperature and sending signals to the GSM module to send text messages in assessing the quality of drinking water [22]. GSM module was also applied in monitoring the water quality parameter in India using the combined capabilities of conductivity sensors, pH sensors, and temperature sensors. The detected water quality parameters from the Arduino Uno send signals to the GSM module and send water quality status to the users [23]. Further, the ability of the Arduino Uno and sensors effectively monitored the pH, temperature, turbidity, and electrical conductivity in identifying possible water contamination and efficiently send text messages and utilized a buzzer to alert that water is not safe for drinking.

To help the salt farmers of Magsaysay, Occidental Mindoro, Philippines, lessen the burden of manual work, the researchers developed a water salinity monitoring system using the combined capability of the microprocessor and water salinity sensor. The detection of water salinity level and automatic manipulation of the gate valve is the focus of this study. The study aims to detect the amount of salt present in the seawater before it allows it to flow into the salt bed. This study aims to present a sensor-based water salinity monitoring system that can automatically open the gate and allow the saline water to enter the salt farms. The researchers tried to incorporate the use of solar panel to maximize the capability of the sun to run the system at all times as many of existing studies are electrically powered. The system relies on the use of text messaging as the medium for monitoring the salinity of the water. This makes the water salinity monitoring system novels as most of the existing monitoring system lies on the AC source and internet-based system. Specifically, the study aims to test and evaluate the system in terms of functionality, durability, and efficiency; to determine the performance efficiency rating of the developed system through salinity reading on the sensors, reaction time of the installed microcontrollers-Arduino Uno and precision of the water gate management system in salt production.

Materials and Methods

A. System Components

The design for the water salinity monitoring system is consists of a 100W solar panel acting as the power source; a solar charger controller for the monitoring of solar energy generation; a 12V battery for the storage of the energy harvested; a DC motor, relay, Arduino Uno, and the GSM Module for the automation process. Fig. 1 show the components of the GSM based water salinity monitoring for salt farm gate operations.

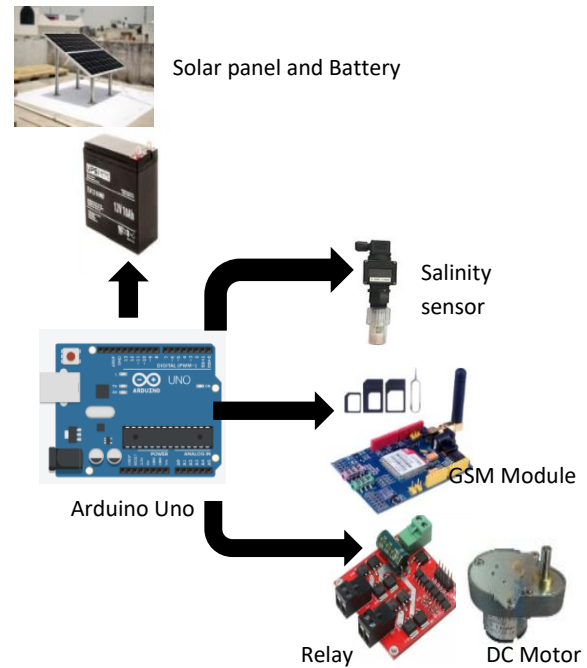


Fig. 1: Components of the GSM based water salinity monitoring system.

Solar Panel and Battery

The developed water salinity monitoring system used 100W solar panel for the energy generation and storage. Specifically, a cell-type monocrystalline solar panel with 36 cells (3 x 12) and cell efficiency of 22.0% was used in the study. The solar panel is installed facing the nearby sea water source to fully capture the solar energy. While the 12V battery was coupled with solar panel to ensure that the system will work at all times thru the energy stored in the battery.

Arduino Uno

Arduino Uno is the microcontroller used in the study and acts as the brain of the system. An ATmega328P Arduino Uno with operating voltage of 5V-20mA (per I/O pin) and 14 digital I/O pins (PWM output) was used as the command controller in the system. The board is connected to the computer where the programming is initialized and adjusted. It sends signal to the relay and GSM Module to perform the default functions.

Salinity Sensor

The salinity sensor ranging from 0 to 50,000 ppm was used in the system to measure the salt content of the sea water. The sensor was initially calibrated and design to have a response time of 90% of full-scale readings in 10 seconds. The sensor is submerged in the seawater for the automatic detection of the saline level. The salinity sensor was programmed to accept the standard salinity value of 34,000-35,000PPM to be considered as the appropriate/good level for salt ponds.

GSM Module

The GSM Module enables the devices to send messages about the default information programmed in the microcontroller. It is the component that links the mobile devices to the system. Specifically, the SIM900 and 1800MHz -dual band GSM (phase 2/2+)/GPRS (multi-slot class 10) was installed in the system. The smallest GSM module (24mm x 24mm x 3mm) was used and can control via AT commands. The module has high compatibility rating with the selected Arduino Uno and sensor used in the system. The module has the keypad and display interface that allows the salt farmers to receive and read the message even in remote areas. The applicability of the GSM module in the study area was the first factor considered in the study since the internet is not viable in the study area.

Relay and DC Motor

The relay coupled with DC Motor is used as an electrically operated switching device that controls the opening of the gate in the salt farm. A 5V reed type standard relay is design to control the load in monitoring the salinity level. While a small DC Motor (8 x 35mm) with speed ranging from 5,000 -14,000rpm and 0.36-160mNm motor torque was installed in the system. When the detected salinity value conforms to the default level, it controls the gate facing the sea water to automatically open and allow the intrusion of water.

B. Fabrication of the Prototype

After procuring all the necessary materials, all parts are connected and assembled based on the design of the prototype. The solar panel, battery and solar charger controller were initially tested to determine charging and discharging time; then the sensors, Arduino Uno, relay and DC Motor were connected and programmed based on required salinity readings; while the GSM module was last to install to ensure that all other system components connected to the module is functioning well. After the components are installed, initial testing was done to ensure the functionality and efficiency of the system. After analyzing the results of the evaluation, modification and adjustments were considered to finalize the system.

C. System Work Flow Diagram

Fig. 2 presented the system workflow diagram of the system. The operations started with the harvesting of solar energy. The 100W solar panel harvests the energy and converted it to electrical energy and is stored in the connected 12V battery. The charger controller connected to the battery ensures the power requirement of the system. Since the sensors are exposed to sea water, it was programmed to read the salinity levels every hour. It will send signals to the GSM Module and interprets the data. When the default salinity level is read by the system, it

will be interpreted by the Arduino Uno and send message to the GSM Module and salt farmer. The other components will work based on the program and will send signal to the other components connected to the gate of salt farms. This enable the automatic openings of the gate and allow the saline water to enter the salt beds.

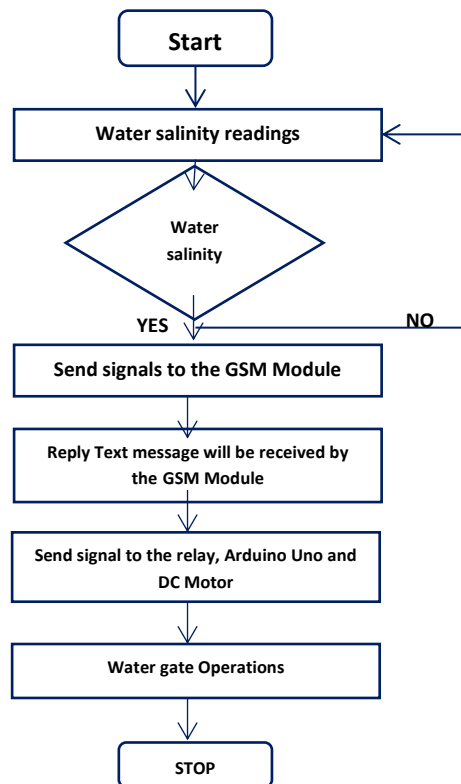


Fig. 2: System work flow diagram.

The operations of the GSM-based water salinity monitoring system in the salt farm start with the readings of the salinity level. The sensors dipped in the seawater monitors the saline level basis, when the required saline level is reached, the sensors send the readings and the data is analyzed by the Arduino Uno and GSM Module. Then, a text message is received by the mobile phone connected in the system indicating the saline level. The system provides an opportunity for the salt farmers to lessen the manual inspection of the salinity level in the gates of the salt ponds. The automatic readings of the salinity level of the sensors, Arduino Uno and GSM Module allowed the real time opening of the gates.

Results and Discussion

Product Description

The GSM-based water salinity monitoring system is designed to lessen the tedious, extensive field work and the regular sampling of saltwater. The system was configured and programmed considering the required salinity level, the volume required to fill the salt beds, and the prompt response of the module to the commands.

A 12V battery is used in the system that stored the energy harvested from the solar sheets while the DC Motor is designed to convert the direct current into mechanical energy. Attached to the battery is the microcontroller: Arduino Uno which serves as the brain that sends the command to all the components attached. The water salinity sensors are programmed to the microcontroller and read the salinity levels of the water.

The GSM Module is programmed to the microcontroller to receive, analyze, and respond to the commands. While the 12V channel relay is programmatically controlled to switch on/off the devices attached.

Fig. 3 illustrates the actual view of the GSM-based water salinity monitoring system facing the saline water source.

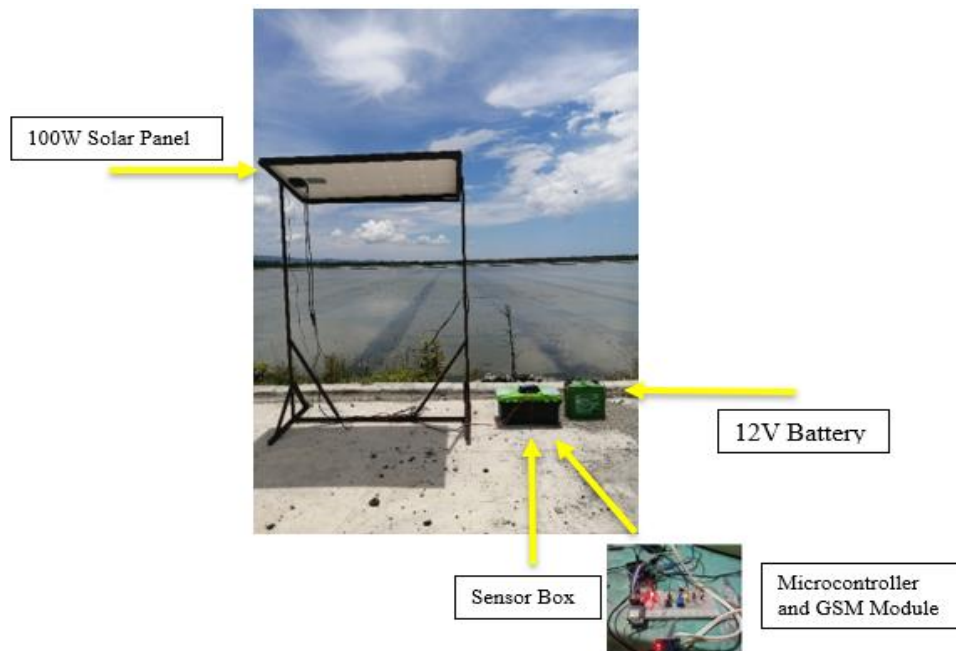


Fig. 3: Actual view of the GSM based water salinity monitoring system.

Testing Results

After the initial testing and checking of the programs, the system is installed in a salt farm in Magsaysay, Occidental Mindoro allowing the selected salt farmers to test and evaluate its efficiency. The system was energized and a demonstration of the function was conducted before the evaluation process.

The functionality of the system which refers to the usefulness and observance of the desired output obtained a mean score of 3.45 interpreted as "Very Good". This means that the system is useful and design objectives are achieved. Durability refers to the strength of the system in terms of materials and design obtained a mean score of 2.86, interpreted as "Good" reflecting that the material and design need some modification as some parts are exposed and should have a protected mechanism to ensure the longevity of the usage as recommended by the evaluators.

Lastly, the efficiency of the products in terms of conformance to the desired output and the absence of human work was rated 3.67 interpreted as "Very Good", describing that the product achieved the desired

objective of lessening the manual operations as the salinity sensors read the salinity level lessening the monitoring process and the on-site observation of salt farmers.

Table 1: Grand mean score of evaluation

Content	Mean	Interpretation
1. Functionality	3.45	Very Good
2. Durability	2.86	Good
3. Efficiency	3.67	Very Good
Grand Mean	3.32	Very Good

Legend: 0.00-1.00-Poor; 1.01-2.00 Fair; 2.01-3.00-Good; 3.01-4.00 Very Good; 4.01-5.00 Excellent

The grand mean of 3.32 interpreted as "Very Good" reflects that the system met the design functions; the system design and materials are appropriate and specifications meet the desired functions though additional protective mechanism must be provided to ensure the safety of the devices and parts; the system is efficient and consistent with its desired objectives of

lessening the manual activities involved in the monitoring of water salinity.

Performance Efficiency of the Product

After testing the functionality, durability, and efficiency of the developed system by the selected salt farmers, testing of the system components followed to determine the precision of salinity sensors, solar sheets, and battery. The salinity sensors were programmed to accept 34,000-35,000PPM as the standard salinity level, once the required level is obtained; it automatically sends a message to the mobile phone through the GSM module. The system was initially installed and observed from 6:00 am to 4:00 pm. Table 2 shows the observed salinity level and the response of the system.

Table 2: Salinity monitoring and reaction time in GSM module

Voltage reading	Salinity (PPM)	Time	Send message to the GSM Module	Reaction time (milliseconds)
13.4 V	32,000	6:00	No	220
13.3 V	33,000	7:00	No	215
13.0 V	33,432	8:00	No	220
13.0 V	33, 445	9:00	No	230
12.5 V	34, 995	10:00	Yes	215
12.4 V	35,112	11:00	Yes	225
12.2 V	33, 435	12:00	No	225
12.1 V	33,128	1:00	No	230
12.3 V	33,000	2:00	No	240
11.9 V	34,123	3:00	Yes	225
11.8 V	24,500	4:00	No	235

The table shows the variation of salinity with time, voltage readings, and the recorded reaction time of the GSM Module. It can be observed that voltage readings drop with time as it is being utilized by the devices attached to the water salinity monitoring system. However, charging of the battery is done every morning before the energization of the system to ensure continuous power. The table also reflected the salinity readings vary with time with maximum salinity measured during noon and decreasing when the sunsets. The salinity sensors upon reading the default salinity value send the message to the GSM module. Three (3) recorded salinity values reach 34,000-35,000ppm and triggered the Arduino Uno to send a signal to the GSM Module. However, it can also observe that below the default salinity value recorded reaction time is still present as the Arduino Uno receives the salinity readings but no

message will be received by the GSM Module.

The reaction time in the Arduino Uno further implies a real-time response as indicated in milliseconds. Moreover, once the message is received in the Arduino Uno, the opening of the gates can be triggered. Results of the performance efficiency suggested that the GSM-based water salinity monitoring system can be a great help in salt production as the tedious monitoring of the salinity is lessened.

Conclusion

The developed GSM-based waster salinity monitoring system is designed to lessen the human activities entailed in ensuring the required salinity in salt farms. The materials used in the developed system suit the design purpose and achieve functionality, durability, and efficiency. The overall rating of 3.32 reflects that the system met the design functions; the design and materials are appropriate and the specifications meet the desired purpose; the system is efficient and consistent with its desired objectives of lessening the manual activities involved in the monitoring of water salinity. The performance efficiency of the system implied that the reaction of the Arduino Uno in triggering the GSM Module is in real-time as the salinity readings are received. The default salinity value of 34,000-35,000ppm triggered the Arduino Uno to send a signal to GSM Module and send a message to the gate operations. The solar panel and the battery installed in the system ensure the continuity of usage and power up the devices, circuit boards, GSM Module and Arduino Uno to operate on its design purpose. The technology developed and tested by selected salt farmers can eventually provide a system for gate operations.

Author Contributions

M. Enriquez, A. Abella designed and constructed the system components. A. Abella, tested the developed system and gather the data. M. Enriquez analyzed the data, interpreted the results and wrote the manuscript.

Acknowledgment

The authors would like to thank the Tamaraw Salt Producers Cooperative for the data provided and for allowing the conduct of testing in their salt farms; and the salty farmers who willingly joined the demonstrations and the evaluation of the system.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>GSM</i>	Global System for Mobile communications
<i>IoT</i>	Internet of Things
<i>IDE</i>	Spectral Angle Mapper
<i>TCP/IP</i>	Transmission Control Protocol/Internet Protocol
<i>pH</i>	Potential of Hydrogen
<i>ppm</i>	Parts per minute
<i>WIFI</i>	Wireless fidelity
<i>CCTV</i>	Closed-circuit television
<i>MCU</i>	Microcontroller
<i>SOC</i>	System on chip

References

- [1] M. Affam, D. N. Asamoah, "Economic potential of salt mining in ghana towards the oil find," *Res. J. Environ. Earth Sci.*, 3(5): 448-456, 2011.
- [2] M. M. Abu-Khader, "Viable engineering options to enhance the NaCl quality from the Dead Sea in Jordan," *J. Cleaner Prod.*, 14: 80-86, 2006.
- [3] Business Mirror, Occidental Mindoro boosting salt industry through new technology, 2017.
- [4] J. Arriola-Morales, J. Batlle-Sales, M. A. Valera, G. Linares, O. Acevedo, "Spatial variability analysis of soil salinity and alkalinity in an endoergic volcanic watershed," *Int. J. Ecol. Dev.* 14(F09): 1–17, 2009.
- [5] A. J. Skinner, M. F. Lambert, "An automatic soil pore-water salinity sensor based on a wetting-front detector," *IEEE Sens. J.* 11(1): 245-254, 2011.
- [6] N. Fahmi, S. Huda, E. Prayitno, M. U. H. Al Raysid, M. C. Roziqin, M. U. Pamenang, "A prototype of monitoring precision agriculture system based on WSN," in *Proc. International Seminar on Intelligent Technology and Its Applications (ISITIA)*: 323-328, 2017.
- [7] S. G. Kannan, G. Thilagavathi, "Online farming based on embedded systems and wireless sensor networks," in *Proc. International Conference on Computation of Power, Energy, Information and Communication (TCCPEIC)*: 71-74, 2013.
- [8] D. Z. Stupar, J. S. Bajić, A. V. Joža, B. M. Dakić, M. P. Slankamenac, M. B. Živanov, E. Cibula, "Remote monitoring of water salinity by using side-polished fiber-optic U-shaped sensor," in *Proc. 15th International Power Electronics and Motion Control Conference (EPE/PEMC), LS4c-4*, 2012.
- [9] Z. Shareef, S. R. N. Reddy, "Design and wireless sensor Network Analysis of Water Quality Monitoring System for Aquaculture," in *Proc. the International Conference on Computing Methodologies and Communication, Erode, India*, 405–408, 2019.
- [10] S. Suryono, S. P. Putro, W. Widowati, S. Sunarno, "A capacitive model of water salinity wireless sensor system based on wifi-microcontroller," in *Proc. 6th International Conference on Information and Communication Technology (ICICT)*: 211-215, 2018.
- [11] M. S. Wibaya, K. Putra, N. Wendri, "Design of salinity content (salinometer) tools digital based on microcontroller ATmega328," *Int. J. Sci. Res. (ISR)*, 8(1): 1678-1680, 2019.
- [12] F. E. Idachaba, J. O. Oloweleni, A. E. Ibhaze, O. O. Oni, "IoT Enabled Real-Time Fishpond Management System," Presented at the World Congress on Engineering and Computer Science Vol. I, San Francisco, USA, 2017.
- [13] J. Y. Lin, H. L. Tsai, W. H. Lyu, "An Integrated Wireless Multi-Sensor System for Monitoring the Water Quality of Aquaculture," *Sensors*, 21(24): 8179, 2021.
- [14] V. A. Wardhany, H. Yuliandoko, M. U. Subono, A. R. Harun, I. G. Puja Astawa, "Fuzzy logic based control system temperature, pH and water salinity on vanammei shrimp ponds," in *Proc. International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*: 145-149, 2017.
- [15] X. Liu, X. Gong, Y. Liu, "Research on salinity detecting based on embedded CAN-ethernet gateway," in *Proc. International Conference on Measuring Technology and Mechatronics Automation*: 257-260, 2009.
- [16] M. Asif, M. Abdullah, M. Arif, J. Nouman, M. Atteq, M. Saad, M. Ghulam, "GSM based advanced water salinity and tds monitoring system," *Global Sci. J.*, 8(8): 1792-1816, 2020.
- [17] T. P. Truong, D. T. Nguyen, T. Huynh, "Design and implementation of an iot-based river water salinity monitoring system using MSP432," *J. Phys. Conf. Ser.*, 1878(1): 12-23, 2021.
- [18] L. K. S. Tolentino, C. P. De Pedro, J. D. Icamina, J. B. Navarro, L. J. D. Salvacion, G. C. D. Sobrevilla, G. A. M. Madrigal, "Development of an IoT-based intensive aquaculture monitoring system with automatic water correction," *Int. J. Comput. Digital Syst*, 10: 1355-1365, 2020.
- [19] M. Koul, P. B. Salunkhe, M. SinghGill, "ZIGBEE and GSM based smart agriculture system," *Int. J. Sci. Res. Dev.*, 3(3): 2321-0613, 2015.
- [20] R. Subalakshmi, A. Anu Amal, S. Arthireena, "GSM based automated irrigation using sensors," *Int. J. Trend Res. Dev.*, 6(11), 2106-2108, 2017.
- [21] N. S. Haron, M. K. Mahamad, I. A. Aziz, M. Mehat, "Remote water quality monitoring system using wireless sensors," in *Proc. the 8th WSEAS International Conference on Electronics, Hardware, Wireless and Optical Communications (EHAC'09)*: 148-154, 2009.
- [22] A. Sowjanya, S. Sai Chandu, D. Lokesh, K. Shiva Shankar Reddy, "Arduino based water quality monitoring and notification system using GSM," *Int. J. Creative Res. Thoughts*, 6(2): 218-221, 2018.
- [23] S. Gokulanathan, P. Manivasagam, N. Prabu, T. Venkatesh, "A GSM based water quality monitoring system using Arduino," *Shanlax Int. J. Arts Sci. Humanit.*, 6(4): 22-26, 2019.

Biographies



Michelle D. Enriquez is presently designated as the Dean of the College of Engineering of Occidental Mindoro State College, Philippines. She is currently finishing her degree in PhD Civil Engineering at De La Salle University, Philippines. Her research interests include water resources engineering; development researches in salt processes; water quality monitoring system and policy directions for

water management.

- Email: michelle_enriquez_d@dlsu.edu.ph
- ORCID: 0000-0002-1618-6335
- Web of Science Researcher ID: ABE-3230-2021.
- Scopus Author ID: NA
- Homepage: NA



Adrian Paul N. Abella is a faculty of College of engineering of occidental Mindoro State College, Philippines. He is currently pursuing his Master of Science in electronics Communication Engineering at Mapua University, Philippines. His research interest includes application and development of system using electronic and electrical devices, sensors, and microcontrollers.

- Email: adrianpaulabella@gmail.com
- ORCID: [0000-0001-9822-2143](https://orcid.org/0000-0001-9822-2143)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

M. Enriquez, A. Abella, "GSM based water salinity monitoring system for water gate management in salt farms," J. Electr. Comput. Eng. Innovations, 11(2): 383-390, 2023.

DOI: [10.22061/jecei.2023.9564.633](https://doi.org/10.22061/jecei.2023.9564.633)

URL: https://jecei.sru.ac.ir/article_1844.html





Research Paper

1WQC Pattern Scheduling to Minimize the Number of Physical Qubits

E. Nikahd^{1,*}, M. Houshmand², M. Houshmand³

¹Computer Engineering Department, Shahid Rajaee Teacher Training University, Tehran, Iran.

²Department of Computer Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran.

³Department of Electrical Engineering, Imam Reza International University, Mashhad, Iran.

Article Info

Article History:

Received 24 November 2022

Reviewed 12 January 2023

Revised 26 February 2023

Accepted 11 March 2023

Keywords:

One-way quantum computing model

Measurement-based quantum computation

Scheduling

Integer-linear programming

Abstract

Background and Objectives: One of the quantum computing models without a direct classical counterpart is one-way quantum computing (1WQC). The computations are represented by measurement patterns in this model. One of the main downsides of the 1WQC model is the much larger number of qubits in a measurement pattern, compared to its equivalent in the circuit model. Therefore, proposing a method for optimally using the physical qubits to implement a measurement pattern is of interest.

Methods: In a measurement pattern, despite a large number of qubits, the measured qubit is not needed after each measurement and can be used as another logical qubit. In this study, by using this feature and presenting an integer linear programming (ILP) model to change the ordering of a standard measurement pattern actions, the number of physical qubits required to implement that measurement pattern is minimized.

Results: In the proposed method, compared to the scheduling based on the standard pattern, the number of required physical qubits on benchmark circuits is reduced by 56.7% on average. Although the proposed method produces the optimal solution, one of the most important limitations of that and ILP-based methods, in general, is their high execution time and memory requirements, which grow exponentially with the increase of the problem size.

Conclusions: In this study, an ILP model is proposed to minimize the number of physical qubits used to realize a measurement pattern by efficiently scheduling the operations and reusing the physical qubits. Due to its exponential complexity, the proposed method cannot be used for large measurement patterns whose solution can be conspired as future works.

*Corresponding Author's Email
Address: nikahd@sru.ac.ir

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Quantum computing is a branch of information processing, which is a combination of three sciences, namely, computer science, information theory, and quantum physics [1]. This science is of great interest due to reaching the end of CMOS technology advancement, its high computing power, and also the significant role it plays in the secure transmission of information [1]-[3].

The most famous model of quantum computing is the

quantum circuit model which is an analogy to the common classical computation model composed of a network of logic gates [1], [4]. One-way quantum computing model (1WQC), first proposed by Raussendorf and Briegel in 2001 [5], is a practically and conceptually different alternative model [5], [6]. This model utilizes unique features of quantum mechanics such as entanglement and measurement and hence it has no classical analogue. In this model, qubits are initialized in

a special highly entangled resource state (namely a cluster or graph state), and the universal quantum computation is driven by performing a sequence of single-qubit measurements in certain basis and post-measurement corrections. Calculations in 1WQC are shown in the form of measurement patterns consisting of four types of instructions: qubits preparation (N), entanglement (E), measurement (M) and correction (C) [7], [8].

1WQC is one of the measurement-based quantum computing models (MBQC) which is promising for physical implementation and has attracted the attention of researchers [9]-[11]. Despite the advantages that this model has [12], [13], one of the main down sides of that is the much larger number of qubits in a measurement pattern, compared to its equivalent in the circuit model. On the other hand, one of the limitations of physical construction, especially in ion-trap technology, is the number of available physical qubits. Therefore, 1WQC will be hard to realize due to its large number of required qubits [9].

Despite the large number of qubits in a measurement pattern, there is no need to construct the whole graph state at the beginning and it is possible to extend it on the fly by reordering the measurement pattern [14]-[16]. Furthermore, after measuring a logical qubit, there is no more required action on it and it can be removed from the computation space. The measured physical qubit can also be reused as another logical qubit to extend the graph state on the fly if there is not a limitation in underlying technology. As a result, by a proper reordering the actions of a pattern, it is possible to minimize the number of required physical qubits to realize a 1WQC measurement pattern. Now, a question is remained to answer: what is the best order of a measurement pattern that minimizes the number of necessary physical qubits and which physical qubit is allocated to a logical qubit? Focusing on this issue, in this paper an integer-linear programming (ILP) model is proposed to schedule a measurement pattern targeting to minimize the number of required physical qubits.

The rest of this study is organized as follows. Section 2 covers basic concepts related to the 1WQC model. Related work is reviewed in Section 3. The proposed approach is described in Section 4. In Section 5, the proposed method is evaluated by some measurement patterns and finally Section 6 concludes the paper.

Background

Computations in the 1WQC model are shown by a measurement pattern which is defined by the set $P = (V, I, O, A)$ [7], [8]. V is the set of all qubits, $I \subseteq V$ is the set of input qubits, $O \subseteq V$ is the set of output qubits, and A is the set of actions that act on V . The pattern is written as a sequence of actions that includes four different types: preparing qubits (N), entanglement (E),

measurement (M) and correction (C) that are applied from right to left as determined in the following.

- **Qubit Preparation N_u** : prepares a qubit u in the state $|+\rangle = \frac{1}{\sqrt{2}}(|0\rangle + |1\rangle)$. Normally this action is applied to all of the non-input qubits.
- **Entanglement action $E(u,v)$** : entangles the qubits u and v by applying CZ gate on them. To visualize a pattern, qubits can be shown by vertices of a graph, namely entanglement graph, where the entanglement between the qubits is represented by the edges of the graph.
- **Single-qubit measurement M_u^α** : measures the qubit u in the orthonormal basis of:

$$|\pm\alpha\rangle = \frac{1}{\sqrt{2}}(|0\rangle + e^{i\alpha}|1\rangle) \quad (1)$$

where, $\alpha \in [0, 2\pi]$ is the measurement angle. Normally all of the non-output qubits will be measured. The measurement result applied to a qubit u is denoted by $s_u \in \mathbb{Z}_2$. If u collapses into the $|+\alpha\rangle$ after the measurement, then $s_u = 1$ and otherwise if it collapses into the $|-\alpha\rangle$, then $s_u = 0$. The measurement outcomes can be summed module 2 to generate a signal. In general, a measurement angle may depend on the other ones through two signals s and t as:

$$t[M_u^\alpha]^s = M_u^{(-1)^s\alpha+t\pi} \quad (2)$$

A measurement that depends on the signals s and t can be done if all the measurement results appeared in s and t are known. That means all those measurements must be done beforehand.

- **Pauli correction X_u^s and Z_u^s** : apply the Pauli X and Z gates on the qubit u , respectively, if $s=1$ and do nothing if $s=0$.

A pattern is called a standard pattern, if the order of actions appeared in it is preparation, entanglement, measurement, and finally corrections, respectively [7]. In a 1WQC pattern, a qubit can be removed from the computation space only after measuring it [15], [16]. Therefore, if all preparation and entanglement operations are performed first, as in a standard pattern, it is necessary to allocate a physical qubit for each qubit of the pattern. This means that the number of physical qubits required to implement a standard pattern will be equal to the number of qubits of that pattern.

Definition 1 [7]: An open graph (G, I, O) has flow if and only if there exists a map $f: O^c \rightarrow I^c$ and a strict partial order $<_f$ over V

such that all of the following conditions hold for all $i \in O^c$

- $i <_f f(i)$
- if $j \in N(f(i))$, then $j = i$ or $i <_f j$, where $N(v)$ contains adjacent vertices of v in G

- $i \in N(f(i))$

In this case, $(f, <_f)$ is called a flow on (G, I, O) .

Definition 2 [8]: An open graph (G, I, O) has generalized flow (gflow) if and only if there exists a map $g: O^c \rightarrow P^{I^c}$ (the set of all subsets of vertices in I^c) and a strict partial order $<_g$ over V such that all of the following conditions hold for all $i \in O^c$.

- if $j \in g(i)$ then $i <_g j$,
- if $j \in \text{Odd}(g(i))$, then $j = i$ or $i <_g j$, where $\text{Odd}(K) = \{k \mid |N(k) \cap K| = 1 \bmod 2\}$,
- $i \in \text{Odd}(g(i))$.

In this case, $(g, <_g)$ is called a gflow on (G, I, O) .

Related Work

The unique features of the 1WQC model has drawn the researchers' attention in many studies after its first proposal in 2001. A number of studies have focused on the fast simulate of the 1WQC model on the classic computers [16], [17].

One of the important applications of the 1WQC model is blind quantum computation [18]-[20]. Blind quantum computation allows a client with limited quantum capabilities to delegate his computational problem to a remote quantum server such that the client's input, output, and algorithm are kept private from the server.

Most of physical design and scheduling work done in quantum computing has been focused on the quantum circuit model [21]-[28]. While, there is only a few researches focused on the physical design of 1WQC [29]-[32]. The studies done in the 1WQC model assume all preparations and entanglements are first done and after that computation is pursued by only single-qubit measurements and post-measurement corrections, as in the standard pattern. For example, [33] proposed a design flow to directly map a 1WQC pattern to a 2D nearest-neighbor architecture, without trying to reduce the number of physical qubits.

The most related work to our study is the work done in [15]. In that study, the minimal number of physical qubits that must be present in a system to directly implement a given measurement pattern has theoretically been proven. It has been shown that to realize a measurement pattern $P = (V, I, O, A)$ with flow [7], the minimum number of physical qubits is $\min(|O| + 1, |V|)$, while for measurement patterns with only gflow¹ [8], the number of needed qubits may be as high as $|V| - 2$. However, that approach does not provide a practical way to reach this minimal number of the physical qubits which is the main concern of this study.

Proposed Approach

In a standard pattern, all preparation operations followed by entanglement are performed first. Therefore, the number of physical qubits needed to realize a standard pattern is equal to the number of total qubits of the pattern, i.e., $|V|$. However, it is possible to reorder the operations of the pattern to minimize the required number of physical qubits [15]. Indeed, we can extend the graph state on the fly by reusing a physical qubit after measuring it as another logical qubit. The problem of finding the best order of operations that minimizes the number of required physical qubits is the subject of this paper. To do so, an ILP model is proposed to schedule a measurement pattern in such a way that it minimizes the needed physical qubits by maximizing qubit reusing. Our approach works for all of the patterns with flow or only gflow.

Theorem 1 [14]: In a measurement pattern, a qubit u can be measured if its dependencies to the other measurement have been resolved and also its entanglements with its neighbor qubits have been applied.

Therefore, based on this theorem, one can select a logical qubit u from the list of qubits with resolved dependencies for measurement. Then, one can allocate physical qubits to it and its neighbor qubits and perform the entanglements between them. After that, the qubit u is measured. Finally, the allocated physical qubit to u is released and can be reused as another logical qubit.

To illustrate the method, an example is provided. Fig. 1 shows the entanglement graph of the SWAP gate. Each node represents a qubit and each edge is an entanglement operation between the corresponding qubits. $\{q1, q2\}$ and $\{q6, q8\}$ are the input and output qubits, respectively. We suppose that the input qubits already exist i.e., physical qubits with proper states have been allocated to them beforehand or may feed into the circuit from outside. All of the non-output qubits must be measured and, in this case, there is no dependency between them. Therefore, in the first step all non-output qubits are candidates to be chosen for measurement. Intuitively, input qubits are the best choices for the measurement in the first step, as they already exist.

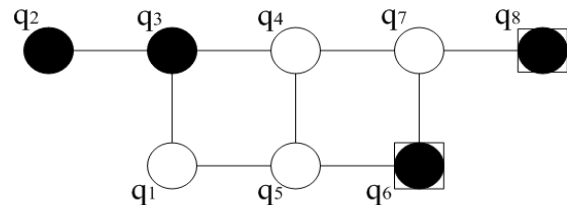


Fig. 1: Graph state of SWAP gate.

¹ Generalized flow

Table 1 shows the best order of measurements that reduces the number of physical qubits from 8 to 3 in comparison with the standard pattern. In this table, logical and physical qubits are denoted by lq and pq respectively and $(lq:pq)$ means that a physical qubit pq is assigned to a logical qubit lq .

0) Physical qubits are allocated to input qubits.

1) lq_2 is selected for measurement. Therefore, the entanglement between lq_2 and its neighbors, i.e. lq_3 , must be performed before measuring lq_2 . As there is no physical qubit allocated to lq_3 and there is no free physical qubit, we need a new one, i.e. pq_3 to be allocated to lq_3 . Now we are ready to perform $E(lq_2, lq_3)$ and then measure lq_2 . After measuring

lq_2 , pq_2 is released and can be reused as another logical qubit.

- 2) lq_3 is chosen for measurement. So, $E(lq_3, lq_1)$ and $E(lq_3, lq_4)$ must be done first. To perform $E(lq_3, lq_4)$ it is needed to allocate a physical qubit to lq_4 . To do so, pq_2 (that was released in the previous step) is used, i.e. $(lq_4:pq_2)$. After that, lq_3 is measured and its corresponding physical qubit, i.e. pq_3 , is released.
- 3) to 6) This process will be continued until all of the measurements are performed using only three physical qubits pq_1 , pq_2 and pq_3 .
- 7) Finally, the output qubits are lq_6 and lq_8 corresponding to pq_2 and pq_3 , respectively.

Table 1: an example of optimal solution for scheduling of SWAP pattern

step	Measurement order	Qubit allocation (Logical Qubit: Physical Qubit)
0	-	{(lq1:pq1), (lq2:pq2), (lq3:-), (lq4:-), (lq5:-), (lq6:-), (lq7:-), (lq8:-)}
1	lq2	{(lq1:pq1), (;q2:pq2), (lq3:pq3), (lq4:-), (lq5:-), (lq6:-), (lq7:-), (lq8:-)}
2	lq3	{(lq1:pq1), (lq3:pq3), (lq4:pq2), (lq5:-), (lq6:-), (lq7:-), (lq8:-)}
3	lq1	{(lq1:pq1), (lq4:pq2), (lq5:pq3), (lq6:-), (lq7:-), (lq8:-)}
4	lq4	{(lq4:pq2), (lq5:pq3), (lq6:-), (lq7:1), (lq8:-)}
5	lq5	{(lq5:pq3), (lq6:pq2), (lq7:pq1), (lq8:-)}
6	lq7	{(lq6:pq2), (lq7:pq1), (lq8:pq3)}
7	-	{(lq6:pq2), (lq8:pq3)}

Note that there is no dependent measurement in the SWAP pattern. In general case with dependent measurement, a qubit can be selected for measurement only if its dependencies have been resolved.

A. ILP Model

In this section, an ILP model is proposed to schedule a measurement pattern $P = (V, I, O, A)$ targeting to minimize the number of required physical qubits to realize that pattern.

I) Parameters

$lq \in \mathbb{N}$: an index to identify a logical qubit

$pq \in \mathbb{N}$: an index to identify a physical qubit

$N \in \mathbb{N}$: the number of non-output qubits of the pattern P , i.e. $|V| - |O|$

$E(lq)$: the set of neighbor qubits of lq along with lq itself

$D(lq)$: the set of lq dependencies, i.e. the set of qubits on which the measurement of lq depends

II) Variables

$TA(lq) \in \mathbb{N}$: the time (step) that a physical qubit is allocated to the logical qubit lq

$TM(lq) \in \mathbb{N}$: the time (step) of measuring the logical qubit lq

$u(pq) \in \mathbb{Z}_2$: a binary variable to determine whether the physical qubit pq is used or not, defined as (3):

$$u(pq) = \begin{cases} 1 & \text{if physical qubit } pq \text{ is used} \\ 0 & \text{o.w.} \end{cases} \quad (3)$$

$x(lq, t) \in \mathbb{Z}_2$: a binary variable to determine whether qubit lq is measured at time t or not, shown in (4):

$$x(lq, t) = \begin{cases} 1 & \text{if } lq \text{ is measured in time } t \\ 0 & \text{o.w.} \end{cases} \quad (4)$$

$y(lq, pq) \in \mathbb{Z}_2$: a binary variable to determine whether the physical qubit pq is assigned to the logical qubit lq or not, provided by (5):

$$y(lq, pq) = \begin{cases} 1 & \text{if } pq \text{ is allocated to } lq \\ 0 & \text{o.w.} \end{cases} \quad (5)$$

$z(lq, t) \in \mathbb{Z}_2$: a binary variable to determine whether at time t a physical qubit is assigned to the logical qubit lq but has not yet been measured or not:

$$z(lq, t) = \begin{cases} 1 & TA(lq) \leq t \leq TM(lq) \\ 0 & \text{o.w.} \end{cases} \quad (6)$$

To set this variable, two auxiliary variables are defined as follows, where $z(lq, t)$ will be equal to their logical AND operation:

$z1(lq, t) \in \mathbb{Z}_2$: a binary variable to determine whether the logical qubit lq is measured after time t or not, defined as (7):

$$z1(lq, t) = \begin{cases} 1 & t \leq TM(lq) \\ 0 & o.w. \end{cases} \quad (7)$$

$z2(lq, t) \in \mathbb{Z}_2$: a binary variable to determine whether a physical qubit as allocated to the logical qubit lq before time t or not, shown in (8):

$$z2(lq, t) = \begin{cases} 1 & TA(lq) \leq t \\ 0 & o.w. \end{cases} \quad (8)$$

III) Objective Function

The objective function is minimizing the number of used physical qubits, as provided by (9):

$$\min \sum_{pq} u(pq) \quad (9)$$

IV) Constraints

- 1) If a physical qubit pq is assigned to any logical qubit, that physical qubit must exist:

$$y(lq, pq) \leq u(pq) \quad \forall lq, \forall pq \quad (10)$$

- 2) Each logical qubit is selected and measured only once:

$$\sum_{t=1}^N x(lq, t) = 1 \quad \forall lq \setminus O \quad (11)$$

- 3) Only one logical qubit can be selected at any time for measurement:

$$\sum_{lq} x(lq, t) = 1 \quad \forall t \quad (12)$$

- 4) Exactly one physical qubit must be allocated to each logical qubit:

$$\sum_{pq} y(lq, pq) = 1 \quad \forall lq \quad (13)$$

- 5) Calculating the measurement time of a logical qubit:

$$TM(lq) = \sum_{t=1}^N t * x(lq, t) \quad (14)$$

- 6) If a logical qubit lq is measured after time t , $z1(lq, t)$ must be set:

$$TM(lq) - t < z1(lq, t) * BigM \quad (15)$$

- 7) If a physical qubit is assigned to the logical qubit lq before time t , $z2(lq, t)$ must be set:

$$t - TA(lq) < z2(lq, t) * BigM \quad (16)$$

- 8) Before measuring a logical qubit lq , physical qubits must be assigned to it and its neighbors:

$$TM(lq) \geq TA(lq') \quad \forall lq \text{ and } \forall lq' \in E(lq) \quad (17)$$

- 9) Determining whether at time t the physical qubit is assigned to the logical qubit lq or not:

$$z(lq, t) = z1(lq, t) \wedge z2(lq, t) \quad (18)$$

- 10) In each time, a physical qubit can be assigned to at most one logical qubit:

$$\sum_{lq} y(lq, pq) \wedge z(lq, t) \leq 1 \quad \forall pq, \forall t \quad (19)$$

- 11) A logical qubit lq can be selected for measurement if and only if its dependencies have been resolved:

$$TM(lq) > TM(lq') \quad \forall lq \text{ and } \forall lq' \in D(lq) \quad (20)$$

There is an implementation note in this model: the maximum and the minimum number of physical qubits needed to realize a pattern $P = (V, I, O, A)$ is equal to $|V|$, and $|O| + 1$ [15], respectively. One can assume that

the number of available physical qubits is equal to $|V|$ and finally the solution determines whether they are used or not. However, the only metric that is minimized in the objective function (9) is the number of used physical qubits and it is not important that which of them are used. This feature causes a large number of optimal solutions, which enlarges the solution space and reduces the convergence speed of the model. Indeed, any composition of the minimum required qubits from $|V|$ is a candidate solution. To limit the solution space, one can set the number of available physical qubits to $|O| + 1$ and increase it one by one until an optimal solution is found.

As an example, the optimal solution of the proposed ILP model for SWAP pattern is shown in Table 2.

Table 2: ILP result for SWAP pattern

Parameter	Value
N	6
lq	{1,2,3, ..., 8}
pq	{1,2,3, ..., 8}
E(lq)	E[1]={1, 3, 5}, E[2]={2, 3} E[3]={1, 2, 3, 4}, E[4]={3, 4, 5, 7} E[5]={1, 4, 5, 6}, E[6]={5, 6, 7} E[7]={4, 6, 7, 8}, E[8]={7, 8}
D(lq)	D[1], D[2], ..., D[8] = {}
Variable	Value
TA(lq)	TA[1]=1, TA[2]=1, TA[3]=1, TA[4]=2, TA[5]=3, TA[6]=5, TA[7]=1, TA[8]=6
TM(lq)	TM[1]=3, TM[2]=1, TM[3]=2, TM[4]=4, TM[5]=5, TM[7]=6
u(pq)	u[1], u[2], u[3]
x(lq, t)	x[1,3], x[2,1], x[3,2], x[4,4], x[5,5], x[7,6]
y(lq, t)	y[1,3], y[2,1], y[3,2], y[4,1], y[5,2], y[6,1], y[7,3], y[8,2]
z(lq, t)	z[1,1], z[1,2], z[1,3], z[2,1], z[3,1], z[3,2], z[4,2], z[4,3], z[4,4], z[5,3], z[5,4], z[5,5], z[6,5], z[6,6], z[7,4], z[7,5], z[7,6], z[8,6]
z1(lq, t)	z1[1,1], z1[1,2], z1[1,3], z1[2,1], z1[3,1], z1[3,2], z1[4,1], z1[4,2], z1[4,3], z1[4,4], z1[5,1], z1[5,2], z1[5,3], z1[5,4], z1[5,5], z1[6,1], z1[6,2], z1[6,3], z1[6,4], z1[6,5], z1[6,6], z1[7,1], z1[7,2], z1[7,3], z1[7,4], z1[7,5], z1[7,6], z1[8,1], z1[8,2], z1[8,3], z1[8,4], z1[8,5], z1[8,6]
z2(lq, t)	z2[1,1], z2[1,2], z2[1,3], z2[1,4], z2[1,5], z2[1,6], z2[2,1], z2[2,2], z2[2,3], z2[2,4], z2[2,5], z2[2,6], z2[3,1], z2[3,2], z2[3,3], z2[3,4], z2[3,5], z2[3,6], z2[4,2], z2[4,3], z2[4,4], z2[4,5], z2[4,6], z2[5,3], z2[5,4], z2[5,5], z2[5,6], z2[6,5], z2[6,6], z2[7,4], z2[7,5], z2[7,6], z2[8,6]
Objective Function	Value
$\min \sum_{pq} u[pq]$	3

Indeed, ILP solver has found the values of variables based on the input model and parameters in such a way that it minimizes the objective function, while it satisfies constraints. One can simply verify the result using the information of this table. Note that, for the binary variables, only variables with a value of the unity are shown.

Results and Discussion

The proposed model was implemented using SCIP solver [33] and was run on a Core-i7 CPU operating at 2.4 GHz with 8 GB of memory. To evaluate the model, we applied it to some benchmark circuits from [16], [34], [35]. To generate the equivalent measurement patterns of the benchmark circuit, they were decomposed into CZ and $J(\alpha)$ gates. Then, the approach presented in [36], [37] was applied in order to produce the corresponding pattern. The optimizations which include standardization, signal shifting and Pauli simplifications [36] were also performed on the patterns.

The runtime of the proposed method as well as the obtained optimal solutions are given in Table 3. As this table shows, the number of the required qubits in the proposed approach ($|O|+1$) is the same as the theoretically proven minimal number of qubits in [16]. It should be recalled that the number of used physical qubits in a standard model is equal to $|V|$. Based on the obtained results, our approach (which find the optimal solution) reduces the number of physical qubits by 56.7% on average.

As shown in Table 3, for small patterns with less than

14 qubits, the model obtains the answer in a few seconds. However, with the increase in the pattern size, the run time grew exponentially and took more than 6 hours for GHZ_23 with 45 qubits. For larger patterns, e.g. GHZ_25, ILP solver was unable to find the answer for up to 12 hours.

Conclusion

1WQC is one of the measurement-based quantum computing models that presents a different approach to build quantum computers and is one of the most promising models for physical realization. However, the number of qubits in a 1WQC measurement pattern is much more than its number in the equivalent circuit model, and this issue makes this model hard to implement.

In this study, an ILP model is proposed to minimize the number of used physical qubits to realize a measurement pattern by efficiently scheduling the operations and reusing the physical qubits. The proposed method is able to find the optimal solution for both patterns with flow or only gflow.

Although the proposed method produces the optimal solution, one of the most important limitations of the proposed method and ILP-based methods in general is their high execution time and memory requirements, which grows exponentially with the increase of the problem size. For this reason, the proposed method cannot be used for large measurement patterns. Providing a suitable heuristic method to solve this problem will be pursued as future works.

Table 3: The runtime (in second) of the proposed ILP model and comparison of the obtained result with the standard pattern

Measurement pattern	$ V $	$ O $	The runtime of the proposed method (s)	The number of required physical qubits	Improvement %
CNOT	4	2	1	3	25.0
SWAP	8	2	1	3	62.5
Toffoli	17	3	5	4	76.5
QECC2_0_2	5	2	1	3	40.0
QECC3_0_2	6	3	1	4	33.3
QECC4_0_2	10	4	3	5	50.0
QECC4_1_2	9	4	3	5	44.4
QECC4_2_2	15	4	720	5	66.6
QECC6_2_2	14	6	405	7	50.0
QFT2	12	2	2	3	75.0
QFT3	30	3	537	4	86.6
Dusch10	29	10	232	11	62.0
Grover3	29	3	897	4	86.2
GHZ_20	39	20	8496	21	46.1
GHZ_23	45	23	21968	24	46.6
GHZ_25	49	25	N/A	NA	NA
Avg. Improvement	-	-	-	-	56.7

Author Contributions

Eesa Nikahd proposed the algorithm and designed the experiments. He also carried out the data analysis. Mahboobeh Houshmand and Monireh Houshmand collected the data. All of the authors interpreted the results and contributed to the writing of the manuscript.

Acknowledgment

The authors are grateful to anonymous reviewers for their valuable and constructive comments on an earlier version of this manuscript.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

<i>1WQC</i>	One-Way Quantum Computing
<i>gflow</i>	Generalized Flow
<i>ILP</i>	Integer Linear Programming
<i>MBQC</i>	Measurement-Based Quantum Computing

References

- [1] M. A. Nielsen, I. L. Chuang, *Quantum Computation and Quantum Information*, 10th Anniversary Edition. Cambridge: Cambridge University Press, 2010.
- [2] M. Nakahara, *Quantum Computing: From Linear Algebra to Physical Realizations*. CRC press, 2008.
- [3] G. Benenti, G. Casati, G. Strini, *Principles of Quantum Computation and Information-Volume I: Basic Concepts*. World scientific, 2004.
- [4] D. McMahon, *Quantum Computing Explained*. John Wiley & Sons, 2007.
- [5] R. Raussendorf, H. J. Briegel, "A one-way quantum computer," *Phys. Rev. Lett.*, 86(22): 5188, 2001.
- [6] R. Jozsa, *An Introduction to Measurement Based Quantum Computation*, NATO Science Series, III: Computer and Systems Sciences. Quantum Information Processing from Theory to Experiment, 199: 137-158, 2006.
- [7] V. Danos, E. Kashefi, P. Panangaden, S. Perdrix, *Extended Measurement Calculus, Semantic Techniques in Quantum Computation*, 235-310, 2009.
- [8] D. E. Browne, E. Kashefi, M. Mhalla, S. Perdrix, "Generalized flow and determinism in measurement-based quantum computation," *New J. Phys.* 9(8): 250, 2007.
- [9] B. Lanyon, P. Jurcevic, M. Zwerger, C. Hempel, E. Martinez, W. D'ur, H. Briegel, R. Blatt, C. F. Roos, "Measurement-based quantum computation with trapped ions," *Phys. Rev. Lett.*, 111(21): 210501, 2013.
- [10] J. E. Bourassa, R. N. Alexander, M. Vasmer, A. Patil, I. Tzitrin, T. Matsuura, D. Su, B. Q. Baragiola, S. Guha, G. Dauphinais, et al., "Blueprint for a scalable photonic fault-tolerant quantum computer," *Quantum*, 5: 392, 2021.
- [11] C. Reimer, S. Sciara, P. Roztocky, M. Islam, L. Romero Cort'es, Y. Zhang, B. Fischer, S. Loranger, R. Kashyap, A. Cino, et al., "High-dimensional one-way quantum processing implemented on d-level cluster states," *Nat. Phys.*, 15(2): 148-153, 2019.
- [12] M. Zwerger, H. Briegel, W. D'ur, "Hybrid architecture for encoded measurement-based quantum computation," *Sci. Rep.*, 4(1): 1-5, 2014.
- [13] M. Zwerger, H. Briegel, W. D'ur, "Measurement-based quantum communication," *Appl. Phys. B*, 122(50): 1-15, 2016.
- [14] E. Nikahd, M. Houshmand, M. S. Zamani, M. Sedighi, "One-way quantum computer simulation," *Microprocess. Microsyst.*, 39(3): 210-222, 2015.
- [15] M. Houshmand, M. Houshmand, J. F. Fitzsimons, "Minimal qubit resources for the realization of measurement-based quantum computation," *Phys. Rev. A*, 98(1): 012318, 2018.
- [16] E. Nikahd, M. Houshmand, M. S. Zamani, M. Sedighi, "OWQS: one-way quantum computation simulator," in *Proc. 15th Euromicro Conference on Digital System Design*: 98-104, 2012.
- [17] E. Nikahd, M. Houshmand, M. S. Zamani, M. Sedighi, "GOWQS: Graph-based one-way quantum computation simulator," in *Proc. 24th Iranian Conference on Electrical Engineering (ICEE)*: 738-744, 2016.
- [18] A. Broadbent, J. Fitzsimons, E. Kashefi, "Universal blind quantum computation," in *Proc. 50th Annual IEEE Symposium on Foundations of Computer Science*: 517-526, 2009.
- [19] J. Fitzsimons, E. Kashefi, "Unconditionally verifiable blind quantum computation," *Phys. Rev. A*, 96(1): 012303, 2017.
- [20] M. Houshmand, M. Houshmand, S. Tan, J. Fitzsimons, "Composable secure multi-client delegated quantum computation," *arXiv preprint arXiv:1811.11929*, 2018.
- [21] A. Farghadan, N. Mohammadzadeh, "Quantum circuit physical design flow for 2D nearest-neighbor architectures," *Int. J. Circuit Theory Appl.*, 45(7): 989-1000, 2017.
- [22] A. Farghadan, N. Mohammadzadeh, "Mapping quantum circuits on 3D nearest-neighbor architectures," *Quantum Sci. Technol.*, 4(3): 035001, 2019.
- [23] A. M. Childs, E. Schoute, C. M. Unsal, "Circuit transformations for quantum architectures," *arXiv preprint arXiv:1902.09102*, 2019.
- [24] G. Li, Y. Ding, Y. Xie, "Tackling the qubit mapping problem for NISQ-era quantum devices," in *Proc. the 24th International Conference on Architectural Support for Programming Languages and Operating Systems*: 1001-1014, 2019.
- [25] A. Zulehner, A. Paler, R. Wille, "An efficient methodology for mapping quantum circuits to the IBM QX architectures," *IEEE Trans. Comput. Aided Des. Integr. Circuits Syst.*, 38(7): 1226-1236, 2018.
- [26] R. Wille, L. Burgholzer, A. Zulehner, "Mapping quantum circuits to IBM QX architectures using the minimal number of SWAP and H operations," in *Proc. 56th ACM/IEEE Design Automation Conference (DAC)*: 1-6, 2019.
- [27] P. Murali, J. M. Baker, A. Javadi-Abhari, F. T. Chong, M. Martonosi, "Noise-adaptive compiler mappings for noisy intermediate-scale quantum computers," in *Proc. the 24th International Conference on Architectural Support for Programming Languages and Operating Systems*: 1015-1029, 2019.
- [28] E. Nikahd, N. Mohammadzadeh, M. Sedighi, M. S. Zamani, "Automated window-based partitioning of quantum circuits," *Phys. Scr.*, 96(3): 035102, 2021.
- [29] E. T. Campbell, J. Fitzsimons, "An introduction to one-way quantum computing in distributed architectures," *Int. J. Quantum Inf.*, 8(1): 219-258, 2010.
- [30] S. C. Benjamin, J. Eisert, T. M. Stace, "Optical generation of matter qubit graph states," *New J. Phys.*, 7(1): 194, 2005.
- [31] J. Chen, L. Wang, E. Charbon, B. Wang, "Programmable architecture for quantum computing," *Phys. Rev. A*, 88(2): 022311, 2013.

- [32] S. Sanaei, N. Mohammadzadeh, "Qubit mapping of one-way quantum computation patterns onto 2D nearest-neighbor architectures," *Quantum Inf. Process.*, 18: 1-19, 2019.
- [33] A. Gleixner, M. Bastubbe, L. Eifler, T. Gally, G. Gamrath, R. L. Gottwald, G. Hendel, C. Hojny, T. Koch, M. E. L'ubbecke, S. J. Maher, M. Miltenberger, B. M'uller, M. E. Pfetsch, C. Puchert, D. Rehfeldt, F. Schl'osser, C. Schubert, F. Serrano, Y. Shinano, J. M. Viernickel, M. Walter, F. Wegscheider, J. T. Witt, J. Witzig, *The SCIP Optimization Suite 8.0*, Technical Report (Optimization Online, 2021).
- [34] V. V. Albert & P. Faist, eds., *The Error Correction Zoo*, 2022.
- [35] D. Maslov, *Reversible logic synthesis Benchmarks* page, 2021. <https://reversiblebenchmarks.github.io>
- [36] A. Broadbent, E. Kashefi, "Parallelizing quantum circuits," *Theor. Comput. Sci.*, 410(26): 2489-2510, 2009.
- [37] M. Houshmand, M. H. Samavatian, M. S. Zamani, M. Sedighi, "Extracting one-way quantum computation patterns from quantum circuits," in *Proc. The 16th CSI International Symposium on Computer Architecture and Digital Systems (CADS 2012)*: 64-69, 2012.

Biographies



Eesa Nikahd received his B.Sc. degree in Computer Engineering from Shiraz University, Shiraz, Iran in 2010. He received his M.Sc. and Ph.D. degrees from Amirkabir University of Technology, Tehran, Iran in 2012 and 2018, respectively. In 2021, he joined the department of computer engineering, Shahid Rajaei Teacher Training University as an assistant professor. His current research interests include fault-tolerant quantum computing, quantum physical design and quantum machine learning.

- Email: nikahd@sru.ac.ir
- ORCID: [0000-0001-5112-1695](https://orcid.org/0000-0001-5112-1695)
- Web of Science Researcher ID: HNJ-2096-2023
- Scopus Author ID: 55570206700
- Homepage: <https://www.sru.ac.ir/nikahd/>



Mahboobeh Houshmand received her B.Sc. and M.Sc. in computer engineering, majoring in software, from Ferdowsi University of Mashhad in 1386 and 1389, respectively, and her Ph.D. in computer engineering, majoring in computer architecture from Amirkabir University of Technology in 1393. From the end of the summer of 2015 to the end of the summer of 2016, she was a post-doctoral researcher in the field of quantum cryptography jointly at the National University of Singapore and Singapore University of Technology and Design. Dr. Houshmand is currently an assistant professor in the computer engineering department of Islamic Azad University of Mashhad. Her research interests include quantum information theory and quantum computing, cryptography, multi-agent systems and data mining.

- Email: ma.houshmand@iau.ac.ir
- ORCID: [0000-0003-2017-4369](https://orcid.org/0000-0003-2017-4369)
- Web of Science Researcher ID: AAA-0000-0000
- Scopus Author ID: 36080655500
- Homepage: <http://ceit.aut.ac.ir/~houshmand/>



Monireh Houshmand received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from Ferdowsi University of Mashhad, in 2005, 2007 and 2011, respectively. In 2011, she joined the department of electrical engineering, Imam Reza International University as an assistant professor. She became an associate professor in 2019. Her research interests include quantum cryptography, quantum error correction, quantum logic synthesis and artificial intelligence.

- Email: m.hooshmmmand@imamreza.ac.ir
- ORCID: [0000-0001-9215-1532](https://orcid.org/0000-0001-9215-1532)
- Web of Science Researcher ID: ABD-4497-2021
- Scopus Author ID: 35752876400
- Homepage: <https://scholar.google.com.sg/citations?user=K5bMQeQAAAAJ&hl=en>

How to cite this paper:

E. Nikahd, M. Houshmand, M. Houshmand, "1WQC pattern scheduling to minimize the number of physical qubits," *J. Electr. Comput. Eng. Innovations*, 11(2): 391-398, 2023.

DOI: [10.22061/jecei.2023.9374.613](https://doi.org/10.22061/jecei.2023.9374.613)

URL: https://jecei.sru.ac.ir/article_1849.html





Research paper

New Platform for IoT Application Management Based on Fog Computing

S. Kalantary¹, J. Akbari Torkestani^{2,*}, A. Shahidinejad¹

¹Department of Computer Science, Qom Branch, Islamic Azad University, Qom, Iran.

²Department of Computer Science, Arak Branch, Islamic Azad University, Arak, Iran.

Article Info

Article History:

Received 14 December 2022

Reviewed 25 January 2023

Revised 01 March 2023

Accepted 11 March 2023

Keywords:

Internet of things

fog computing

perfect difference graph

layered architecture

*Corresponding Author's
Email Address:
ja.akbari@iau.ac.ir

Abstract

Background and Objectives: With the great growth of applications sensitive to latency, and efforts to reduce latency and cost and to improve the quality of service on the Internet of Things ecosystem, cloud computing and communication between things and the cloud are costly and inefficient; Therefore, fog computing has been proposed to prevent sending large volumes of data generated by things to cloud centers and, if possible, to process some requests. Today's advances in 5G networks and the Internet of Things show the benefits of fog computing more than ever before, so that services can be delivered with very little delay as resources and features of fog nodes approach the end user.

Methods: Since the cloud-fog paradigm is a layered architecture, to reduce the overall delay, the fog layer is divided into two sub-layers in this paper, including super nodes and ordinary nodes in order to use the coverage of super peer networks to use the connections between fog nodes in addition to taking advantage of the features of that network and improving the performance of large-scale systems. It causes fog nodes to interact with each other in processing requests and fewer data will be sent to the cloud, resulting in a reduction in overall latency. To reduce the cost of bandwidth used among fog nodes, we have organized a sub-layer of super nodes in the form of a Perfect Difference Graph (PDG). The new platform proposed for aggregation of fog computing and Internet of Things (FOT) is called the P2P-based Fog supported Platform (PFP).

Results: We evaluate the utility of our proposed method by applying ifogsim simulator and the results achieved are as follows: (1) power consumption parameter in our proposed method 24% and 38% have improved compared to the structure three-layer fog computing architecture and without fog layer respectively; (2) network usage parameter in our proposed method 26% and 32% have improved compared to the structure three-layer fog computing architecture and without fog layer respectively; (3) average response time parameter in our proposed method 17% and 58% have improved compared to the structure three-layer fog computing architecture and without fog layer respectively; and (4) delay parameter in our proposed method 1% and 0.4% have improved compared to the structure three-layer fog computing architecture and without fog layer respectively.

Conclusion: Numerical results obtained from the simulation show that the delay and cost parameters are significantly improved compared to the structure without fog layer and three-layer fog computing architecture. Also, the results show that increasing number of things has the same effect in all cases.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Constrained systems are not able to interact with the vast amount of data on the Internet of Things (IoT); Therefore, cloud computing has been widely used and is an integral part of IoT. IoT creates unimaginable amounts of different types of data. This data is distributed throughout the environment, and is called big data. Some of the most important challenges to be considered for big data include processing, preparation and storage [1], [2], [3] and [6].

Cloud computing is a computing framework which is a good solution in this regard. Data is sent to cloud data centers for processing and storage and will be available after analysis and processing [7]. Using cloud computing in IoT applications has the following advantages: cost saving, reliability, manageability. The strength of IoT is that it permeates people's daily lives in terms of personal and home issues such as smart city, smart home, smart health, life assistance and work issues such as industry and factory automation and smart transportation. However, delays due to the cloud being away from end-users challenge the usefulness of IoT systems in many applications. Despite this, fog computing has been recommended to deal with many cloud processing problems such as unreliable latency, lack of proper mobility support and lack of location-awareness support, and reduced processing speed due to increased data transfer size and consequent reduced bandwidth. Fog processing is used to prevent the transmission of this large amount of data to cloud data centers and also to perform a series of necessary pre-processing on them. Fog computing is a distributed computing paradigm that acts as an intermediate layer between cloud data centers and IoT devices [4]. This concept was first defined by Cisco as the development of cloud computing, from the core to the edge of the network. In fact, fog calculations were introduced as mini clouds [5]. The fog computing environment consists of traditional network components such as routers, switches, proxy servers, base stations, and so on. These components enable fog computing to geographically distribute cloud-based services at the edge of the network. Therefore, fog computing can support data location, scalability, interoperability, and mobility. The architecture of cloud-fog-thing layer paradigm is shown in Fig. 1.

In this research, a new platform is proposed for fog of things (FOT), which defines a fog computing structure for IoT. The P2P-based Fog-supported Platform (PFP) utilizes the features of super peer networks to use communications and interactions among fog nodes in request processing as well as improving performance and quality of services (QoS); therefore, the nodes in the fog layer are organized into two sub-layers of super and ordinary nodes.

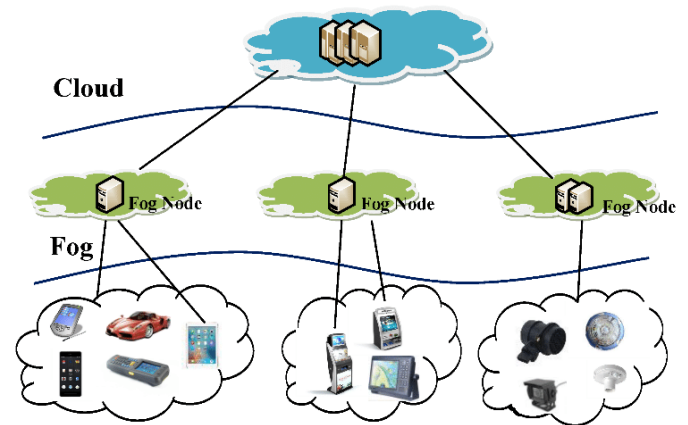


Fig. 1: Architecture of cloud-fog-thing layer.

Higher level nodes are called super fog nodes that are connected to each other in the form of a Perfect Difference Graph (PDG). Since the diameter parameter in these graphs is equal to 2, we expect a reduction in the communication delay and the volume of messages sent and consequently reduced available bandwidth. This paradigm is simulated by iFogSim simulator to manage IoT applications like smart city [29], [30]. Despite the numerous benefits of fog computing, research in this field is still emerging and immature, and many researchers are still studying it.

The rest of this article is organized as follows:

Section A provides an overview of research and studies on the smart city and the platforms provided by fog computing; Section B presents the proposed method; In the evaluation section, the results obtained from the simulation of the proposed method are stated along with its comparison with some of the performed projects, and finally the conclusion and suggestions for further research in this field are mentioned in the last section.

Technical Work Preparation

A. Related Work

Padova Smart City project, discussing the urban features of the IoT system, such as the services and items required to implement the Smart City [8]. In [9] the authors stated that coordinated distributed plans are required to create IoT applications in the smart city; Although, more attention has been paid to the integration of cloud computing and IoT in smart city applications. In [33] a cloud-based framework for creating a smart city through IoT capabilities is expressed. A framework called FOCAN (A Fog-supported Smart City Network Architecture for Management of Applications on the Internet of Everything Environments) in which components and services of the smart city communicate with each other and with fog computing [10]. The FOCAN architecture consists of two levels: IOE and fog nodes,

which manage IoT applications. FOCAN is an efficient computing and communication structure which minimizes average energy consumption. A framework introduces in [11] for IoT in which data collected from medical sensors are stored on fog servers, and uses a centralized platform for end-to-end communication between end users and medical sensors. This structure has been able to support mobility well. In [12] the design of an open stack platform is proposed that considers scalability in smart city applications using fog computing. The authors in [13] have proposed a platform that uses fog computing to improve the performance of traffic and driving issues in VANET (Vehicular ad hoc network) networks to satisfy the need for location-awareness. In [14] the authors have designed a three-layer architecture for smart buildings based on fog computing, the results of which show that the fog layer is very effective in using network resources and reducing bottlenecks in cloud computing. In [15], a three-layer architecture is proposed for big data analysis in smart cities, including intermediate computational fog nodes, edge computing nodes, and fog nodes specifically with sensing power. Authors in [16] proposed a three-layer architecture is proposed for IoT applications called cloud, fog, and dew, in which the dew layer refers to edge devices such as sensors and video cameras. In [17], the authors have proposed the soft-IoT paradigm, in which they provide protocols using fog computing to facilitate the processing of local data and service delivery on small servers by virtual entities. In [18] a new architecture has been collected and updated, and has presented a process of real-time and heterogeneous information from different sources, and then it has been tested by providing a smart parking service in a smart city. In [19], a platform is provided based on fog data that uses fog computing to reduce cloud storage and transmission delays in smart health applications. The addressed issues in [20], include theoretical modeling of fog computing architecture, in particular, service delays, power consumption, and cost; But there is no specific policy to reduce service delays. In [21], the authors have suggested "task distribution" to minimize overall cost and service quality requirements in fog computing-based medical Cyber-Physical System (CPS). In [22] a service distribution strategy in the cloud-fog scenario is proposed so that the services are subdivided into sub-services and their parallel experiments are run on edge devices to minimize service delays. The optimization of data transfer from IoT sensors to the cloud by the enhanced learning technique to predict the data which will be transferred from the sensors in the future, then the amount of data transfer is reduced by determining the data which are not to be transferred, and the service quality is increased subsequently [23]. The distributed cloud storage replication method and the greedy exploratory method

are proposed to minimize latency in order to counteract the amount of data transmitted between the sensor and the cloud [24]. However, it should be noted that fog computing has not been used in the last two cases. In [36] authors present a new platform integrating big data streaming processing with machine learning (ML)-based applications. And they provide a comprehensive IoT data processing workflow, including data access and transfer, big data processing, online ML, long-term storage, and monitoring. In [37] considers possible fog computing applications and potential enabling technologies towards sustainable smart cities in the IoT environments. In addition, different caching techniques and the use of Unmanned Aerial Vehicles (UAVs), and various Artificial Intelligence (AI) and Machine Learning (ML) techniques in caching data for fog-based IoT systems are comprehensively discussed. Finally, the potential and challenges of such systems are also highlighted. FogFrame is a framework for IoT application that use multi-tier fog computing and create fog colony [38]. A new delay tolerant network for IoT data processing introduces in [39] that uses multi-layer fog servers.

It is worth mentioning that in all the works reviewed in this section, the fog layer isn't divided into two different layers of fog nodes, and they have a similar structure and the same capabilities and capacities. And a fog layer has been used and investigated in research.

B. Proposed Method

Since PDG is used in the proposed platform, before starting the details of the proposed method, we will briefly introduce and describe its features.

Definition 1: A Perfect Difference Set (PDS) is a set of residues $\{S_0, S_1, \dots, S_{\delta} + 1\} \bmod n$, so that any non-zero residue can be uniquely expressed in $\{S_i - S_j\}$ format.

Definition 2: A PDG is a graph with n vertices where $n = \delta^2 + \delta + 1$, δ is the power of graph and at least equals 2. In PDG, node i is connected to $1 \leq j \leq \delta$, $(i \pm S_j) \bmod n$, and S_j is an element of PDS (Perfect Difference set) from δ order.

According to [25] and [26], 4 edges can be defined in each PDG as follows;

Ring edge: The edge connecting consecutive nodes i and $i + S_1 \bmod n$.

Chord edge: The edge connecting the non-consecutive nodes i and $i + S_j \bmod n$ where $2 \leq j \leq \delta$,

Forward edge: For node, including chord edges connecting nodes i and $i + S_j \bmod n$, and ring edge connecting nodes i and $i + S_1 \bmod n$.

Backward edge: For node i , including the chord edges connecting nodes i and $i - S_j \bmod n$, and the ring edge connecting nodes i and $i - S_1 \bmod n$.

Table 1 shows 10 initial values of PDS.

Table 1: Ten initial value of PDS

n	δ	Perfect difference sets
7	2	0, 1, 3
13	3	0, 1, 3, 9
21	4	0, 1, 4, 14, 16
31	5	0, 1, 3, 8, 12, 18
57	7	0, 1, 3, 13, 32, 36, 43, 52
73	8	0, 1, 3, 7, 15, 31, 36, 54, 63
91	9	0, 1, 3, 9, 27, 49, 56, 61, 77, 81
133	11	0, 1, 3, 12, 20, 34, 38, 81, 88, 94, 104, 109
183	13	0, 1, 3, 16, 23, 28, 42, 76, 82, 86, 119, 137, 154, 175
273	16	0, 1, 3, 7, 15, 63, 90, 116, 127, 136, 181, 194, 204, 233, 238, 255

a. Describing the Proposed Method

The fog node is the key component of fog computing used in this research, which provides the resources and activities requested by services and has the computing, storage, and networking ability necessary to run IoT applications [27]. On the other hand, each node must be able to communicate with the other fog node, cloud data centers, and things; because fog nodes are the intermediate layer between clouds and things [28]. Therefore, it must also have mechanisms to communicate with heterogeneous components, data collection, control, and analysis.

In this research, fog nodes are divided into two categories of super fog nodes and ordinary fog nodes based on performance, storage and computational capabilities, and location in the proposed PFP structure. Therefore, the proposed PFP architecture consists of three levels: the lowest layer includes the things in the smart city, which are connected to the Internet using various communication technologies such as 4.5G, Wi-Fi, or ZigBee, and need to be clustered according to their location; Fog nodes are in the middle level, and cloud data centers are located at the highest level. Fig. 2 shows a general framework of the proposed architecture. There are three layers in this architecture including thing, fog, and cloud. Cloud servers are located at the cloud layer and consist of several processing and storage units. The strength of this architecture goes back to the second layer, the fog layer. Fog nodes are divided into two categories based on processing, storage, and networking capabilities; therefore, the fog layer consists of two sub-layers: the upper layer is called Super Fog layer (SFL) and the bottom layer is called Ordinary Fog Layer (OLF).

In PFP, things in the things layer must be clustered; it is performed based on the location of things. In addition to distance, speed and direction of movement are also

used as criteria in clustering, as some things have a very high mobility. Each thing must be associated with an ordinary fog node. Things clustered in a group will be associated with the nearest ordinary fog node. In this study, it is assumed that each active thing in IoT is associated with only one ordinary node and registers itself to only one ordinary fog.

Based on [32], things in the IoT can be classified into 3 categories. The first category will be things whose destruction causes severe irreparable physical, economic or social damage, such as a wireless pacemaker or car brake system controller. The second category includes things whose absence or breakdown has severe physical or economic effects, such as the misusing the air conditioning. Finally, the third category of things consist of those whose absence, deterioration or withdrawal from the system is not a serious threat to living beings and also economic or social conditions. Therefore, the type of each thing is determined based on this classification in the smart city. Because each fog node contains a local database, the information needed for the applications and the data in this classification can be stored in the fog nodes to store things information and the data they generate; storage and using different recovery policies to access the data also creates a prioritization to respond to requests. Storage can be used in analyzing and storage of big data.

In the past IoT platforms, all data obtained from different sources such as sensor devices, IoT devices, and websites were sent to the cloud, leading to reduced processing speed and using large quantity of bandwidth due to the high size of data transfer. Now, ordinary fog nodes perform an initial analysis of the received data, and an index of analytical data in ordinary fog nodes is stored in the super fog node instead of sending data to the cloud. Due to the mentioned storage and using the resources available in fog nodes, the need for communication with cloud data centers is minimized and there is less delay.

The upper sub layer, SFL, contains super fog nodes that are more powerful in terms of processing, storage, and networking capabilities than the nodes in the bottom layer, OLF; although they are still a long way from the capabilities of cloud servers, they can be considered mini-clouds. The nodes in this layer are connected with each other according to PDG, order δ . Each SFN is associated with several OFNs. For example, all OFNs defined in a settlement are associated with an SFN. In this structure, it must be noted that each thing is associated with only one ordinary fog node (of course, to maintain the information network, an alternative ordinary fog node is available is always available) and each OFN is connected to only one SFN and they are not in direct contact with each other.

As mentioned before, fog nodes, also have networking facilities and equipment in addition to storage. Therefore,

part of the activities of ordinary fog nodes is intended to meet the immediate needs based on the demand for things and the provision of resources. Based on this, and considering the types of connections and how to send queries in PDG-based super peer-to-peer networks, 5 types of connections can be defined in this structure. Thing-to-thing relationship (t2t), thing-to-ordinary fog node relationship (t2o), super node connection to ordinary node (s2o), super node connection with another super node fog (s2s) and super node fog connection to cloud data centers (s2c) These connections are shown separately in Fig. 2, except for the t2t connection.

Requests that cannot be processed by things can be divided into 3 categories based on the amount of resources required and the estimated processing time, and like the traffic class field in the IPv6 header, the type of request can be specified, so that there is less delay in sending requests. ; Thus, the first category called low-res requests (which require low resource) that can be processed in ordinary fog nodes and are compatible with the resources and capabilities of ordinary fog nodes. The second categories (which require middle resource) includes requests that must be processed by super fog nodes, so they can be sent to super fog nodes immediately after reception to be processed by one of the super fog nodes. They can be introduced as semi-heavy processing; finally, the last category is called high-res requests (which require high resource) that must be processed by the cloud, so they are sent to the cloud immediately after being received by the super fog node, and are referred to as heavy-processing requests. Although, it should be noted that in situations such as non-acceptance in the queue, each request can be processed by another processing node, for example, low-res requests may also be processed in super fog nodes or clouds.

The strength of this method is in sending messages in the super layer of fog. Anyway, the message reaches this layer, the transmission of the message between the nodes follows the PDG algorithm. If the request is of mid-res type, it can be accepted and processed in super fog nodes; therefore, according to the PDG-based communications in SFL, requests are sent to other super fog nodes to find a suitable fog node to run. If no super fog node accepts the request at this step, it must be sent to the cloud. The following describes how PDG-based super fog nodes are connected [25] and [26]. According to [34] sending requests from fog nodes to other fog nodes (either ordinary fog nodes or super fog nodes) or sending them to the cloud is called request offloading or load sharing. In our proposed structure, deciding to offload a request to other fog nodes depends on the type of request, the fog node response time and, the conditions of the fog node in terms of available space.

This will happen if the response time of the fog is longer than the maximum allowed delay of the request and the type of request also allows offload.

In this paper, our purpose is to examine whether the proposed structure is less delayed in responding to requests; Note that according to [35], we define a delay the time required to service a request sent from a thing, that is, the time interval between the moment a request is sent by an thing until a response is received.

b. Communicating and Sending Messages in PDG-Based SFL Layer

The system proposed in this paper is based on graph is called G , $G = \langle V, E \rangle$ where V is the total set of fog nodes including ordinary and super fog nodes. G' is a subgraph from G , $G' = \langle V', E' \rangle$ can be considered as a directionless graph where $V' \subset V$ is the super fog node and $E' \subset E$ is the connections among the super fog node, called interconnection; It should be noted that the connections between ordinary fog nodes and super fog nodes are called intraconnection. Graph G' is a PDG of order δ and logical topology of super fog node is PDG. If the $i \in V'$ node wants to send a message to the system in order to find the appropriate node for accepting and then processing the request, a two-step process occurs (PDG-Algorithm):

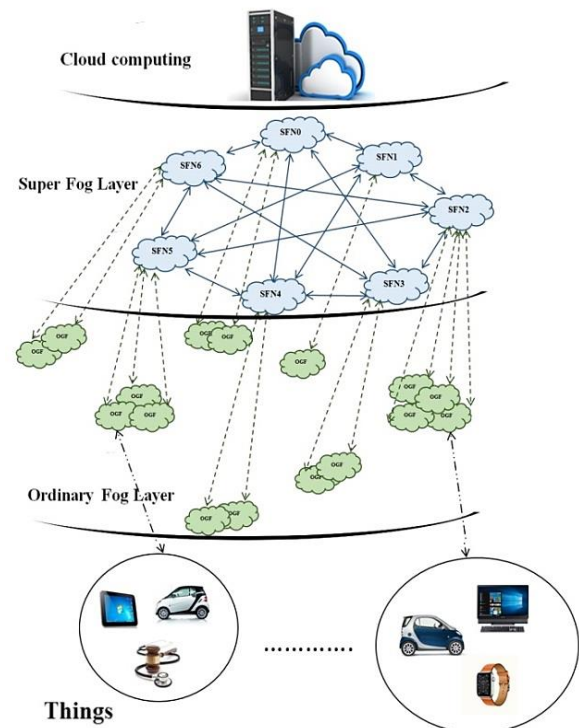


Fig. 2: Architecture of the proposed structure (PFP)

S2S connection \longleftrightarrow , O2S connection \dashrightarrow and T2O connection $\cdots \rightarrow$

Step 1: The message with the possibility of moving 2 steps (TTL = 2) is sent to the neighbors with whom it is

connected according to the forward edge, and also the same message is sent to the neighbors with whom it is connected according to the backward edge ($TTL = 1$), and the receiving intermediate nodes reduce the TTL by one unit as soon as the message is received.

Step 2: If the intermediate node receives a message, it sends it to all nodes associated with the backward edge except the node from which it received the main message.

Since the diameter of the graph in this structure is 2, the communications are established with very little delay. Also, the number of messages moving and imposed on the super fog nodes equals $\delta_2 + \delta$, which is used in calculating the consumed bandwidth and makes the system cost less.

The procedure of the proposed method can be presented in Algorithm 1.

Algorithm 1: Pseudo code for proposed method

Input: rth request for uth user($Req\{u,r\}$)
Output: Assigned node for request processing
Begin
1. for all IoT users
2. for all IoT requests
3. Add $Req\{u,r\}$ to fog-queue
4. end for
5. end for
6. while ($Req\{u,r\}$ is in the fog-queue)
7. for all $Req\{u,r\}$ in associated fog-queue according to Geographical clustering
8. if (possible assignment OFN) return OFN id
9. else
10. call(PDG-based SFL layer algorithm)
11. end of while
End

Fig. 3 presents proposed method in flowchart format.

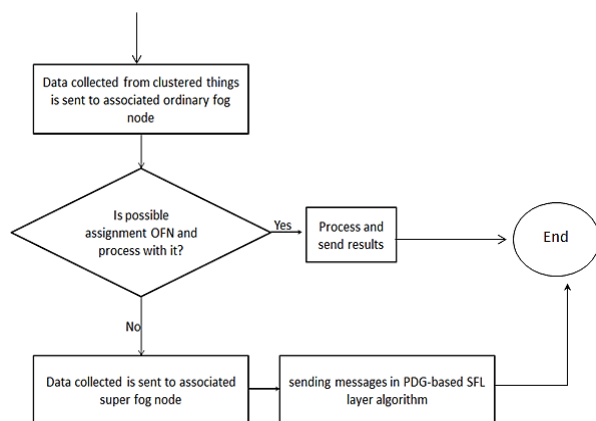


Fig. 3: Flowchart of proposed method.

Results and Discussion

In this section, we examine the proposed approach through simulation. In the following, we will explain how

to set the simulator and performance criteria. Then, the simulation results will be discussed.

A. Simulation Settings

The simulations presented in this section are performed using the iFogSim library [29] and [30], which is an extension of CloudSim [31]. This simulator is used for modeling the cloud computing infrastructure and IoT services. The iFogSim toolkit allows the user to describe fog nodes and provide resource management mechanisms for performing IoT services, as well as the possibility to evaluate performance metrics related to fog environments. In the present study, the simulation was performed using the iFogSim library on a computer with an Intel Corei5 CPU, a 250 GB disk, a 4 GB RAM and Windows 10. The iFogSim emulator consists of a set of classes, including the FogDevice, Sensor, and Actuator classes for fog modeling, and a set of classes, including AppModule, AppEdge, and Tuple for modeling IoT services. Note that the FogDevice class is one of the most basic iFogSim simulator classes used to simulate fog nodes, which has memory, network, and computing resource features. This class specifies the hardware specifications of the fog nodes as well as their connections. In the proposed approach, FogDevices are designed on several levels. At the lowest level are things and IoT devices, and at the highest level are VMs, which connect to gateways using links. To configure cloud layer infrastructure and fog layers, fog devices are assumed with the characteristics listed in Table 2. The Sensor class is used to simulate IoT sensors and can be used to generate tuples that are equivalent to tasks in a cloud computing environment. The Actuator class is used to implement the output operation.

Table 2: cloud layer and fog layer specification

		MIPS	RAM	storage	Down-BW	Up-BW
Cloud Layer	Host	48800	60000	1000000	100	10000
	VM	3800	8000	100000	1000	10000
Fog Layer	SFN	200	4000	30000	1000	10000
	OFN	500	1000	10000	50	5000

B. Performance Criteria and Simulation Results

In this section, the results of the simulated scenario discussed above are presented and the results obtained are compared with the other two modes. In one mode, there is only one layer of fog in the system and all fog nodes have the same structure and are in the same state in response to requests.

In the other mode, the considered system employs a cloud layer but no fog layer. The efficiency parameters studied in the research include energy consumption,

average response time and the amount of network consumption. The computing power of fog devices is given based on millions of instructions per second (MIPS). RAM and storage are specified in MB and bandwidth is measured in Mbps. In this section, we use notations that describe in Table 3.

In the present study, the experiments were repeated 4 times; each time we changed the number of things in the system; the number of sensors was considered 10, 20, 50 and 100. The number of ordinary and super fog nodes has not changed. Fig. 3 shows the amount of energy consumption that we compute it based on (1) for proposed algorithm. According to the obtained results and as we expected, the amount of energy consumption increased following an increase in the number of things. Since we used PDG network and two layers of fog with different characteristics in the proposed approach, the least amount of energy was consumed in all cases compared to the other two modes (the system with one layer of fog, and the system without fog layer). The information of fog devices is received more accurately and quickly due to the presence of PDG network; therefore, we face a reduction in energy consumption.

$$\text{Energy consumption} = CEC + (CT - UT) * LHP \quad (1)$$

Table 3: Notations and definitions

Notation	Definition
CEC	Current Energy Consumption
CT	Current Time
UT	Update Time
LHP	Last Host Power
TL	Total latency
TS	Total size of tuple
MST	Maximum simulation Time
EST	Estimated Service Time
EET	Ending Execution Time
N	Total number of executed tuple
ST	Service Time

In Fig. 4, network usage is investigated and compared in 3 modes: the system with only cloud layer and without the fog layer, the system with a fog layer with similar fog nodes, and the proposed model. As expected, the network consumption increases sharply as the number of things in the system increases in the absence of the fog layer. In the other two modes, we observe less network consumption than in the first mode. Comparing the conditions in the fog layer system, it can be stated that the network consumption in the proposed algorithm is of

the lowest value in all conditions. This parameter for proposed method computes based on (2).

$$\text{Network usage} = \frac{TL * TS}{MST} \quad (2)$$

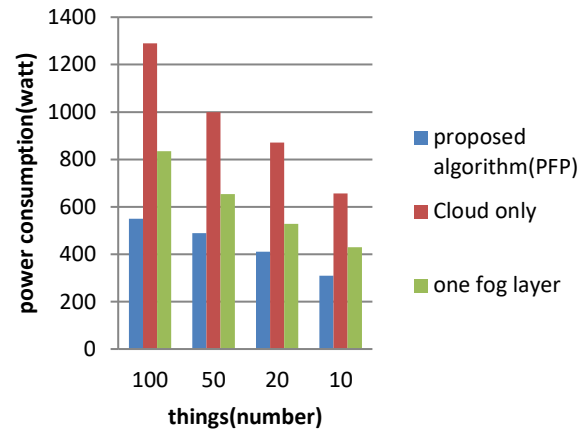


Fig. 3: Power consumption.

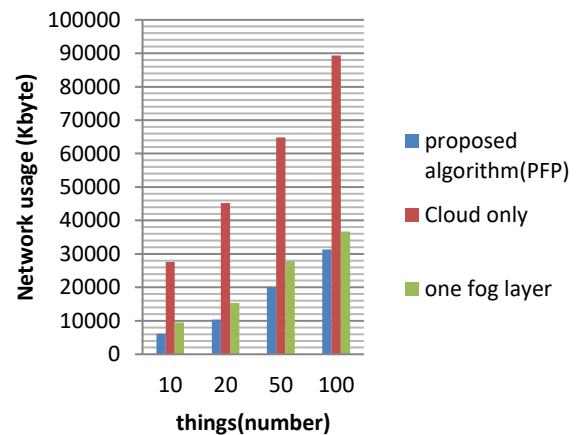


Fig. 4: Network usage.

Fig. 5 shows the average response time in the system. As shown in the figure, the average response time increases with increasing the number of sensors, which is predictable; because the number of requests in the system increases following the increase in the number of sensors and consequently there will be an increase in the average response time of the system. But what is noteworthy is that the average response time in the proposed algorithm has improved compared to the other two modes. Because request response management is more appropriate in the proposed algorithm.

Fig. 6 shows the delay in different time intervals. The proposed method has the least delay compared to other methods. The reason for that is the use of pdg communication, which leads to the selection of fog node with higher accuracy, and as a result, the response time will be reduced and the amount of delay will be less. We

considered the time intervals as 10 second intervals and repeated the work up to 100 seconds. In the situation where the fog layer is a single layer, because the load distribution happens with a longer delay, the delay should be increased compared to the proposed method. This parameter is computed based on (3).

$$Delay = \frac{EST * N + (EET - ST)}{N} \quad (3)$$

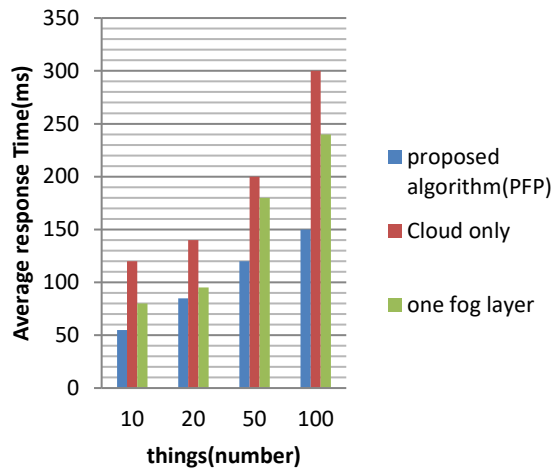


Fig. 5: Average response time.

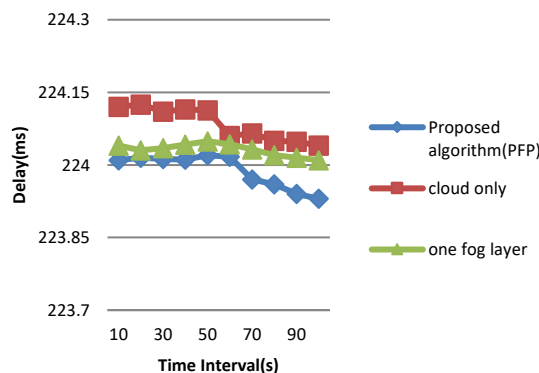


Fig. 6: Delay.

Conclusion

In this paper, a new platform is proposed for fog of things (FoT) that defines a fog computing structure for IoT. The P2P-based fog supported platform (PFP) utilizes the features of super peer networks to use communications and interactions between fog nodes in request processing as well as improving performance and QoS; therefore, it organizes the nodes in the fog layer into two sub-layers of super and ordinary nodes. Higher level nodes are super fog nodes that are connected to each other in the form of a PDG (Perfect Difference Graph). Since the diameter parameter in these graphs is equal to 2, examining the results, it was observed that the performance evaluation parameters including energy consumption, average response time, network

consumption and delay have been significantly improved. This paradigm is simulated by ifogsim simulator for managing IoT applications like smart city. The numbers of super and ordinary fog nodes are constant and have not changed in this research. It is suggested for the future researches to investigate the effect of it. The structure of these nodes also follows the PDG; it is suggested to consider and compare other structures in the future research.

Author Contributions

S.Kalantary designed the experiments and collected the data. All authors carried out the data analysis. J. Akbari and A.Shahidinejad were supervisor. All authors interpreted the results and wrote the manuscript.

Acknowledgment

The authors would like to thank both the Editor and anonymous reviewers of JECEI for their valuable and constructive Comments and feedbacks.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

IoT	Internet of Things
PDG	Perfect Difference Graph
FoT	Fog of Things
QoS	Quality of Service
P2P	Peer to Peer
CPS	Cyber-Physical System
FoCAN	Fog-supported Smart City Network Architecture
UAV	Unmanned Aerial Vehicles
ML	Machine Learning
AI	Artificial Intelligence
SFL	Super Fog layer
OFL	Ordinary Fog layer
TTL	Time to Live

References

- [1] Y. Hajjaji, W. Boulila, I. R. Farah, I. Romdhani, A Hussain, "Big data and IoT-based applications in smart environments: A systematic review," *Compu. Sci. Rev.*, 39: 100318, 2021.
- [2] F. E. F. Samann. S. R. M. Zeebaree, S. Askar, "IoT provisioning QoS based on cloud and fog computing," *J. App. Sci. Tech. Trends*, 2(01): 29-40, 2021.

- [3] I. Lee, K. Lee, "The Internet of Things (IoT): Applications, investments, and challenges for enterprises," *Bus. Horiz.*, 58(4): 431-440, 2015.
- [4] T. Wang, Y. Liang, W. Jia, M. Arif, A. Liu, M. Xie, "Coupling resource management based on fog computing in smart city systems," *J. Network Compu. Appl.*, 135: 11-19, 2019.
- [5] M. Etemadi, M. Ghobaei-Arani, A. Shahidinejad, "Resource provisioning for IoT services in the fog computing environment: An autonomic approach," *Comput. Commun.*, 161: 109-131, 2020.
- [6] S. Kalantary, J. Akbari Torkestani, A. Shahidinejad, "Resource discovery in the Internet of Things integrated with fog computing using Markov learning model," *J. Supercompu.*, 77: 13806-13827, 2021.
- [7] M. Ghobaei, A. Shahidinejad, M. Torabi, "Resource elasticity management using fuzzy controller based on threshold changes in the cloud computing environment," *Electron. Cyber Def.*, 8(3): 63-81, 2020.
- [8] A. Cenedese, A. Zanella, L. Vangelista, M. Zorzi "Padova smart city: An urban internet of things experimentation," in *Proc. IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, 2014.
- [9] H. A. Khattak, H. Farman, B. Jan, I. Ud Din, "Toward integrating vehicular clouds with IoT for smart city services," *IEEE Network*, 33(2): 65-71, 2019.
- [10] P. G. V. Naranjo, Z. Pooranian, M. Shojafar, M. Conti, R. Buyya, "FOCAN: A Fog-supported smart city network architecture for management of applications in the Internet of Everything environments," *J. Parallel Distrib. Comput.*, 132: 274-283, 2019.
- [11] C. Thota, R. Sundarasekar, G. Manogaran, R. Varatharajan, M. K. Priyan, "Centralized fog computing security platform for IoT and cloud in healthcare system, in *Fog computing: Breakthroughs in research and practice*," *IGI global*: 365-378, 2018.
- [12] D. Bruneo, S. Distefano, F. Longo, G. Merlino, A. Puliafito, V. D'Amico, M. Sapienza, G. Torrisi, "Stack4Things as a fog computing platform for Smart City applications," in *2016 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2016.
- [13] J. Grover, A. Jain, S. Singhal, A. Yadav, "Real-time vanet applications using fog computing," in *Proc. First International Conference on Smart System, Innovations and Computing*, 2018.
- [14] F. J. Ferrández-Pastor, H. Mora, A. Jimeno-Morenilla, B. Volckaert, "Deployment of IoT edge and fog computing technologies to develop smart building services," *Sustainability*, 10(11): 3832, 2018.
- [15] B. Tang, Z. Chen, G. Hefferman, T. Wei, H. He, Q. Yang, "A hierarchical distributed fog computing architecture for big data analysis in smart cities," in *Proc. the ASE BigData & SocialInformatics*: 1-6, 2015.
- [16] K. Skala, D. Davidovic, E. Afghan, I. Sovic, Z. Šojat, "Scalable distributed computing hierarchy: Cloud, fog and dew computing," *Open J. Cloud Comput. (OJCC)*, 2(1): 16-24, 2015.
- [17] C. Prazeres, M. Serrano, "Soft-iot: Self-organizing fog of things," in *Proc. 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 2016.
- [18] R. Tandon, P. Gupta, "Optimizing smart parking system by using fog computing," in *Proc. International Conference on Advances in Computing and Data Sciences*, 2019.
- [19] A. M. Rahmani, T. Nguyen Gia, B. Negash, A. Anzanpour, I. Azimi, M. Jiang, P. Liljeberg, "Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach," *Future Gener. Comput. Syst.*, 78: 641-658, 2018.
- [20] X. Masip-Bruin, X. Masip-Bruin, E. Marín-Tordera, G. Tashakor, A. Jukan, G. J. Ren, "Foggy clouds and cloudy fogs: a real need for coordinated management of fog-to-cloud computing systems," *IEEE Wireless Commun.*, 23(5): 120-128, 2016.
- [21] G. M. Dias, Prediction-based strategies for reducing data transmissions in the IoT, *Universitat Pompeu Fabra*, 2016.
- [22] V. B. Souza Xavi Masip-Bruin, E. Marín-Tordera, W. Ramirez, S. Sanchez, "Towards distributed service allocation in fog-to-cloud (f2c) scenarios," in *Proc. 2016 IEEE global communications conference (GLOBECOM)*, 2016.
- [23] M. Ramachandra, "Optimization of the data transactions and computations in IoT sensors," in *Proc. 2016 International Conference on Internet of Things and Applications (IOTA)*, 2016.
- [24] A. Kumar, N. C. Narendra, U. Bellur, "Uploading and replicating internet of things (IoT) data on distributed cloud storage," in *Proc. 2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*, 2016.
- [25] J. S. Li, C. H. Chao, "An efficient superpeer overlay construction and broadcasting scheme based on perfect difference graph," *IEEE Trans. Parallel Distrib. Syst.*, 21(5): 594-606, 2009.
- [26] N. Singh et al., Study of Topological Analogies of Perfect difference Network and Complete Graph, 2019.
- [27] M. Ghobaei-Arani, A. Souri, A. A. Rahmanian, "Resource management approaches in fog computing: a comprehensive review," *J. Grid Comput.*, 18: 1-42, 2019.
- [28] E. M. Tordera, E. Marín Tordera, X. Masip-Bruin, J. Garcia-Alminana, A. Jukan, G. Jie Ren, J. Zhu, J. Farre, "What is a fog node a tutorial on current concepts towards a common definition," *arXiv preprint arXiv:1611.09193*, 2016.
- [29] H. Gupta, A. Vahid Dastjerdi, S. K. Ghosh, R. Buyya, "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments," *Softw. Pract. Exper.*, 2017. 47(9): 1275-1296, 2017.
- [30] R. Mahmud, R. Buyya, "Modelling and simulation of fog and edge computing environments using iFogSim toolkit," *Fog and edge computing: Principles and paradigms*: 433-465, 2019.
- [31] R. N. Calheiros, R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exper.*, 41(1): 23-50, 2011.
- [32] P. Maiti, J. Shukla, B. Sahoo, A. Kumar Turuk, "Mathematical modeling of qos-aware fog computing architecture for iot services," in *Proc. Emerging Technologies in Data Mining and Information Security*: 13-21, 2019.
- [33] T. H. Kim *et al.*, Smart city and IoT, *Elsevier*: 159-162, 2017.
- [34] A. Yousefpour, C. Fung, T. Nguyen, K. Kadiyala, F. Jalali, A. Niakanlahiji, J. Kong, J. P. Jue, "All one needs to know about fog computing and related edge computing paradigms: A complete survey," *J. Syst. Archit.*, 98: 289-330, 2019.
- [35] A. Yousefpour, G. Ishigaki, J. P. Jue, "Fog computing: Towards minimizing delay in the internet of things," in *Proc. 2017 IEEE international conference on edge computing (EDGE)*, 2017.
- [36] Z. Wan, Z. Zhang, R. Yin, G. Yu, "KFIML: Kubernetes-based fog computing iot platform for online machine learning," *IEEE Internet Things J.*, 9(19): 19463-19476, 2022.
- [37] H. Zahmatkesh, F. Al-Turjman, "Fog computing for sustainable smart cities in the IoT era: Caching techniques and enabling technologies-an overview," *Sustainable Cities Soc.*, 59: 102139, 2020.
- [38] O. Skarlat, S. Schulte, "FogFrame: a framework for IoT application execution in the fog," *PeerJ Comput. Sci.*, 7(22): e588, 2021.
- [39] D. A. Chekired, L. Khoukhi, "Multi-tier fog architecture: A new delay-tolerant network for IoT data processing," in *Proc. 2018 IEEE International Conference on Communications (ICC)*, 2018.

Biographies



Samira Kalantary received her B.S. and M.S. degree from Islamic Azad University, Arak, Iran, in 2003 and 2007 respectively, both in Computer Engineering. She is currently studying her Ph.D. in the Department of Computer Engineering, Islamic Azad University, Qom branch, Iran. Her research interests are in the area of Internet of things, fog computing, resource management and network.

- Email: kalantarysamira@gmail.com
- ORCID: [000-0002-1641-8626](https://orcid.org/000-0002-1641-8626)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Javad Akbari Torkestani received the Ph.D. degree in computer science from the Islamic Azad University in 2010. He is Associate professor in Islamic Azad university, Arak branch currently. His publication topics and his research interests are Intelligent system, Optimization, Grid computing and Web Engineering.

- Email: ja.akbari@iau.ac.ir
- ORCID: [0000-0002-6075-4889](https://orcid.org/0000-0002-6075-4889)
- Web of Science Researcher ID: NA
- Scopus ID: 25222831700
- Homepage: NA



Ali Shahidinejad received the Ph.D. degree in computer science from the University Technology Malaysia, Johor Bahru, Malaysia, in 2015. He is currently working toward the second Ph.D. degree with Deakin University, Melbourne, Australia. His publication topics are Internet, Internet of Things, and distributed processing.

- Email: a.shahidinejad@gmail.com
- ORCID: [0000-0003-4856-9119](https://orcid.org/0000-0003-4856-9119)
- Web of Science Researcher ID: NA
- Scopus ID: 54586030700
- Homepage: NA

How to cite this paper:

S. Kalantary, J. Akbari Torkestani, A. Shahidinejad, "New platform for IoT application management based on fog computing," J. Electr. Comput. Eng. Innovations, 11(2): 399-408, 2023.

DOI: [10.22061/jecei.2023.9489.626](https://doi.org/10.22061/jecei.2023.9489.626)

URL: https://jecei.sru.ac.ir/article_1850.html





Research paper

DPRSMR: Deep learning-based Persian Road Surface Marking Recognition

S. H. Safavi^{1,*}, M. Sadeghi², M. Ebadpour²

¹ Faculty of Advanced Technologies, University of Mohaghegh Ardabili, Namin, Iran.

² Faculty of Engineering, University of Mohaghegh Ardabili, Ardabil, Iran.

Article Info

Article History:

Received 16 Dec 2022
Reviewed 28 Jan 2023
Revised 01 March 2023
Accepted 11 March 2023

Keywords:

Self-driving technology
Scene understanding
Persian Road Surface Marking Recognition
Deep Learning
Alex-Net
VGG

*Corresponding Author's Email
Address: h.safavi@uma.ac.ir

Abstract

Background and Objectives: Persian Road Surface Markings (PRSMs) recognition is a prerequisite for future intelligent vehicles in Iran. First, the existence of Persian texts on the Road Surface Markings (RSMs) makes it challenging. Second, the RSM could appear on the road with different qualities, such as poor, fair, and excellent quality. Since the type of poor-quality RSM is variable from one province to another (i.e., varying road structure and scene complexity), it is a very essential and challenging task to recognize unforeseen poor-quality RSMs. Third, almost all existed datasets have imbalanced classes that affect the accuracy of the recognition problem.

Methods: To address the first challenge, the proposed Persian Road Surface Recognizer (PRSR) approach hierarchically separates the texts and symbols before recognition. To this end, the Symbol Text Separator Network (STS-Net) is proposed. Consequently, the proposed Text Recognizer Network (TR-Net) and Symbol Recognizer Network (SR-Net) respectively recognize the text and symbol. To investigate the second challenge, we introduce two different scenarios. Scenario A: Conventional random splitting training and testing data. Scenario B: Since the PRSM dataset include few images of different distance from each scene of RSM, it is highly probable that at least one of these images appear in the training set, making the recognition process easy. Since in any province of Iran, we may see a new type of poor quality RSM, which is unforeseen before (in training set), we design a realistic and challengeable scenario B in which the network is trained using excellent and fair quality RSMs and tested on poor quality ones. Besides, we propose to use the data augmentation technique to overcome the class imbalanced data challenge.

Results: The proposed approach achieves reliable performance (precision of 73.37% for scenario B) on the PRSM dataset. It significantly improves the recognition accuracy up to 15% in different scenarios.

Conclusion: Since the PRSMs include both Persian texts (with different styles) and symbols, prior to recognition process, separating the text and symbol by a proposed STS-Net could increase the recognition rate. Deploying new powerful networks and investigating new techniques to deal with class imbalanced data in the recognition problem of the PRSM dataset as well as data augmentation would be an interesting future work.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

The world is advancing toward a driverless future. However, self-driving car technology is still in its infancy stage, especially in Iran, and it cannot be deployed on urban traffic-filled roads yet. Although the fully autonomous car is not released yet in Iran, but the Advanced Driver Assistance Systems (ADAS) can help to achieve some levels of automation. Road intelligence in ADAS can be achieved by computer vision techniques. It enables self-driving vehicles to identify obstacles, traffic signs, and road markings, which avoid collisions and accidents. Road Surface Markings (RSMs) refer to the symbols or texts painted on the road surface with the aim of traffic guidance for drivers and pedestrians. Standard RSMs include lane indication arrows, crosswalks, caution words, speed limits, etc. These markings are as important as traffic signs at the side or on top of the roads, as they enable a better understanding of autonomous vehicles about their surrounding environments.

On the other hand, if one takes, for example, Google's self-driving car (developed for the U.S.) and tries to drive it in other (Europe or Asian) countries such as Iran, it will end up in an accident since there are lots of unforeseen scenes that are needed to be learned by the recognition algorithm. Road markings mainly include texts and symbols. Although the symbol markings in different countries are similar, text markings vary between countries and depend on the country's language. In Iran, road markings include some texts in the Persian language, and there is a need to address the recognition problem of them as well as symbol markings. To this end, and to facilitate the advent of self-driving car technology on the streets of all countries, this paper takes a small step toward it.

A. Related Datasets

Recently, various datasets have been released, which include RSMs. Most of them provide only RGB camera images such as ROMA [1], Road Marking Dataset [2], Reading the road dataset [3], Tsinghua Road Marking (TRoM) [4], BDD100K [5], and PRSM [6]. ROMA (ROad MARKings) image database [1] was collected in 2008. It comprises more than 100 original images of various road scenes. Moreover, the authors in [2] gathered a new dataset for road marking detection and classification. It consists of over 1200 labeled images of road markings with bounding boxes showing the location of the markings.

Furthermore, the authors in [3] created a benchmark ground truth class annotated dataset containing 2068 images spanning the city, residential, and motorway roads and over 13099 unique annotations. This dataset contains seven symbol-type categories and does not include texts. Tsinghua Road Markings (TRoM) dataset [4] is proposed for the recognition of road marking. This

dataset has collected in Beijing municipality, China. It covers a diversity of traffic and weather conditions. In the current version of TRoM, the authors annotated 19 categories of road markings for recognition use. The BDD100K dataset [5] provides a large-scale, diverse driving video dataset with rich labels that reflects the challenges of street scene understanding. In addition to frames, it consists of GPS/IMU information to record the trajectories. Persian road surface marking (PRSM) dataset consists of 18 popular classes (6 text markings and 12 symbol markings) with the option of labeling different qualities such as excellent, fair, and poor. The whole dataset includes more than 68 thousand labeled images of RSMs. Moreover, the authors consider the rotation above 30 degrees of each road surface marking.

On the other hand, a few multimodal datasets use different sensors like the KITTI vision benchmark [7], Malaga Urban Dataset [8], and Oxford RobotCar dataset [9]-[10]. The KITTI Vision Benchmark Suite is the Karlsruhe Institute of Technology and Toyota technological institute (KITTI) dataset [7]. Also, they provide a benchmark for various autonomous vehicle applications. The KITTI suite includes images and other information for different tasks such as stereo, optical flow, visual odometry, 3D object detection, and 3D tracking. Malaga urban dataset [8] was gathered entirely in urban scenarios with a car equipped with several sensors, including one stereo camera and five laser scanners. Furthermore, the Oxford RobotCar dataset [9]-[10] contains over 100 repetitions of a consistent route through Oxford, UK, captured over a year. The dataset captures different combinations of weather, traffic, and pedestrians, along with longer-term changes such as construction and roadworks.

B. Related Works

Recently, vision-based techniques for road scene understanding such as lane detection [11]-[14], road surface marking detection and recognition [11], [15]-[17], road type classification [18], pedestrian action recognition [19], etc. have achieved great interest. Lane detection is an initial and important task to guide the car to be between lines. In [11], the authors propose a real-time integrated framework to perform lane-detection and tracking, road surface marking detection, and recognition on various datasets. In [13], the authors implement a real-time lane detection based on conventional edge features and Hough transform. Noise removal is applied using Gaussian filter and then the binary image is extracted using the Otsu algorithm. Then the Canny edge detection algorithm is followed by the Hough transform to perform lane detection. In [14], the authors propose an approach for lane detection called fully convolutional neural network (FCNN) that consists of nine convolutional layers. The runtime of implemented FCNN on Raspberry Pi reported as 3.75 seconds that is not

suitable for real-time applications. Hence, the authors suggest accelerating the processing using FPGA or neural processing units. The authors in [15] investigate the effect of illumination on road surface marking recognition, and they present a real-time method that tries to find an illumination-free representation of road surfaces. The authors in [16] benefited from the YOLOv3 object detector [20] to detect 25 classes of road surface marking over 25 thousand images collected from Google Images.

C. Limitations

However, the vision-based techniques provide the details of the scene, but the reliability of them are affected by many challenges such as different weather condition (fog, haze, rain), different lighting condition (sunny, sunset, nighttime), sudden change of lighting (in and out of the tunnel), occlusion, etc., [21]. Besides cameras, other sensors like Radar and Light Detection and Ranging (LiDAR) could enhance the reliability of detection and recognition. Furthermore, the advent of Mobile Laser Scanning (MLS) technology assists the detection task [22]. Currently, not only the available multimodal datasets are not large enough to achieve higher accuracy, but also they do not have accurate ground-truth labels. Therefore, a weakly supervised learning system for real-time lane and road marking detection using multimodal data was proposed in [12].

D. Key Contributions

In this paper, the recognition of road surface marking on the PRSM dataset [6] presented. Fig. 1, shows the different classes of the PRSM dataset. Also, the first row of Table 1, shows the basic class distribution of the PRSM dataset. The contributions of the paper can be summarized as follows:

- We propose a network architecture for recognizing PRSMs inspired by VGG16 [15] and Alex-Net [16]. We call it Persian Road Surface Recognizer-Net (PRSR-Net).
- We investigate different challenging scenarios on the

PRSM dataset. We design a realistic and interesting scenario to recognize unforeseen poor-quality road surface markings.

- To deal with the class imbalanced challenge, we propose to use the data augmentation technique.
- To achieve higher recognition accuracy, we propose to separate the text and symbol markings. The whole framework called Persian Road surface Recognizer (PRSR).

The rest of this paper is organized as follows: The second section introduces the recognition framework and describes the proposed approach. The third section describes the different scenarios considered in this paper. The fourth section gives simulation results. Finally, the fifth section concludes the paper.

Proposed Approach: Persian Road Surface Recognizer

A. Network Architecture

Training a deep neural network often takes a substantial amount of time and needs powerful hardware. Regarding speed of the training procedure and overall accuracy, we propose to use the advantages of both Alex-Net [23] and VGG [24], respectively. Although the VGG and Alex-Net are not currently state-of-the-art methods, they are still used in the core of most recent neural networks [17], [25]-[30]. Therefore, these two architectures inspired us to use the advantages of each to get acceptable accuracy in Persian road surface marking recognition. Fig. 2 illustrates the proposed network architecture for road surface marking recognition.

The proposed architecture is composed of four essential stages. The max-pooling layers are used at the end of each stage. Similar to VGG16, we apply more than one convolutional layer before each max-pooling layer. Hence, the network captures more details. Accordingly, there are two convolutional layers to get enough features in each stage. Besides, the depth of the model was essential for its high performance.

Table 1: Class distribution of the used PRSM dataset with data augmentation

	Caution Symbol	Caution Text	Crosswalk	Crosswalk Caution Symbol	Crosswalk Caution Text	Forward	Forward and Turn Left	Forward and Turn Right	School	Slow	Speed Bump	Speed Limit	Stop	Stop line	Strain Speed	Turn Left	Turn Right	Yield Line
Basic	1824	4915	27893	1180	233	5747	963	2089	1085	3893	5368	193	1773	5615	938	203	625	3519
Augmented	3897	-	-	2888	1568	-	2988	3252	2472	-	-	1600	2721	-	2272	1793	1995	-
Used	3000	2844	3000	2888	1568	3000	2988	3000	2472	2213	3000	1500	2500	3000	2200	1600	1700	2291

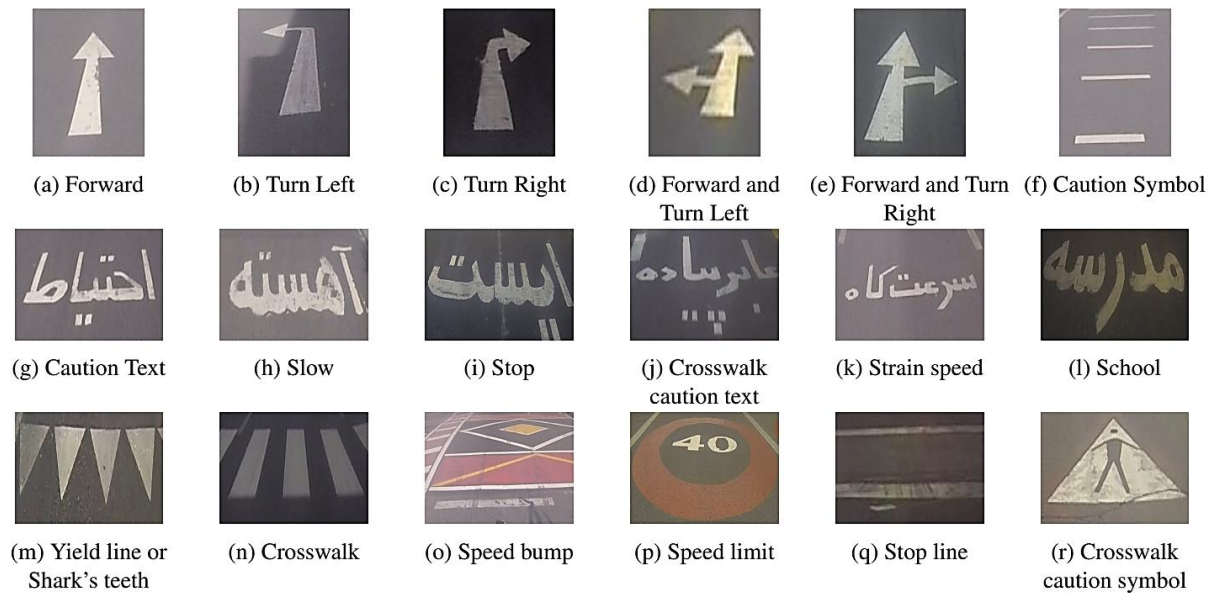


Fig. 1: Classes of PRSM dataset [6].

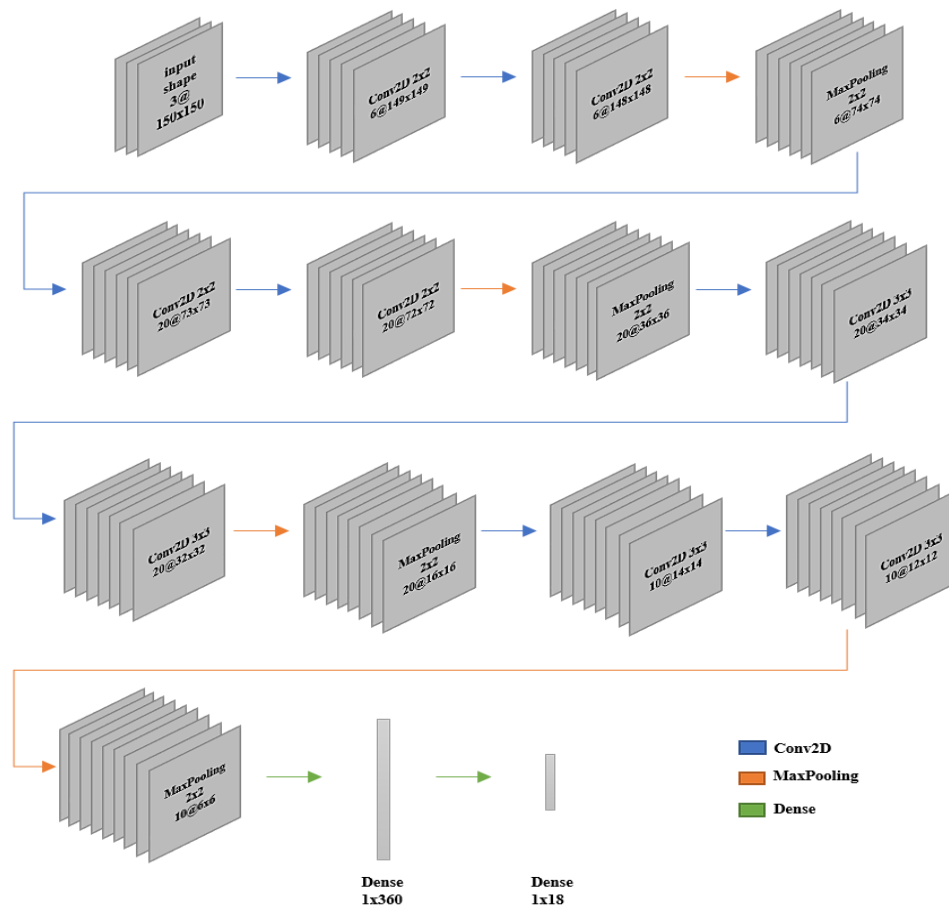


Fig. 2: Proposed network architecture for road surface marking recognition.

The Alex-Net architecture uses less depth number at the beginning and the end of the convolutional layers compared to the middle ones. Therefore, unlike VGG16, the depth number of each CNN layer is chosen based on the Alex-Net. Hence, as can be seen from Fig. 2, we use two CNN layers with a depth of six at the beginning and

two CNN layers with a depth of ten at the end. The choice of these numbers is due to making the number of nodes in the next flattened layer less to make it faster; otherwise, we could use more than ten layers as depth. The depth of the middle four CNN layers is 20 to get more details. The numbers are not exactly like the numbers in

Alex-net or the VGG16. We just used the patterns of both architectures.

B. Different Scenarios

The road markings could be partially visible, occluded, or even faded. In addition, diverse lighting conditions (sunny, shadows, and bright) can affect RSMs recognition. The PRSM dataset includes road markings with three kinds of labeled quality: Poor, Fair, and Excellent. Fig. 3 and Fig. 4 show some road markings with different qualities in this dataset. It can be seen that it is hard to recognize the poor-quality markings as an unforeseen image in the test set. Besides, environmental factors (i.e., varying road structure and scene complexity) are variable from one province to another. Moreover, the available datasets are not large enough to capture comprehensive structural variations of RSMs. Therefore, it can lead to unforeseen scenes for self-driving cars. Hence, these challenges motivated us to design a different interesting scenario in which the model should recognize the unforeseen road markings with acceptable accuracy.



Fig. 3: “Forward” Symbol class with different qualities: First row includes “Excellent” quality markings. The second row includes “Fair” quality markings. The third row includes “Poor” quality markings. All of them are selected from the PRSM dataset [6].

The following subsections include two main scenarios regarding how we choose the train and test set.

Scenario A: We train our proposed model using images with different qualities in the first scenario. We randomly choose 70% of the images as training data and the rest for the test part in each class. This splitting scenario is conventional in different machine-learning tasks. The PRSM dataset contains various quality images (excellent, fair, and poor) in the learning phase. Therefore, the highest accuracy is expected to achieve in this scenario

compared with scenarios that are not learned using images with all kinds of quality.



Fig. 4: “Caution Text” class with different qualities: First row includes “Excellent” quality markings. The second row includes “Fair” quality markings. The third row includes “Poor” quality markings. All of them are selected from the PRSM dataset [6].

Scenario B: In this scenario, we train our model with excellent and fair-quality images, and then the model is validated with poor-quality images. Compared to Scenario A, it is expected to have a low accuracy due to the difficulty of the scenario. Training a model with high-quality images and testing them on poor images that the model has never seen would have been less accurate. However, we tried several approaches in this scenario, containing practical and new techniques, aiming to get better results. In this regard, we propose the following methods:

1. Training the model with a balanced dataset (by eliminating extra images).
2. Data Augmentation.
3. Separating text and symbols.

Training the model with a balanced dataset: In real life, the RSMs do not appear equally. We expect to see the “CrossWalk” symbol more than, e.g., the “School” symbol. Table 1 shows the class distribution of road markings in the PRSM dataset. We observe that the “CrossWalk” class contains more than 27000 images, however, the “Caution Symbol” includes only about 1800 images. We believe these differences could lead to a model which learned more in the “CrossWalk” class rather than the “Caution Symbol” class. Therefore, first, we prefer to choose a limited number of images from each to create a more balanced dataset. Hence, we could create a model which learned equally over classes. It is also worth noting that [6] used only 10000 images of the “CrossWalk” class.

Data Augmentation: Balanced dataset is preferred as long as it provides enough information for recognition.

Although the class imbalanced problem can be avoided by elimination of extra images, the overall recognition accuracy would be decreased. Therefore, Data Augmentation technique could be used to produce additional new images for classes that have less than 2100 images. To augment these classes, we used an image generator in Keras. The following parameters were applied to create new images: Height shift of 0.1, width shift of 0.1, rotation range of 7, shear range of 0.15, zoom range of 0.1, and some brightness alters. After generating new images, we fit them into PRSR-Net. Table 1 represents the number of produced images for the mentioned classes. In addition, we still use limited numbers of images from some classes to prevent the model from over-learning in some classes.

Separating text and symbols: Similar to other datasets, the PRSM dataset includes two main categories: symbolic classes and text classes. For example, “Stopline” is symbolic, while “slow” is text-based. Fig. 1 illustrates an example of them. All images of the second row in this figure are from text classes. Keep in mind that text classes in the PRSM dataset are Persian. Inspecting the experimental results of previous techniques on the PRSM dataset, we observe some misclassified symbolic classes with text classes and vice versa. Therefore, we propose a novel hierarchical approach. If the recognition process is decomposed into two deep learning models, one for text recognition and one for symbol recognition, we expect an improvement in accuracy. In summary, we propose three deep models:

1. Symbol Text Separator (STS-Net)
2. Text Recognizer (TR-Net)
3. Symbol Recognizer (SR-Net)

Fig. 5 demonstrates the block diagram of the proposed approach. Because we need to train three different models with different outputs, it is clear that we will need different architectures, varying from the basic one represented in previous techniques. However, they are still taken from the basic model. STS-Net has only two outputs. So training this can be easier than the other ones. Because of output reductions, we eliminate two last Conv layers ($20 \times 20 \times 3$) from our basic model. Hence, it leads to less time and resources used for training.

Table 2: Different Scenarios defined in this paper

	Train Set	Test Set	Network Architecture	Num. of epochs	Data augmentation	Overall Accuracy
Scenario A	Poor, Fair, Excellent	Poor, Fair, Excellent	Fig. 2	35	No	99.15
Scenario B	Fair, Excellent	Poor	Fig. 5, Fig. 6	35	Yes	65
Scenario C	Fair, Excellent	Poor	Fig. 5, Fig. 6	35	No	73.37

Fig. 7 represents the accuracy changes in different epochs and loss function variation. The confusion matrix is also shown in Fig. 8. As it shows, the most challenging classes for the model were Turn Left and Crosswalk

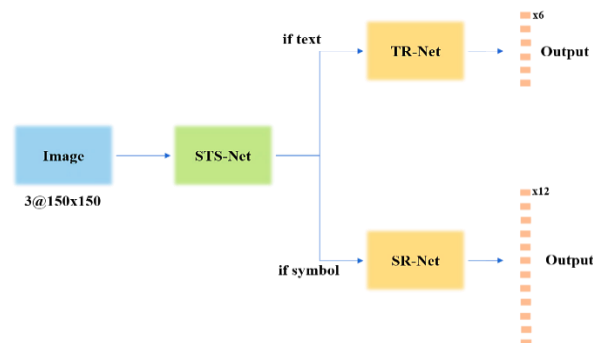


Fig. 5: Block Diagram of the proposed approach.

Fig. 6 represents the STS-Net architecture. About other models, based on the expected number of their outputs, we just altered the last dense layer. In TR-Net, the number of outputs should be six (the number of classes that are text classes) rather than eighteen in basic architecture. Similarly, in SR-Net we changed the last dense layer of the basic model and reduced it from 18 to 12 (the number of classes that are symbolic).

Results and Discussion

In this section, we present the results of the proposed methods. We also compare the results with earlier work [6]. For all of our four presented models, we apply the Adam optimizer. The back-propagation process uses the cross-entropy loss function. Training is done using different epochs, which we describe in the following sections. To make the procedure fast enough, we used the GPU version of Tensorflow, using Keras-GPU as API in python 3.5. GPU used in this article was NVIDIA GeForce 930MX with its 2G RAM. CPU was Intel corei7 and a DDR4 RAM with 8G of capacity.

In the following, different simulation scenarios are investigated and Table 2 summarizes them.

A. Scenario A

As it was explained before, this scenario was more convenient for the model to learn. Although all images were used, random images choosing and learning from poor-quality images helped us get better output, classifying and predicting unforeseen poor-quality images. In 35 epochs, we could gain almost 99.15% learning accuracy and 97% validation accuracy.

Caution Text. PRSR-Net misclassified almost 20% of their images. Training more epochs or using data augmentation on the most misclassified classes would be a better solution to improve the model.

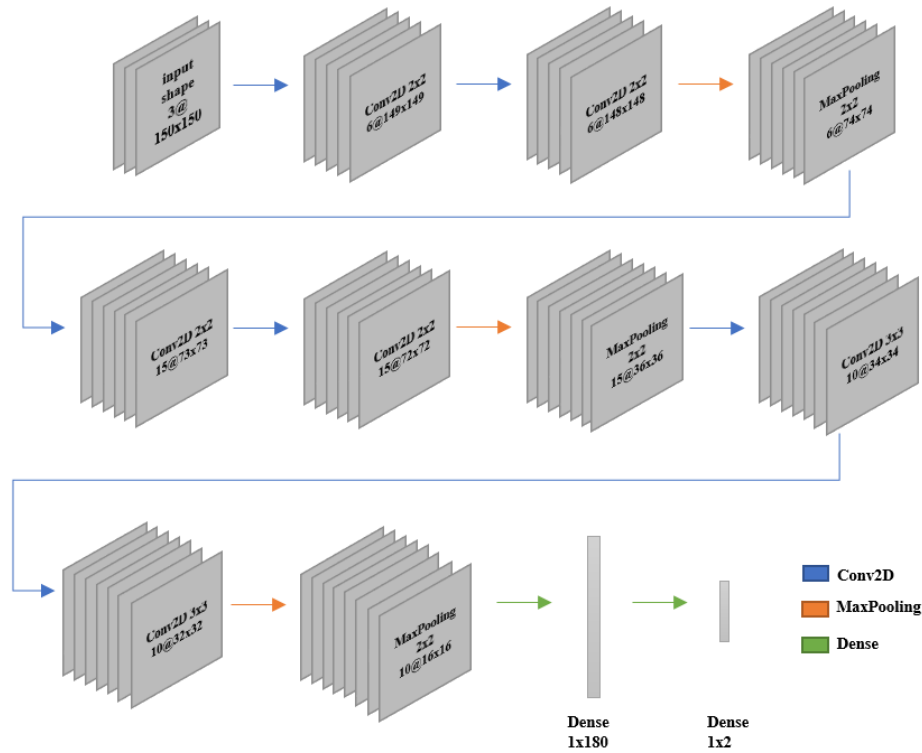


Fig. 6: Proposed Network Architecture for Symbol Text Separator (STS-Net).

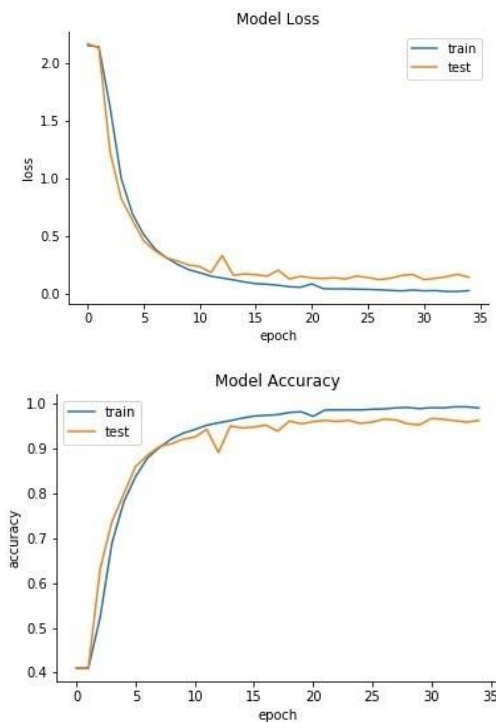


Fig. 7: Model Loss and Accuracy for Scenario A.

Fig. 9 represents some of the misclassified images. Obviously, they are hardly classified even by a human. Compared to the previous works, there are some advantages here. Since we see the “Crosswalk” class more in real life, rather than, e.g., the “School” class, it is supposed to have a large number of images in the dataset. Our model’s validation accuracy for this class was

about 98%, while earlier work classified only 88% of this class correctly. “Speed Limit” is another example. The presented model reached 97% accuracy in this class, while previous work did only 91%. For those with less accuracy, we should take this point into account how deep neural networks work. As we all know, deep learning models need more data than other classifiers. Having said this, a lack of enough data could lead our model to misclassify some classes more than others. As we suggested, data augmentation on these classes can solve the problem and raise accuracy.

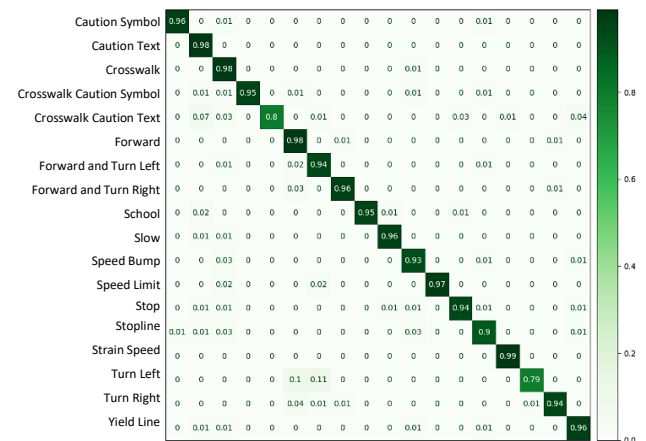


Fig. 8: Confusion matrix for the scenario A.

B. Scenario B

Training a balanced dataset: As we mentioned earlier, we created a balanced training dataset, and we restricted the number of images in classes that the proportion of

them was unacceptable, intending to reduce consumed time and fewer computations required.



Fig. 9: misclassified examples.

Pointing out again, we trained PRSR-Net (our basic model) with excellent and fair-quality images and tested them with poor-quality images. With these in mind, we got almost 64% validation accuracy in 35 epochs. Although it is still not the desirable accuracy, it is 8% more than the best result of previous work [6], which was 58%. In addition, note that we achieved this accuracy using less number of images compared to [6]. Moreover, we used data augmentation to add more images to classes that hadn't as many as other classes. The general points of this part are described earlier.

During 35 epochs, we got almost 65% accuracy. This accuracy is noteworthy due to the less-used number of images compared to the last technique. In other words, using all images, we would have achieved a more accurate model. However, it increased our accuracy by about 1%. Fig. 10 shows the confusion matrix of this method.

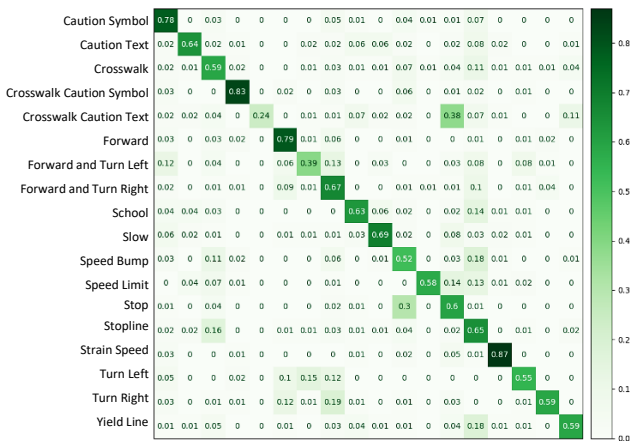


Fig. 10: Confusion matrix for the proposed method: scenario B.

C. Scenario C

Separating text and symbols: In this approach, we trained three different models. Starting from STS-Net, we

trained it in 20 epochs, getting almost 99% validation accuracy. SR-Net and TR-Net were both trained in 35 epochs, achieving validation accuracy of almost 73.5% and 75.2%, respectively. Overall accuracy was about 74%, being higher than every presented technique in scenario B and also 16% more accurate than [6]. The block diagram of this approach is already shown in the earlier section. Fig. 11 shows the confusion matrix of this method.

“Stop Line” and “Forward and Turn Right” with 21% and 28% accuracies, respectively, were misclassified the most and had not satisfying results. Regardless of these classes, the others gained better outputs. E.g., Yield Line, Crosswalk, and Forward are classes that have almost perfect results. Take “Turn Left” as an example. The validation accuracy of this class in [6] is about 28%, while the proposed method could achieve a significant performance which is 80% validation accuracy.

Finally, Table 3, summarizes the number of parameters used in the proposed PRSR Network. Moreover, each frame is processed at about 15 msec.

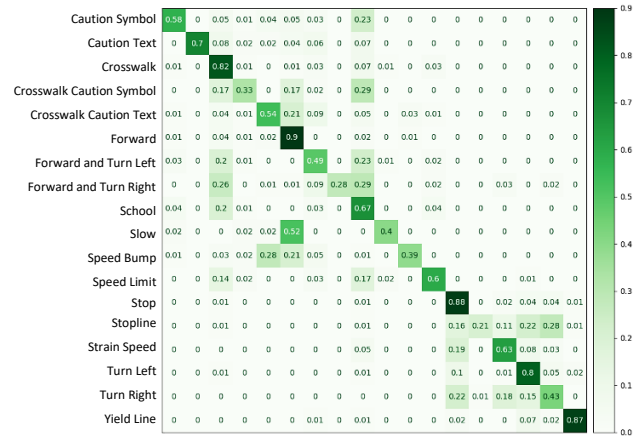


Fig. 11: Confusion matrix for the proposed method: scenario C.

Table 3: Number of parameters in the PRSR Network of Fig. 2

Layer	Output Shape	Number of Parameters
Conv_2D	(149,149,6)	78
Conv_2D	(148,148,6)	150
Max_Pooling_2D	(74,74,6)	0
Conv_2D	(73,73,20)	500
Conv_2D	(72,72,20)	1620
Max_Pooling_2D	(36,36,20)	0
Conv_2D	(34,34,20)	3620
Conv_2D	(32,32,20)	3620
Max_Pooling_2D	(16,16,20)	0
Conv_2D	(14,14,10)	1810
Conv_2D	(12,12,10)	910
Max_Pooling_2D	(6,6,10)	0
Flatten	360	0
Dense	360	129960
Dense	18	6498
Total Trainable Parameters		148766

Conclusion

In this paper, inspired by Alex-Net and VGG, a deep-learning approach was developed to overcome the recognition challenge of the PRSM dataset. Since the PRSMs include both Persian texts (with different styles) and symbols, prior to recognition process, separating the text and symbol by a proposed STS-Net could increase the recognition rate. Moreover, we design a realistic and challengeable scenario in which the network is trained using excellent and fair quality RSMs and tested on poor quality ones. The proposed approach achieves reliable performance (precision of 73.37%) on this dataset. It significantly improves the recognition accuracy by up to 15% in different scenarios compared to [6]. Deploying new powerful networks and investigating new techniques to deal with class imbalanced data in the recognition problem of the PRSM dataset as well as data augmentation would be an interesting future work.

Author Contributions

S. H. Safavi raise the idea. All authors equally contribute in designing the experiments. M. Sadeghi and M. Ebadpour run the simulations. All authors involve in data analysis and interpreting the results and S. H. Safavi wrote the initial draft of the manuscript and revised the manuscript. This paper was a research collaboration of M. Sadeghi and M. Ebadpour as overplus university activities of them during their B.Sc.

Acknowledgment

The authors would like to thank both the Editor and anonymous Reviewers for their insightful comments and suggestion to improve the quality of the paper.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

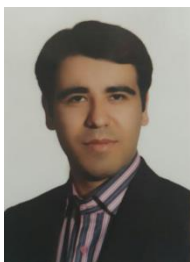
<i>DPRSMR</i>	Deep learning-based Persian Road Surface Marking Recognition
<i>CNN</i>	Convolutional Neural Networks
<i>PRSM</i>	Persian road surface marking
<i>PRSR</i>	Persian Road Surface Recognizer
<i>STS-Net</i>	Symbol Text Separator Network
<i>TR-Net</i>	Text Recognizer Network
<i>SR-Net</i>	Symbol Recognizer Network

References

- [1] T. Veit, J. P. Tarel, P. Nicolle, and P. Charbonnier, "Evaluation of road marking feature extraction," in Proc. 11th International IEEE Conference on Intelligent Transportation Systems: 174–181, 2008.
- [2] T. Wu, A. Ranganathan, "A practical system for road marking detection and recognition," in Proc. IEEE Intelligent Vehicles Symposium: 25–30, 2012.
- [3] B. Mathibela, P. Newman, I. Posner, "Reading the road: Road marking classification and interpretation," IEEE Trans. Intell. Transp. Syst., 16(4): 2072–2081, 2015.
- [4] X. Liu, Z. Deng, H. Lu, L. Cao, "Benchmark for road marking detection: Dataset specification and performance baseline," in Proc. IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), 2017.
- [5] F. Yu *et al.*, "BDD100K: A diverse driving video database with scalable annotation tooling," arXiv preprint arXiv:1805.04687, 2018.
- [6] S. H. Safavi *et al.*, "Image dataset for persian road surface markings," in Proc. IEEE 10th Iranian Conference on Machine Vision and Image Processing (MVIP): 258–264, 2017.
- [7] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, "Vision meets Robotics: The KITTI Dataset," Int. J. Rob. Res., 32(11): 1231–1237, 2013.
- [8] J. L. Blanco, F. A. Moreno, J. Gonzalez-Jimenez, "The Málaga urban dataset: High-rate stereo and lidars in a realistic urban scenario," Int. J. Rob. Res., 33(2): 207–214, 2014.
- [9] W. Maddern, G. Pascoe, C. Linegar, P. Newman, "1 Year, 1000km: The Oxford RobotCar Dataset," Int. J. Rob. Res. (IJRR), 36(1): 3–15, 2017.
- [10] W. Maddern, G. Pascoe, M. Gadd, D. Barnes, B. Yeomans, P. Newman, "Real-time kinematic ground truth for the oxford robotcar dataset," in arXiv preprint arXiv: 2002.10152, 2020.
- [11] A. Gupta, A. Choudhary, "A framework for camera-based real-time lane and road surface marking detection and recognition," IEEE Trans. Intell. Veh., 3(4): 476–485, 2018.
- [12] T. Bruls, W. Maddern, A. A. Morye, P. Newman, "Mark yourself: Road marking segmentation via weakly-supervised annotations from multimodal data," in Proc. IEEE International Conference on Robotics and Automation (ICRA): 1863–1870, 2018.
- [13] A. Fallah, A. Soliemani, H. Khosravi, "Real-time lane detection based on image edge feature and hough transform," J. Electr. Comput. Eng. Innovations (JECEI), 9(2): 193–202, 2021.
- [14] N. S. Danishevskiy, I. A. Ershov, D. O. Budanov, "Computer vision system for road surface marking recognition," in Proc. IEEE International Conference on Electrical Engineering and Photonics (EEEPolytech): 130–133, 2022.
- [15] B. A. Maxwell *et al.*, "Real-time physics-based removal of shadows and shading from road surfaces," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): 1277–1285, 2019.
- [16] E. S. Dawam, X. Feng, "Smart city lane detection for autonomous vehicle" in Proc. IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech): 334–338, 2020.
- [17] M. R. Bachute, J. M. Subhedar, "Autonomous driving architectures: insights of machine learning and deep learning algorithms," Elsevier, Machine Learning with Applications, 6(100164): 1–25 2021.
- [18] D. K. Dewangan, S. P. Sahu, "RCNet: road classification convolutional neural networks for intelligent vehicle system," Intell. Serv. Rob., 14(2): 199–214, 2021.

- [19] R. D. Brehar, M. P. Muresan, T. Marița, C. C. Vancea, M. Negru, S. Nedevschi, "Pedestrian street-cross action recognition in monocular far infrared sequences," *IEEE Access*, (9): 74302-74324, 2021.
- [20] J. Redmon, A. Farhadi, "Yolov3: An incremental improvement", *arXiv preprint arXiv:1804.02767*, 2018.
- [21] Z. Feng, M. Li, M. Stolz, M. Kunert, W. Wiesbeck, "Lane detection with a high-resolution automotive radar by introducing a new type of road marking," *IEEE Trans. Intell. Trans. Syst.*, 20(7): 2430-2447, 2018.
- [22] S. Chen, Z. Zhang, H. Ma, L. Zhang, R. Zhong, "A content-adaptive hierarchical deep learning model for detecting arbitrary-oriented road surface elements using MLS point clouds," *IEEE Trans. Geosc. Remote Sens.*, 61(5700516): 1-16, 2023.
- [23] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in neural information processing systems (NIPS 2012)*: 1097–1105, 2012.
- [24] K. Simonyan. A. Zisserman, "Very deep convolutional networks for largescale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [25] R. Girshick, "Fast r-cnn," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*: 1440-1448, 2015.
- [26] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems (NIPS)*, (28):1-9, 2015.
- [27] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 39(6):1137-1149, 2017.
- [28] X. Ding, X. Zhang, N. Ma, J. Han, G. Ding, J. Sun, "RepVGG: Making VGG-style ConvNets great again," in *Proc. IEEE/CVF conference on computer vision and pattern recognition (CVPR)*: 13733-13742, 2021.
- [29] L. Chen *et al.*, "Deep integration: A multi-label architecture for road scene recognition," *IEEE Trans. Image Proc.*, 28(10): 4883–4898, 2019.
- [30] K. Lis *et al.*, "Detecting the unexpected via image resynthesis" in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*: 2152-2161, 2019.

Biographies



Seyed Hamid Safavi is an Assistant Professor at University of Mohaghegh Ardabili (UMA). He received his PhD in electrical engineering from Shahid Beheshti University (SBU), Tehran, Iran in 2017. He also received his B.Sc. and M.Sc. degrees both in electrical engineering from K. N. Toosi University of Technology (KNTU), Tehran, Iran in 2010 and 2012, respectively. He was also a visiting fellow at Singapore University of Technology and Design (SUTD) from 2015 to 2016. He received the best Ph.D.

thesis award from IEEE SBU Student Branch. He also received the IEEE travel grant to participate at the 1st IEEE ComSoc Summer School in 2015. His research interests include Signal and Image Processing, Machine Learning, Compressive Sensing, and Convex Optimization.

- Email: h.safavi@uma.ac.ir
- ORCID: [0000-0001-7833-7381](https://orcid.org/0000-0001-7833-7381)
- Web of Science Researcher ID: B-7187-2014
- Scopus Author ID: 53864204500
- Homepage: https://uma.ac.ir/cv.php?slc_lang=fa&sid=1&mod=scv&cv=2513



Mahyar Sadeghi received his B.Sc. from the University of Mohaghegh Ardabili (UMA) in computer engineering. He is currently a M.Sc. student in computer science -Data Engineering and Artificial Intelligence in university of Genova, Italy. He is also a research fellow in this university, working on "AI agent-based model oriented to traffic control using CAVs". His research interests include Reinforcement Learning, Machine Learning, Multi Agent Systems and brain inspired Artificial Intelligence.

- Email: mahyarsadeghi.ga@gmail.com
- ORCID: [0000-0003-0388-8669](https://orcid.org/0000-0003-0388-8669)
- Web of Science Researcher ID: HDM-4187-2022
- Scopus Author ID: NA
- Homepage: <https://sites.google.com/view/mahyarsadeghi-ga/home?pli=1>



Mohsen Ebadpour received his B.Sc. from the University of Mohaghegh Ardabili (UMA) in computer engineering. He is currently a M.Sc. student in artificial intelligence at Amirkabir University of Technology (AUT) and a research fellow at the Image Processing and Pattern Recognition (IPPR) laboratory. Currently, he is working on "Few-Shot Learning in Multi-Label Classification." His research interests include Image Processing, Computer Vision, Deep Learning and Generative models, Graph Theory, and Complex Networks.

- Email: m.ebadpour@aut.ac.ir
- ORCID: [0000-0003-2745-7329](https://orcid.org/0000-0003-2745-7329)
- Web of Science Researcher ID: HDN-2945-2022
- Scopus Author ID: NA
- Homepage: <https://ce.aut.ac.ir/~ebadpour/>

How to cite this paper:

S. H. Safavi, M. Sadeghi, M. Ebadpour, "DPRSMR: Deep learning-based persian road surface marking recognition," *J. Electr. Comput. Eng. Innovations*, 11(2): 409-418, 2023.

DOI: [10.22061/jecei.2023.9496.627](https://doi.org/10.22061/jecei.2023.9496.627)

URL: https://jecei.sru.ac.ir/article_1851.html





Research paper

Determination of the Maximum Dynamic Range of Sinusoidal Frequencies in A Wireless Sensor Network with Low Sampling Rate

A. Maroosi^{1,*}, H. Khaleghi Bizaki²

¹Department of Computer Engineering, University of Torbat Heydarieh, Torbat Heydarieh, Iran.

²Department of Electrical and Computer Engineering, Malek Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:

Received 06 January 2023

Reviewed 05 March 2023

Revised 20 March 2023

Accepted 11 May 2023

Keywords:

Chinese Remainder Theorem (CRT)

Multi-sensor system

Under sampling

Signal reconstruction

*Corresponding Author's Email Address:

ali.maroosi@torbath.ac.ir

Abstract

Background and Objectives: Subsampling methods allow sampling signals at rates much lower than Nyquist rate by using low-cost and low-power analog-to-digital converters (ADC). These methods are important for systems such as sensor networks that the cost and power consumption of sensors are the core issue in them. The Chinese remainder theorem (CRT) reconstructs a large integer (input frequency) from its multiple remainders (aliased or under-sampled frequencies), which are produced from under-sampling or integer division by several smaller positive integers. Sampling frequencies can be reduced by approaches based on CRT.

Methods: The largest dynamic range of a generalized Chinese remainder theorem for two integers (input frequencies) has already been introduced in previous works. This is equivalent to determine the largest possible range of the frequencies for a sinusoidal waveform with two frequencies which the frequencies of the signal can be reconstructed uniquely by very low sampling frequencies. In this study, the largest dynamic range of CRT for any number of integers (any number of frequencies in a sinusoidal waveform) is proposed. It is also shown that the previous largest dynamic range for two frequencies in a waveform is a special case of our proposed procedure.

Results: A procedure for multiple frequencies detection from reminders (under-sampled frequencies) is proposed and maximum tolerable noises of under-sampled frequencies for unique detection is obtained. The numerical examples show that the proposed approach, in some cases, can gain 11.5 times higher dynamic range than the conventional methods for a multi-sensor under-sampling system.

Conclusion: Other studies introduced the largest dynamic range for the unique reconstruction of two frequencies by CRT. In this study, the largest dynamic ranges for any number of frequencies are investigated. Moreover, tolerable noise is also considered.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

The Chinese Remainder Theorem (CRT) is a well-known research topic, which reconstructs a large positive integer from its remainders [1]-[3]. Nowadays, CRT is widely used in different applications including signal processing, image

processing, etc. In our previous works, the high range of frequency is estimated by sensors with a very low sampling rate [1], [2]. In [4], the CRT algorithm is used to achieve range estimation of multiple targets in a pulse Doppler radar when the measured ranges are overlapped with noise error. An approach based on CRT is introduced

in [5] to estimate frequencies when a signal is under sampled by multiple under-sampling frequencies.

In [6], the statistical model of CRT-based multiple parameter estimation is investigated, and two approaches are introduced to address the problems of ambiguity resolution in parameter estimation.

A method based on CRT is introduced in [7] to estimate the direction of arrival (DOA) of the signal. This algorithm has less complexity with similar precision in comparison with other algorithms for DOA estimation.

In [8], CRT and non-orthogonal multiple access (NOMA) techniques are introduced for unmanned aerial vehicle (UAV) relay networks to improve communication between transmitter and receiver.

A combination of CRT with Haar Wavelet Transform was proposed as a watermarking technique in [9] to hide information.

The Haar Wavelet Transform has been used for imperceptibility, and CRT provides the security of the watermarked image. A reversible sketch data structure based on CRT was proposed in [10] to compress and fuse big data network traffic. In [11], a novel CRT-based conditional privacy was introduced to keep an authentication scheme for securing vehicular authentication.

In this work, the CRT can help the trusted authorities to generate and broadcast group keys to the network vehicles.

In [12], the authors proposed a multiple secret image-sharing scheme by CRT and Boolean exclusive-OR operation.

A robust and secure data-hiding method in the Tchebichef domain is presented based on CRT [13]. The efficiency of the algorithm was confirmed by implementing the algorithm over different images.

Power efficiency is one of the critical design factors in wireless sensor network systems. In such systems, it is possible to digitalize received analog signal by sensors with very low frequencies and use CRT for manipulation and reconstruction of the frequencies of the main signal. In [14] a low-frequency power efficient digital signal processing architecture for mathematic operations based on CRT was designed and implemented.

A packet forwarding scheme based on CRT was developed for wireless sensor networks in [15]. The advantages of this scheme are energy efficiency, low computational complexity and high reliability.

In [16], an approach based on the frequency domain sparse common support and CRT was developed for frequency determination of multiple sinusoidal signals when the sampling rate even less than Nyquist rate. Authors in [17] proposed an approach to reconstruct the multiple frequencies of a sinusoidal waveform from aliased frequencies by the CRT approach.

In all these researches the dynamic range for unambiguously reconstruction integers (e.g. frequencies), which are divided by a set of modules (e.g. sampling frequencies) from their remainders (e.g. aliased frequencies) is important.

The higher dynamic range for a set of modules means the possibility to reconstruct the larger range of integers unambiguously by remainder of integers from those modules. Thus, any improvement in the dynamic range will lead to more efficient schemes in many applications [3].

The dynamic range for the unique determination of an integer (frequency) N_1 with modules (sampling frequencies) $\Gamma = \{m_1, m_2, \dots, m_\gamma\}$ is the least common multiple (lcm) of modules i.e., $d = \text{lcm}(m_1, m_2, \dots, m_\gamma)$ [5], [18]. A dynamic range for the unique determination of two integers (frequencies) N_1 and N_2 can be obtained as $d = \min_{I_1, I_2} \{\max\{I_1, I_2\}\}$ where $I_1 \cup I_2 = \Gamma$ [19]. The first generic dynamic range for reconstruction of multiple integers (more than two integers (frequencies)) from their modules was introduced in [20].

A sharpened dynamic range for ρ integers ($\rho = 1, 2, \dots$) was presented as $d = \min_{I_1, \dots, I_\rho} \{\max\{I_1, \dots, I_\rho\}\}$ where

$\bigcup_{i=1}^{\rho} I_i = \Gamma$ in [19]. Dynamic ranges for multiple integers when there are conditions over integers are presented in [21], [22].

The largest dynamic range for two integers is obtained as $d = \min_{I_1, I_2} \{I_1 + I_2\}$ where $I_1 \cup I_2 = \Gamma$ in [23] and the maximum tolerable error for two integers was discussed in [24] and it is applied in [25] for the direction of arrival (DOA) of two sources and in [26], [27] was used for secret image sharing by the modular operation.

Most of the previous studies discussed the unambiguous dynamic range for two integers (frequencies) or assumed conditions for integers (frequencies) [21], [22], [28] that will be discussed with details in the background section while we present a close form relationship of the largest dynamic range for multiple integers (frequencies) without condition on them. Furthermore, we show that the largest dynamic range for reconstruction of two integers (frequencies) is a special case of our work.

The presentation is organized as follows. Related works with theoretical background is discussed in Background section.

A proposition for finding the maximum possible range for unique reconstruction of any number of input frequencies from under-sampled frequencies is introduced in Proposed Approach section. Furthermore, the proposed proposition is specified for two and three input frequencies in corollaries, the maximum tolerable noise for the maximum possible range is obtained and a

procedure for reconstruction is also introduced in the Proposed Approach section.

Different numeric examples to verify the effectiveness of the proposed approaches are introduced in the Simulation Results section.

Finally, the work is concluded in the Conclusion section.

Background

Consider a complex waveform without noise as follows [23]:

$$x(t) = \sum_{l=1}^{\rho} A_l e^{i(2\pi F_l t + \phi_l)} + w(t) \quad (1)$$

where A_l 's are unknown nonzero complex coefficients and F_l 's; $1 \leq l \leq \rho$ are multiple unknown frequencies in Hz that should be determined. The $w(t)$ is additive white Gaussian noise.

Consider γ sensors in a wireless sensor network with $\Gamma = \{f_{s1}, f_{s2}, \dots, f_{s\gamma}\}$; $\gamma \geq 2$ sampling rates as Fig. 1 in which all may be much less than the unknown frequencies i.e. $f_{si} \ll F_l$; $i = 1, \dots, \gamma$; $l = 1, \dots, \rho$ [1], [2].

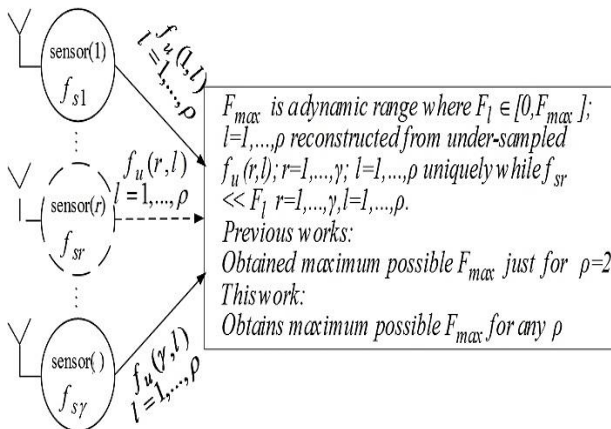


Fig. 1: A multi-sensor system for the determination of frequencies by information fusion from sensors.

Assume these sampling frequencies are co-prime i.e. $M = \text{lcm}(f_{s1}, f_{s2}, \dots, f_{s\gamma}) = f_{s1}f_{s2}\dots f_{s\gamma}$ and without loss of generality assume $f_{s1} < f_{s2} < \dots < f_{s\gamma}$ that lcm is the least common multiplier.

Then, the multiple under-sampled waveforms by sampling frequency f_{sr} ; $r = 1, \dots, \gamma$ are given by [23]:

$$x_{f_{sr}}(n) = \sum_{l=1}^{\rho} A_l e^{2\pi j F_l n / f_{sr}}; \quad n \in \mathbb{Z} \quad (2)$$

Using the f_r -point discrete Fourier transform (DFT) to $x_{f_{sr}}(n)$, relation (2) can be written as:

$$DFT_{f_{sr}}(x_{f_{sr}}(n))[k] = \sum_{l=1}^{\rho} A_l \delta(k - f_{u(l,r)}), \quad (3)$$

$$1 \leq r \leq \gamma$$

where $\delta(k)$ is equal to 1 when $k = 0$ and others $\delta(k) = 0$. The $f_{u(l,r)}$ is remainder (under-sampled frequency) of F_l with module (sampling frequency) f_{sr} i.e. $f_{u(l,r)} = F_l \bmod f_{sr}$.

Thus, following under-sampled frequencies set $S_r(F_1, \dots, F_\rho)$ can be written.

$$S_r(F_1, \dots, F_\rho) = \bigcup_{l=1}^{\rho} \{f_{u(l,r)}\}, \quad (4)$$

$$r = 1, \dots, \gamma$$

Consider F_{max} be an upper bound of input frequencies when all input frequencies less than F_{max} (i.e., $F_l \leq F_{max}$, $l = 1, \dots, \rho$) can be uniquely reconstructed from their remainders. Some works have been done to determine F_{max} for unambiguous reconstruction of multiple integers (multiple frequencies) from their remainders sets where we briefly review them in the sequel.

Proposition 1 [29], [30]: A large dynamic range (F_{max}) for unique determination F_l , $l = 1, \dots, \rho$ when under-sampled with f_{si} , $i = 1, \dots, \gamma$ is $F_{max} = \max(v, f_{s\gamma})$ where $v = \min_{1 \leq r_1 \leq \dots \leq r_\eta \leq \gamma} \text{lcm}\{f_{s(r_1)}, \dots, f_{s(r_\eta)}\}$ that $\gamma = \eta\rho + \theta$ for some $0 \leq \theta < \rho$.

A proposed majority method for the determination multiple integers from their moduli introduced in [19] as follows:

Proposition 2 [19]: A large dynamic range for multiple integers (multiple frequencies) is $F_{max} = \min_{I_1 \cup \dots \cup I_\rho = \Gamma} \max\{\prod_{f_{si} \in I_1} f_{si}, \dots, \prod_{f_{si} \in I_\rho} f_{si}\}$ where I_i , $i = 1, \dots, \rho$ is the partition of set $\Gamma = \{f_{s1}, \dots, f_{s\gamma}\}$ in to ρ disjoint set where $I_1 \cup \dots \cup I_\rho = \Gamma$ and $I_i \cap I_j = \emptyset$ for $1 \leq i \neq j \leq \rho$ and I_i can be an empty set.

Proposition 3 [23]: For two integers $\rho = 2$ as $\{F_1, F_2\}$ with moduli $\Gamma = \{f_{s1}, \dots, f_{s\gamma}\}$ the largest dynamic range for unambiguous reconstruction from remainders is $F_{max} = \min_{I_1, I_2} \{\text{lcm}(I_1) + \text{lcm}(I_2)\} = \min_{I_1, I_2} \{\prod_{f_{sr} \in I_1} f_{sr} + \prod_{f_{sr} \in I_2} f_{sr}\}$ where $I_1 \cup I_2 = \Gamma$.

The largest dynamic range for single integer ($\rho = 1$) is lcm of all moduli i.e. $F_{max} = \text{lcm}(f_{s1}, \dots, f_{s\gamma})$. It can be inferred from Proposition 1-3 that the dynamic range for multiple integers, in general, is less than lcm of all modules.

Thus, some works [22], [28] tried to achieve maximal possible range similar to single integer i.e. lcm of all

moduli.

To do this, they used some conditions on the multiple integers (input frequencies) or /and moduli (under-sampling frequencies) that was reviewed briefly at the following.

Proposition 4: If $F_p - F_1 < f_{s1}/2$ then $F_{max} = lcm(f_{s1}, \dots, f_{s\gamma}) = \prod_{i=1}^{\gamma} f_{si}$.

Note that this paper achieves lcm of all moduli while the condition $F_p - F_1 < f_{s1}/2$ is admitted.

Proposition 5 [22]: If $F_p - F_1 < f_{s1}$, $GCD(\rho, f_{si}) = 1$; $1 \leq i \leq \gamma - 1$ and $\rho^2 - \rho(f_{s1} + l) + (l - 1)f_{sp} > 0$ for $2 \leq l \leq \rho$ then $F_{max} = lcm(f_{s1}, \dots, f_{s\gamma}) = \prod_{i=1}^{\gamma} f_{si}$ where $f_{s1} < \dots < f_{s\gamma}$ and GCD is the greatest common division.

Note that in this case, the difference between two disjoint integers (input frequencies) should be less than the minimum sampling frequency f_{s1} (i.e. $F_p - F_1 < f_{s1}$) while for Proposition 4, it is $f_{s1}/2$.

A multiple frequencies determination for narrow bandwidth signals when the maximum difference between input frequencies (multiple integers) are less than the maximum sampling frequency (moduli) i.e. $F_p - F_1 < f_{s\gamma}$ was also proposed in [21].

Proposed Approaches

Lemma 1 [23]: If two input frequencies sets $X = \{F_1, \dots, F_p\}$ and $Y = \{F'_1, \dots, F'_\rho\}$ have the same remainder sets, i.e. $S_r(X) = S_r(Y)$, then the minimum of these integers would be zero, i.e. $\min\{X \cup Y\} = 0$, and the maximum of these integers would be $\max\{X \cup Y\} = F_{max}$ that F_{max} is a large dynamic range.

The order of remainders of each modulus is not known from output of DFT [19].

In other words if the ordered remainders set of multiple integers for r^{th} modulus be $S_r(F_1, \dots, F_p) = \bigcup_{l=1}^{\rho} \{f_{u(l,r)}\}, r = 1, \dots, \gamma$ then received the remainders set from output of DFT is $S'_r(F'_1, \dots, F'_\rho) = \bigcup_{l=1}^{\rho} \{t_{u(\delta_r(l),r)}\}, r = 1, \dots, \gamma$ where δ_r is an arbitrarily chosen onto mapping from index set $T = \{1, \dots, \rho\}$ to the indices of elements in $S_r(F_1, \dots, F_p)$. Note that when $\delta_r(l) = i$ we have $t_{u(\delta_r(l),r)} = t_{u(i,r)} = f_{u(i,r)}$.

It means l^{th} remainder of modulus r in ordered set of DFT, i.e. S'_r , corresponding to the i^{th} integer, i.e. F_i , (See Table. 1).

Table 1: Assigning the remainders from remainder set of each modulus to integers

Integer s	Mod f_{s1}	Mod f_{s2}	...	Mod $f_{s\gamma}$
F'_1	$t_{u(\delta_1(1),1)}$	$t_{u(\delta_2(1),2)}$...	$t_{u(\delta_\gamma(1),\gamma)}$
F'_2	$t_{u(\delta_1(2),1)}$	$t_{u(\delta_2(2),2)}$...	$t_{u(\delta_\gamma(2),\gamma)}$
...
F'_ρ	$t_{u(\delta_1(\rho),1)}$	$t_{u(\delta_2(\rho),2)}$...	$t_{u(\delta_\gamma(\rho),\gamma)}$

We try to find integers F'_l based on its remainders $t_{u(\delta_r(l),r)}, r = 1, \dots, \gamma$ (see Table. 1). The relationship between an integer F_i and its remainders is as follows:

$$f_{u(i,r)} = F_i \bmod f_{sr} = F_i - k_{i,r} f_{sr} \quad (5)$$

where $k_{i,r} \in \{0, 1, \dots, \lfloor F_{max} / f_{sr} \rfloor\}$.

The relationship between moduli $f_{sr}; r = 1, \dots, p$ and $f'_{u(l,r)}$ as the remainder of F'_l is as follows:

$$f'_{u(l,r)} = F'_l \bmod f_{sr} = F'_l - k'_{l,r} f_{sr} \quad (6)$$

Proposition 6: Assume two frequencies sets $X = \{F_1, \dots, F_p\}$ and $Y = \{F'_1, \dots, F'_\rho\}$ have the same under-sampled (remainder) sets i.e. $S_r(X) = S_r(Y)$ with sampling frequencies (moduli) $f_{sr}, r = 1, \dots, \gamma$. Now assume from all γ remainders for each $F_i \in X$ and $F'_l \in Y$ there are $\alpha_{(l,i)}; l = 1, \dots, \rho; i = 1, \dots, p$ common remainders (same remainders) with moduli $f_{sr_h^{(l,i)}}; h = 1, \dots, \alpha_{(l,i)}$ between $F_i \in X$ and $F'_l \in Y$. Then the difference value between $F_i \in X$ and $F'_l \in Y$ is as follows:

$$F'_l - F_i = k_{l,i} lcm\left(\bigcup_{h=1}^{\alpha_{(l,i)}} f_{sr_h^{(l,i)}}\right);$$

$$k_{l,i} \in \{0, \pm 1, \dots, \pm \left\lfloor \frac{F_{max}}{lcm(f_{sr_1^{(l,i)}}, \dots, f_{sr_{\alpha_{(l,i)}}^{(l,i)}})} \right\rfloor\} \quad (7)$$

Proof of Proposition 6: There are $\alpha_{(l,i)}$ same remainders between F'_l and F_i thus difference between these $\alpha_{(l,i)}$ remainders should be zero i.e. $f'_{u(l,r_h^{(l,i)})} - f_{u(i,r_h^{(l,i)})} = 0; h = 1, \dots, \alpha_{(l,i)}$. By considering (5) and (6) we have $f'_{u(i,r_h^{(l,i)})} - f_{u(i,r_h^{(l,i)})} = F'_l - F_i - k''_{l,r_h} f_{sr_h^{(l,i)}} = 0$ where $k''_{l,r} \in \{0, \pm 1, \dots, \pm \lfloor F_{max} / f_{sr} \rfloor\}$. So, it is possible to have following relationships:

$$F'_l - F_i = k''_{l,r_1^{(l,i)}} f_{sr_1^{(l,i)}} \\ = \dots = k''_{l,r_h^{(l,i)}} f_{sr_h^{(l,i)}} = \dots = k''_{l,r_{\alpha(l,i)}^{(l,i)}} f_{sr_{\alpha(l,i)}^{(l,i)}} = \Lambda \quad (8)$$

From (8) it is obvious that Λ should be multiple of α_{li} moduli frequencies, i.e. $\Lambda/f_{sr_h^{(l,i)}} = k''_{l,r_h^{(l,i)}}; h = 1, \dots, \alpha_{(l,i)}$. Therefore, the smallest value that dividable to all moduli frequencies $f_{sr_1^{(l,i)}}, \dots, f_{sr_h^{(l,i)}}, \dots, f_{sr_{\alpha(l,i)}^{(l,i)}}$ is the least common multiple (lcm) of them i.e. $\text{lcm}(f_{sr_1^{(l,i)}}, \dots, f_{sr_h^{(l,i)}}, \dots, f_{sr_{\alpha(l,i)}^{(l,i)}})$. Thus, Λ is multiple of lcm of α_{li} moduli frequencies i.e. $\Lambda = k_{l,i} \text{lcm}(f_{sr_1^{(l,i)}}, \dots, f_{sr_h^{(l,i)}}, \dots, f_{sr_{\alpha(l,i)}^{(l,i)}})$. From Lemma 1 it is clear that $F'_l, F_i \in [0, F_{\max}]$ thus $\Lambda = F'_l - F_i = k_{l,i} \text{lcm}(\bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}); k_{l,i} \in \{0, \pm 1, \dots, \pm \lfloor F_{\max} / \text{lcm}(f_{sr_1^{(l,i)}}, \dots, f_{sr_{\alpha(l,i)}^{(l,i)}}) \rfloor\}$.

Proposition 7: Assume a set of ρ frequencies (integers) as $X = \{F_1, \dots, F_\rho\}$ from under-sampled frequencies (remainders) with sampling frequencies (moduli) $f_{sr}, r = 1, \dots, \gamma$ can be reconstructed unambiguously when $\max(X) < F_{\max}$ where F_{\max} is called the largest dynamic range.

The largest dynamic range for ρ integers from remainders (frequencies) with moduli (sampling frequencies) $f_{sr}, r = 1, \dots, \gamma$ can be obtained as follows:

$$F_{\max} = \max(\{F_1, \dots, F_\rho\}) \\ \sum_{l=1}^{\rho} \sum_{i=1}^{\rho} \left| F'_l - F_i - k_{l,i} \text{lcm}(\bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}) \right| = 0 \quad (9) \\ \bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}} = \Gamma, l = 1, \dots, \rho$$

where $Y = \{F'_1, \dots, F'_\rho\}$ have the same remainders sets as X with moduli $f_{sr}, r = 1, \dots, \gamma$, $Y \neq X$ and based on proposition 6 for each $F_i \in X$ and $F'_l \in Y$ there are $\alpha_{(l,i)}; l = 1, \dots, \rho; i = 1, \dots, \rho$ common remainders (same remainders) with moduli $f_{sr_h^{(l,i)}}; h = 1, \dots, \alpha_{(l,i)}$ between $F_i \in X$ and $F'_l \in Y$. Since, X and Y have the same reminders.

Thus, common remainder between $F_i \in X$ and all $F'_l \in Y$ and compartment sampling frequencies $f_{sr_h^{(l,i)}}; h =$

$$1, \dots, \alpha_{(l,i)} \text{ should be } \Gamma \text{ i.e. } \bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}} = \Gamma. \text{ Similar relation is existing between } F'_l \in Y \text{ and all } F'_i \in X \text{ i.e. } \\ \bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}} = \Gamma.$$

Proof of Proposition 7: Consider two different sets $X = \{F_1, \dots, F_\rho\}$ and $Y = \{F'_1, \dots, F'_\rho\}$ have the same remainder sets with moduli $f_{sr}, r = 1, \dots, \gamma$ where $\max(X) \leq F_{\max}$ and $\max(Y) \leq F_{\max}$. The $\alpha(l, i)$ is the number of common remainders between F'_l and F_i and $k_{l,i} \in \{0, \pm 1, \dots, \pm \lfloor F_{\max} / \text{lcm}(f_{sr_1^{(l,i)}}, \dots, f_{sr_{\alpha(l,i)}^{(l,i)}}) \rfloor\}$ and Γ is

the set of all moduli $\Gamma = \bigcup_{r=1}^{\gamma} f_{sr}$. Thus each F'_l has $\alpha(l, i)$ common remainder sets with F_i that $\sum_{i=1}^{\rho} \alpha(l, i) = \gamma$ or $\bigcup_{i=1}^{\rho} \bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}} = \Gamma$. Based on Proposition 6 the difference value is equal to (7). Thus, we can say

$$F'_l - F_i - k_{l,i} \text{lcm}(\bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}) = 0. \text{ This relation must be}$$

fulfilled for a F'_l and all F_i 's $i = 1, \dots, \rho$ i.e. $\sum_{i=1}^{\rho} \left| F'_l - F_i - \right.$

$$\left. k_{l,i} \text{lcm}(\bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}) \right| = 0. \text{ Furthermore, this relationship}$$

should be satisfied for all F'_l 's $l = 1, \dots, \rho$ i.e.

$$\sum_{l=1}^{\rho} \sum_{i=1}^{\rho} \left| F'_l - F_i - k_{l,i} \text{lcm}(\bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}) \right| = 0.$$

In the following the proposed procedure is introduced to obtain the largest dynamic range F_{\max} from (9). Note that when all under-sampling frequencies multiplied by constant c (increased c times) the lcm of under-sampling frequencies are also multiplied by c . Then, the maximum possible frequencies that satisfied (9) i.e. F_{\max} will also be multiplied by c .

Procedure 1: The procedure for determination of the largest dynamic range can be summarized as follows:

Step 0: Initialize the largest dynamic range as $F_{\max} = F_{\max}^{\text{Ini}}$ in which F_{\max}^{Ini} is greater than (e.g. ten times of) conventional dynamic range mentioned in proposition 2 i.e. $F_{\max}^{\text{Ini}} \gg \min_{i_1 \cup \dots \cup i_\rho = \Gamma} \max\{\prod_{f_{si} \in i_1} f_{si}, \dots, \prod_{f_{si} \in i_\rho} f_{si}\}$.

Step 1: Categorize moduli of F'_l (i.e. $\Gamma = \{f_{s1}, \dots, f_{s\gamma}\}$) to

ρ disjoint subsets as $\alpha(l, i) = \bigcup_{h=1}^{\alpha(l,i)} f_{sr_h^{(l,i)}}, i = 1, \dots, \rho$. The

$\alpha(l, i)$ is a set of common moduli between F'_l and F_i and $\alpha(l, i)$ is common moduli between F'_l 's and all F_i 's. Since $\alpha(l, i)$ is obtained by categorizing γ moduli of F'_l , it is

$$\bigcup_{i=1}^{\rho} \alpha(l, i) = \Gamma, l = 1, \dots, \rho.$$

Step 2: Common compartment modules between F'_l 's and F_i 's can be considered as a matrix:

$$\begin{matrix} & F_1 & \dots & F_i & \dots & F_\rho \\ F'_1 & a_{(1,1)} & \dots & a_{(1,i)} & \dots & a_{(1,\rho)} \\ \vdots & \vdots & & \vdots & & \vdots \\ F'_l & a_{(l,1)} & \dots & a_{(l,i)} & \dots & a_{(l,\rho)} \\ \vdots & \vdots & & \vdots & & \vdots \\ F'_\rho & a_{(\rho,1)} & \dots & a_{(\rho,i)} & \dots & a_{(\rho,\rho)} \end{matrix} \quad (10)$$

Based on Step 1 of the procedure each row is related to all F'_l 's moduli. Thus, each row is chosen such that $\bigcup_{l=1}^{\rho} a_{(l,i)} = \Gamma, l = 1, \dots, \rho$. Each column related to all F_i 's moduli. Thus, each column should check to be sure that $\bigcup_{l=1}^{\rho} a_{(l,i)} = \Gamma, i = 1, \dots, \rho$. If this condition is met, go to Step 3; otherwise, return to Step 1 and produce other possible moduli from Γ .

Step 3: Based on (9) and representation $a_{(l,i)}$ in (10) following relationship can be written

$$F'_l - F_i = k_{(l,i)} lcm(a_{(l,i)}), \quad l = 1, \dots, \rho \text{ and } i = 1, \dots, \rho \quad (11)$$

According to lemma 1, the F_1 should be zero. By considering $i = 1$ in (11) and $F_1 = 0$ we have the following relation:

$$F'_l = k_{(l,1)} lcm(a_{(l,1)}), l = 1, \dots, \rho \quad (12)$$

Now, by substituting (12) in (11), the F_i 's for $i > 2$ can be obtained as below:

$$F_i = k_{(l,1)} lcm(a_{(l,1)}) - k_{(l,i)} lcm(a_{(l,i)}), \quad l = 1, \dots, \rho \text{ and } i = 2, \dots, \rho \quad (13)$$

that $k_{(l,i)} \in \{0, \pm 1, \dots, \pm [F_{max} / lcm(a_{(l,i)})]\}$, $a_{(l,i)} = \bigcup_{h=1}^{h=1} f_{sr_h^{(l,i)}}, i = 1, \dots, \rho$ and also it is assumed that $F_{max} = F_{max}^{ini}$.

Step 4: Based on (11), each F'_l with each F_i 's that has common moduli should meet $F'_l = k_{(l,i)} lcm(a_{(l,i)}) + F_i$, similar relations should be met for each F_i . When two sets $X = \{F_1, \dots, F_\rho\}$ and $Y = \{F'_1, \dots, F'_\rho\}$ are found so that satisfy conditions in (11), we can consider $\max(X)$ as final F_{max} and finish the process.

Otherwise, choose a bigger F_{max}^{ini} e.g. double of previous F_{max}^{ini} and go to step 1. It is notable, Proposition 7 presents a relationship to find the largest dynamic range (F_{max}) numerically by procedure1 for any ρ that not presented in the previous studies. However, procedure 1 can be simplified for some cases include $\rho = 2$ and $\rho = 3$. By considering two integers, i.e. $\rho = 2$, we show in Corollary 1 that the close form relationship for the largest dynamic range of two integers in [23] is a special case of proposed proposition 7.

Corollary 1: The largest dynamic range (maximum

possible range of frequency for unique detection) for proposed proposition 7 when $\rho = 2$ (two frequencies) is $F_{max} = \min_{I_1 \cup I_2 = \Gamma} \{lcm(I_1) + lcm(I_2)\}$.

Proof of Corollary 1: For this case, the condition in (9) can be written as

$$\sum_{l=1}^2 \sum_{i=1}^2 \left| F'_l - F_i - k_{l,i} lcm \left(\bigcup_{h=1}^{h=1} f_{sr_h^{(l,i)}} \right) \right| = 0. \text{ Thus, there}$$

are the following relationships:

$$\begin{aligned} F'_1 - F_1 &= k_{1,1} lcm \left(\bigcup_{h=1}^{h=1} f_{sr_h^{(1,1)}} \right), \\ F'_1 - F_2 &= k_{1,2} lcm \left(\bigcup_{h=1}^{h=1} f_{sr_h^{(1,2)}} \right), \\ F'_2 - F_1 &= k_{2,1} lcm \left(\bigcup_{h=1}^{h=1} f_{sr_h^{(2,1)}} \right), \\ F'_2 - F_2 &= k_{2,2} lcm \left(\bigcup_{h=1}^{h=1} f_{sr_h^{(2,2)}} \right) \end{aligned} \quad (14)$$

Let us show common compartment modules between F'_l 's and F_i 's as a matrix:

$$\begin{matrix} & F_1 & F_2 \\ F'_1 & a_{(1,1)} & a_{(1,2)} \\ F'_2 & a_{(2,1)} & a_{(2,2)} \end{matrix} \quad (15)$$

where $a_{(l,i)}$ is the common disjoint moduli between F'_l and F_i . Let's consider community of all moduli as $\Gamma = \bigcup_{r=1}^{r=1} f_{sr}$. Thus, community between subsets of F'_1 i.e. $a_{(1,1)}$ and $a_{(1,2)}$ in a row of mentioned matrix should be Γ i.e. $a_{(1,1)} \cup a_{(1,2)} = \Gamma$ similar for F'_2 , F_1 and F_2 there are $a_{(1,1)} \cup a_{(1,2)} = a_{(2,1)} \cup a_{(2,2)} = a_{(1,1)} \cup a_{(2,1)} = a_{(1,2)} \cup a_{(2,2)} = \Gamma$. These conditions are satisfied when $a_{(1,1)} = a_{(2,2)}$ and $a_{(1,2)} = a_{(2,1)}$. In other words, if all modules Γ are divided to two disjoint subsets I_1 and I_2 where $I_1 \cup I_2 = \Gamma$ then the matrix in (15) can be rewritten as:

$$\begin{matrix} & F_1 & F_2 \\ F'_1 & I_1 & I_2 \\ F'_2 & I_2 & I_1 \end{matrix} \quad (16)$$

Now, according to (14) and (16) and based on Lemma 1 by considering $F_1 = 0$ it can be written:

$$\begin{aligned} F'_1 &= k_{1,1} lcm(I_1), \\ F'_2 &= k_{2,1} lcm(I_2), \\ F'_1 - F_2 &= k_{1,2} lcm(I_2), \\ F'_2 - F_2 &= k_{2,2} lcm(I_1) \end{aligned} \quad (17)$$

To satisfy both relations $F_2 = k_{1,1}lcm(I_1) - k_{1,2}lcm(I_2)$ and $F_2 = k_{2,1}lcm(I_2) - k_{2,2}lcm(I_1)$ in (14) it should satisfy $k_{1,1} = k_{2,1} = 1$ and $k_{1,2} = k_{2,2} = -1$. Thus, $F_2 = lcm(I_1) + lcm(I_2)$ and the set of integers would be $X = \{0, F_2\}$ and F_{max} is the minimum possible of F_2 , i.e. $F_{max} = \min_{I_1 \cup I_2 = \Gamma} \max(X) = \min_{I_1 \cup I_2 = \Gamma} \{lcm(I_1) + lcm(I_2)\}$.

This, shows proposition 3 is a special case of proposed proposition 7 when $\rho = 2$.

Corollary 2: The common moduli between F_i 's and F_l , i.e. $a_{(l,i)} = \bigcup_{h=1}^{h=1} f_{sr_h^{(l,i)}}$, for $\rho = 3$ that admit the conditions in proposition 7 can be simplified as follows:

$$\begin{array}{ccc} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} \end{bmatrix} \\ F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} I_1 \cup I_2 & I_3 \cup I_6 & I_4 \cup I_5 \\ I_3 \cup I_4 & I_2 \cup I_5 & I_1 \cup I_6 \\ I_5 \cup I_6 & I_1 \cup I_4 & I_2 \cup I_3 \end{bmatrix} \end{array} \quad (18)$$

where $I_i, i = 1, \dots, 6$ are disjoint subsets and $\bigcup_{i=1}^6 I_i = \Gamma = \bigcup_{r=1}^r f_{sr}$.

Note that, corollary 2 substitute the steps 1 and 2 of procedure 1 for the calculation $a_{(l,i)}$'s when $\rho = 3$.

Proof of Corollary 2: By considering the condition in (9), i.e. $\sum_{l=1}^3 \sum_{i=1}^3 \left| F_l' - F_i - k_{l,i}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(l,i)}}\right) \right| = 0$, it is possible to write the following relations:

$$\begin{array}{l} F_1' - F_1 = k_{1,1}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(1,1)}}\right), \\ F_1' - F_2 = k_{1,2}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(1,2)}}\right), \\ F_1' - F_3 = k_{1,3}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(1,3)}}\right), \\ F_2' - F_1 = k_{2,1}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(2,1)}}\right), \end{array} \quad (19)$$

$$F_2' - F_2 = k_{2,2}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(2,2)}}\right),$$

$$F_2' - F_3 = k_{2,3}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(2,3)}}\right),$$

$$F_3' - F_1 = k_{3,1}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(3,1)}}\right),$$

$$F_3' - F_2 = k_{3,2}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(3,2)}}\right),$$

$$F_3' - F_3 = k_{3,3}lcm\left(\bigcup_{h=1}^{h=1} f_{sr_h^{(3,3)}}\right)$$

Similar to corollary 1, the common compartment modules between F_i 's and F_l 's can be shown as below matrix:

$$\begin{array}{ccc} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} \end{bmatrix} \end{array} \quad (20)$$

The community of moduli in all rows and columns should be admitted $\Gamma = \bigcup_{r=1}^r f_{sr}$, as:

$$\begin{aligned} & a_{(1,1)} \cup a_{(1,2)} \cup a_{(1,3)} \\ & = a_{(3,1)} \cup a_{(3,2)} \cup a_{(3,3)} \\ & = a_{(3,1)} \cup a_{(3,2)} \cup a_{(3,3)} \\ & = a_{(1,1)} \cup a_{(2,1)} \cup a_{(3,1)} \\ & = a_{(1,2)} \cup a_{(2,2)} \cup a_{(3,2)} \\ & = a_{(1,3)} \cup a_{(2,3)} \cup a_{(3,3)} \\ & = \Gamma \end{aligned} \quad (21)$$

To satisfy (21), the common moduli in matrix (20) can be expressed as follows:

$$\begin{array}{ccc} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} I_1 \cup I_2 & I_3 \cup I_6 & I_4 \cup I_5 \\ I_3 \cup I_4 & I_2 \cup I_5 & I_1 \cup I_6 \\ I_5 \cup I_6 & I_1 \cup I_4 & I_2 \cup I_3 \end{bmatrix} \end{array} \quad (22)$$

where $I_i, i = 1, \dots, 6$ are disjoint subsets $I_i \subset \Gamma$ and $\bigcup_{i=1}^6 I_i = \Gamma = \bigcup_{r=1}^r f_{sr}$. Now, the value $a_{(l,i)}$ that satisfied (21) can be used in steps 3 and 4 of Procedure 1 to find F_{max} . In fact, steps 1 and 2 of procedure 1 can be replaced by corollary 2 for calculation $a_{(l,i)}$ s when $\rho = 3$.

Procedure 2: The procedure of determination input

frequencies from their under-sampled frequencies for complex waveform. There are some similarities between the procedure of determination of frequencies from under sampled frequencies of a sinusoidal complex waveform (i.e. $\sum_{l=1}^q A_l e^{i(2\pi F_l t)} + w(t)$) and the determination of frequencies of real sinusoidal waveform (i.e. $\sum_{i=1}^q A_i \cos(2\pi F_i t + \phi_i) + w(t)$) in [2]. However, should consider the fact that the under-sampled frequencies of the real waveform and the complex waveform are different as follows [2]:

$$f_{u(k,j)} = \begin{cases} (-1)^{\bar{v}_k} (F_j - m_{kj} f_{sk}); & \bar{v}_k \in \{1,2\} \\ \text{Real waveform} \\ F_j - m_{kj} f_{sk} & \text{Complex waveform} \end{cases} \quad (23)$$

Thus, as can see in Fig. 2 (b) the under-sampled frequency curve for complex waveforms is not continues and a few noises or changes in frequency (F_j) can cause big change in $f_{u(k,j)}$ and reduce $f_{u(k,j)}$ from maximum to zero or vice versa.

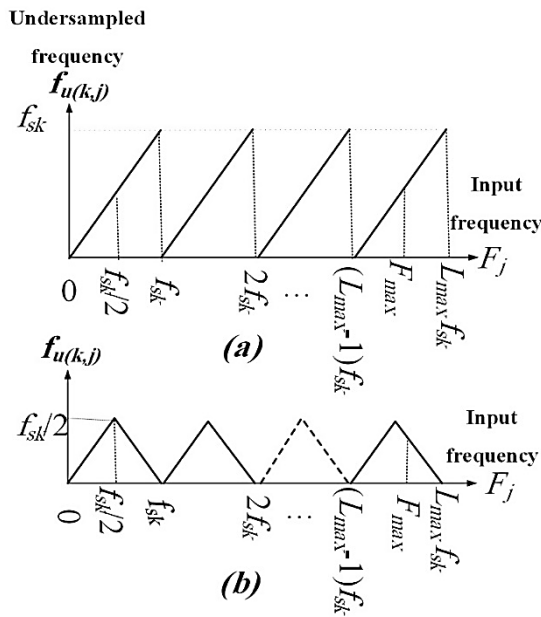


Fig. 2: Under-sampled frequency $f_{u(k,j)}$ as a function of j^{th} analog input frequency $F_j \in [0, F_{max}]$ after sampling with the k^{th} sampling frequency f_{sk} from (a) real signal waveform and (b) complex signal waveform.

Step 1: There are q frequencies that are sampled with p ADC's, thus there are $p \times q$ under-sampled frequencies as $\tilde{f}_{u(k,j)}$; $k = 1, \dots, p, j = 1, \dots, q$. However, the correspondences between q input frequencies and the q outputs under-sampled frequencies are unknown.

Thus, these $p \times q$ noisy under-sampled frequencies should be divided into q groups with p elements in each group as $\{S_1, \dots, S_j, \dots, S_q\} = \{\{\hat{f}_{u(i,1)}, i = 1, \dots, p\}, \dots, \{\hat{f}_{u(i,j)}, i = 1, \dots, p\}, \dots, \{\hat{f}_{u(i,q)}, i =$

$1, \dots, p\}\}$ in which the set $S_j = \{\hat{f}_{u(i,j)}, i = 1, \dots, p\}$; $j = 1, \dots, q$ denotes a noisy under-sampled frequencies set that corresponding to the j^{th} input frequency.

Step 2: Determines the distance $DIS_{\hat{f}_{u(i,j)}}$ for each set of $S_j = \{\hat{f}_{u(i,j)}, i = 1, \dots, p\}, j = 1, \dots, q$ as follows:

$$DIS_j = \max(DIS_{\hat{f}_{u(1,j)}}, \dots, DIS_{\hat{f}_{u(i,j)}}, \dots, DIS_{\hat{f}_{u(p,j)}}) \quad (24)$$

where the procedure for the computing of $DIS_{\hat{f}_{u(k,j)}}$ and $F_{est}(\hat{f}_{u(k,j)})$ for each set of S_j is described as follows:

Step 1: Calculate the frequencies \hat{F}_k^t s in the band $\hat{F}_k^t \in [0, F_{max}]$ from $\hat{f}_{u(k,j)}$, when sampling frequency is f_{sk} as below:

$$\begin{aligned} \hat{F}_k^t &= \hat{k}_k^t f_{sk} + \hat{f}_{u(k,j)}; \\ 0 &\leq \hat{k}_k^t f_{sk} < F_{max}; \quad \hat{k}_k^t = 0, 1, \dots \end{aligned} \quad (25)$$

Step 2: Determine under-sampled frequencies \hat{f}_{ui}^t ; $i = 1, 2, \dots, p, i \neq k$ related to \hat{F}_k^t s when are sampled with sampling frequencies other than sampling frequency in step 1 i.e. f_{si} ; $i = 1, 2, \dots, p, i \neq k$.

$$\hat{F}_k^t = \hat{k}_i^t f_{si} + \hat{f}_{ui}^t \quad (26)$$

Step 3: Substitute \hat{f}_{ui}^t ; $i = 1, 2, \dots, p, i \neq k$ with their noisy under-sampled $\hat{f}_{u(i,j)}$; $i = 1, 2, \dots, p, i \neq k$ in (26). Then calculate the following relationship:

$$\begin{aligned} \tilde{F}_i^t &= \left\{ \tilde{F}_i^t \mid \text{minimize} |\tilde{F}_i^t - \hat{F}_k^t| \right\}, \\ \tilde{F}_1^t &= \hat{k}_i^t f_{si} + \hat{f}_{u(i,j)}, \\ \tilde{F}_2^t &= (\hat{k}_i^t + 1) f_{si} + \hat{f}_{u(i,j)} \\ \tilde{F}_3^t &= (\hat{k}_i^t - 1) f_{si} + \hat{f}_{u(i,j)} \end{aligned} \quad (27)$$

Note that unlike the under-sampled frequencies of the real waveform in the complex waveform small changes caused by noise can make a big change in the under-sampled frequencies as can be seen in Fig.2 (b). Thus, to substitute \hat{f}_{ui}^t s by $\hat{f}_{u(i,j)}$ s should consider the $\hat{k}_i^t f_{si}$, $(\hat{k}_i^t + 1) f_{si}$ and $(\hat{k}_i^t - 1) f_{si}$.

Step 4: Find the \hat{k}_i^t s that minimize the following relationship and name them as \hat{k}_i^{t*} :

$$\begin{aligned} \hat{k}_1^{t*}, \dots, \hat{k}_i^{t*}, \dots, \hat{k}_p^{t*} &= \{\hat{k}_1^t, \dots, \hat{k}_i^t, \dots, \hat{k}_p^t\}; \\ \min_{\hat{k}_1^t, \dots, \hat{k}_i^t, \dots, \hat{k}_p^t} &\max\{|\tilde{F}_1^t - \tilde{F}_2^t|, \dots, \\ &|\tilde{F}_1^t - \tilde{F}_p^t|, \dots, |\tilde{F}_i^t - \tilde{F}_{i+1}^t|, \dots, \\ &|\tilde{F}_i^t - \tilde{F}_p^t|, \dots, |\tilde{F}_{p-1}^t - \tilde{F}_p^t|\}; \end{aligned} \quad (28)$$

Definition 1: The maximum distance (DIS) between the

frequencies \tilde{F}_i^t in (27) related to $\hat{f}_{u(k,j)}$; $k = 1, \dots, p$ is called $DIS_{\hat{f}_{u(k,j)}}$; $\hat{f}_{u(k,j)} \in S_j$ and defined as follows:

$$DIS_{\hat{f}_{u(k,j)}} \triangleq \max\{|\tilde{F}_1^t - \tilde{F}_2^t|, \dots, |\tilde{F}_1^t - \tilde{F}_p^t|, \dots, |\tilde{F}_i^t - \tilde{F}_{i+1}^t|, \dots, |\tilde{F}_i^t - \tilde{F}_p^t|, \dots, |\tilde{F}_{p-1}^t - \tilde{F}_p^t|\}; \quad (29)$$

$$\hat{k}_1^t = \hat{k}_1^{t*}, \dots, \hat{k}_p^t = \hat{k}_p^{t*}, i = 1, 2, \dots, p$$

Step 5: The estimated input frequency is obtained by mean of the frequencies (\tilde{F}_i^t s) that minimize (28) as below:

$$F_{est}(\hat{f}_{u(k,j)}) = \sum_{i=1}^p \tilde{F}_i^t / p; \quad (30)$$

$$\hat{k}_1^t = \hat{k}_1^{t*}, \dots, \hat{k}_p^t = \hat{k}_p^{t*}, i = 1, 2, \dots, p,$$

Step 3. Obtain the possible input frequencies for state of each set in S_j ; $j = 1, \dots, q$ as $F_{state(n)} = \{F_{est}(\hat{f}_{u(i_1,j)}^*), \dots, F_{est}(\hat{f}_{u(i_j,j)}^*), \dots, F_{est}(\hat{f}_{u(i_q,j)}^*)\}$ where $F_{est}(\hat{f}_{u(i_j,j)})$ was calculated in (30) and $\hat{f}_{u(i_j,j)}$ is ith under-sampled frequency of jth input frequency which minimizes the defined distance in (29) i.e. $DIS_{\hat{f}_{u(i_j,j)}}^* = \min_{\hat{f}_{u(i,j)}} (DIS_{\hat{f}_{u(1,j)}}, \dots, DIS_{\hat{f}_{u(p,j)}})$.

Step 4. Repeat steps 1 to 3 for all different states of dividing $p \times q$ under-sampled frequencies in to q groups with p elements in each group. In other words, steps 1 to 3 should be carried out for $n = 1, \dots, (p!)^{q-1}$ different states.

Find the state that has the minimum value of $DIS_{state(n)}$ as below:

$$DIS_{state(n^*)} = \min_n (DIS_{state(1)}, \dots, DIS_{state(n)}, \dots) \quad (31)$$

The correct input analog frequencies are obtained based on n^* in (31) as $F_{state(n^*)}$.

Proposition 8: The maximum tolerable noise that multiple input frequencies $F_i \in [0, F_{max}]$; $i = 1, \dots, \rho$ from noisy under-sampled frequencies $\tilde{f}_{u(r,i)} = f_{u(r,i)} + \varepsilon_{(r,i)}$, $r = 1, \dots, \gamma$; $i = 1, \dots, \rho$, with sampling frequencies f_{sr} , $r = 1, \dots, \gamma$ is $\varepsilon_{max(Tolerable)} = \chi_{min} / 4$. It is notable F_{max} is the largest possible range that obtained in proposition 7 and not a large range as the previous works.

The noise of each under-sampled frequency ($\varepsilon_{(r,i)}$) and the maximum noise of all under-sampled frequencies (ε_{max}) should be less than the maximum tolerable noise as $\varepsilon_{(r,i)} \leq \varepsilon_{max} \leq \varepsilon_{max(Tolerable)} = \chi_{min} / 4$. Where, in this proposition, the frequency $f_{u(r,i)}$ is a noiseless under sampled frequency, $\varepsilon_{(r,i)}$ is an additive noise, ε_{max} is the maximum value of all $\varepsilon_{(r,i)}$'s and,

$$\chi_{min} = 4\varepsilon_{max(Tolerable)} = \min_{\substack{\{f_{s1}, f_{s2}\} \subseteq \{f_{s1}, f_{s2}, \dots, f_{s\gamma}\}; \\ k_{i1}', k_{i2}', \\ |k_{i1}' f_{s1} - k_{i2}' f_{s2}| \neq 0}} |k_{i1}' f_{s1} - k_{i2}' f_{s2}|; \quad (32)$$

$$1 \leq i_1 < i_2 \leq \gamma$$

Note that, the k_{i_t}' , $t \in \{1, 2\}$ are some integers in (32) can be selected as $k_{i_t}' \in \{0, \pm 1, \dots, \pm k_{i_t}^{max}\}$; $(k_{i_t}^{max} - 1)f_{s_{i_t}} < F_{max} < k_{i_t}^{max} f_{s_{i_t}}$ or $k_{i_t}^{max} = \lceil F_{max} / f_{s_{i_t}} \rceil$ where F_{max} is defined as (4).

Proof of Proposition 8: Consider a frequency F under-sampled with $r = 1, \dots, p$ sampling frequency as follows:

$$F = \bar{k}_1 f_{s1} + f_{u(1,j)} = \dots = \bar{k}_r f_{sr} + f_{u(r,j)} = \dots = \bar{k}_p f_{sp} + f_{u(p,j)} \quad (33)$$

where \bar{k}_r is correct integer that relates noiseless under-sampled frequency $f_{u(r,j)}$ to F .

Based on (29) the $DIS_{\hat{f}_{u(k,j)}}$ is the distance between p estimations of F_j from p available under-sampled frequencies i.e. $S_j = \{\hat{f}_{u(r,j)}, r = 1, \dots, p\}$.

Consider the distance $|\tilde{F}_i^t - \tilde{F}_l^t|$ in $DIS_{\hat{f}_{u(k,j)}}$ in (29) as $D_{il} = |\tilde{F}_i^t - \tilde{F}_l^t|$. We prove that D_{il} , for not incorrect chosen of is greater than $\bar{D}_{il} = |\tilde{F}_i^t - \tilde{F}_l^t|$ where \tilde{F}_i^t and \tilde{F}_l^t are the correct estimated frequency of \tilde{F}_i^t and \tilde{F}_l^t , respectively. In other words \tilde{F}_i^t ; $i = 1, \dots, p$ are $\tilde{F}_i^t = k_i f_{si} + \hat{f}_{u(i,j)}$ that $\hat{f}_{u(i,j)}$ is noisy under-sampled frequencies and k_i are equal to the correct one i.e. \bar{k}_i in (33) or $\tilde{F}_i^t = \bar{k}_i f_{si} + \hat{f}_{u(i,j)}$.

We have $\hat{f}_{u(i,j)} = f_{u(i,j)} + \varepsilon_{(i,j)}$, $\hat{f}_{u(l,j)} = f_{u(l,j)} + \varepsilon_{(l,j)}$, $\tilde{F}_i^t = k_i f_{si} + \hat{f}_{u(r,j)} = k_i f_{si} + f_{u(l,j)} + \varepsilon_{(l,j)}$, $\tilde{F}_l^t = k_l f_{sl} + f_{u(l,j)} + \varepsilon_{(l,j)}$ and substituting $f_{u(i,j)} - f_{u(l,j)} = \bar{k}_j f_{sj} - \bar{k}_l f_{sl}$ from (33) have the following equation:

$$D_{il} = |\tilde{F}_i^t - \tilde{F}_l^t| = |(k_i - \bar{k}_i) f_{si} - (k_l - \bar{k}_l) f_{sl} + \varepsilon_{(i,j)} - \varepsilon_{(l,j)}| = |k_i' f_{si} - k_l' f_{sl} + \varepsilon_{(i,j)} - \varepsilon_{(l,j)}| \quad (34)$$

Now, there are two states. For the correct estimation we have $k_i = \bar{k}_i$, $k_l = \bar{k}_l$ thus $k_i' = 0$ and $k_l' = 0$ and D_{il} in (34) can be rewritten:

$$D_{il} = |\varepsilon_{(i,j)} - \varepsilon_{(l,j)}| \leq 2\varepsilon_{max} \quad ; \quad \text{for the correct estimation} \quad (35)$$

For incorrect estimation $k_i' \neq 0$ and $k_l' \neq 0$. Thus, for the incorrect estimation can write:

$$D_{il} = |k_i' f_{si} - k_l' f_{sl} + \varepsilon_{(i,j)} - \varepsilon_{(l,j)}| \geq |k_i' f_{si} - k_l' f_{sl}| - |\varepsilon_{(i,j)} - \varepsilon_{(l,j)}| \geq |k_i' f_{si} - k_l' f_{sl}| - 2\varepsilon_{max} \quad (36)$$

Based on (32) we have $|k_r' f_{sr} - k_s' f_{ss}| > 4\varepsilon_{\max}$. Thus, for the incorrect estimation can write:

$$D_{rs} \geq 4\varepsilon_{\max} - 2\varepsilon_{\max} = 2\varepsilon_{\max} ; \quad (37)$$

for the incorrect estimation

When one of F_i 's is estimated incorrectly $D_{il} = |\tilde{F}_i^t - \tilde{F}_l^t| \geq 2\varepsilon_{\max}$ in $DIS_{\hat{f}_{u(k,j)}}$ then $DIS_{\hat{f}_{u(k,j)}} \geq 2\varepsilon_{\max}$ or can write:

$$\begin{cases} DIS_{\hat{f}_{u(k,j)}} \leq 2\varepsilon_{\max} & \text{for the correct} \\ & \text{estimation} \\ DIS_{\hat{f}_{u(k,j)}} \geq 2\varepsilon_{\max} & \text{for the incorrect} \\ & \text{estimation} \end{cases} \quad (38)$$

It means by minimizing $DIS_{\hat{f}_{u(k,j)}}$ in (29) as (31) when noises are less than $\varepsilon_{\max(\text{Tolerable})}$ in (32) the frequencies can be determined uniquely.

Results and Discussion

A. The Maximum Possible Dynamic Range of Under-Sampling Frequency Detection

To demonstrate the proposed approach, consider the largest dynamic range for $\rho = 2$ input frequencies (integers) with $\gamma = 6$ sensors and sampling frequencies (moduli) $\Gamma = \bigcup_{r=1}^{\gamma} f_{sr} = \{3,5,7,11,13,17\}$ Hz. According to Proposition 1, the dynamic range is $F_{\max} = \min_{1 \leq r_1 \leq \dots \leq r_3 \leq \gamma} lcm\{f_{s(r_1)}, \dots, f_{s(r_3)}\} = lcm\{3,5,7\} = 105$. From Proposition 2 the dynamic range is $F_{\max} = \min_{I_1 \cup \dots \cup I_\rho = \Gamma} \max\{\prod_{f_{si} \in I_1} f_{si}, \dots, \prod_{f_{si} \in I_\rho} f_{si}\} = 516$ for $I_1 = \{3,11,17\}$ and $I_2 = \{5,7,13\}$. Based on (17) $F_1' = k_{1,1} lcm(I_1)$ and $F_2' = k_{2,1} lcm(I_2)$, and F_2 can be obtained from two formulas $F_2 = k_{1,1} lcm(I_1) - k_{1,2} lcm(I_2)$ and $F_2 = k_{2,1} lcm(I_2) - k_{2,2} lcm(I_1)$. Based on Procedure 1 the minimum possible value for F_2 that satisfies both formulas are obtained when $k_{1,1} = 1$, $k_{2,1} = 1$, $k_{1,2} = -1$, $k_{2,2} = -1$, $I_1 = \{3,11,17\}$, and $I_2 = \{5,7,13\}$. Thus, $F_1' = 561$ Hz, $F_2' = 455$ Hz and $F_2 = 1016$ Hz. Two sets $X = \{F_1, F_2\} = \{0, 1016\}$ and $Y = \{F_1', F_2'\} = \{561, 455\}$ have the same remainders and $F_{\max} = \max(X) = 1016$ Hz. Similarly, based on Corollary 1 we have $F_{\max} = \min_{I_1 \cup I_2 = \Gamma} \{lcm(I_1) + lcm(I_2)\}$ where $I_1 = \{3,11,17\}$, $I_2 = \{5,7,13\}$, $F_{\max} = 1016$ Hz. The dynamic range of two integers without conditions on them for the proposed approach and previous works has been shown in Table. 2.

As discussed previously Proposition 3 [23] which is just for two integers is a special case of proposed proposition 7 when $\rho = 2$.

Table. 2: The dynamic range for two frequencies (integers)

Approach	Dynamic range
Proposition 1 [29], [30]	105
Proposition 2 [19]	561
Proposition 3 [17], [23], available just for two integers	1016
Proposition 7 (proposed approach)	1016

Now, consider the largest dynamic range for $\rho = 3$ input frequencies (integers) for moduli $\Gamma = \{3,5,7,11,13,17\}$. Based on corollary 2 each six disjoint partitions of Γ i.e. $\bigcup_{i=1}^6 I_i = \Gamma$ in matrix form in (18) will satisfy (21) or, equivalently, admit the conditions $\bigcup_{l=1}^{\rho} a_{(l,i)} = \bigcup_{l=1}^{\rho} [\bigcup_{h=1}^{\rho} f_{sr_h(l,i)}] = \Gamma, l = 1, \dots, \rho; \bigcup_{l=1}^{\rho} a_{(l,i)} = \bigcup_{l=1}^{\rho} [\bigcup_{h=1}^{\rho} f_{sr_h(l,i)}] = \Gamma, i = 1, \dots, \rho$ in proposition 7. Now, the obtained $a_{(l,i)}$'s by corollary 2 can be used in steps 3 and 4 of Procedure 1 to find F_{\max} in the following. By considering initial largest dynamic range as $F_{\max}^{ini} = 1000$ Hz in procedure 1, we have $I_1 = \{f_{s4}\}$, $I_2 = \{f_{s1}, f_{s3}\}$, $I_3 = \{f_{s5}\}$, $I_4 = \emptyset$, $I_5 = \{f_{s2}\}$, and $I_6 = \{f_{s6}\}$. Thus, based on (18) there is following relation:

$$\begin{matrix} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} a_{(1,1)} & a_{(1,2)} & a_{(1,3)} \\ a_{(2,1)} & a_{(2,2)} & a_{(2,3)} \\ a_{(3,1)} & a_{(3,2)} & a_{(3,3)} \end{bmatrix} \end{matrix} \quad (39)$$

$$\begin{matrix} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} \{f_{s4}, f_{s1}, f_{s3}\} & \{f_{s5}, f_{s6}\} & \{f_{s5}, f_{s6}\} \\ \{f_{s5}\} & \{f_{s1}, f_{s3}, f_{s2}\} & \{f_{s4}, f_{s6}\} \\ \{f_{s2}, f_{s6}\} & \{f_{s4}\} & \{f_{s1}, f_{s3}, f_{s5}\} \end{bmatrix} \end{matrix}$$

The $k_{(i,j)}$'s that satisfy (11) and (19) are as:

$$\begin{matrix} F_1 & F_2 & F_3 \\ \begin{bmatrix} F_1' \\ F_2' \\ F_3' \end{bmatrix} \begin{bmatrix} k_{(1,1)} & k_{(1,2)} & k_{(1,3)} \\ k_{(2,1)} & k_{(2,2)} & k_{(2,3)} \\ k_{(3,1)} & k_{(3,2)} & k_{(3,3)} \end{bmatrix} = \begin{bmatrix} 1 & 1 & -131 \\ 25 & 3 & -3 \\ 4 & 30 & -2 \end{bmatrix} \end{matrix} \quad (40)$$

Based on Step 3 of Procedure 1 the $F_1 = 0$ and $F_1' - F_1 = k_{(1,1)} lcm(a_{(1,1)}) = k_{(1,1)} lcm(\{f_{s4}, f_{s1}, f_{s3}\}) = 1 \times$

$lcm(\{11,3,7\}) = 231$ and $F'_1 = 231$. Other frequencies also is obtained based on Procedure 1. For the obtained $X = \{F_1, F_2, F_3\} = \{0, 10, 886\}$ and $Y = \{F'_1, F'_2, F'_3\} = \{231, 325, 340\}$, the largest dynamic range is $F_{max} = \max(X) = 886\text{Hz}$ which for all F'_l 's and F_i 's should satisfies (11), i.e. $F'_l - F_i = k_{(l,i)} lcm(a(l,i))$; $l = 1, \dots, \rho, i = 1, \dots, \rho$. As an example, for $F'_2 - F_3 = k_{(2,3)} lcm(a(2,3))$ we have $325 - 886 = -3 \times lcm(\{11,17\})$. It is notable that, the large dynamic range by previous studies based on proposition 2 is $F_{max} = \min_{I_1 \cup \dots \cup I_\rho = \Gamma} \max\{\prod_{f_{si} \in I_1} f_{si}, \dots, \prod_{f_{si} \in I_\rho} f_{si}\} = \max\{lcm(\{3,17\}), lcm(\{5,13\}), lcm(\{7,11\})\} = 77$. The dynamic range of three integers without conditions on them for the proposed approach and previous works has been shown in Table. 3.

Table. 3: The dynamic range for three frequencies (integers)

Approach	Dynamic range
Proposition 1 [19], [30]	17
Proposition 2 [19]	77
Proposition 7 (proposed approach)	886

For previous works, the large dynamic range for unambiguous reconstruction of input frequencies is $F_{max} = 77\text{Hz}$ while the largest dynamic range obtained by proposed approach is $F_{max} = 886\text{ Hz}$ that is 11.5 times greater than the previous works.

Assume a digital instance frequency measurement (DIFM) equipped to ADCs with sampling rates $\Gamma = \bigcup_{r=1}^{\gamma} f_{sr} = \{3, 5, 7, 11, 13, 17\} \times 10^7 = \{30, 50, 70, 110, 130, 170\} \times \text{MHz}$ what is the maximum possible range when 3 input frequencies come simultaneously? Before our work designer could claim designed DIFM guarantees reconstruction 3 simultaneous input frequencies uniquely until $77 \times 10^7 \text{Hz} = 770\text{MHz}$ now based on maximum upper bound obtained by proposed Proposition 7 can claim DIFM can reconstruct frequencies uniquely until $886 \times 10^7 \text{Hz} = 8.86\text{GHz}$. For the user of DIFM is also important to know for a higher range of frequency can guarantee to reconstruct frequencies.

B. The Under-Sampling Frequency Estimation for Noisy Waveform

This section, simulates the effect of noises on frequency estimations when sampling frequencies are very low.

The simulations are conducted for appropriate and non-appropriate under-sampling frequencies. For the first simulation, the maximum bound of frequencies is considered as a large bound obtained in the previous

works for $\rho = 3$ input frequencies and low sampling frequencies $\Gamma = \{3, 5, 7, 11, 13, 17\}\text{Hz}$.

A large bound as shown in Table. 3 for the previous works is 77 Hz. The maximum tolerable noise for this bound based on Theorem 2 of [2] for complex waveform (not real waveform) and for three input frequencies is 0.6.

In this work, we could find the maximum possible range for unique detection of multiple frequencies when sampling with very low sampling frequencies in Proposition 7 that for mentioned sampling frequencies (i.e. Γ) is obtained 886Hz in the previous section. Simulations have been done for 100000 random frequencies per each upper bound of noise for under-sampled frequencies. For the previous works a large obtained dynamic range that guarantee of unique detection was 77 Hz. Thus, random input frequencies are chosen in the range $[0, 77]$ in Fig. 3. The newly obtained upper bound frequency for unique detection of input frequencies is 886Hz. Consequently, random input frequencies are chosen in the range $[0, 886]$ in Fig. 4.

The procedure for detection frequencies was introduced in Procedure 2. The maximum tolerable noise for the proposed approach for three input frequencies and Γ under-sampling frequencies when $F_{max} = 886$ based on proposed Proposition 8 is $\epsilon_{\max(\text{Tolerable})} = \frac{\chi_{\min}}{4} = \frac{2}{4} = 0.25$.

Thus, for the proposed approach the maximum unique detectable frequencies and the maximum tolerable frequency noises are 886Hz and 0.25Hz against 77 Hz and 0.6 Hz for the previous works. For non-appropriate under-sampling frequencies like $\Gamma_{\text{non-appropriate}} = \{5, 6, 8, 12, 15, 18\}\text{Hz}$ that are greater than their counterpart sampling frequencies in Γ but the maximum tolerable noise for this set and $F_{max} = 886$ based on proposed Proposition 8 is $\epsilon_{\max(\text{Tolerable})} = \frac{\chi_{\min}}{4} = \frac{0}{4} = 0$. Thus, for non-appropriate low sampling frequencies even without the noise we cannot detect frequencies uniquely as shown in Fig. 5.

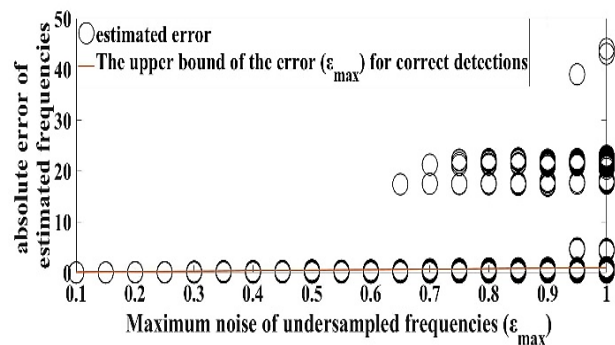


Fig. 3: Under-sampling frequency detection for noisy under-sampled frequencies of multiple input frequencies within range $[0, 77]$ and appropriate sampling frequencies Γ .

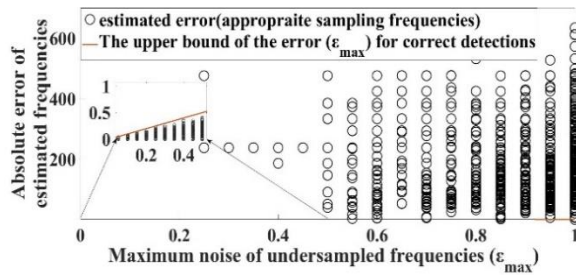


Fig. 4: Under-sampling frequency detection for noisy under-sampled frequencies of multiple input frequencies with range $[0, 886]$ (more than 11 times greater range than previous studies) and appropriate sampling frequencies Γ .

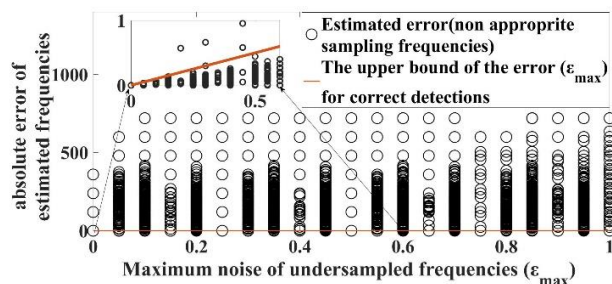


Fig. 5: Under-sampling frequency detection for noisy under-sampled frequencies for multiple input frequencies with range $[0, 886]$ and non-appropriate sampling frequencies $\Gamma_{non-appropriate}$.

Conclusion

This study proposed propositions and a procedure to find the largest possible dynamic range for frequencies in a sinusoidal waveform with any number of frequencies for the unambiguous reconstruction of the frequencies of the waveform with very low sampling rates.

Furthermore, the proposed propositions were specified and simplified for waveforms with two and three frequencies and showed that the previous works for the maximum possible range for reconstruction frequencies of waveforms with two frequencies are a special case of our work.

It has been shown that for some cases the proposed approach could achieve 11.5 times greater dynamic range for the unambiguous reconstruction the frequencies of an under-sampled waveform with very low sampling rates.

A procedure for multiple frequencies detection from reminders (under-sampled frequencies) was proposed and maximum tolerable noises of under-sampled frequencies for unique detection were obtained. There are two main disadvantages when using of under-sampling approaches.

When error in under-sampled frequencies is more than tolerable noise the origin frequencies cannot be reconstructed uniquely.

It is also necessary to have a computation unite to

reconstruct origin frequencies from under-sampled frequencies. However, using the under-sampling approaches is obligatory in some situations such as the sampling rates of ADCs are very less than the range of frequencies or because of energy consumption or price cannot use more ADCs to cover a high range of frequencies.

In this study, the maximum upper bound for any number of the input frequencies for complex waveform was investigated. In some applications, direct sampling from the real waveform is needed because of hardware limitations.

The relation between actual frequencies and under-sampled frequencies from under-sampled waveform different for complex sampling and real (direct) sampling. Finding the maximum upper bound for any number of input frequencies from directly under-sampled waveform (none complex waveform) is suggested for future work.

Author Contributions

Dr. Ali Maroosi. and Dr. Hossein Khaleghi Bizaki suggested the model and innovation of the problem. Simulation has been done by A. Maroosi. The first draft was written by A. Maroosi and reviewed by H. Khaleghi Bizaki.

Acknowledgment

This work has been financially supported by the University of Torbat Heydarieh. The grant number is UTH:1399/09/2284.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

ADC	Analog-to-Digital Converter
CRT	Chinese Remainder Theorem
DOA	Direction of Arrival
NOMA	non-orthogonal multiple access
UAV	unmanned aerial vehicle
lcm	least common multiple
DFT	discrete Fourier transform

References

- [1] A. Maroosi, H. K. Bizaki, "Digital frequency determination of real waveforms based on multiple sensors with low sampling rates," *IEEE Sens. J.*, 12: 1483-1495, 2012.
- [2] A. Maroosi, H. K. Bizaki, "Multiple frequencies determination of sinusoidal real waveform by multiple sensors with low sampling rate," *IEEE Sens. J.*, 17: 8404-8411, 2017.
- [3] H. Xiao, G. Xiao, "On solving ambiguity resolution with robust chinese remainder theorem for multiple numbers," *IEEE Trans. Veh. Technol.*, 68: 5179-5184, 2019.
- [4] C. Cao, Y. Zhao, "Range estimation based on symmetry polynomial aided Chinese remainder theorem for multiple targets in a pulse Doppler radar," *Front. Inf. Technol. Electron. Eng.*, 23: 304-316, 2022.
- [5] X. Li, H. Liang, X. G. Xia, "A robust Chinese remainder theorem with its applications in frequency estimation from undersampled waveforms," *IEEE Trans. Signal Process.*, 57: 4314-4322, 2009.
- [6] H. Xiao, N. Du, Z. Wang, G. Xiao, "Wrapped ambiguity Gaussian mixed model with applications in sparse sampling based multiple parameter estimation," *Signal Process.*, 179: 107825, 2021.
- [7] C. Chenghu, Z. Yongbo, P. Xiaojiao, X. Baoqing, C. Sheng, "A method based on Chinese remainder theorem with all phase DFT for DOA estimation in sparse array," *J. Syst. Eng. Electron.*, 31: 1-11, 2020.
- [8] G. Jinyuan, G. Xiaohui, Z. Guoan, D. Wei, "UAV-Relaying cooperation for internet of everything with CRT-Based NOMA," *Wireless Commun. Mobile Comput.*, 2021: 1-7, 2021.
- [9] U. Sudibyo, F. Eranisa, E. H. Rachmawanto, C. A. Sari, "A secure image watermarking using Chinese remainder theorem based on haar wavelet transform," in *Proc. 2017 4th International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*: 208-212, 2017.
- [10] X. Jing, Z. Yan, X. Jiang, W. Pedrycz, "Network traffic fusion and analysis against DDoS flooding attacks with a novel reversible sketch," *Inf. Fusion*, 51: 100-113, 2019.
- [11] J. Zhang, J. Cui, H. Zhong, Z. Chen, L. Liu, "PA-CRT: Chinese remainder theorem based conditional privacy-preserving authentication scheme in vehicular ad-hoc networks," *IEEE Trans. Dependable Secure Comput.*, 18(2): 722-735, 2019.
- [12] H. Prasetyo, J. M. Guo, "A note on multiple secret sharing using Chinese remainder theorem and exclusive-OR," *IEEE Access*, 7: 37473-37497, 2019.
- [13] A. Ghaemi, H. Danyali, K. Kazemi, "Simple, robust and secure data hiding based on CRT feature extraction and closed-loop chaotic encryption system," *J. Real-Time Image Process.*, 18: 221-232, 2021.
- [14] A. Ananthalakshmi, P. Rajagopalan, "VLSI implementation of residue number system based efficient digital signal processor architecture for wireless sensor nodes," *Int. J. Inf. Technol.*, 11: 829-840, 2019.
- [15] E. Indhumathi, V. Babydeepa, "Energy efficient and reliable CRT based packet forwarding technique in wireless sensor networks," *ADALAY J.*, 9: 377-380, 2020.
- [16] N. Fu, Z. Wei, L. Qiao, Z. Yan, "Short-observation measurement of multiple sinusoids with multichannel sub-nyquist sampling," *IEEE Trans. Instrum. Meas.*, 69: 6853-6869, 2020.
- [17] L. Xiao, X.-G. Xia, "Frequency determination from truly sub-Nyquist samplers based on robust Chinese remainder theorem," *Signal Process.*, 150: 248-258, 2018.
- [18] J. M. Junior, J. P. C. da Costa, F. Römer, R. K. Miranda, M. A. Marinho, G. Del Galdo, "M-estimator based Chinese Remainder Theorem with few remainders using a Kronecker product-based mapping vector," *Digital Signal Process.*, 87: 60-74, 2019.
- [19] H. Liao, X. G. Xia, "A sharpened dynamic range of a generalized Chinese remainder theorem for multiple integers," *IEEE Trans. Inf. Theory*, 53: 428-433, 2006.
- [20] X. G. Xia, "An efficient frequency-determination algorithm from multiple undersampled waveforms," *IEEE Signal Process. Lett.*, 7: 34-37, 2000.
- [21] H. Xiao, G. Xiao, "Notes on CRT-based robust frequency estimation," *Signal process.*, 133: 13-17, 2017.
- [22] L. Xiao, X.-G. Xia, H. Huo, "New conditions on achieving the maximal possible dynamic range for a generalized Chinese remainder theorem of multiple integers," *IEEE Signal Process. Lett.*, 22: 2199-2203, 2015.
- [23] W. Wang, X. Li, X. G. Xia, W. Wang, "The largest dynamic range of a generalized Chinese remainder theorem for two integers," *IEEE Signal Process. Lett.*, 22: 254-258, 2014.
- [24] X. Li, X. G. Xia, W. Wang, W. Wang, "A robust generalized Chinese remainder theorem for two integers," *IEEE Trans. Inf. Theory*, 62: 7491-7504, 2016.
- [25] X. Li, Y. Cao, B. Yao, F. Liu, "Robust generalized Chinese-remainder-theorem-based DOA estimation for a coprime array," *IEEE Access*, 6: 60361-60368, 2018.
- [26] X. Li, Y. Liu, H. Chen, C. C. Chang, "A secret image restoring scheme using threshold pairs of unordered image shares," *IEEE Access*, 7: 118249-118258, 2019.
- [27] X. Li, Y. Liu, H. Chen, C. C. Chang, "A novel secret sharing scheme using multiple share images," *Math. Biosci. Eng.: MBE*, 16: 6350-6366, 2019.
- [28] G. Zhou, X. G. Xia, "Multiple frequency detection in undersampled complex-valued waveforms with close multiple frequencies," *Electron. Lett.*, 33: 1294-1295, 1997.
- [29] X. G. Xia, "On estimation of multiple frequencies in undersampled complex valued waveforms," *IEEE Trans. Signal process.*, 47: 3417-3419, 1999.
- [30] X. G. Xia, K. Liu, "A generalized Chinese remainder theorem for residue sets with errors and its application in frequency determination from multiple sensors with low sampling rates," *IEEE Signal Process. Lett.*, 12: 768-771, 2005.

Biographies



Ali Maroosi is an assistant professor at Department of Computer Engineering, University of Torbat Heydarieh. His research interests are parallel and distributed computing, signal and data processing, networks, intelligence algorithms, etc.

- Email: ali.maroosi@torbath.ac.ir
- ORCID: [0000-0001-6078-655X](https://orcid.org/0000-0001-6078-655X)
- Web of Science Researcher ID: NA
- Scopus Author ID: 16550228100
- Homepage: https://www.torbath.ac.ir/sp-a-maroosi#tabcontrol_12



Hossein Khaleghi Bizaki received his Ph.D. degree in communication engineering from Iran University of Science & Technology (IUST), Tehran, Iran, in 2008. Dr. Bizaki is author or co-author of more than 30 publications. His research interests include Information Theory, Coding Theory, Wireless Communication, and other topics on communication systems and signal processing.

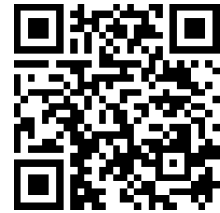
- Email: bizaki@yahoo.com
- ORCID: [0000-0001-9458-8287](https://orcid.org/0000-0001-9458-8287)
- Web of Science Researcher ID: NA
- Scopus Author ID: 15061014800
- Homepage: NA

How to cite this paper:

A. Maroosi, H. Khaleghi Bizaki, "Determination of the maximum dynamic range of sinusoidal frequencies in a wireless sensor network with low sampling rate," J. Electr. Comput. Eng. Innovations, 11(2): 419-432, 2023.

DOI: [10.22061/jecei.2023.9290.602](https://doi.org/10.22061/jecei.2023.9290.602)

URL: https://jecei.sru.ac.ir/article_1877.html





Research paper

A New Low-Stress Boost Converter with Soft-Switching and Using Coupled-Inductor Active Auxiliary Circuit

M. A. Latifzadeh¹, P. Amiri^{2,*}, H. Allahyari³, H. Faezi¹

¹Faculty of Electrical and Computer Engineering, Malek-Ashtar University of Technology, Tehran Iran.

²Faculty of Electrical Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

³Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran.

Article Info

Article History:

Received 28 January 2023

Reviewed 17 March 2023

Revised 06 April 2023

Accepted 15 May 2023

Keywords:

Boost converter

ZVS soft-switching

DC/DC converter

PWM converter

*Corresponding Author's Email
Address: pamiri@sru.ac.ir

Abstract

Background and Objectives: Many applications use boost converters as front-end circuits, including power factor correction (PFC), solar power generation, fuel cell power conversion, battery chargers, and uninterruptible power supply. In addition, boost converters have a simple structure with low component counts, which makes them a convenient choice.

Methods: This article proposes a coupled-inductor active auxiliary circuit to create a new low-stress boost converter with soft-switching. The proposed auxiliary circuit supplies the main switch and diode with soft-switching ZVC turn-on and ZCS turn-off states. The main switch and diode are not deal with any extra stress of voltage or current. Furthermore, the soft switching condition is also provided for auxiliary circuit components.

Results: The proposed auxiliary circuit also has a simple structure, low circulating current losses, low cost, and simplicity in control. The operation state and performance of the proposed soft-switching boost converter are examined, and the design procedure is presented. Finally, a 200W prototype is implemented and tested to validate the theoretical results. The offered experimental data verified the theoretical analysis.

Conclusion: This paper provides a new low-stress soft-switching boost converter using a simple coupled-inductor in the auxiliary circuit. Moreover, the auxiliary part consists of two diodes, one switch, one resonance capacitor, and a coupled inductor. The suggested auxiliary circuit provides soft switching condition for the main switch, which provides ZVS in the turn-on transient and ZCS in the turn-off transient, while in this situation, the soft-switching condition is provided for the auxiliary switch, which turns on under ZCS and also turns off with practically ZVS conditions. The auxiliary circuit does not impose additional voltage or current stress on the main switch. A 200 W prototype is implemented to validate the performance of this snubber cell. The experimental data reported here support the theoretical analysis. The best point of efficiency is 95.9% which is occurred at maximum load, and is 6.3% greater than the traditional counterparts.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Many applications use boost converters as front-end circuits, including power factor correction (PFC), solar

power generation, uninterruptible power supply (UPS), fuel cell power conversion, battery chargers, and car high-intensity discharge (HID) headlamps [1]-[7]. Regarding

boost, converters have a simple structure with low component counts, straightforward to implement, design, and control [8]-[10]. However, the main switch of a boost converter is hard switched, and switching losses and reverse recovery problems of the diode limit the efficiency, power density, and switching frequency. Furthermore, hard switching intensifies electromagnetic interferences (EMI) [11]-[13]. On the other hand, by introducing and developing soft-switching techniques which remove switching losses, these converters can operate at high switching frequencies. The significant advantages of a soft-switching converter are small size, lightweight, low EMI, and high-power density [14]-[15]. Meanwhile, soft-switching PWM converters, like PWM hard-switching converters, use a simple control circuit to design and experiment.

Recently, many passive and active auxiliary circuits have been proposed for the boost converter [16]-[30]. Although passive auxiliary circuits do not need any auxiliary switches and have a simple control method, their construction is typically complex and causes extra voltage and also current stress over the switch and diode of the converter [16]-[19]. Furthermore, this solution can only provide a turn-on ZCS condition and cannot recover capacitive turn-on losses. On the other hand, the active auxiliary circuits are proposed in [20]-[21] provide soft switching conditions for both the main switch and the main diode, although they operate with additional current stress on the main switch. The converter proposed in [22] provides the ZCS condition for the main switch. The primary switch voltage stress is increased in [23]-[24] by the series inductor that is used in the power path. In [25], the soft switching condition is achieved by way of having two auxiliary switches and increasing the converter's cost. [26]-[29] suggest an auxiliary circuit with a large number of components to decrease the current and voltage stress on the primary switch, however this increases cost and lowers efficiency and reliability. Higher conduction losses are caused by [20]-[22] and [25]-[29]'s significant circulation current losses and current stress through the auxiliary switches. The auxiliary switch in [30] also functions when being stressed by high voltage.

In this research, a novel coupled-inductor auxiliary circuit-based low stress soft-switching boost converter is suggested. The recommended auxiliary circuit corrects the issues listed above. The recommended auxiliary circuit turns on and off the main switch and main diode in ZVS and ZCS scenarios. Additionally, when ZCS requirements are met, the auxiliary switch activates, and when switching losses are minimal, it deactivates. The primary switch or main diode are not subjected to any additional current or voltage stress from the suggested auxiliary circuit. The recommended auxiliary circuit also features a straightforward design, little circulating

current, and low losses.

Operating Statuses and Performance Analysis

Fig. 1 depicts the schematic of the proposed converter. Generally, this converter is divided into two sections: a) the boost converter and b) the auxiliary circuit. The input inductor L_1 , the main switch S , the output capacitor C_o , and the main diode D_o comprise the boost converter.

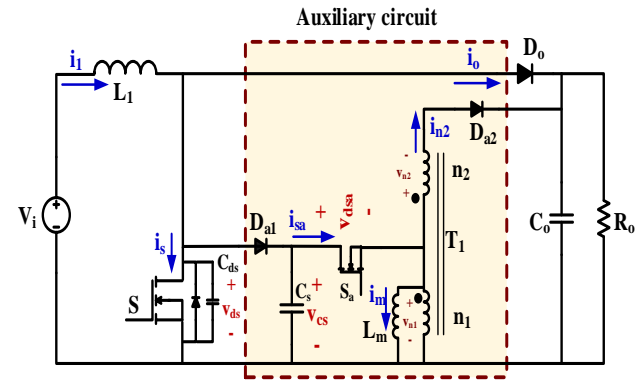


Fig. 1: The schematic of the proposed converter.

The auxiliary circuit is made up of the following components: the auxiliary input diode D_{a1} , the auxiliary capacitor C_s , the auxiliary switch S_a , the coupled inductors T_1 , and the auxiliary output diode D_{a2} . T_1 is modeled by an ideal transformer with the turns ratio $n_1:n_2$ and magnetic inductance L_m , as shown in Fig. 1. Furthermore, C_{ds} is the drain-source parasitic capacitance of the primary switch S . The following assumptions simplify the analysis of the proposed circuit:

- The suggested converter works in a steady-state mode.
- Because the output capacitor C_o and the input inductor L_1 are both large enough, the output capacitor voltage and the input inductor current will be constant over a switching cycle.
- it was supposed that all components of this converter operate without losses, with the exception of the drain-source capacitance of the main switch.

During each switching cycle, seven operating states are identified to illustrate the operating principle of the suggested boost converter. Fig. 2 depicts the corresponding sub-intervals of the proposed structure. The key waveforms of the operational states are also illustrated in Fig. 3. The voltage is expected to be the same as the output voltage V_o before the initial status.

Operating Statuses

Status 1 [t_0, t_1]: Previous to this status, D_o is conducting and S is turned off. This mode is initiated by applying the gate signal to the auxiliary switch S_a , and it is turned on under ZCS conditions. When the output voltage is applied across the magnetizing inductor of the L_m , the current i_m increases linearly to I_b , and the current I_{D_o} falls linearly to

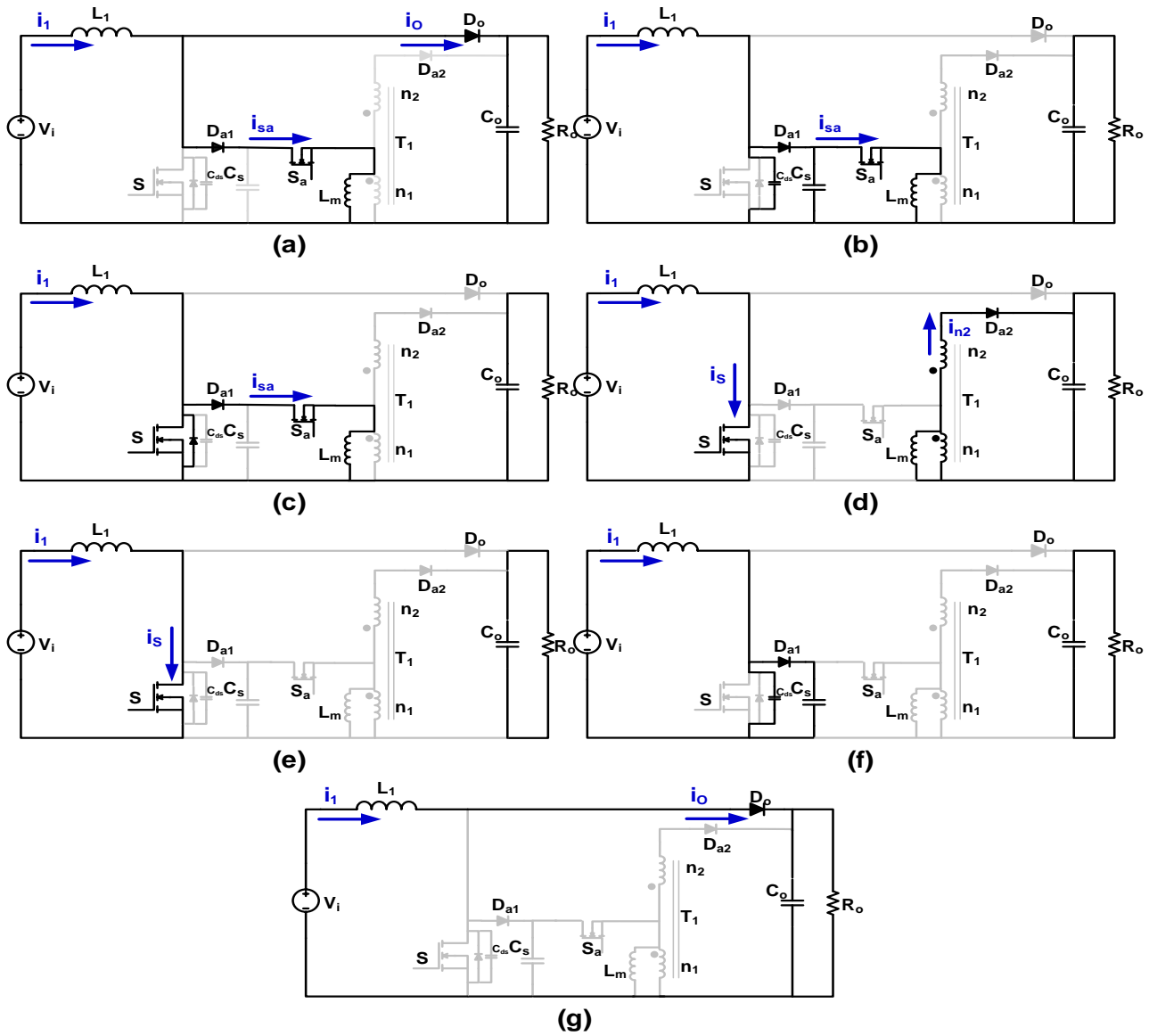


Fig. 2: The equivalent configurations of the proposed converter at operating subintervals (a) Status1, (b) Status2, (c) Status3, (d) Status4, (e) Status5, (f) Status6, (g) Status7.

zero. Under ZCS conditions, the output diode D_o turns off, and the following status begins. The corresponding configuration of this status is depicted in Fig. 2a. The equations of this mode are derived as:

$$i_m = i_{sa} = \frac{V_o}{L_m}(t - t_0) \quad (1)$$

$$v_{cs} = V_o \quad (2)$$

L_m represents the magnetizing inductance of T_1 , and V_o is the output voltage.

This is how the duration of this mode is expressed:

$$\Delta t_1 = \frac{I_b L_m}{V} \quad (3)$$

Status 2 [t_1 , t_2]: When the output diode turns off, this subinterval starts. The magnetizing inductor L_m and the resonant capacitor C_s combine to create a resonant

circuit. In a resonance, as depicted in Fig. 3, the voltage of the capacitor C_s starts to fall from the output voltage to zero. When the voltage of the resonant capacitor is fully depleted, this status is finished. The corresponding configuration of this state is depicted in Fig. 2b. The voltage of the capacitor C_s and the current passes through the auxiliary switch S_a are calculated using the following equations.

$$v_{cs} = V_o \cos(\omega_2(t - t_1)) \quad (4)$$

$$i_{sa} = i_m = I_b + V_o \sqrt{\frac{C_s}{L_m}} \sin(\omega_2(t - t_1)) \quad (5)$$

where ω_2 is the status 2 auxiliary circuit resonance frequency.

The duration of this status can be obtained as follows:

$$\Delta t_2 = \pi \sqrt{C_s L_m} \quad (6)$$

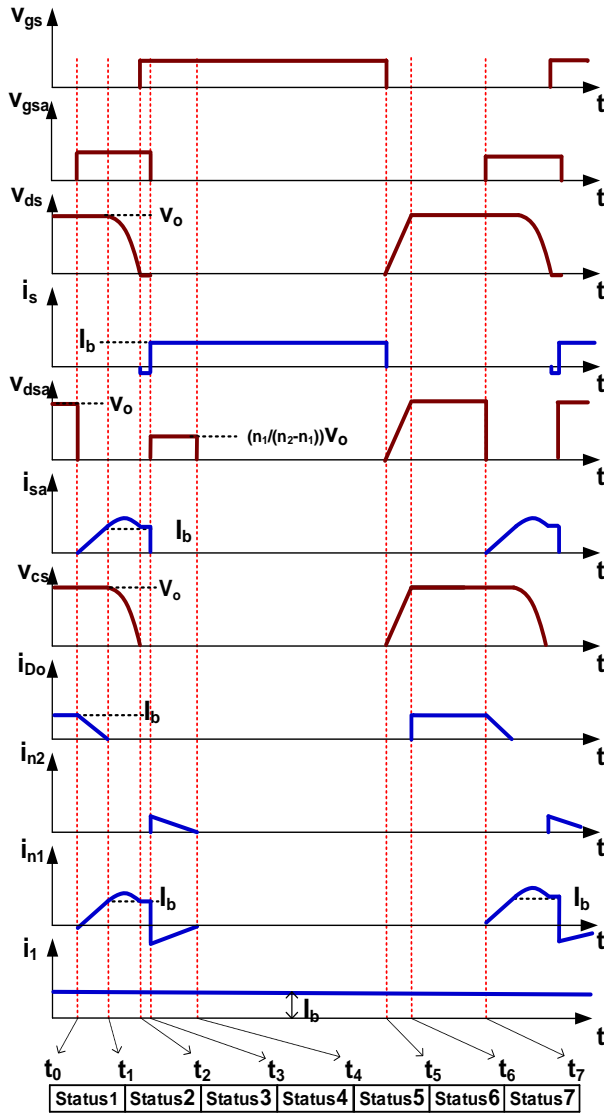


Fig. 3: The key waveforms of the proposed active snubber cell boost converter.

Status 3 [t2, t3]: The main switch S turns on at the start of this mode. Moreover, the input auxiliary diode D_{a1} and the anti-parallel diode of the main switch are conducting. Fig. 2c depicts the equivalent circuit of this mode. The time duration of this subinterval is short and negligible. At the end of this mode, the current flowing through the auxiliary diode D_{a1} and the switch's body diode are zero, and the auxiliary switch Sa is off with zero current flowing through it. The main switch and input inductance function as a straightforward boost circuit at the mode's conclusion, while the coupled inductors discharge the energy stored to the output similar to the flyback circuit.

Status 4 [t3, t4]: At the start of this status, the auxiliary switch S_a is off. The windings n_1 , n_2 , and the output auxiliary diode D_{a2} discharge the energy held in the magnetizing inductor L_m . The auxiliary switch turns off at practically ZVS condition by choosing the appropriate turns ratio for the coupled inductors. The primary switch

S also carries the input current I_b . The equivalent circuit of this status is shown in Fig. 2d. At the end of this status, the auxiliary diode D_{a2} turns off and the energy stored in the magnetizing inductor L_m is fully discharged. The voltages across the windings n_1 and n_2 and the currents flowing through the winding n_2 and the output auxiliary diode D_{a2} are stated as follows:

$$i_{n2} = i_{D_{a2}} = \frac{n_1}{n_2 - n_1} I_b \quad (7)$$

$$v_{n1} = \frac{n_1}{n_1 - n_2} V_o \quad (8)$$

$$v_{n2} = \frac{n_2}{n_1 - n_2} V_o \quad (9)$$

where n_1 is the primary and n_2 is the secondary winding of the coupled inductor T1.

Following is the description of timeframe for this status:

$$\Delta t_4 = \left(\frac{n_2 - n_1}{n_1} \right) \frac{L_m I_b}{V_o} \quad (10)$$

Status 5 [t4, t5]: In this state, the primary switch S is conducting, the output diode of the boost converter is off, and the magnetizing inductances of coupled inductors are fully discharged in this state. The corresponding configuration of this state is shown in Fig. 2e. This condition is comparable to the typical boost converter.

Status 6 [t5, t6]: The main switch S does not conduct at the start of this condition. The input current i_1 parallels the parasitic drain-source capacitance C_{ds} of the main switch by passing via the resonant capacitors C_s . As a result, under ZVS conditions, the voltage of the primary switch rises linearly, and the switch turns off. The output diode turns on under ZVS when the voltage of the main switch reaches the output voltage. The voltage of the resonant capacitor and the voltage of the main switch are thereby constrained to the output voltage. The equivalent configuration of this status is shown in Fig. 2f. Following is the equation for this status:

$$v_{Cs} = v_{ds} = v_{dsa} = \frac{I_b}{C_s} (t - t_5) \quad (11)$$

which the period of this mode is derived as follows:

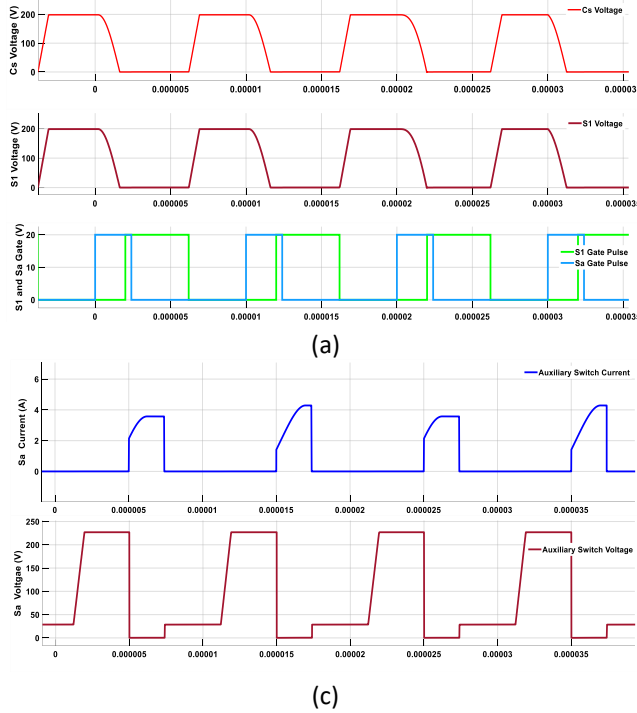
$$\Delta t_6 = C_s \frac{V_o}{I_b} \quad (12)$$

Status 7 [t6, t7]: The input inductor is being discharged at the start of this condition, and the output diode D_o becomes forward-biased. The suggested converter functions similarly to a standard boost converter in this state. The corresponding configuration of this state is shown in Fig. 2g.

Simulation Results

To verify the performance of the proposed active

snubber cell and compare the experimental results with simulation results, the waveforms of the proposed converter are extracted. The gate commands of switches and the voltage of the resonance capacitor and main switches are shown in Fig. 4(a), and it can be seen the main switch turns on just after its voltage reaches zero by resonance occurring between the resonance and body capacitor of the main switch and magnetizing inductance.



The voltage and current waveforms of the main switch are shown in Fig. 4(b).

Moreover, to monitor the performance of the auxiliary switch, its current and voltage waveform are illustrated in Fig. 4(c), and the input and output current of the coupled inductor are depicted in Fig. 4(d) by measuring the current of the auxiliary switch S_a and current of the auxiliary diode D_{a2} .

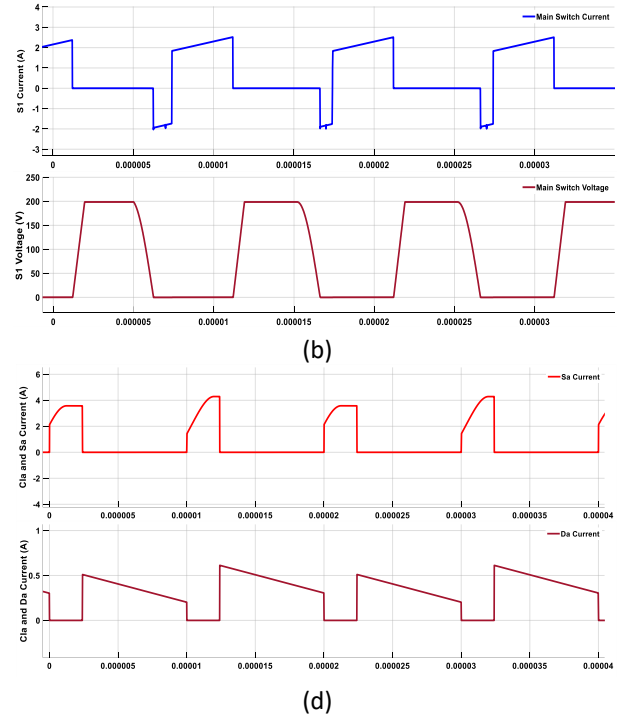


Fig. 4: A 200 W simulation result of the suggested active snubber cell, (a) the gate commands of switches and resonance and main switch voltage, (b) the main switch's voltage and current, (c) the auxiliary switch's voltage and current, (d) measured current of coupled inductors which passed the auxiliary snubber switch and the auxiliary output diode.

Stresses Analysis of the Proposed Converter

At the end of status 6, the peak voltage across the switches S , S_a , and can be calculated as follows:

$$V_S = V_o \quad (13)$$

$$V_{S_a} = \left(\frac{n_1}{n_2 - n_1} \right) V_o \quad (14)$$

The voltage stress across S_a is equal to V_o and the peak current passes through the main switch S can be determined as follow:

$$i_S = \frac{P_{out}}{V_{in}} + \frac{\Delta I}{2} \quad (15)$$

Where ΔI is the current ripple of the input inductor I_b . Also, the peak current which passes through the auxiliary switch is obtained from mode 2:

$$i_{S_a} = I_b + V_o \sqrt{\frac{C_s}{L_m}} \quad (16)$$

The voltage and current stresses of the main diode D_o is calculated as follows:

$$V_{D_o} = V_o \quad (17)$$

$$i_{D_o} = \frac{P_{out}}{V_{in}} + \frac{\Delta I}{2} \quad (18)$$

The voltage stress and the current peak of the input and output auxiliary diodes D_{a1} , D_{a2} are calculated using:

$$V_{D_{a2}} = \frac{n_2}{n_1} V_o \quad (19)$$

$$V_{D_{a1}} \cong 0 \quad (20)$$

$$i_{D_{a1}} = I_b + V_o \sqrt{\frac{C_s}{L_m}} \quad (21)$$

$$i_{D_{a2}} = \frac{n_1}{n_2 - n_1} I_b \quad (22)$$

Design Method

When the input current I_b at the end of status 1 exceeds the current flowing through magnetizing inductance of the coupled inductor T_1 , the main switch S is in the soft-switching condition. As a result, as long as the following equation is established, the main switch works under soft-switching conditions:

$$I_b < i_m(t_1) = \frac{V_o}{L_m} \Delta t_1 \quad (23)$$

Where the Δt_1 is obtained from (3).

By considering $i_m(t_1)$ 15% higher than I_b , the following relation is obtained:

$$1.15 I_b = \frac{V_o}{L_m} \Delta t_1 \quad (24)$$

Additionally, compared to the main switching cycle, the operating time of the auxiliary circuit must be insignificant. This duration is therefore chosen to be ten times shorter than the switching period.

$$\Delta t_1 + \Delta t_2 + \Delta t_4 + \Delta t_6 = \frac{T_{switching}}{10} \quad (25)$$

Where Δt_i : $i = 1, 2, 4, 6$ are the time duration of statuses 1, 2, 4, and 6 and are given from (3), (6), (10), and (12) and $T_{switching}$ is the switching cycle of the proposed converter.

From (24), the inductance L_m is calculated by taking into consideration a specific value for the time duration. The ratio between the peak current of the auxiliary switch and the main switch is used to determine the value of the resonant capacitor C_s .

$$\left(\frac{I_{sa}}{I_s}\right)_{peak} = 1 + \frac{V_o}{I_b} \sqrt{\frac{C_s}{L_m}} = \lambda \quad (26)$$

The longer the time period for λ is obtained, the larger the value for Δt_2 is. The turn ratio of the primary winding of coupled inductor depends on the value of the inductance L_m . The minimum voltage stress on the auxiliary components S_a and D_{a2} is used to calculate the turns ratio $n_1:n_2$ of the coupled inductor. The condition (26) should be examined following the design of the resonant capacitor C_s and the magnetizing inductance L_m . The equations in "Stresses analysis of the proposed converter" section is used to choose all active components. Similar to a traditional PWM boost converter, the values for L_b and C_o are obtained based on the desired current and voltage ripple.

Furthermore, to compare the proposed snubber cell with its counterparts in terms of voltage gain, output power, switching frequency, efficiency, and component counts, such as the number of auxiliary switches and diode, capacitor, inductor, and coupled inductors, a comprehensive comparison is made, and the results are presented in Table 1.

Experimental Results

A 200 W implementation with supply voltage of 90 V, output voltage of 200 V, and switching frequency of 100 kHz has been constructed and evaluated to corroborate the findings of the theoretical analysis of the suggested circuit. The experimental specifications for this converter are shown in Table 2, and the actual converter's circuit. The experimental specifications for this converter are shown in Table 2, and the actual converter's circuit schematic is shown in Fig. 7. The magnetizing inductance L_m 100 μ H and the resonant capacitor C_s 10 nF are both meant for the auxiliary circuit. The linked inductor is also selected with a 2:16 turns ratio. The input inductor is set to have a value of 400 μ H in the interim.

Table 1: proposed converter and its counterpart's comparison

Refs	Efficiency [%]	Experimental Parameters			Number of Auxiliary components				
		V_{out}/V_{in}	P_{out} [W]	f_{sw} [kHz]	Switch	Diode	C	L	Coupled Ind.
[12], [13]	96.23	2.67	600	30	1	2	2	2	0
[16]	96	2	200	100	0	4	1	3	0
[17]	97	1.5	300	180	0	2	2	1	0
[20]	93.9	-	300	20	1	4	2	2	1
[23]	>97	1.07	1200	80	1	1	2	1	0
[24]	98	1.29	1000	100	1	4	2	2	0
[25]	96	1.6	2000	30	2	2	1	0	0
[27]	97	2	100	100	0	3	1	3	0
[30]	98.28	2	200	32.2	1	3	1	1	0
Proposed	95.9	2.2	200	100	1	2	1	0	1

Table 2: The parameters of the experimental prototype

Description	Symbol	Value
Input Voltage	V_{in}	90 V
Output Voltage	V_O	200 V
Max. Output Power	P_O	200 W
Frequency of main switch	f_s	100 kHz
Frequency of auxiliary switch	f_{sa}	100 kHz
Input Inductor	L_1	400 μ H
Resonance Capacitor	C_s	10 nF
Magnetizing Inductance	L_m	100 μ H
Output Capacitor	C_O	100 μ F
Coupled Inductor Core	T_1	E34/14/9
Turn Ratio of CI	n_1/n_2	2/16

Fig. 5a shows the experimental voltage and current of the primary switch. The primary switch turns on under ZVS conditions, and switches off under ZCS situations. As previously mentioned, the snubber circuit also prevents any voltage or current strains from being applied to the primary switch. Fig. 5b displays the auxiliary switch's voltage and current.

As can be observed, the auxiliary switch has an output voltage-equivalent voltage stress, operates under ZCS, and shuts off with low switching losses. Additionally, the peak current of the auxiliary switch is a little bit higher than that of the main switch.

The charge and discharge situation of the resonant capacitor in terms of voltage to achieve the soft switching condition are shown in Fig. 5c. Fig. 5d demonstrates the waveforms of coupled inductors, where the core is charged first, then is discharged when the auxiliary switch is turned off.

To depict the volume and dimensions of this converter, the implemented prototype of the proposed circuit is shown in Fig. 6.

Furthermore, the part number of all active devices and the values of passive devices are noted and presented in Fig. 7.

In Fig. 8, the measured efficiency of this converter is provided. The nominal load of 200 W yields a maximum efficiency of 95.9%, an increase of 6.3% in efficiency over a traditional boost converter.

Notably, all circumstances, including input voltage, switching frequency, and voltage gain, are considered the same for a fair comparison of the efficiency between this article and the conventional boost converter.

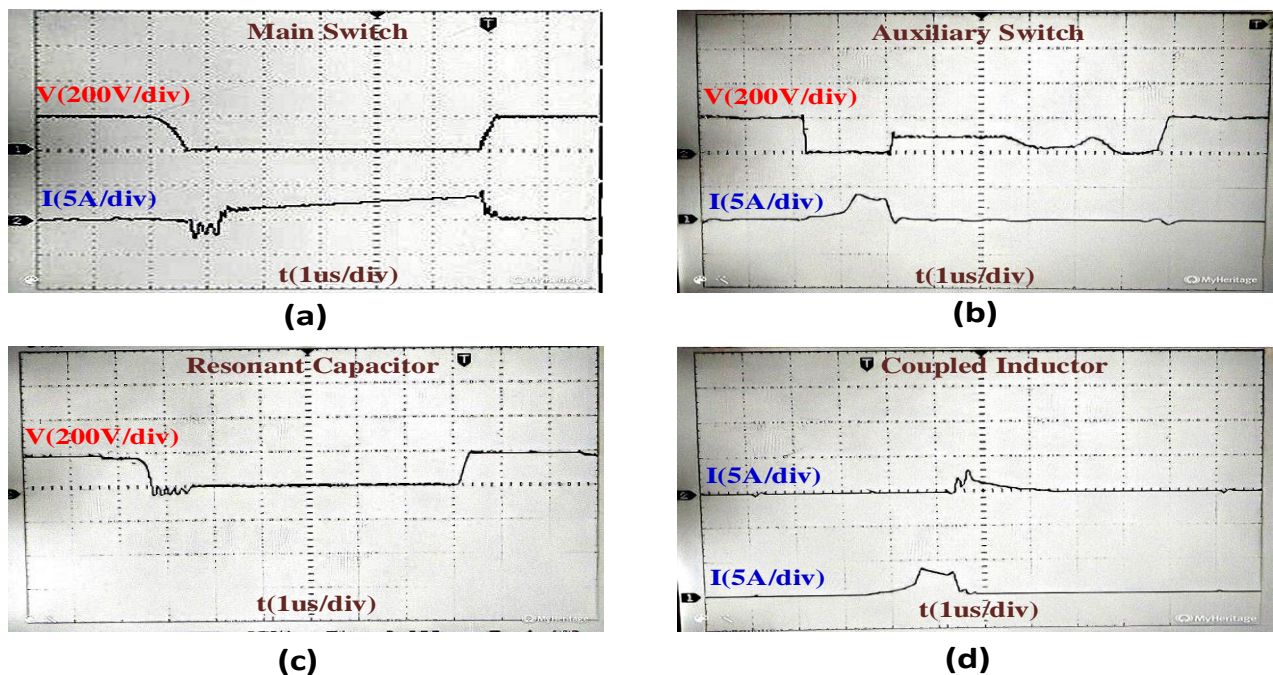


Fig. 5: A 200 W experimental result of the suggested active snubber cell, (a) the main switch's voltage and current, (b) the auxiliary switch's voltage and current, (c) voltage of resonant capacitor, (d) measured current of coupled inductors which passed the auxiliary snubber switch and the auxiliary output diode.

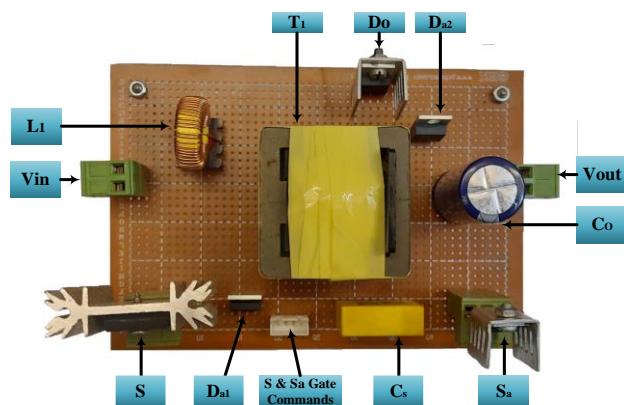


Fig. 6: The implemented figure of the suggested circuit.

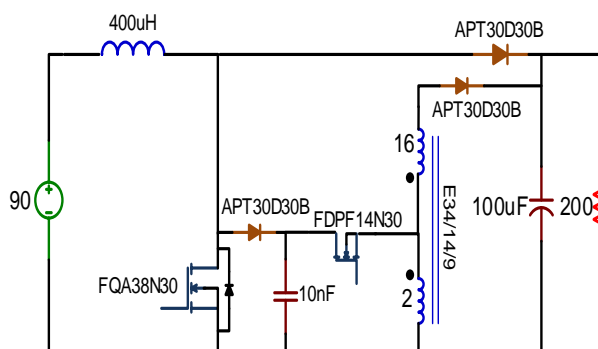


Fig. 7: Parameters of the implemented prototype.

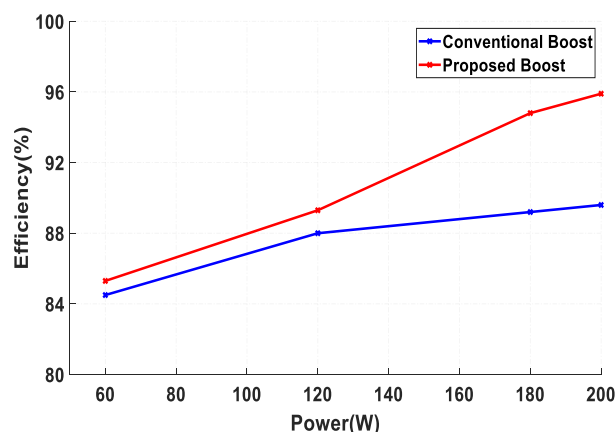


Fig. 8: The measure efficiency of the proposed converter.

Conclusion

This paper provides a new low-stress soft-switching boost converter using a simple coupled-inductor in the auxiliary circuit. Moreover, the auxiliary part consists of two diodes, one switch, one resonance capacitor, and a coupled inductor. The suggested auxiliary circuit provides soft switching condition for the main switch, which provides ZVS in turn-on transient and ZCS in turn-off

transient, while in this situation, the soft-switching condition is provided for the auxiliary switch in which turns on under ZCS and also turns off with practically ZVS conditions. The auxiliary circuit does not impose additional voltage or current stress to the main switch. A 200 W prototype is implemented to validate the performance of this snubber cell. The experimental data reported here support the theoretical analysis. The best point of efficiency is 95.9% which is occurred at maximum load, and is 6.3% greater than the traditional counterparts. The number of components in this topology is more efficient than other counterparts. Still, the auxiliary switch, a high-side switch, could be considered one of the main challenging issues in this topology. On the other hand, this auxiliary switch operates under ZCS, which is suitable for IGBT switches, not MOSFET.

Author Contributions

Mohammad Ali Latifzadeh in collaboration with Hesamodin Allahyari and Hadi Faezi, designed, simulated, carried out the data analysis, implemented the prototype and wrote the manuscript thoroughly under supervising of dr. Parviz Amiri.

Acknowledgment

The editors of JECEI and the anonymous reviewers are sincerely appreciated for their insightful criticism and recommendations, which the authors acknowledge in advance.

Conflict of Interest

Regarding the publishing of this study, the authors state that there are no possible conflicts of interest. The authors have also fully observed all ethical difficulties, such as plagiarism, informed consent, misconduct, data fabrication or falsification, duplicate publishing or submission, and redundancy.

Abbreviations

ZVT	Zero Voltage Transient
ZVS	Zero Voltage Switching
ZCS	Zero Current Switching
PWM	Pulse Width Modulation
EMI	Electro-Magnetic Interference
CIs	Coupled Inductors

References

- [1] H. Soltani Gohari, K. Abbaszadeh, "Bidirectional Buck-Boost integrated converter for plug-in hybrid electric vehicles," *J. Electr. Comput. Eng. Innovations (JECEI)*, 8(1): 109-124, 2020.
- [2] P. Amiri, M. Sharafi, "A high efficiency low-voltage soft switching DC-DC converter for portable applications," *J. Electr. Comput. Eng. Innovations (JECEI)*, 1(2): 73-81, 2013.
- [3] V. K. Goyal, A. Shukla, "Isolated DC-DC boost converter for wide input voltage range and wide load range applications," *IEEE Trans. Ind. Electron.*, 68(10): 9527-9539, 2021.

- [4] A. Shoaee, K. Abbaszadeh, H. Allahyari, "A single-inductor multi-input multi-level high step-up DC-DC converter based on switched-diode-capacitor cells for PV applications," *IEEE J. Emerging Sel. Top. Ind. Electron.*, 4(1): 18-27, 2022.
- [5] R. S. Inomoto, J. R. B. A. Monteiro, A. J. Sguarezi Filho, "Boost converter control of PV system using sliding mode control with integrative sliding surface," *IEEE J. Emerging Sel. Top. Power Electron.*, 10(5): 5522-5530, 2022.
- [6] F. M. Shahir, E. Babaei, M. Farsadi, "Extended topology for a boost DC-DC Converter," *IEEE Trans. Power Electron.*, 34(3): 2375-2384, 2019.
- [7] N. Molavi, E. Adib, H. Farzanehfard, "Soft-switched non-isolated high step-up DC-DC converter with reduced voltage stress," *IET Power Electron.*, 9(8): 1711-1718, 2016.
- [8] H. Bahrami, H. Iman-Eini, B. Kazemi, A. Taheri, "Modified step-up boost converter with coupled-inductor and super-lift techniques," *IET Power Electron.*, 8(6): 898-905, 2015.
- [9] M. Delzende Sarfejo, H. Allahyari, H. Bahrami, A. Afifi, M. Latif Zadeh, E. Yavari, M. Ghavidel Jalise, "A passive compensator for imbalances in current sharing of parallel-SiC MOSFETs based on planar transformer," *IET Power Electron.*, 14(14): 2400-2412, 2021.
- [10] M. Nikbakht, S. Abbasian, M. Farsijani, K. Abbaszadeh, "An ultra high gain double switch quadratic boost coupled inductor based converter," in *Proc. 2022 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC)*: 167-172, 2022.
- [11] M. Fazeli-Hasanabadi, A. Shoaee, K. Abbaszadeh, H. Allahyari, "An interleaved high step-up dual-input single-output DC-DC converter for electric vehicles," in *Proc. 2022 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC)*: 145-149, 2022.
- [12] F. Falahi, E. Babaei, S. Bagheri, "Soft-switched interleaved high step-up non-isolated DC-DC converter with high voltage gain ratio," in *Proc. 2022 13th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC)*: 128-133, 2022.
- [13] S. Park, G. Cha, Y. Jung, C. Won, "Design and application for PV generation system using a soft-switching boost converter with SARC," *IEEE Trans. Ind. Electron.*, 57(2): 515-522, 2010.
- [14] H. Bahrami, H. Allahyari, E. Adib, "An improved wide ZVS soft-switching range PWM bidirectional forward converter for low power applications with simple control circuit," *IET Power Electron.*, 15(15): 1652-1663, 2022.
- [15] H. Bahrami, H. Allahyari, E. Adib, "A self-driven synchronous rectification ZCS PWM two-switch forward converter with minimum number of components," *IEEE Trans. Ind. Electron.*, 69(12): 12842-12850, 2022.
- [16] M. Mohammadi, E. Adib, M. R. Yazdani, "Family of soft-switching single-switch PWM converters with lossless passive snubber," *IEEE Trans. Ind. Electron.*, 62(6): 3473-3481, 2015.
- [17] H. Choe, Y. Chung, C. Sung, J. Yun, B. Kang, "Passive snubber for reducing switching-power losses of an IGBT in a DC-DC boost converter," *IEEE Trans. Power Electron.*, 29(12): 6332-6341, 2014.
- [18] T. Zhan, Y. Zhang, J. Nie, Y. Zhang, Z. Zhao, "A novel soft-switching boost converter with magnetically coupled resonant snubber," *IEEE Trans. Power Electron.*, 29(11): 5680-5687, 2014.
- [19] M. T. Kejani, S. H. Aleyasin, A. Safaeinasab, K. Abbaszadeh, "A new non-isolated single switch high step-up DC/DC converter based on inductor cells," in *Proc. 2021 12th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC)*: 1-5, 2021.
- [20] W. Qian, X. Zhang, Z. Li, "Design and operation analysis of a novel coupled-inductor based soft switching boost converter with an auxiliary switch," in *Proc. 2016 IEEE 8th Int. Power Electron. and Motion Control Conf. (IPEMC-ECCE Asia)*: 2534-2537, Hefei, 2016.
- [21] M. Ghavidel Jalise, H. Allahyari, A. Shoaee, E. Adib, H. Bahrami, M. A. Latif Zadeh, M. H. Fahimifar, "A self-driven synchronous rectification wide-range ZVS single-switch forward converter controlled using the variable inductance," *IET Power Electron.*, 16(2): 320-332, 2022.
- [22] H. Bahrami, E. Adib, S. Farhangi, H. Iman-Eini, R. Golmohammadi, "ZCS-PWM interleaved boost converter using resonance-clamp auxiliary circuit," *IET Power Electron.*, 10(3): 405-412, 2017.
- [23] Y. Jang, M. M. Jovanovic, "A new, soft-switched, high-power factor boost converter with IGBTs," *IEEE Trans. Power Electron.*, 17(4): 469-476, 2002.
- [24] S. Cetin, "Power-Factor-Corrected and fully soft-switched pwm boost converter," *IEEE Trans. Ind. Appl.*, 54(4): 3508-3517, 2018.
- [25] J. Bauman, M. Kazerani, "A novel capacitor-switched regenerative snubber for DC/DC boost converters," *IEEE Trans. Ind. Electron.*, 58(2): 514-523, 2011.
- [26] Y. Chen, Z. Li, R. Liang, "A novel soft-switching interleaved coupled-inductor boost converter with only single auxiliary circuit," *IEEE Trans. Power Electron.*, 33(3): 2267-2281, 2018.
- [27] J. A. Lambert, J. B. Vieira, L. Carlos de Freitas, L. dos Reis Barbosa, V. J. Farias, "A boost PWM soft-single-switched converter with low voltage and current stresses," *IEEE Trans. Power Electron.*, 13(1): 26-35, 1998.
- [28] N. Ting, I. Aksoy, Y. Sahin, "ZVT-PWM DC-DC boost converter with active snubber cell," *IET Power Electron.*, 10(2): 251-260, 2017.
- [29] C. Wang, C. Lin, C. Lu, J. Li, "Analysis, design, and realisation of a ZVT interleaved boost DC/DC converter with single ZVT auxiliary circuit," *IET Power Electron.*, 10(14): 1789-1799, 2017.
- [30] A. Mondzik, R. Stala, S. Piróg, A. Penczek, P. Gucwa, M. Szarek, "High efficiency DC-DC boost converter with passive snubber and reduced switching losses," *IEEE Trans. Ind. Electron.*, 69(3): 2500-2510, 2022.

Biographies



Mohamad Ali Latifzadeh received his B.Sc. degree in electrical engineering from Malek Ashtar University of Technology, Isfahan, Iran, in 2002, and the M.Sc. degree in electrical engineering from the Malek Ashtar University of Technology, Tehran, Iran, in 2007 and PhD degree in electrical engineering from the Malek Ashtar University of Technology Tehran, Iran, in 2018. His research interests include design, modeling in DSP embedded module, tracking, mono-pulse tracking and satellite ground station systems.

- Email: mal_1358@yahoo.com
- ORCID: [0009-0001-6909-6261](https://orcid.org/0009-0001-6909-6261)
- Web of Science Researcher ID: NA
- Scopus Author ID: 57258099200
- Homepage: NA



Parviz Amiri was born in 1970. He received his B.Sc. degree from university of Mazandaran 1994, the M.Sc. from Khaje Nasir toosi University (KNTU Tehran, Iran) in 1997, and his Ph.D. from University of Tarbiat Modares (TMU Tehran, Iran) in 2010, all degrees in Electrical Engineering (Electronics). His main research includes electronic circuit design in industries. His primary research interest is in RF and

power electronics circuits, with focus on highly efficient and high linear power circuit design. He is currently with the Faculty of Electrical and Computer Engineering at Shahid Rajaee teacher Training University in Tehran, Iran.

- Email: pamiri@sru.ac.ir
- ORCID: [0009-0001-6909-6261](https://orcid.org/0009-0001-6909-6261)
- Web of Science Researcher ID: NA
- Scopus Author ID: 57258099200
- Homepage: NA



Hesamodin Allahyari was born in Tehran, Iran, in 1994. He received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2017, and the M.Sc. degree in electrical engineering from the K.N. Toosi University of Technology, Tehran, Iran, in 2020. He is currently with High Voltage Power Electronics Laboratory, K.N. Toosi University of Technology, Tehran, Iran. His research interests include dc-dc power converter, resonant converter, and pulse power supplies.

- Email: h.allahyari@email.kntu.ac.ir
- ORCID: [0000-0002-7098-6456](https://orcid.org/0000-0002-7098-6456)
- Web of Science Researcher ID: ABD-3635-2021
- Scopus Author ID: 57257874800
- Homepage: NA



Hadi Faezi was born in Sabzevar, Iran, in 1977. He received the B.Sc. degree in electrical engineering from Malek Ashtar University of Technology, Isfahan, Iran, in 2000, and the M.Sc. degree in electrical engineering from the School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran, in 2003 and the PhD degree in electrical engineering from the Iran University of Science and Technology (IUST), Tehran, Iran, in 2019. His research interests include Microwave, antennas and electromagnetic radiation, pulsed power supplies and power converters.

- Email: faezi@mut.ac.ir
- ORCID: [0000-0002-6303-3454](https://orcid.org/0000-0002-6303-3454)
- Web of Science Researcher ID: HNS-9988-2023
- Scopus Author ID: 24724346300
- Homepage: NA

How to cite this paper:

M. A. Latifzadeh, P. Amiri, H. Allahyari, H. Faezi "A new low-stress soft-switching boost converter using coupled-inductor active auxiliary circuit" J. Electr. Comput. Eng. Innovations, 11(2): 433-442, 2023.

DOI: [10.22061/jecei.2023.9304.607](https://doi.org/10.22061/jecei.2023.9304.607)

URL: https://jecei.sru.ac.ir/article_1878.html





Research paper

A New Hybrid NMF-based Infrastructure for Community Detection in Complex Networks

M. Ghadirian, N. Bigdeli*

Department of Control Engineering, Faculty of Technical and Engineering, Imam-Khomeini International University, Qazvin, Iran.

Article Info

Article History:

Received 09 January 2023

Reviewed 28 March 2023

Revised 21 April 2023

Accepted 07 May 2023

Keywords:

Complex networks

Nonnegative matrix factorization

Modularity

General modularity density

Graph clustering

*Corresponding Author's Email Address:
n.bigdeli@eng.ikiu.ac.ir

Abstract

Background and Objectives: Community detection is a critical problem in investigating complex networks. Community detection based on modularity/general modularity density are the popular methods with the advantage of using complex network features and the disadvantage of being NP-hard problem for clustering. Moreover, Non-negative matrix factorization (NMF)-based community detection methods are a family of community detection tools that utilize network topology; but most of them cannot thoroughly exploit network features. In this paper, a hybrid NMF-based community detection infrastructure is developed, including modularity/ general modularity density as more comprehensive indices of networks. The proposed infrastructure enables to solve the challenges of combining the NMF method with modularity/general modularity density criteria and improves the community detection methods for complex networks.

Methods: First, new representations, similar to the model of symmetric NMF, are derived for the model of community detection based on modularity/general modularity density. Next, these indices are innovatively augmented to the proposed hybrid NMF-based model as two novel models called 'general modularity density NMF (GMDNMF) and mixed modularity NMF (MMNMF)'. In order to solve these two NP-hard problems, two iterative optimization algorithms are developed. **Results:** it is proved that the modularity/general modularity density-based community detection can be consistently represented in the form of SNMF-based community detection. The performances of the proposed models are verified on various artificial and real-world networks of different sizes. It is shown that MMNMF and GMDNMF models outperform other community detection methods. Moreover, the GMDNMF model has better performance with higher computational complexity compared to the MMNMF model.

Conclusion: The results show that the proposed MMNMF model improves the performance of community detection based on NMF by employing the modularity index as a network feature for the NMF model, and the proposed GMDNMF model enhances NMF-based community detection by using the general modularity density index.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Networks are used to model complex interconnected

data. Common examples of this type of modeling are biological networks, social networks, and citation networks [1]. One standard method for representing a

network is using a graph data structure consisting of nodes and edges. Exploring and understanding network structures can provide useful information about interconnections. For instance, in social networks, the edges between the nodes depict the interaction between users.

Community detection is one of the powerful analysis methods that help understand the organization of network structures [1]. In social networks, a community (also called a cluster, a module, or a group) comprises users or nodes with close relations or connections but lost connections with others. Over the past years, various measurement criteria have been proposed to evaluate the quality of graph partitioning, including modularity and normalized mutual information (NMI) [2]. Some of these criteria such as modularity and general modularity density [3] have been applied to cluster complex networks as well. By the introduction of modularity, many modularity-based community detection algorithms have been suggested, including integer programming [4], genetic algorithm [5], greedy algorithm [6], and vector partition problem [7]. However, the modularity index has some restrictions in community detection [3], [8]. For example, the modularity depends on the total size of the links, and small communities tend to be merged into large communities. Therefore, new optimization criteria have been offered for community detection algorithms, including general modularity density maximization [3] and localized modularity optimization [8]. Detecting communities based on general modularity density is superior in cases such as resolving most modular networks, detecting communities of different sizes, and not dividing a clique into two parts. In recent years, various methods have been proposed for optimizing modularity density maximization as their index. For instance, one may refer to mixed integer linear programming [9], genetic algorithm [10], memetic algorithm [11], and linear mathematical programming, which consisted of two models based on mixed-integer linear programming and two models based on binary decomposition [12].

In the literature, other algorithms with different approaches have also been presented for clustering complex networks. They include the label propagation algorithm [13], [14], random walk algorithm [15], [16], greedy and weight-balanced algorithm [17], and nonnegative matrix factorization (NMF) algorithm [18]. Among them, NMF-based clustering methods have attracted much attention and wide applications in various fields such as speech separation [19], image processing [20], hyperspectral unmixing [21], document clustering [22], community detection in graph mining and data mining [23], and detection of fake news in social media [24]. On the other hand, one of the recent challenges in

community detection is to use network features or its a priori information to improve the performance of NMF. For instance, a priori information was innovatively integrated into NMF and used for community detection in some studies [25], [26]. The GNMF (graph regularized NMF) model was improved using hypergraph regularization in previous research [27]. The modularized deep NMF (MDNMF) extended DNMF (Deep NMF) [28] by preserving topology information and instinct community structure properties [28], [29]. Community detection was developed by integrating the tri-NMF model with the modularized information called the 'Mtrinmf method' [30]. Community detection was developed by integrating the tri-NMF model with the modularized information called the 'Mtrinmf method' [30]. In the Mtrinmf method, the tri-NMF model is linearly combined with a modularity optimization. However, the linear combination method could not be generalized to the NMF model due to its modularity structure. This is a main drawback in the existing literature in this area, as, to the authors' best knowledge, the NMF model has not been customized using the modularity index or general modularity density index as network features for community detection. Using this customization can improve data clustering based on the NMF method to detect communities on complex networks.

Based on the provided discussions, NMF-based and modularity/general modularity density-based community detection methods are the most important community detection methods which have both advantages and limitations. That is, modularity/general modularity density is specific to the clustering of complex networks, while, it suffers from the NP-hard problem. On the other hand, the NMF-community detection benefits from the initial knowledge, having iterative solution and generality for all types of data, but, it is not specified for complex networks. These properties motivated the authors to consolidate the advantages and refine the properties of the NMF-based and modularity/general modularity density-based community detection methods via introducing a new hybrid NMF-based community detection infrastructure. The new hybrid NMF-based community detection infrastructure includes modularity or general modularity density as more comprehensive indices of the complex networks. In this way, various features such as prior information and community structure are simultaneously employed to cluster the networks. However, the model structures of the NMF-based community detection method and modularity/modularity density-based community detection methods are not consistent. Therefore, to develop the unified infrastructure, first, new representations would be derived for the model of community detection based on modularity/general

modularity density, which are similar to the model of community detection based on symmetric NMF (SNMF). SNMF clustering is one of the symmetric types of NMF methods that has been mentioned in a previous study [31]. These consistent representations would be later employed to extend NMF for community detection in complex networks. Next, two novel hybrid community detection methods are proposed due to the difference between community detection based on modularity/modularity density and the NMF and considering the derived equivalent representations of modularity/modularity density indices. These methods are called mixed modularity nonnegative matrix factorization (MMNMF) and general modularity density nonnegative matrix factorization (GMDNMF). MMNMF and GMDNMF would innovatively integrate modularity/general modularity density indices into the NMF model, respectively, to improve community detection performance by utilizing proper network features. However, these methods are NP-hard problems, which should be solved numerically. Therefore, iterative update rules would be derived and proved as optimal solutions for MMNMF and GMDNMF models. The performance of the two models is verified on two practical artificial networks and ten real-world networks. It is indicated that although MMNMF and GMDNMF have better performance respecting other community detection methods, these methods slightly differ in precision and computational complexity, which would be calculated and compared in this study. The final preference should be therefore performed based on problem requirements. The remaining sections of this paper are organized as follows:

Section 2 studies modularity and general modularity density optimizations and reviews related works on NMF-based community detection methods. Section 3 presents our proposed methods and analyzes iterative optimization algorithms. The computational complexity of our methods is calculated in Section 4, followed by studying the effect of parameters and presenting several comparative experimental results. Finally, Section 5 summarizes the proposed procedures, achievements, and discusses future works.

Related Works

Community detection based on the NMF method is one of the efficient methods for clustering various types of data such as audio, text, image, graph, and the like. This paper proposes a combination of NMF-based community detection and community detection based on network features. Therefore, first, a study of community detection based on modularity and general modularity density is presented, and then NMF-based community detection methods are reviewed and discussed in this section.

Modularity Maximization

A graph can represent a complex network without loss of network features. $G = (V, E)$ is a representation of directed (undirected) and unweighted graphs where V denotes the set of n nodes and E is a set of m edges between the two nodes. Modularity maximization is a popular community detection method for understanding the structure of networks. It considers the strength of the relationship density of each edge within each community and can regard the related nodes between communities. An algorithm based on modularity maximization clustering a network in two-by-two communities was first proposed by Newman [1]. Next, to cluster more than two-by-two communities, the generalized modularity index (Q) was presented as follows [1], [30]:

$$Q = \frac{1}{2m} \text{tr}(X^T B X) \quad (1a)$$

$$B = A - B_1 \quad (1b)$$

$$(B_1)_{ij} = \frac{k_i k_j}{2m} \quad (1c)$$

where A and B are adjacency and modularity matrices, respectively. In addition, $k_i, X \in R^{n \times k}$ and m are the degree of the i^{th} nodes, the community membership matrix and the number of edges, respectively. Besides, k denotes the number of communities in the complex network. Moreover, the maximization problem can be rewritten as [30]:

$$\max_X Q = \max_X \frac{1}{2m} \text{tr}(X^T B X) \quad (2)$$

where $X^T X = I$ satisfies the orthogonal condition [32].

Since this method is a NP-hard problem, finding a way to solve it has been one of the challenges over the past years.

General Modularity Density Maximization

Community detection based on general modularity density maximization is efficient for clustering complex networks [3]. The general modularity density index considers each cluster's average inner degree and outer degree. The inner degree refers to the sum of the edges of interval nodes in each cluster, and the outer degree is the sum of edges between the nodes inside the cluster with the nodes of another cluster. It can be rewritten for k number of partitions ($\{V_r\}_{r=1}^k$) as follows [3]:

$$D_\lambda(\{V_r\}_{r=1}^k) = \sum_{r=1}^k \frac{2\lambda l(V_r, V_r) - 2(1-\lambda)l(V_r, \bar{V}_r)}{|V_r|} \quad (3)$$

where $l(V_1, V_2) = \sum_{i \in V_1, j \in V_2} A_{ij}$, $l(V_1, \bar{V}_1) = \sum_{i \in V_1, j \in \bar{V}_1} A_{ij}$, $\bar{V}_1 = V \setminus V_1$ and V_r is the set of vertices in the r th community. Furthermore, D_λ evaluates small and large clusters by using ratio association and ratio cut for $\lambda < 0.5$ and $\lambda > 0.5$, respectively. Therefore, D_λ equals modularity density when $\lambda = 0.5$. Advantages such as

selecting the best communities with different sizes, not dividing cliques, and resolving graph types are obtained by selecting different λ values.

Lemma 2.1 [33]. D_λ can be written as a trace of the similarity matrix of a complex network as follows:

$$D_\lambda(\{V_r\}_{r=1}^k) = \text{tr}(\tilde{X}^T(2A - 2(1 - \lambda)C)\tilde{X}), \tilde{X} = XD \quad (4)$$

where D is a diagonal matrix with $D_{ii} = 1/\sqrt{\sum_{j=1}^n X_{ji}^2}$ values. Additionally, X and C represent a community relation matrix and a diagonal matrix with $C_{ii} = \sum_{j=1}^n A_{ii}$ values, respectively.

Proof: Given that X_{ir} indicates the existence probability of node i which belongs to the community and $X_r = (X_{1r}, \dots, X_{nr})$ represents the probability of each node which belongs to V_r , D_λ in (3) can be rewritten according to $l(V_r, \tilde{V}_r) = l(V_r, V) - l(V_r, V_r)$ as:

$$D_\lambda(\{V_r\}_{r=1}^k) = \sum_{r=1}^k \frac{2l(V_r, V_r) - 2(1 - \lambda)l(V_r, V)}{|V_r|} \quad (5)$$

Moreover, $|V_r|$, $l(V_r, V_r)$ and $l(V_r, V)$ can be written based on X_r as $|V_r| = X_r^T X_r$, $l(V_r, V_r) = X_r^T A X_r$, $l(V_r, V) = X_r^T C X_r$. Thus (5) can be reformulated as:

$$D_\lambda(\{V_r\}_{r=1}^k) = \sum_{r=1}^k \frac{2X_r^T A X_r - 2(1 - \lambda)X_r^T C X_r}{X_r^T X_r} = \sum_{r=1}^k \tilde{X}_r(2A - 2(1 - \lambda)C)\tilde{X}_r \quad (6)$$

where $\tilde{X}_r = X_r/\sqrt{X_r^T X_r}$ or $\tilde{X} = XD$ if $D_{ii} = 1/\sqrt{\sum_{j=1}^n X_{ji}^2}$. Therefore, (3) can be represented as (4), and the proof is completed accordingly.

□ **Corollary:** Considering that D_λ can be formulated as a trace form of (4), community detection based on general modularity density is a maximization problem with D_λ as its cost function, which can be rewritten as follows [12], [33]:

$$\max_{\tilde{X}} D_\lambda(\{V_r\}_{r=1}^k) = \max_{\tilde{X}} \text{tr}(\tilde{X}^T B_2 \tilde{X}) \quad (7)$$

$$\text{s.t. } \tilde{X} > 0, \tilde{X}^T \tilde{X} = I_k$$

where, $B_2 = 2A - 2(1 - \lambda)C$, C , $\tilde{X}^T \tilde{X} = I$ are modularity density matrix, a diagonal matrix with $C_{ii} = \sum_{j=1}^n A_{ii}$ and a relaxing condition for orthogonality, respectively.

Nonnegative Matrix Factorization (NMF)

NMF models factorize a given similarity matrix $Y \in R^{n \times n}$ into two new matrices $W \in R^{n \times k}$ and $H \in R^{n \times k}$: $Y \simeq WH^T$ where W and H are called the community indicator feature matrix and called community relation matrix, respectively. The error between Y and WH^T is measured by a cost function of $J_{NMF}(W, H)$. W and H can be found by minimizing $J_{NMF}(W, H)$ as follows:

$$\min_{W, H} J_{nmf}(W, H) = \|Y - WH^T\|_F^2 \quad (8)$$

where $\|\cdot\|_F$ stands for the Frobenius norm. If Y is assumed a symmetric matrix (such as undirected graph),

then all the characteristics of the clustering index can be aggregated in one matrix ($W = H$). Therefore, as an extension to NMF, SNMF can drastically improve community detection. The objective function of SNMF model would be rewritten as follows [31]:

$$\min_H J_{SNMF}(H) = \|Y - HH^T\|_F^2 \quad (9)$$

Using network features or prior information in the NMF-based methods has been a challenging topic in recent years. Lu et al. [26] have recently used prior information to improve community detection and proposed two semi-supervised NMF-based methods named SVDCNMF and SVDCSNMF. In this method, the adjacency matrix A is considered as the similarity matrix (i.e., $Y = A$, $H = X$), and its objective function is represented as:

$$\min_X J_{SVDCSNMF}(X) = \|A - XX^T\|_F^2 + 2\lambda \text{tr}(X^T L X) \quad (10)$$

where L is the graph Laplacian of prior information. Since the Laplacian matrix is specific to the graph structure, this method will not provide the best clustering for other types of network features (such as modularity).

He et al. [25] suggested a robust semi-supervised NMF method named RSSNMF to enhance the robustness of semi-supervised NMF for uncertainties and errors in prior information. The cost function of the RSSNMF method is as follows:

$$\min_X J_{RSSNMF}(X) = \|A - XX^T\|_2 + \alpha \text{tr}(X^T P X Q) + \beta \text{tr}(X^T R X) \quad (11)$$

where α and β are semi-supervised tuning parameters

$$\text{and } Q = \begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & \dots & \vdots \\ \vdots & \vdots & \ddots & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}. \text{ Prior information is applied in}$$

the following matrix:

$$P_{ij} = \begin{cases} 1 & \text{if } x_i, x_j \text{ have same labels or } i = j \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

$$R_{ij} = \begin{cases} 1 & \text{if } x_i, x_j \text{ have different labels} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

In addition, the terms $\text{tr}(X^T P X Q)$ and $\text{tr}(X^T R X)$ are derived from must- and cannot-link pairwise constraints among nodes, respectively. This method only uses the prior information and did not consider other features of the network structure.

Similarly, Wu et al. [27] proposed a mixed hypergraph NMF named MHGNMF by combining NMF with hypergraph regularization, which encodes the higher-order information into NMF by hypergraph. The objective function of MHGNMF is defined as:

$$\min_X J_{MHGNMF}(X) = \|A - XX^T\|_F^2 + \beta \text{tr}(X^T L_h X) \quad (14)$$

where L_h is hyperlaplacian matrix. Likewise, Yan et al. [30] combined modularity optimization and tri-NMF-based

community detection named Mtrnmf, which uses network features to enhance the performance of tri-NMF. The cost function of the Mtrnmf method can be written as (15):

$$\min_{U, X} J_{\text{Mtrnmf}}(U, X) = \|A - XUX^T\|_F^2 - \beta \text{tr}(X^T BX) \quad (15)$$

where, $\text{tr}(X^T BX)$, U and $\|A - XUX^T\|_F^2$ are the modularity optimization, the community indicator feature matrix and triNMF optimization terms, respectively. The Mtrnmf model innovatively combines the modularity criterion linearly with the triNMF model and has produced the best clustering ever.

Additionally, Huang et al. [29] suggested a new model named modularized deep nonnegative matrix factorization (MDNMF), which combined modularity and DNMF-based community detection. Deep NMF (DNMF) is another extension of NMF model that acquires additional levels of abstraction of the similarity between the nodes of each levels [29] and factorizes a given adjacency matrix A into $p + 1$ nonnegative factors.

The MDNMF model has been composed as follows:

$$\min_{U_i, X, M, C} J_{\text{MDNMF}} = \|A - U_1 U_2 \dots U_p X^T\|_F^2 + \alpha \|M - X^T C^T\|_F^2 - \beta \text{tr}(M^T B M) + \lambda \text{tr}(X L X^T) \quad (16)$$

$$\text{st. } U_i \geq 0, X \geq 0, \forall i = 1, 2, \dots, p$$

where, L , M , C and λ denote the graph Laplacian matrix, the modularity cluster term, the final cluster term and the regularization parameter, respectively. $\lambda \text{tr}(X L X^T)$ utilizes a regularized graph and $\|A - U_1 U_2 \dots U_p X^T\|_F^2$ refers to DNMF-based community detection model. This is one of the methods that has combined the graph features such as modularity criterion with DNMF-based community detection and illustrated a suitable clustering, but due to the DNMF-based community detection model, this method will have a high computational complexity and a high dependence on the correct selection of parameters in DNMF model.

From the above-mentioned discussion, it can be concluded that the modularity maximization in (1) suffers from the NP-hard problem [1], [3], [8]. Moreover, The NMF models have attracted much attention and have wide applications in various fields [9]-[12]. On the other hand, general modularity density maximization in (2) has not been used to improve NMF-based community detection, yet. Therefore, combining the NMF models and the graph features such as modularity/ general modularity density criterion will be considered, in this paper. The main achievements of the proposed methods are development of an iterative solution for modularity optimization with NMF model, specialization of NMF model for complex networks, and utilizing general modularity density criterion for NMF.

The Proposed Methods

According to the provided discussions about various NMF-based community detection methods, one could conclude that modularity and general modularity density indices as the network features have not been employed with NMF in a unified community detection method. To extend the NMF-based method to a hybrid method containing these indices, first, we derive the new representation of modularity-based and general modularity density-based community detection methods that are similar to the model of community detection based on SNMF. Next, these consistent representations help combine NMF-based community detection with modularity/general modularity density indices. It leads us to propose general modularity density nonnegative matrix factorization (GMDNMF) and mixed modularity nonnegative matrix factorization (MMNMF) models. Finally, proper iterative optimization methods for solving MMNMF and GMDNMF problems are developed accordingly.

New Representations of Modularity/General Modularity Density Maximization

In this section, new models of the modularity/general modularity density optimization are derived, which are similar to the model of symmetric nonnegative matrix factorization optimization problem summarized as:

Theorem 3.1. The modularity optimization in (2) and general modularity density optimization in (7) for complex networks can be represented in a similar form of the SNMF model, respectively, as follows:

$$\max_X Q = \min_X \|B - X X^T\|_F^2 \quad (17a)$$

and

$$\max_{\tilde{X}} D_\lambda(\{V_r\}_{r=1}^k) = \min_{W, \tilde{X}} \|B_2 - \tilde{X} \tilde{X}^T\|_F^2 \quad (17b)$$

As shown, the new formulation in (17) is similar to the model of SNMF in (9), while in (17a) and (17b), Y is replaced with B and B_2 , respectively; in addition, H is replaced with X and \tilde{X} , respectively.

Proof: Modularity optimization of (2) is re-formulated as follows:

$$\max_X \frac{1}{2m} \text{tr}(X^T B X) \propto - \frac{1}{2m} \min_X \text{tr}(X^T B X) \propto - \min_X \text{tr}(X^T B X) \quad (18a)$$

If $X^T X = I$ and B is constant, (18a) is re-written as:

$$\max_X \frac{1}{2m} \text{tr}(X^T B X) \propto \min_X (\text{tr}(X^T X X^T X) - 2 \text{tr}(X^T B X) + \text{tr}(B B^T)) \quad (18b)$$

According to trace properties, namely, $\text{tr}(X^T B X) = \text{tr}(B X X^T)$, $\text{tr}(X^T X X^T X) = \text{tr}(X X^T X X^T)$, (18b) is re-written as:

$$\max_X \frac{1}{2m} \text{tr}(X^T B X) \propto \min_X \text{tr}(X X^T X X^T - 2 B X X^T + B B^T) \propto \min_X \|B - X X^T\|_F^2 \quad (19)$$

As a result, the new representation of modularity optimization is consistent with SNMF. The consistent representation of general modularity density optimization and SNMF can be similarly derived, completing the proof.

GMDNMF and MMNMF Models

The NMF model is an efficient method for clustering data types. However, it is not the best method for clustering complex networks because it may ignore some useful information and characteristics such as general modularity density and modularity indices. Motivated by this observation, in this section, general modularity density and modularity indices are augmented to NMF-based community detection to improve its performance. For this purpose, we refer to Theorem 3.1, when, it was shown that community detection based on modularity optimization (2) can be represented as an SNMF optimization problem with similarity matrix B , and community detection based on general modularity density optimization (7) is similar to community detection based on SNMF with similarity matrix B_2 . Therefore, according to Theorem 3.1, for improving community detection based on NMF via modularity or general modularity density indices, it is necessary to combine SNMF-based community detection and NMF-based community detection methods. However, due to different structures of community detection based on NMF (NMF optimization with similarity matrix A) and community detection based on modularity/general modularity density optimization (SNMF optimization with similarity matrix B/B_2), NMF-based community detection cannot be linearly combined with modularity/general modularity density-based community detection as in previous research. Thus, multi-view clustering via joint NMF such as the methods presented in other studies [23] and [34] would be exploited in this paper. In this context, MMNMF and GMDNMF models are proposed to integrate modularity and general modularity density into the NMF model to improve community detection. These models are devised for complex networks as:

$$\min_{W, X, \tilde{X}, X^*} J_{\text{MMNMF}} = \|A - W X^T\|_F^2 - \text{tr}(\tilde{X}^T B \tilde{X}) + \frac{1}{2} \|\tilde{X} - X^*\|_F^2 + \frac{1}{2} \|X - X^*\|_F^2 \quad (20)$$

$$s. t. W, X, \tilde{X}, X^* > 0, \sum_{r=1}^k X_{ir} = 1, \tilde{X}^T \tilde{X} = I_k$$

and

$$\min_{W, X, \tilde{X}, X^*} J_{\text{GMDNMF}} = \|A - W X^T\|_F^2 - \text{tr}(\tilde{X}^T B_2 \tilde{X}) + \frac{1}{2} \|\tilde{X} \tilde{D} - X^*\|_F^2 + \frac{1}{2} \|X - X^*\|_F^2 \quad (21)$$

$$s. t. W, X, \tilde{X}, X^* > 0, \sum_{r=1}^k X_{ir} = 1, \tilde{X}^T \tilde{X} = I_k$$

where X^* is the result of community detection models and \tilde{D} denotes a diagonal matrix with $\tilde{D}_{ii} = \sqrt{\sum_{j=1}^n X_{ji}^2}$ or $\tilde{D} = D^{-1}$ values. In (20) and (21), $\|A - W X^T\|_F^2$ represents NMF-based community detection, and $\text{tr}(\tilde{X}^T B \tilde{X})$ and $\text{tr}(\tilde{X}^T B_2 \tilde{X})$ refer to community detection based on modularity/general modularity density indices, respectively. Here, X^* is an interface parameter for combining NMF-based community detection with modularity-based and general modularity density-based community detection methods. GMDNMF and MMNMF models are NP-hard problems due to the orthogonal constraint. Many methods exist for extending the orthogonal constraint to a nonnegative term [2], [35]. For this purpose, we use the presented method in previous studies [35]. This method adds an orthogonal constraint to the objective model and can preserve clustering performance. Accordingly, the new objective functions for GMDNMF and MMNMF models are formulated as:

$$\min_{W, X, \tilde{X}, X^*} J_{\text{MMNMF}} = \|A - W X^T\|_F^2 - \text{tr}(\tilde{X}^T B \tilde{X}) + \frac{1}{2} \|\tilde{X} - X^*\|_F^2 + \frac{1}{2} \|X - X^*\|_F^2 + \eta \|\tilde{X}^T \tilde{X} - I\| \quad (22)$$

$$s. t. W, X, \tilde{X}, X^* > 0, \sum_{r=1}^k X_{ir} = 1$$

$$\min_{W, X, \tilde{X}, X^*} J_{\text{GMDNMF}} = \|A - W X^T\|_F^2 - \text{tr}(\tilde{X}^T B_2 \tilde{X}) + \frac{1}{2} \|\tilde{X} \tilde{D} - X^*\|_F^2 + \frac{1}{2} \|X - X^*\|_F^2 + \eta \|\tilde{X}^T \tilde{X} - I\| \quad (23)$$

$$s. t. W, X, \tilde{X}, X^* > 0, \sum_{r=1}^k X_{ir} = 1$$

where η and $\|\tilde{X}^T \tilde{X} - I\|$ are the orthogonal condition control parameter and the orthogonal condition control cost, respectively. It is worth mentioning that the value of parameter λ in B_2 , which was first introduced in (16), is chosen via a simple rule presented in [30]. That is, this parameter is selected in the interval of $[0, 1]$ in small steps (e.g., 0.1). Then, the best community and its relating value of λ is selected by the best-obtained modularity index value.

Iterative Optimization Algorithms for MMNMF and GMDNMF Models

This section will develop an iterative method to solve the proposed MMNMF model of (22) and GMDNMF model of (23). This iterative method is performed via an alternative updating strategy (i.e., the model

updates one variable via the Lagrange method while the other variables are constant). Finally, the update variable process is repeated until the convergence or reaching the final iteration number.

Iterative Optimization Algorithm for the GMDNMF Model

The trace form (23) can be rewritten as:

$$\min_{W, X, \tilde{X}, X^*} J_{\text{GMDNMF}} = (\text{tr}(A A^T) - 2 \text{tr}(A X W^T) + \text{Tr}(W X^T X W^T)) - \text{tr}(\tilde{X}^T B \tilde{X}) + \frac{1}{2} (\text{tr}(\tilde{X} \tilde{D} \tilde{D}^T \tilde{X}^T) -$$

$$2tr(\tilde{X}\tilde{D}X^{*T}) + tr(X^*X^{*T}) + \frac{1}{2}(tr(XX^T) - 2tr(XX^{*T}) + tr(X^*X^{*T})) + \eta(tr(\tilde{X}^T\tilde{X}\tilde{X}^T\tilde{X}) - 2tr(\tilde{X}^T\tilde{X}) + k) \quad (24)$$

Given that \tilde{D}_{ii} equals $\sqrt{\sum_{j=1}^n X_{ji}^2}$, calculating iterative updating rules for X is more complex compared to the other variables. Accordingly, first, updating rules are formulated for W , \tilde{X} , and X^* , followed by deriving the iterative rules for updating X .

Updating rules for W , \tilde{X} , and X^* :

the Lagrange cost function for (24) can be resulted as:

$$L_{\text{GMDNMF}} = J_{\text{GMDNMF}} + tr(\psi W) + tr(\phi \tilde{X}) + tr(\varphi X^*) \quad (25)$$

where ψ , ϕ , and φ are the Lagrangian multipliers for constraints $W > 0$, $\tilde{X} > 0$, and $X^* > 0$, respectively. Then, the derivatives of L_{GMDNMF} would be then derived as follows:

$$\begin{aligned} \frac{\partial L_{\text{GMDNMF}}}{\partial W} &= \psi + 2WX^T X - 2AX \\ \frac{\partial L_{\text{GMDNMF}}}{\partial \tilde{X}} &= \phi - 2B\tilde{X}^T + \tilde{X}\tilde{D}\tilde{D}^T - X^*\tilde{D}^T + 4\eta\tilde{X}\tilde{X}^T\tilde{X} - 4\eta\tilde{X} \\ \frac{\partial L_{\text{GMDNMF}}}{\partial X^*} &= \varphi - \tilde{X}\tilde{D} + X^* - X + X^* \end{aligned} \quad (26)$$

where $B = 2A - 2(1 - \lambda)C$. According to Karush-Kuhn-Tucker (KKT) conditions (i.e., $\psi_{ir}W_{ir} = 0$, $\phi_{ir}\tilde{X}_{ir} = 0$, and $\varphi_{ir}X_{ir}^* = 0$), the solution can be formulated as follows:

$$\begin{aligned} (WX^T X)_{ir}W_{ir} - (AX)_{ir}W_{ir} &= 0 \\ -4(A\tilde{X}^T)_{ir}\tilde{X}_{ir} + 4(1 - \lambda)(C\tilde{X}^T)_{ir}\tilde{X}_{ir} + (\tilde{X}\tilde{D}\tilde{D}^T)_{ir}\tilde{X}_{ir} - (X^*\tilde{D}^T)_{ir}\tilde{X}_{ir} + 4\eta(\tilde{X}\tilde{X}^T\tilde{X})_{ir}\tilde{X}_{ir} - 4\eta(\tilde{X})_{ir}\tilde{X}_{ir} &= 0 \\ -(\tilde{X}\tilde{D})_{ir}X_{ir}^* - (X)_{ir}X_{ir}^* + 2(X^*)_{ir}X_{ir}^* &= 0 \end{aligned} \quad (27)$$

Finally, iterative updating rules are formulated as:

$$\begin{aligned} W_{ir} &= W_{ir} \cdot \frac{(AX)_{ir}}{(WX^T X)_{ir}} \\ \tilde{X}_{ir} &= \tilde{X}_{ir} \cdot \frac{4(A\tilde{X}^T)_{ir} + (X^*\tilde{D}^T)_{ir} + 4\eta(\tilde{X})_{ir}}{4(1 - \lambda)(C\tilde{X}^T)_{ir} + (\tilde{X}\tilde{D}\tilde{D}^T)_{ir} + 4\eta(\tilde{X}\tilde{X}^T\tilde{X})_{ir}} \\ X_{ir}^* &= X_{ir}^* \cdot \frac{(\tilde{X}\tilde{D})_{ir} + (X)_{ir}}{2(X^*)_{ir}} \end{aligned} \quad (28)$$

Due to the probability of having zero values in the denominator of X^* according to updating rules in (28), a modified updating function is:

$$X_{ir}^* = X_{ir}^* \cdot \frac{(\tilde{X}\tilde{D})_{ir} + (X)_{ir} + 10^{-9}}{2(X^*)_{ir} + 10^{-9}} \quad (29)$$

Furthermore, to satisfy $\sum_{r=1}^k X_{ir}^* = 1$ condition, we use the same method in previous research [30] as:

$$X_{ir}^* := \frac{X_{ir}^*}{\sum_{r=1}^k X_{ir}^*} \quad (30)$$

Updating rule for X :

The Lagrange cost function where Ω is the Lagrangian multiplier for constraint $X > 0$ can be rewritten as:

$$L_{\text{GMDNMF}} = (tr(WX^T XW^T) - 2tr(AXW^T)) + \frac{1}{2}(tr(XX^T) - 2tr(XX^{*T})) + \frac{1}{2}R + tr(\Omega X) \quad (31)$$

where $R = tr(\tilde{X}\tilde{D}\tilde{D}^T\tilde{X}) - 2tr(\tilde{X}\tilde{D}X^{*T})$ and \tilde{D} is the diagonal matrix with $\tilde{D}_{ii} = \sqrt{\sum_{j=1}^n X_{ji}^2}$ values. One can rewrite R as follows [36]:

$$R = \sum_{j=1}^n \sum_{r=1}^k \tilde{X}_{jr} \sum_{i=1}^n X_{ir}^2 \tilde{X}_{jr} - 2 \sum_{j=1}^n \sum_{r=1}^k \tilde{X}_{jr} \sqrt{\sum_{i=1}^n X_{ir}^2} X_{jr}^* \quad (32)$$

The partial derivative of R with respect to X_{ir} is as follows:

$$P_{ir} = \frac{\partial R}{\partial X_{ir}} = 2 \left(X_{ir} \sum_{j=1}^n \tilde{X}_{jr}^2 - \frac{X_{ir}}{\sqrt{\sum_{i=1}^n X_{ir}^2}} \sum_{j=1}^n \tilde{X}_{jr} X_{jr}^* \right) \quad (33)$$

Therefore, the derivatives of L_{GMDNMF} can be derived as (34):

$$\frac{\partial L_2}{\partial X} = \Omega + g_1(2XW^T W - 2A^T W) + \frac{1}{2}(2X - 2X^*) + \frac{1}{2}P \quad (34)$$

By using KKT conditions (i.e., $\Omega_{ir}X_{ir} = 0$), one has:

$$(2XW^T W - 2A^T W)_{ir}X_{ir} + (X - X^*)_{ir}X_{ir} + \frac{1}{2}P_{ir}X_{ir} = 0 \quad (35)$$

Considering (31) and (33), iterative updating rules would be as follows:

$$X_{ir} = X_{ir} \cdot \frac{X_{ir}^* + 2(A^T W)_{ir} + \frac{X_{ir}}{\sqrt{\sum_{i=1}^n X_{ir}^2}} \sum_{j=1}^n \tilde{X}_{jr} X_{jr}^*}{X_{ir} + 2(XW^T W)_{ir} + X_{ir} \sum_{j=1}^n \tilde{X}_{jr}^2} \quad (36)$$

Eventually, according to (28), (29), (30), and (36), the GMDNMF model is proposed as Algorithm 1.

Algorithm1: GMDNMF model

Inputs:

- Adjacency matrix A
- Number of communities k
- General density parameters λ
- Orthogonal condition control parameter η
- Maximum number of iterations I_t

Output:

Clustering label of each node

- 1: **Initialize** W, X, \tilde{X} and X^*
 - 2: **For** $t = 1: I_t$ **do**
 - 3: $W_{ir} = W_{ir} \cdot \frac{(AX)_{ir}}{(WX^T X)_{ir}}$
 - 4: $\tilde{X}_{ir} = \tilde{X}_{ir} \cdot \frac{4(A\tilde{X}^T)_{ir} + (X^*\tilde{D}^T)_{ir} + 4\eta(\tilde{X})_{ir}}{4(1 - \lambda)(C\tilde{X}^T)_{ir} + (\tilde{X}\tilde{D}\tilde{D}^T)_{ir} + 4\eta(\tilde{X}\tilde{X}^T\tilde{X})_{ir}}$
 - 5: $X_{ir} = X_{ir} \cdot \frac{X_{ir}^* + 2(A^T W)_{ir} + \frac{X_{ir}}{\sqrt{\sum_{i=1}^n X_{ir}^2}} \sum_{j=1}^n \tilde{X}_{jr} X_{jr}^*}{X_{ir} + 2(XW^T W)_{ir} + X_{ir} \sum_{j=1}^n \tilde{X}_{jr}^2}$
 - 6: $X_{ir}^* = X_{ir}^* \cdot \frac{(\tilde{X}\tilde{D})_{ir} + (X)_{ir} + 10^{-9}}{2(X^*)_{ir} + 10^{-9}}$
 - 7: $X_{ir}^* := \frac{X_{ir}^*}{\sum_{r=1}^k X_{ir}^*}$
 - 8: **End for**
 - 9: **Return** $(v_t, I_t) = \text{argmax}_{r \leq k} X_{ir}^*$
-

Iterative Optimization Algorithm for the MMNMF Model

To the iterative optimization algorithm for the MMNMF model, first, one can define the trace form of (22) as follows:

$$\min_{W, X, \tilde{X}, X^*} J_{\text{MMNMF}} = (tr(AA^T) - 2tr(AXW^T) + tr(WX^T XW^T)) - tr(\tilde{X}^T B \tilde{X}) + \frac{1}{2}(tr(\tilde{X}\tilde{X}^T) - 2tr(\tilde{X}X^{*T}) + tr(X^*X^{*T})) + \frac{1}{2}(tr(XX^T) - 2tr(XX^{*T}) + tr(X^*X^{*T})) + \eta(tr(\tilde{X}^T \tilde{X} \tilde{X}^T \tilde{X}) - 2tr(\tilde{X}^T \tilde{X}) + k) \quad (37)$$

$$s. t. W, X, \tilde{X}, X^* > 0, \sum_{r=1}^k X^*_{ir} = 1$$

Following the same procedure of Section 3.3.1 (Appendix A), the iterative updating rules are derived (38):

$$\begin{aligned} W_{ir} &= W_{ir} \cdot \frac{(AX)_{ir}}{(WX^T X)_{ir}} \\ X_{ir} &= X_{ir} \cdot \frac{2(A^T W)_{ir} + (X^*)_{ir}}{2(XW^T W)_{ir} + (X)_{ir}} \\ \tilde{X}_{ir} &= \tilde{X}_{ir} \cdot \frac{2(A\tilde{X}^T)_{ir} + (X^*)_{ir} + 4\eta(\tilde{X})_{ir}}{2(B_1 \tilde{X}^T)_{ir} + (\tilde{X})_{ir} + 4\eta(\tilde{X}\tilde{X}^T \tilde{X})_{ir}} \\ X^*_{ir} &= X^*_{ir} \cdot \frac{(\tilde{X})_{ir} + (X)_{ir} + 10^{-9}}{2(X^*)_{ir} + 10^{-9}} \\ X^*_{ir} &:= \frac{X^*_{ir}}{\sum_{r=1}^k X^*_{ir}} \end{aligned} \quad (38)$$

Finally, the iterative optimization algorithm for the MMNMF model is suggested as Algorithm 2.

Algorithm2: MMNMF model

Inputs:

- Adjacency matrix A
- Number of communities k
- Orthogonal condition control parameter η
- Maximum number of iterations I_t

Output:

Clustering label of each node

- 1: **Initialized** W, X, \tilde{X} and X^*
 - 2: **For** $t = 1: I_t$ **do**
 - 3: $W_{ir} := W_{ir} \cdot \frac{(AX)_{ir}}{(WX^T X)_{ir}}$
 - 4: $\tilde{X}_{ir} := \tilde{X}_{ir} \cdot \frac{2(A\tilde{X}^T)_{ir} + (X^*)_{ir} + 4\eta(\tilde{X})_{ir}}{2(B_1 \tilde{X}^T)_{ir} + (\tilde{X})_{ir} + 4\eta(\tilde{X}\tilde{X}^T \tilde{X})_{ir}}$
 - 5: $X_{ir} := X_{ir} \cdot \frac{2(A^T W)_{ir} + (X^*)_{ir}}{2(XW^T W)_{ir} + (X)_{ir}}$
 - 6: $X^*_{ir} := X^*_{ir} \cdot \frac{(\tilde{X})_{ir} + (X)_{ir} + 10^{-9}}{2(X^*)_{ir} + 10^{-9}}$
 - 7: $X^*_{ir} := \frac{X^*_{ir}}{\sum_{r=1}^k X^*_{ir}}$
 - 8: **End for**
 - 9: **Return** $(v_i, I_i) = \text{argmax}_{r \leq k} X^*_{ir}$
-

Experiments and Analysis

In this section, the computational complexities of the proposed models are computed and compared, followed by discussing assessment standards for performance

evaluation. Finally, other community detection methods are introduced to compare some popular network sets, and the results and capabilities of the proposed methods will be demonstrated accordingly.

Assessment Standards

In this paper, NMI and modularity index (Q) are used to evaluate the performance of different community detection methods. NMI information is widely applied to compare the similarity between partition labels and the ground truth partition labels. NMI information was adopted as:

$$NMI(C, C') = \frac{-2 \sum_{i=1}^{|C|} \sum_{j=1}^{|C'|} n_{C_i \cap C'_j} \log\left(\frac{n_{C_i \cap C'_j}}{n_{C_i} n_{C'_j}}\right)}{\sum_{i=1}^{|C|} n_{C_i} \log\left(\frac{n_{C_i}}{n}\right) + \sum_{j=1}^{|C'|} n_{C'_j} \log\left(\frac{n_{C'_j}}{n}\right)} \quad (39)$$

where n_{C_i} and $|C|$ indicate the number of members in partitions C_i and number of partitions in C , respectively. If NMI tends to one, the partition labels will be closer to ground truth partition labels, and if it tends to zero, the partition labels will be dissimilar to these labels.

Performance Analysis

The employed real-world and artificial networks are described in this subsection. The comparative results of our GMDNMF and MMNMF methods with other methods such as Mtrnmf, NMF, LPA, CNM, Infomap, MHGNMF, and LRSCD are illustrated on the networks. For more explanation, the Mtrnmf is one of the efficient methods that has been able to exploit network features such as modularity index to improve the performance of community detection based on the tri-NMF method [35]. The LPAM method modified the LPA method by adding the modularity index [13]. The CNM method is a fast-greedy optimization for directly solving the modularity index [37]. Infomap is a popular community detection method based on flow running dynamic by random walk [18]. MHGNMF is a new mixed hypergraph regularized NMF method which makes use of structure similarity information and topological connection information [27]. The MHGNMF method is divided into MHGNMF_kl and MHGNMF_sq algorithms based on the type of the community detection function. The MHGNMF_sq algorithm is selected due to the use of the Frobenius norm in the optimization function. According to [38], LRSCD is a community detection method based on a low-rank decomposition strategy for decomposing each node vector in a new space (the geometric space). The NMF, CNM, and Infomap are common community detection methods, and MHGNMF, LRSCD, and Mtrnmf methods are the recent community detection algorithms that have been considered for comparison.

Performance Analysis on Ten Real-World Networks

Ten real-world networks have been chosen to evaluate different community detection methods. The information

of these real-world networks has been tabulated [Table 1](#). Here, \bar{c} is the number of ground-truth communities. The Karate, Jazz, Political books, Dolphins, Football, and Polbooks are small real-world networks, while the Polblogs, Cora, Citeseer, and Pubmed are large real-world networks [\[27\]](#). [Table 2](#) lists the best results of eight methods on ten real-word networks based on the modularity index (Q). The following results are concluded based on data in [Table 2](#):

- MMNMF, GMDNMF, and GMDNMF ($\lambda = \frac{1}{2}$) methods have better clustering capability compared to the NMF method based on the modularity index.
- In the Pubmed network, fast methods such as CNM and LPAM have better clustering in comparison with

other NMF-based methods. Due to the computational errors in large-scale networks, NMF-based community detection usually has clustering errors. However, the GMDNMF offers better clustering when compared to other methods.

- Due to different values of λ , the GMDNMF method and some methods can offer the best community detection compared to other ones in other networks. For example, GMDNMF in the Cora network, and GMDNMF and MHGNMF_sq in the Polbooks network are the best methods for community detection.
- In addition to the GMDNMF method, MMNMF and MHGNMF_sq methods outperform other methods.

Table 1: Real-world network information

Networks	N	m	\bar{c}	Description
Karate	34	78	2	Zachary karate club network (Karate) [39]
Jazz	198	2742	4	Jazz network (Jazz) [40]
Political books	105	441	3	Political books network (Political books) [41]
Dolphins	62	159	4	Lusseau's bottlenose dolphins social network (Dolphins) [42]
Football	115	613	12	American college football network [43]
Polblogs	1490	16718	2	Blogs about politics [44]
Cora	2708	5429	7	A Cora citation network [45]
Citeseer	3312	4732	6	A Citeseer citation network [46]
Pubmed	19717	44338	3	A Pubmed citation network [47]

Table 2: Modularity index (Q) for different methods and real-world networks

	GMDNMF	MMNMF	GMDNMF (λ = 0.5)	MHGNMF_sq [27]	LRSCD [38]	NMF [12]	Mtrnmf [30]	Infomap [19]	LPAM [11]	CNM [37]
Karate	0.419 ($0.5 < \lambda < 0.74$)	0.419	0.419	0.419	0.419	0.142	0.419	0.403	0.397	0.383
Jazz	0.444 ($\lambda = 0.57$)	0.444	0.439	0.444	0.442	0.436	0.442	0.442	0.444	0.444
Political books	0.526 ($\lambda = 0.39$)	0.520	0.520	0.526	0.520	0.513	0.526	0.526	0.520	0.508
Dolphins	0.528 ($\lambda = 0.32, 0.28$)	0.526	0.520	0.526	0.526	0.514	0.526	0.520	0.518	0.498
Football	0.605 ($\lambda = 0.73, 0.79$)	0.603	0.600	0.605	0.603	0.588	0.603	0.603	0.603	0.556
Polblogs	0.427 ($\lambda = 0.7, 0.9$)	0.425	0.425	0.425	0.425	0.424	0.425	0.423	0.425	0.427
Cora	0.604 ($\lambda = 0.2$)	0.564	0.582	0.601	0.564	0.548	0.590	0.231	0.526	0.600
Citeseer	0.798 ($\lambda = 0.3, 0.2$)	0.776	0.691	0.712	0.629	0.576	0.621	0.798	0.551	0.724
Pubmed	0.641 ($\lambda = 0.8$)	0.594	0.567	0.581	0.473	0.523	0.438	0.726	0.44	0.751

Analysis for Two Artificial Networks

As a further performance investigation, MMNMF and GMDNMF methods are applied to Lancichinetti-Fortunato-Radicchi (LFR) [26] and Girvan-Newman (GN) [1] networks. The GN network is divided into four non-overlapping communities with 32 nodes in each community. The average degree of each node equals $Z_{in} + Z_{out} = 16$ where Z_{in} and Z_{out} denote the internal and external degrees of the nodes, respectively. LFR networks have some essential characteristics of networks, including power, low distribution of node degrees, and community size. The parameters of the generated LFR network are defined as follows:

The number of nodes is 700, and the average degree and max degree of the network are 20 and 50, respectively. Power law exponent for degree distributions is considered -3, and power-law distribution of community size is -1.

In addition, the community size ranges from 20 to 60 nodes, and the mixing parameter μ varies from 0.1 to 0.9. Moreover, ten independent experiments were executed for comparison.

The GMDNMF method clusters the networks according to the λ parameter. Thus, one way to improve the performance of this method is to choose the correct λ value.

To show the effect of choosing λ on the performance of the GMDNMF method, the networks were formed in terms of various Z_{out} and μ values. Each time, keeping Z_{out} and μ constant for choosing the best λ , we ran the algorithms with various λ values (0-1) in the steps of 0.01

and steps of 0.1 for GN and LFR networks, respectively. The NMI and Q information are depicted in Figs. 1 and 2. As shown, for a given Z_{out} and μ , different λ values cause different NMI and Q information evaluations. This procedure was repeated for different values of Z_{out} and μ , where Z_{out} and μ changed in steps 1 and 0.1, respectively.

Additionally, to demonstrate the effect of λ more clearly, an epigraph of Figs. 1 and 2 has been shown in Figs. 3 and 4. Based on Fig. 3, $\lambda = 0.87$ and $\lambda = 0.98$ maximize both Q ($Q = 0.222$) and NMI ($NMI = 0.942$) information on the GN network with $Z_{out} = 8$. Additionally, in Fig. 4, parameters $\lambda = 0.4$ and $\lambda = 0.9$ maximize Q ($Q = 0.124$) and NMI ($NMI = 0.132$) information on the LFR network with $\mu = 0.8$, respectively. Finally, the experimental results of the mentioned methods on GN and LFR networks are provided in Tables 3 and 4. Based on the obtained data, the following conclusions could be drawn:

- Compared to the NMF and Mtrnmf methods on LFR and GN networks, the proposed GMDNMF outperforms other methods for all μ and Z_{out} values. Moreover, the MMNMF method beats the NMF and Mtrnmf methods for the upper and middle values of μ and Z_{out} .
- In general, for lower values of μ and Z_{out} , the Infomap method outperforms other methods, but GMDNMF and MMNMF would be better for the upper and middle values of μ and Z_{out} compared to other methods.

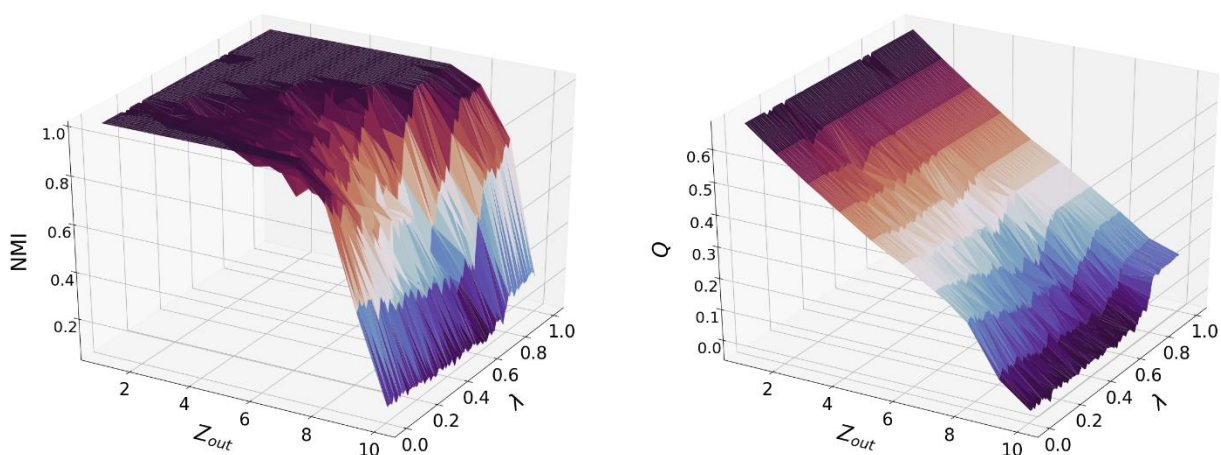


Fig. 1: The modularity index (Q) and NMI information for the GN network for different values of λ with the step of 0.01 and Z_{out} with the step of 1.

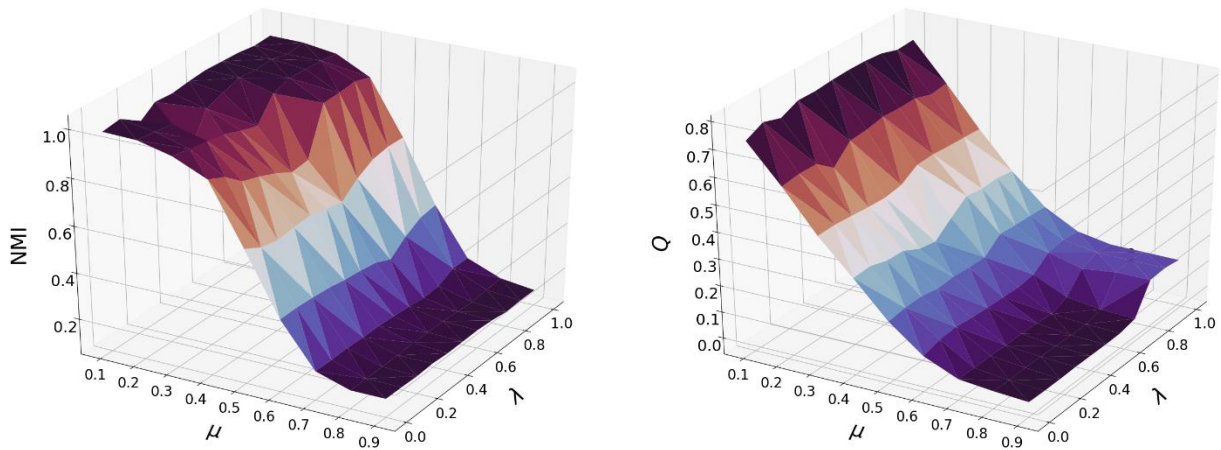


Fig. 2: The modularity index (Q) and NMI information for the LFR network for different values of λ and μ with the steps of 0.1.

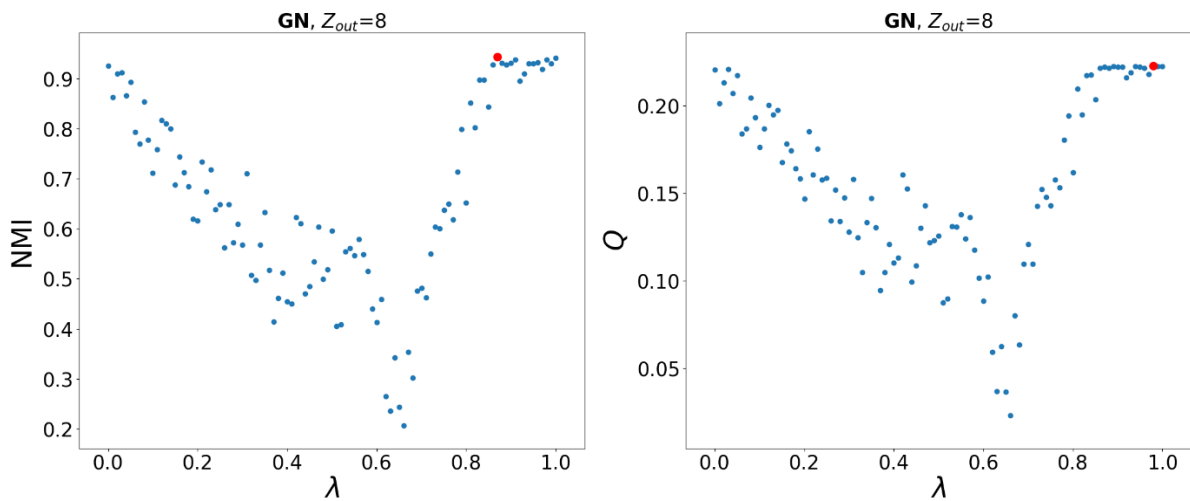


Fig. 3: The modularity index (Q) and NMI information for the GN network for different values of λ with the step of 0.01 and Z_{out} with the step of 1.

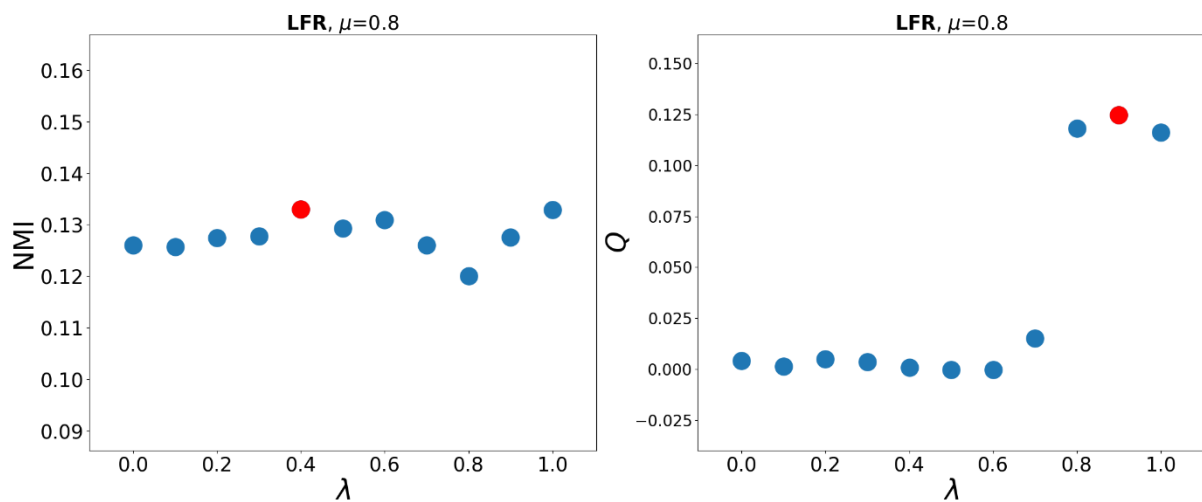


Fig. 4: The modularity index (Q) and NMI information for the LFR network for different values of λ and μ with the steps of 0.1.

Table 3: Comparison of different methods by mean of NMI information for 10 independent experiments on the GN network

Z_{out}	NMI								
	GMDNMF (variable λ)	MMNMF	GMDNMF ($\lambda = 0.5$)	MHGNNMF_sq	LRSCD	NMF	Mtrnmf	Infomap	LPAM
4	1	1	0.98	1	1	1	1	1	1
5	1	1	1	0.99	1	1	1	1	1
6	1	1	0.99	1	1	1	1	0	1
7	1	1	0.925	0.99	1	1	1	0	0.98
8	0.97	1	0.925	0.64	0.89	0.89	0.64	0	0.96
9	0.73	0.704	0.47	0.41	0.47	0.41	0.33	0	0.38
10	0.2	0.1	0.18	0.054	0.16	0.04	0.054	0	0.16

Table 4: Comparison of different methods by mean of NMI information for 10 independent experiments on the LFR network

μ	NMI								
	GMDNMF	MMNMF	GMDNMF ($\lambda = 0.5$)	MHGNNMF_sq	LRSCD	NMF	Mtrnmf	Infomap	LPAM
0.1	0.99	0.95	0.98	0.99	0.99	0.98	0.98	1	1
0.2	0.99	0.94	0.96	0.98	0.99	0.98	0.98	1	1
0.3	0.98	0.93	0.93	0.95	0.98	0.83	0.95	1	0.98
0.4	0.92	0.88	0.85	0.91	0.91	0.70	0.88	0.92	0.91
0.5	0.66	0.66	0.62	0.38	0.62	0.57	0.2	0	0.54
0.6	0.39	0.39	0.37	0.20	0.33	0.33	0.11	0	0.30
0.7	0.19	0.19	0.17	0.2	0.15	0.17	0.06	0	0.15
0.8	0.13	0.13	0.13	0.09	0.12	0.12	0.05	0	0.1
0.9	0.12	0.11	0.12	0.05	0.12	0.12	0.05	0	0.09

Complexity Analysis and Comparison

In this subsection, the order of computational complexity would be computed for the proposed methods using O notation. For this purpose, first, the order of complexity of the main component would be evaluated, and the computational complexity of the two algorithms would be obtained accordingly.

In the MMNMF model, the computational complexity of updating rules is of $O(n^2k) + O(k^2n)$, $O(n^2k) + O(k^2n)$, $O(n^2k) + O(k^2n)$, and $O(n)$ for W , X , \tilde{X} and X^* , respectively. Further, the computational complexity of MMNMF models is $O(n^2k) + O(k^2n) + O(n)$. Given that k is a small constant (i.e., $k \ll N$), the total complexity order is $O(n^2k)$.

The analysis of the GMDNMF model is similar to the MMNMF model. Therefore, the complexity orders of updating rules for W , \tilde{X} , and X^* are $O(n^2k) + O(k^2n)$, $O(n^2k) + O(kn) + O(k^2n)$, and $O(kn)$, respectively. Moreover, the updating rule for X has the order of $O(3n^2k) + O(k^2n) + O(k^2)$. Ultimately, the complexity order for all parameters in the GMDNMF model is

$O(n^2k) + O(k^2n) + O(kn) + O(k^2)$. Considering that $k \ll N$, the total complexity for both models is $O(n^2k)$. If different λ values are examined seeking for the best results, the computational complexity will be $O(n^2kK_\lambda)$ where K_λ indicates the number of selected values for λ . The K_λ value can be selected with respect to a tradeoff between computational complexity and performance improvement. As the value of K_λ increases (e.g., $K_\lambda = 100$), better performance would be achieved, but with more computational complexity. Therefore, K_λ should be determined based on a compromise between performance and complexity.

Finally, if both algorithms converge after I_t iteration, the total complexity computation is $O(I_t n^2 k)$ in the MMNMF model and GMDNMF model with $\lambda = 0.5$. Furthermore, the total complexity for the K_λ values of different λ values in the GMDNMF model is $O(I_t n^2 k K_\lambda)$. Hence, the computational complexity of the GMDNMF model with respect to different λ values ($O(I_t n^2 k K_\lambda)$) is higher than that of the MMNMF model ($O(I_t n^2 k)$). Consequently, as pointed in previous studies [18], [25],

[4], [30] for other NMF-based community detection methods, the proposed methods are unsuitable for large-scale complex networks due to their complexity orders.

To compare the speed of different models, the run times of nine models on six real-world and two synthetic networks are recorded and brought in Tables 5 and 6. For this purpose, in Table 5 the proposed methods have been compared with other methods being NMF [12], LPAM [45], CNM [46] and Infomap [19]. From the results, it is clear that while the run times of all methods are of the same order of magnitude, but, the CNM model is faster than other models and LPAM model has more execution time. Hence, from the run time perspective, our algorithms are inferior to CNM but better than LPAM and Infomap models. Next, in Table 6 the run times of GMDNMF, MHGNMF_sq and Mtrinmf models are compared. For these methods, in contrast to methods in Table 5, parameter tuning is required, which is done generally by trial and error.

Therefore, Table 6 presents the run times of these models for the selection of 20 different values for the internal parameters.

The results in Table 6 show that GMDNMF is faster than the Mtrinmf, but slower than MHGNMF_sq models for 20 different values of the internal parameters. Hence, from NMF-based community detection models, the run time of MMNMF model is quite near to the NMF model and GMDNMF model is near to MHGNMF_sq and Mtrinmf methods. In summary, according to the results of Tables 1 to 6, our models are more flexible, less sensitive with better performance respecting to other models by utilizing the information and characteristics of the networks such as general modularity density and modularity indices, while their run times remain acceptable.

The machine used for the present study is powered Intel Core i7-6770 CPU and 16 GB RAM with 64-bit Windows 10, and Python (version 3.8) as the selected software.

Table 5: Comparison of run times (seconds) for different models and sets

	Karate	Jazz	Political books	Dolphins	Football	Polblogs	Cora	Citeseer	Pubmed	GN	LFR
MMNMF	0.181	0.93	0.85	0.27	0.83	3.78	13.76	15.09	3427	0.297	4.75
GMDNMF ($\lambda = 0.5$)	0.196	0.87	0.83	0.27	0.89	4.02	13.89	18.09	3731	0.354	5.65
NMF [12]	0.175	0.89	0.82	0.27	0.73	3.32	13.06	13.29	2953	0.290	5.05
Infomap [19]	0.203	0.91	0.86	0.28	0.97	4.13	14.7	16.6	4030	0.449	5.10
LPAM [45]	0.428	2.18	0.98	0.71	2.08	13.12	66.8	76.2	5837	0.891	24.84
CNM [46]	0.118	0.80	0.63	0.18	0.67	2.56	9.3	11.7	1916	0.127	3.12

Table 6: Comparison of run times (seconds) of models for different sets with 20 various internal parameter values

	Karate	Jazz	Political books	Dolphins	Football	Polblogs	Cora	Citeseer	Pubmed	GN	LFR
GMDNMF	4.5	20.2	27.1	6.3	14.7	97.1	189.1	365.3	28875	7.3	102.9
MHGNMF_sq [27]	3.9	15.3	14.8	4.4	15.3	76.9	178.0	347.1	26901	6.8	93.8
Mtrinmf [30]	4.8	21.8	28.7	6.6	16.6	112.7	200.5	420.4	32510	8.2	120.4

Conclusion

In this paper, MMNMF and GMDNMF were presented as two novel NMF-based community detection methods to identify the best communities in complex networks. To this end, it was proved that the modularity/general modularity density-based community detection could be

consistently represented in the form of SNMF-based community detection. This consistent representation helped combine NMF-based community detection with modularity/general modularity density-based community detection approaches. The proposed MMNMF model improved the performance of community detection

based on NMF by employing the modularity index as the network feature for the NMF model. The proposed GMDNMF model could enhance NMF-based community detection using the general modularity density index. Iterative update rules were derived as an optimal solution for solving MMNMF and GMDNMF optimization models. The performances of the two models were verified on various artificial and real-world networks of different sizes. According to the results, MMNMF and GMDNMF performed better than the other community detection methods. Additionally, the GMDNMF model had higher computational complexity compared to the MMNMF model, but it outperformed this model.

As future works, the proposed MMNMF and GMDNMF can be extended for NMF-based community detection in multi-layer networks. These proposed models may improve the performance of the multi-view clustering method for community detection by combining link and content information.

Author Contributions

M. ghadirian designed and simulated the proposed method and wrote the manuscript. N. Bigdeli chose strategies, analyzed the results, edited the manuscript, and managed the entire process.

Acknowledgment

The author would like to thank the editor and reviewers for their helpful comments

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

NMI	Normalized Mutual Information
IP	Integer Programming
NMF	Nonnegative Matrix Factorization
GNMF	Graph Regularized Nonnegative Matrix Factorization
DNMF	Deep Nonnegative Matrix Factorization
MDNMF	Modularized Deep Nonnegative Matrix Factorization
tri-NMF	Tri-factor Nonnegative Matrix Factorization
Mtrinmf	Modularized Tri-factor Nonnegative Matrix Factorization

SNMF	Symmetric Nonnegative Matrix Factorization
MMNMF	Mixed Modularity Nonnegative Matrix Factorization
GMDNMF	General Modularity Density Nonnegative Matrix Factorization
SVDCNMF	Singular-value Decomposition Community Detection Nonnegative Matrix Factorization
RSSNMF	Robust Semi-supervised Nonnegative Matrix Factorization
CNM	Clauset-Newman-Moore
MHGNNMF	Mixed hypergraph Nonnegative Matrix Factorization
LRSCD	Low-rank Subspace Learning-based Network Community Detection
LFR	Lancichinetti–Fortunato–Radicchi
GN	Girvan–Newman

References

- [1] M. E. J. Newman, "Networks," OUP, 2018.
- [2] P. Bedi, C. Sharma, "Community detection in social networks," *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery*, 6 (3): 115-135, 2016.
- [3] Z. Li, S. Zhang, R. S. Wang, X. S. Zhang, L. Chen, "Quantitative function for community detection," *Phys. Rev. E*, 77 (3): 036109, 2008.
- [4] L. H. N. Lorena, M. G. Quiles, L. A. N. Lorena, "Improving the performance of an integer linear programming community detection algorithm through clique filtering," in *Proc. International Conference on Computational Science and Its Applications (ICCSA)*: 757–769, 2019.
- [5] M. Sathyakala, M. A. Sangeetha, "Weak clique based multi objective genetic algorithm for overlapping community detection in complex networks," *J. Ambient Intell. Humaniz. Comput.* 12: 6761–6771, 2021.
- [6] M. Mohammadi, M. Fazeli, M. Hosseinzadeh, "Parallel louvain community detection algorithm based on dynamic thread assignment on graphic processing unit", *J. Electr. Comput. Eng. Innovations (JECEI)*, 10(1): 75-88, 2022.
- [7] C. K. Tsung, S. L. Lee, H. J. Ho, S. Chou, "A modularity-maximization-based approach for detecting multi-communities in social networks," *Ann. Oper. Res.*, 303: 381–411, 2021.
- [8] S. Muff, F. Rao, A. Cafilisch, "Local modularity measure for network clusterizations", *Phys. Rev. E*, 72: 056107, 2005.
- [9] K. Sato, Y. Izunaga, "An enhanced MILP-based branch-and-price approach to modularity density maximization on graphs," *Comput. Oper. Res.*, 106: 236–245, 2019.
- [10] J. Liu, J. Zeng, "Community detection based on modularity density and genetic algorithm," in *Proc. 2010 International Conference on Computational Aspects of Social Networks*: 29-32, 2010.
- [11] M. Li, J. Liu, "A link clustering based memetic algorithm for overlapping community detection," *Phys. A: Stat. Mech. Appl.*, 503: 410–423, 2018.

- [12] A. Costa, "MILP formulations for the modularity density maximization problem," *Eur. J. Oper. Res.*, 245(1): 14–21, 2015.
- [13] M. J. Barber, J. W. Clark, "Detecting network communities by propagating labels under constraints," *Phys. Rev. E*, 80: 026129, 2009.
- [14] Q. Wu, R. Chen, L. Wang, K. Guo, "A label propagation algorithm for community detection on high-mixed networks," *Concurr. Comput. Pract. Exp.*, 33 (9): e6141, 2020.
- [15] M. Rosvall, C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Natl. Acad. Sci. U.S.A.*, 105 (4): 1118–1123, 2008.
- [16] J. Zhou, L. Li, A. Zeng, Y. Fan, Z. Di, "Random walk on signed networks," *Phys. A: Stat. Mech. Appl.*, 508: 558–556, 2018.
- [17] C. Liu, F. Huang, R. Li, Q. Yang, Y. Li, S. Yu, "Community detection using multitopology and attributes in social networks," *Concurr. Comput. Pract. Exp.*, 34 (12): e6028, 2020.
- [18] R. S. Wang, S. Zhang, Y. Wang, X. Zhang, L. Chen, "Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures," *Neurocomputing* 72 (1-3): 134–141, 2008.
- [19] L. Xu, T. Ming, W. Xiaofei, W. Chao, F. Qiang, Y. Yonghong, "Single-channel speech separation based on non-negative matrix factorization and factorial conditional random field," *Chin. J. Electron.*, 27 (5): 1063–1070, 2018.
- [20] S. Peng, W. Ser, B. Chen, Z. Lin, "Robust semi-supervised nonnegative matrix factorization for image clustering," *Pattern Recognit.*, 111: 107683, 2021.
- [21] S. Zhang, G. Zhang, F. Li, C. Deng, S. Wang, A. Plaza, J. Li, "Spectral-spatial hyperspectral unmixing using nonnegative matrix factorization," *IEEE Geosci. Remote. Sens.*, 60: 5505713, 2021.
- [22] E. L. Lydia, P. K. Kumar, K. Kumar, S. K. Lakshmanaprabu, R. M. Vidhyavathi, "Charismatic Document clustering through novel K-Means Non-negative Matrix Factorization (KNMF) Algorithm using key phrase extraction," *Int. J. Parallel Program.*, 48: 496–514, 2020.
- [23] C. He, Y. Tang, K. Liu, H. Li, S. Liu, "A robust multi-view clustering method for community detection combining link and content information," *Phys. A: Stat. Mech. Appl.*, 514: 396–411, 2018.
- [24] K. Shu, S. Wang, H. Liu, "Beyond news contents: the role of social context for Fake news detection," in *Proc. Twelfth ACM International Conference on Web Search and Data Mining*: 312–320, 2019.
- [25] C. He, Q. Z. Y. Tang, S. Liu, J. Zheng, "Community detection method based on robust semi-supervised nonnegative matrix factorization," *Phys. A: Stat. Mech. Appl.*, 523: 279–291, 2019.
- [26] H. Lu, X. Sang, Q. Zhao, J. Lu, "Community detection algorithm based on nonnegative matrix factorization and pairwise constraints," *Phys. A: Stat. Mech. Appl.*, 522: 205–214, 2019.
- [27] W. Wu, S. Kwong, Y. Zhou, Y. Jia, W. Gao, "Nonnegative matrix factorization with mixed hypergraph regularization for community detection," *Inf. Sci.*, 435: 263–281, 2018.
- [28] M. Zhang, Z. Zhou, "Structural deep nonnegative matrix factorization for community detection," *Appl. Soft Comput.*, 97: 106846, 2020.
- [29] J. Huang, T. Zhang, W. Yu, J. Zhu, E. Cai, "Community detection based on modularized deep nonnegative matrix factorization," *Int. J. Pattern Recognit. Artif. Intell.*, 2 (35): 2159006, 2021.
- [30] C. Yan, Z. Chang, "Modularized tri-factor nonnegative matrix factorization for community detection enhancement," *Phys. A: Stat. Mech. Appl.*, 533: 122050, 2019.
- [31] X. Ma, L. Gao, L. Fu, X. Yong, "Semi-supervised clustering algorithm for community structure detection in complex networks," *Phys. A: Stat. Mech. Appl.*, 389(1): 187–197, 2010.
- [32] X. Wang, P. Cui, J. Wang, J. pei, W. Zhu, S. Yang, "Community preserving network embedding," in *Proc. Thirty-First AAAI Conference on Artificial Intelligence*: 203–209, 2017.
- [33] X. Ma, D. Dong, Q. Wang, "Community detection in multi-layer networks using joint nonnegative matrix factorization," *IEEE Trans. Knowl. Data. Eng.*, 31 (2): 273–286, 2019.
- [34] L. Zong, Z. Zhang, L. Zhao, H. Yu, Q. Zhao, "Multi-view clustering via multi-manifold regularized non-negative matrix factorization," *Neural Netw.*, 88: 74–89, 2017.
- [35] S. Peng, W. Ser, B. Chen, Z. Lin, "Robust orthogonal nonnegative matrix tri-factorization for data representation," *Knowl-Based Syst.*, 201–202: 106054, 2020.
- [36] J. Liu, C. Wang, J. Gao, J. Han, "Multi-view clustering via Joint nonnegative matrix factorization," in *Proc. the 2013 SIAM International Conference on Data Mining*: 252–260, 2013.
- [37] A. Clauset, M. E. J. Newman, C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, 70(6): 066111, 2004.
- [38] Z. Ding, Z. Shang, D. Sun, B. Luo, "Low-rank subspace learning based network community detection," *Knowl-Based Syst.*, 155: 71–82, 2018.
- [39] W. W. Zachary, "An information flow model for conflict and fission in small groups," *J. Anthropol. Res.*, 33 (4): 452–473, 1977.
- [40] P. M. Gleiser, L. Danon, "Community structure in jazz," *Adv. Compl. Syst.*, 6 (4): 565–573, 2003.
- [41] J. Kunegis, "KONECT: The koblenz network collection," in *Proc. 22nd International Conference on World Wide Web*: 1343–1350, 2013.
- [42] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson, "The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations," *Behav. Ecol. Sociobiol.*, 54 (4): 396–405, 2003.
- [43] A. Lancichinetti, S. Fortunato, F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Phys. Rev. E*, 78 (4): 046110, 2008.
- [44] L. Yang, X. Cao, D. Jin, X. Wang, D. Meng, "A unified semi-supervised community detection framework using latent space graph regularization," *IEEE Trans. Cyber.*, 45(11): 2585–2598, 2015.
- [45] L. A. Adamic, N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in *Proc. 3 the 3rd international workshop on Link discovery*: 36–43, 2005.
- [46] D. He, Z. Feng, D. Jin, X. Wang, W. Zhang, "Joint identification of network communities and semantics via integrative modeling of network topologies and node contents," in *Proc. the Thirty-First AAAI Conference on Artificial Intelligence*: 116–124, 2017.
- [47] G. Namata, B. London, L. Getoor, B. Huang, U. EDU, "Query-driven active surveying for collective classification," in *Proc. 10th Workshop on Mining and Learning with Graphs*, 8, 2012.

Biographies



Mohammad Ghadirian was born in Iran, in 1992, He received M.Sc. degree in Electrical Engineering Majoring in Control from the Sharif University of Technology, Tehran, Iran in 2016. He is already Ph.D. candidate in Electrical Engineering Department of Imam Khomeini International University, Qazvin, Iran. His research interest includes in graph mining, data mining and medical image processing.

- Email: s956191004@edu.ikiu.ac.ir
- ORCID: ID: 0000-0002-4106-2406
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Nooshin Bigdeli was born in 1977 in Iran, and completed her Ph.D. degree in Electrical Engineering majoring in Control at Sharif University of Technology, Tehran, Iran in 2007. She is currently professor of Electrical Engineering Department of Imam Khomeini International University, Qazvin, Iran. Her research interests include control systems, applied optimization, intelligent systems, model predictive control as well as model

order reduction in high order systems.

- Email: n.bigdeli@eng.ikiu.ac.ir
- ORCID: [0000-0001-5536-4491](https://orcid.org/0000-0001-5536-4491)
- Web of Science Researcher ID: AAT-8622-2021
- Scopus Author ID: 8528681600
- Homepage: <http://www.ikiu.ac.ir/members/?id=23&lang=0>

How to cite this paper:

M. Ghadirian, N. Bigdeli, "A new hybrid nmf-based infrastructure for community detection in complex networks," J. Electr. Comput. Eng. Innovations, 11(2): 443-458, 2023.

DOI: [10.22061/jecei.2023.9150.577](https://doi.org/10.22061/jecei.2023.9150.577)

URL: http://jecei.sru.ac.ir/article_1879.html





Research paper

Revised Estimations for Cost and Success Probability of GNR-Enumeration

G. R. Moghissi*, A. Payandeh

Department of ICT, Malek-Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:

Received 12 January 2023
Reviewed 07 March 2023
Revised 25 April 2023
Accepted 07 May 2023

Keywords:

BKZ simulation
Enumeration cost
Success probability
Optimal enumeration radius
Bounding function generator

*Corresponding Author's Email
Address: fumoghissi@chmail.ir

Abstract

Background and Objectives: Since exact manner of BKZ algorithm for higher block sizes cannot be studied by practical running, therefore simulation of BKZ is used to predict the total cost of BKZ and quality of output basis. This paper revises some main components of BKZ-simulation for better predictions.

Methods: At first, by definition of full-enumeration success probability, the optimal enumeration radius is formally defined. Next, this paper defines three more pruning types, besides the well-known pruning by bounding function in GNR-enumerations, and consequently uses these four pruning types collectively in revision of success probability estimation. Also, by using these four pruning types and the process of updating-radius, this paper revises the estimation of enumeration cost. Finally, this paper introduces a simple technique to generate partially better bounding functions.

Results: For block sizes of $50 \leq \beta \leq 240$, better domains of radius parameters in GNR enumeration are formally introduced. Also, our revised estimation of success probability (for GNR bounding function) in our test results shows non-negligible gap from former estimations in some main former studies. Moreover, our results show that the cost results by our proposed estimator of GNR-enumeration cost are closer to the cost results determined in experimental running of enumeration, than the cost results by Chen-Nguyen estimator.

Conclusion: This paper revises the estimators of cost and success probability for GNR-Enumeration, and justifies the value of these revised estimators by sufficient test results (in actual running and simulation of BKZ). Also, our novel definition of optimal enumeration radius can be used effectively in actual running and simulation of BKZ.

This work is distributed under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)



Introduction

Lattice reduction is the main part of most lattice security attacks. BKZ algorithm is one of the main practical lattice reductions. The security parameters in lattice-based cryptographic primitives are estimated by determining the total cost and output quality of BKZ algorithm in high block sizes. For predicting the manner of BKZ in higher block sizes, practical running is not the way, therefore BKZ-simulators are introduced. There are some claimant BKZ-simulations in former studies, such as the one is

introduced by Chen and Nguyen [1], the simulation by Shi Bai et al. [2], and the simulation by Aono et al. [3]; The outputs of BKZ-simulation are divided by two main parts as total cost and output quality which can be used in lattice based security analysis.

The cost of enumeration function on the lattice block of $\mathcal{L}_{[1 \dots \beta]}$ can be estimated by old version of [1] as $N = 2^{0.00405892 \beta^2 - 0.337913 \beta + 34.9018}$ or by [4] as $N = 2^{0.000784 \beta^2 - 0.366 \beta - 0.9}$; This is obvious that using exact versions of "success probability estimators",

“enumeration cost estimators”, “minimum effective (optimal) enumeration radius” and “bounding function generator”, which are revised in this paper, can make such these cost models more exact and more close to the practical estimations. To the best of our knowledge, the technique of GNR-enumeration and corresponding concepts studied in [1], [13], is considered yet in current studies and security estimations of Lattice based cryptography, however other techniques such as sieve algorithm, discrete pruning and RSR algorithm may show better results in practical attacks. In fact, the significance of our contributions in this paper for bit-security estimation of lattice-based cryptographic primitives can be more justified by the massive efforts of Albrecht et al., in estimation of the LWE/NTRU schemes [5] (see the user-friendly scripts for these estimations in [6]). For example, the cost of primal attack against “Falcon-1024-2.87-12289” (with claimed bit-security of 230) by using enumeration with four different cost models in Table 10 from [5] is estimated as 2^{418} , 2^{474} , 2^{836} and 2^{1118} ! The authors of this paper believe that using non-exact components and definitions on enumeration functions lead to these gaps in bit-security estimations, while our contributions in this paper try to fix the problem of such non-exactness.

It is worthy of noting that the results of [5] are massively used in “Post-Quantum Cryptography Standardization Project” (see the corresponding information in [7]), also in bit-security estimations of current cryptography researches, such as [8]-[11].

In other view on this work, designing a BKZ-simulation with GNR-pruned enumeration needs to some necessary building-blocks which include enumeration radius, generation of bounding function, estimation of success probability, LLL simulation, estimation of GNR enumeration cost, sampling method for enumeration solution, simulation of updating GSO. Our previous study in [12] focuses on design of sampling method for enumeration solution (as solution norm and coefficient vectors). This paper introduces some main revisions for following components: optimal enumeration radius, generation of bounding function, estimation of success probability and GNR enumeration cost. The components which are studied in this paper (except estimation of enumeration cost) can be used in actual running of BKZ algorithm (besides the simulation of BKZ) too! Our contributions in this paper are described briefly as follows:

- By definition of full-enumeration success probability in this paper, the optimal value for radius parameter \sqrt{Y} (as initial radius factor r_{FAC}) and corresponding bound for solution norm of full-enumeration are defined exactly in average-case. This definition can be used dynamically to compute optimal enumeration

radius in BKZ simulation and even actual running of BKZ algorithm. In other sides, former studies on BKZ-simulation [1]-[3] don't use optimal version of the radius parameter of r_{FAC} .

- The former studies [1]-[3] use the efficient idea by [1] to estimate the success probability of GNR-enumerations which only consider the pruning type by cylinder-intersection of bounding function; This paper proposes three more types of pruning in estimation of success probability;
- The former studies [1]-[3] use the efficient idea by [1] to estimate the cost of GNR-enumerations which only consider the pruning type by cylinder-intersection of bounding function; Similar to our revision for success probability, this paper considers all of four types of pruning along with the process of updating enumeration radius in our estimation of GNR-enumeration cost;
- This paper introduces a generator of bounding function including cutting point of $\text{Cut} = d$ [12]; In former studies [1]-[3], if the simulation tries to generate bounding functions with much small success probability, this is possible that the success probability of this bounding functions unintentionally becomes much less than intended value or even zero!

The remainder of this paper is organized as follows.

Second section is dedicated to essential background for our contributions in this paper. In **third section**, we describe our contributions as follows:

- In **third section (Part A)**, the optimal enumeration radius is defined exactly;
- In **third section (Part B)**, our estimation of success probability is introduced;
- In **third section (Part C)**, our estimation of GNR-enumeration cost is introduced;
- In **third section (Part D)**, a simple technique for forcing $\text{Cut} = \beta$ in generation of bounding function is introduced.

Also, our test results for these contributions are introduced in **fourth section**. Finally, in **fifth section**, the conclusion for this work is expressed.

Background

In this section, the needed preliminaries on theory of lattice, BKZ-reduction and other corresponding concepts for this work are introduced.

A. Basic Definitions and Notations

In this section, some basic concepts, needed in this paper, are defined.

Lattices. For n -linearly independent vectors of $b_1, \dots, b_n \in \mathbb{R}^m$, the lattice generated by these vectors is defined as following set:

$$\mathcal{L}(b_1, \dots, b_n) = \{\sum_{i=1}^n x_i b_i : x_i \in \mathbb{Z}\}. \quad (1)$$

The set of vectors $[b_1, \dots, b_n]$ is known as a lattice basis which is usually shown by a column-matrix B where $b_i \in \mathbb{Z}^m$ for cryptographic applications. Also, the rank and dimension of lattice $\mathcal{L}(B)$ are respectively shown by n and m . In this paper, the notation of \mathcal{L}_i is defined as follows:

$$\mathcal{L}_i = [b_1, b_2, \dots, b_i]. \quad (2)$$

Euclidean norm. The length of a lattice vector $v = (v_1, \dots, v_m)$ is measured by $\|v\| = \sqrt{v_1^2 + \dots + v_m^2}$. In this paper, the phrases of “norm” and “length” refer to Euclidean norm.

Volume of Lattices. The volume of a lattice $\mathcal{L}(B)$ is defined by determinant of basis matrix:

$$\text{Vol}(\mathcal{L}(B)) = |\det B|. \quad (3)$$

First Successive-Minima of lattice \mathcal{L} . The norm of shortest nonzero vector in lattice \mathcal{L} is first successive-minima of that lattice and is shown by $\lambda_1(\mathcal{L})$.

In the worst-case of the SVP solver, the optimal (smallest) value of Hermite-factor for all n -dimensional input lattice bases are defined formally as follows:

Hermite’s constant. Hermite’s constant γ_n is supremum of the ratio $(\lambda_1(\mathcal{L})/\text{Vol}(\mathcal{L})^{1/n})^2$ over all n -dimensional lattices.

By sterling approximation for high dimensional space, volume of a n -dimensional sphere (ball) is computed as follows:

$$\begin{aligned} V_n(R) &= \text{Vol}(\text{Ball}_n(R)) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)} R^n \\ &\approx \frac{1}{\sqrt{n\pi}} \left(\frac{2\pi e}{n}\right)^{\frac{n}{2}} R^n. \end{aligned} \quad (4)$$

In this paper, $V_l(R)$ refers to the volume of a l -dimensional ball with radius R . The gamma function $\Gamma(x)$ is defined for $x > 0$ by $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$, where by using sterling approximation, the gamma function $\Gamma(n/2 + 1)$ is defined as $\Gamma(n/2 + 1) \approx \sqrt{n\pi} \left(\frac{n}{2e}\right)^{n/2}$;

One of the main heuristics in lattice theory is **Gaussian Heuristic** which estimates the number of points in a set S . This heuristic is used massively in our analysis. This heuristic is defined as follows:

Heuristic 1 (Gaussian Heuristic). “Given a lattice \mathcal{L} and a set S , the number of points in $S \cap \mathcal{L}$ is approximated by $\text{Vol}(S)/\text{Vol}(\mathcal{L})$ ” [13];

By using Gaussian Heuristic, if a lattice \mathcal{L} is limited in a centered ball with radius of $R = \lambda_1(\mathcal{L})$, then it is expected that there is at least one lattice vector in $\text{Ball}_n(R)$ with radius R , which is the shortest vector. Therefore, the value of $\lambda_1(\mathcal{L})$ can be estimated by **Gaussian Heuristic** of this lattice as follows (by using sterling approximation):

$$\text{GH}(\mathcal{L}) = \left(\frac{\text{Vol}(\mathcal{L}(B))}{\text{Vol}(\text{Ball}_n(1))}\right)^{\frac{1}{n}} \approx \sqrt{\frac{n}{2\pi e}} (\det B)^{\frac{1}{n}}. \quad (5)$$

Gram-Schmidt Orthogonal basis (GSO basis). For a given lattice basis $B = (b_1, b_2, \dots, b_n)$, the Gram-Schmidt orthogonal basis $B^* = (b_1^*, b_2^*, \dots, b_n^*)$ is defined as follows:

$$\pi_i(b_i) = b_i^* = b_i - \sum_{j=1}^{i-1} \mu_{i,j} b_j^*, \quad (6)$$

where $\mu_{i,j} = \frac{b_i b_j^*}{\|b_j^*\|^2}$ and $1 \leq j < i \leq n$.

The parameter of $\mu_{i,j} \in \mathbb{R}$ is named a GSO coefficient and b_i^* refers to i -th vector of GSO basis of B^* . For an input lattice basis B , the volume of the lattice can be computed by the norm of GSO vectors as follows:

$$\text{Vol}(\mathcal{L}(B)) = \prod_{i=1}^n \|b_i^*\|. \quad (7)$$

Other important heuristic in lattice theory is Schnorr’s **Geometric Series Assumption (GSA)** which is defined as follows:

Geometric Series Assumption (GSA). The geometric series of $\|b_i^*\| = r^{i-1} \|b_1^*\|$ with the **GSA** constant $r \in [3/4, 1)$ can be assumed for a BKZ-reduced basis [3].

Gamma distribution. The Gamma distribution, which is a two-parameter and continuous probability distribution, is defined as follows (for input shape-parameter of k and scale-parameter of θ):

$$\text{Gamma}(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k) \theta^k}, \quad \text{where } x > 0. \quad (8)$$

Exponential distribution. The Exponential distribution, which is a one-parameter and continuous probability distribution, is defined as follows (for input parameter λ):

$$\text{Expo}(x; \lambda) = \lambda e^{-\lambda x}, \quad \text{where } x > 0 \text{ and } \lambda > 0. \quad (9)$$

The mean and variance in Exponential distribution respectively are determined by $1/\lambda$ and $1/\lambda^2$.

Note: The notation of 1^β represents a vector with length of β as a bounding function with entries of 1.

B. Enumeration and GNR-Pruning

In this paper, for each lattice block of $\mathcal{L}_{[j,k]} = \mathcal{L}(b_j, b_{j+1}, \dots, b_k)$, the block size $\beta = k - j + 1$ is assumed sufficiently big. Also since these lattice blocks are assumed to be used in BKZ algorithms, in fact, the notation of $\mathcal{L}(b_j, b_{j+1}, \dots, b_k)$ refers to the projected form of $\pi_j(b_j, b_{j+1}, \dots, b_k)$, as a lattice block from index j to k , while its vectors are projected on the vectors of $(b_1, b_2, \dots, b_{j-1})$.

Full-enumeration. For fixed enumeration radius R (by no updating radius), the tree of full-enumeration includes all lattice points in n -dimensional ball of radius R .

Full-enumeration Cost. For fixed enumeration radius R (by no updating radius), the number of total nodes of the full-enumeration tree can be estimated as follows [1]:

$$N \approx \sum_{l=1}^{k-j+1=d} H_l, \quad (10)$$

where

$$H_l = \frac{1}{2} \frac{V_l(R)}{\prod_{i=d-l+1}^d \|b_i^*\|} = \frac{1}{2} \frac{R^l V_l(1)}{\prod_{i=d-l+1}^d \|b_i^*\|}. \quad (11)$$

The value of H_l represents the **Gaussian Heuristic** prediction of the number of nodes at the level l (see [1], [13]).

Note: In this paper, “enumeration cost”, “total nodes of GNR-enumeration” and “number of enumeration tree nodes” are referred to N as the number of total nodes in the tree [13].

The concepts of cylinder-intersection, bounding function and GNR-pruning formally are defined as follows:

Cylinder-intersection. The l -dimensional cylinder-intersection with radius of (R_1, \dots, R_l) is defined as follows [13]:

$$C_{R_1 \dots R_l} = \{(x_1, \dots, x_l) \in \mathbb{R}^l, \forall 1 \leq i \leq l, \sum_{t=1}^i x_t^2 \leq R_i^2\}. \quad (12)$$

Bounding function. The vector of $\mathcal{R} = [\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_\beta]$ where $0 \leq \mathcal{R}_1 \leq \mathcal{R}_2 \leq \dots \leq \mathcal{R}_\beta = 1$, when multiplied by initial radius of R , defines a bounded cylinder-intersections with radius $(R_1, \dots, R_l) = (R \times \mathcal{R}_1, \dots, R \times \mathcal{R}_l)$ for $1 \leq l \leq \beta$, and consequently can be used to prune the enumeration tree [13].

GNR-pruning (Sound pruning). For a lattice block of $B_{[j,k]} = (b_j, b_{j+1}, \dots, b_k)$ and the coefficient vector $x \in \mathbb{Z}^\beta$, GNR-pruning replaces the inequalities of $\|\pi_{k+1-i}(x \cdot B_{[j,k]})\| \leq R$ for $1 \leq i \leq k-j+1$ as a bounded ball (in full-enumeration) by $\|\pi_{k+1-i}(x \cdot B_{[j,k]})\| \leq \mathcal{R}_i \times R$, where $0 \leq \mathcal{R}_1 \leq \dots \leq \mathcal{R}_{k-j+1} = 1$ as a cylinder-intersection [1].

The pseudo-code of the GNR pruned enumeration is shown in Appendix B from [13]. Based on the definition of GNR-pruning, this paper uses the concepts of final solution vector (usually referred to solution vector) and partial solution candidate as follows:

Final solution vector. For a lattice block of $B_{[j,k]} = (b_j, b_{j+1}, \dots, b_k)$ and the coefficient vector $x \in \mathbb{Z}^\beta$, the projected vector of $\pi_j(v) = \pi_j(x \cdot B_{[j,k]})$ which satisfies all of the conditions of $\|\pi_{k+1-i}(x \cdot B_{[j,k]})\| \leq \mathcal{R}_i \times R$ for $1 \leq i \leq k-j+1$, is a final solution vector.

Note: In this paper, $\pi_j(v)$ is shown by the notation of v for simplicity.

Fact 1 is an obvious proposition on GNR pruned enumeration.

Fact 1. If there are several solution vectors in cylinder-intersection of a GNR pruned enumeration over \mathcal{L}_β , the shortest solution among them never be eliminated by updating radius and finally is returned as the final response of this enumeration;

Partial solution candidate. For a lattice block of $B_{[j,k]} = (b_j, b_{j+1}, \dots, b_k)$, the coefficient vector $x \in \mathbb{Z}^\beta$ and the projection level of ℓ for $1 \leq \ell \leq k-j+1$, the projected vector of $\pi_{k+1-i}(v) = \pi_{k+1-i}(x \cdot B_{[j,k]})$ which satisfies all of the conditions of $\|\pi_{k+1-i}(x \cdot B_{[j,k]})\| \leq \mathcal{R}_i \times R$ for $1 \leq i \leq k-\ell+1$, is a partial solution candidate at the level of ℓ in enumeration tree;

The success probability is one of the main features of bounding function which can be defined as follows [13]:

Success probability of bounding function: For any lattice block of $\mathcal{L}_{[j,k]} = [b_j, b_{j+1}, \dots, b_k]$, initial enumeration radius R and bounding function \mathcal{R} , if there is just one lattice vector v in n -dimensional ball with radius of R (i.e., $\|v\| \leq R$), the probability of finding solution vector v after GNR pruning by \mathcal{R} in enumeration tree is defined as the success probability of \mathcal{R} , which is shown by $p_{succ}(\mathcal{R})$.

For analysis of the success probability of GNR bounding function, Gama et al. use following heuristic [13]:

Heuristic 2. “The distribution of the coordinates of the target vector v , when written in the normalized Gram-Schmidt basis $(b_1^*/\|b_1^*\|, \dots, b_n^*/\|b_n^*\|)$ of the input basis, look like those of a uniformly distributed vector of norm $\|v\|$ ”;

The coefficient of orthonormal basis vector $z = (z_1, z_2, \dots, z_{k-j+1=d})$ in **Heuristic 2** which corresponds with the target lattice vector of v can be formulated as follows [13]:

$$v = [z_1, \dots, z_d] \begin{bmatrix} b_k^*/\|b_k^*\| \\ \vdots \\ b_j^*/\|b_j^*\| \end{bmatrix} = (v_1, \dots, v_m). \quad (13)$$

where $b_i^*/\|b_i^*\|$ is i -th vector of the orthonormal basis of $b_1^*/\|b_1^*\|, \dots, b_n^*/\|b_n^*\|$ [13]. Also, the solution vector v can be written by the coefficient vector $w = (z_d/\|b_1^*\|, \dots, z_2/\|b_{d-1}^*\|, z_1/\|b_d^*\|)$ on the GSO block basis as follows [12]:

$$v = (v_1, \dots, v_m) = (w_1, \dots, w_d) \begin{bmatrix} b_1^* \\ \vdots \\ b_d^* \end{bmatrix}. \quad (14)$$

The coordinates of the coefficient vector z are reversed (i.e., z_i corresponds to $b_{k-i+1}^*/\|b_{k-i+1}^*\|$), and it is clear that $\|z\| = \|v\|$ [13]. Also, the vector $u = (u_1, u_2, \dots, u_{k-j+1=d}) = (z_1/R, z_2/R, \dots, z_d/R)$ is chosen to be uniformly distributed from the d -dimensional ball of the radius 1 (by the notation of $u \sim \text{Ball}_d$). By using these formulations, success probability of a GNR bounding function \mathcal{R} can be defined as follows [13]:

$$p_{succ}(\mathcal{R}) = \Pr_{u \sim \text{Ball}_d} \left(\forall i \in [1, d], \sum_{l=1}^i u_l^2 \leq \frac{R_i^2}{R_d^2} \right) = \Pr_{u \sim \text{Ball}_d} \left(\forall i \in [1, d], \sum_{l=1}^i u_l^2 \leq \mathcal{R}_i^2 \right). \quad (15)$$

Note: Since in last block of BKZ, the size of blocks become less than initial block size of β , so the variable size of $d = k - j + 1$ is used to emphasize this fact.

C. Cost of GNR-enumeration

The estimation of total nodes in GNR pruned enumeration tree is the same as the full-enumeration (Schnorr-Euchner enumeration), except that instead of using balls of radius R , GNR pruned enumeration employs the cylinder-intersections of radius $(R_1, \dots, R_l) = (R \times \mathcal{R}_1, \dots, R \times \mathcal{R}_l)$ for $1 \leq l \leq \beta$. In reminder of this paper, the enumeration radius is determined by parameter of r_{FAC} , as follows:

$$R = r_{\text{FAC}} \times GH(\mathcal{L}). \quad (16)$$

By using **Heuristic 1 (Gaussian Heuristic)**, the number of nodes at the level l of the GNR pruned enumeration tree can be estimated as follows:

$$H'_l = \frac{1}{2} \frac{V_{R_1, \dots, R_l}}{\prod_{i=k-l+1}^k \|b_i^*\|} = \frac{1}{2} \frac{R^l V_{\mathcal{R}_1, \dots, \mathcal{R}_l}}{\prod_{i=k-l+1}^k \|b_i^*\|}. \quad (17)$$

The volume of cylinder-intersection of C_{R_1, \dots, R_l} can be defined as follows:

$$V_{R_1, \dots, R_l} = \text{Vol}(C_{R_1, \dots, R_l}) = V_l(R) \times \Pr_{u \sim \text{Ball}_l}(\forall j \in [1, l], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2). \quad (18)$$

Therefore, the total number of nodes in the GNR pruned enumeration tree can be estimated as follows:

$$N'(\mathcal{L}_{[j,k]}, \mathcal{R}', R) \approx \sum_{l=1}^{k-j+1} H'_l \approx \sum_{l=1}^{\beta} \Pr_{u \sim \text{Ball}_l}(\forall j \in [1, l], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2) \times H_l. \quad (19)$$

Note: In this paper, the total number of nodes in full-enumeration tree and number of nodes in level l of full-enumeration tree are shown by N and H_l , while the total number of nodes in GNR pruned enumeration tree and number of nodes in level l of GNR pruned enumeration tree are shown by N' and H'_l .

As shown in Section 3.3 from [13], the most populated level of full-enumeration tree is middle-level. By assuming the populated level in middle-level, paper [13] concludes the approximation of $N'(\mathcal{L}_{[1,\beta]}, \mathcal{R}', R) \approx H'_{l=\beta/2}$ for all GNR-enumeration (not just for full-enumeration). To the best of our knowledge, by using this approximation for well-defined bounding function of Linear pruned, paper [13] concludes that the total cost of enumeration pruned by an optimal bounding function with success probability $\approx 100\%$ tends to $\frac{1}{2^{\beta/4}}$ times of total cost of full-enumeration; In other side, by using this approximation for some well-defined bounding functions of Piecewise-Linear and Step bounding function, paper [13] tries to show that for extremely small success probability, the total cost of enumeration pruned by these two bounding functions (as extreme-pruning) tends to $\frac{1}{2^{\beta/2}}$ times of total cost of full-enumeration. Our analysis in [14] proves the

assumption of most populated level of $l = \beta/2$ and speedup of $\frac{1}{2^{\beta/4}}$ for an optimal bounding function with success probability $\approx 100\%$, but rejects the assumption of most populated level of $l = \beta/2$ and speedup of $\frac{1}{2^{\beta/2}}$ for Piecewise-Linear bounding function for extremely small success probability.

The success probability of the bounding function \mathcal{R} can be estimated by Monte-Carlo simulation (see Algorithm 8 in [1]) which is used in some test results of this paper, but it is not efficient since the number of samples required for this estimation is proportional to $\frac{1}{p_{\text{succ}}(\mathcal{R})}$ [13]. The Monte-Carlo estimation of success probability is defined by the number of $\frac{1}{p_{\text{succ}}(\mathcal{R})}$ samples of random vector $u \sim \text{Ball}_d$, and counting the success of each sample satisfying the bounding function constraints which is defined as follows [1]:

$$\begin{aligned} \forall i \in [1, d], \sum_{l=1}^i z_l^2 &\leq \mathcal{R}_i^2 R_d^2 \equiv \\ \forall i \in [1, d], \sum_{l=1}^i u_l^2 &\leq \frac{\mathcal{R}_i^2}{R_d^2} = \mathcal{R}_i^2 \equiv \\ \forall i \in [1, d], \sum_{l=1}^i \frac{\omega_{d-l+1}}{\sum_{t=1}^d \omega_t} &\leq \mathcal{R}_i^2, \end{aligned} \quad (20)$$

where $\omega_i \leftarrow \text{Gamma}(1/2, 2)$ and $R = R_d$ is the enumeration radius. Some speedup for Monte-Carlo estimation of $p_{\text{succ}}(\mathcal{R})$ can be introduced by replacing the ball with a smaller containing body whose volume is known and also the vector u can be sampled uniformly from it [13]. Moreover, some cases are noted in [13], where the volume $V_{\mathcal{R}_1, \dots, \mathcal{R}_l}$ can be computed exactly. In these cases, when vector u is sampled from Ball_l , the distribution of vector $(u_1^2 + u_2^2, u_3^2 + u_4^2, \dots, u_{l-1}^2 + u_l^2)$ can be given by a Dirichlet distribution with the parameters of $\frac{l}{2} + 1$ ones, which are simply a uniform distribution over the set of all vectors whose coordinates are non-negative and summed to at most 1 (see page 593 of [15]).

Accordingly, in this particular case, some conditions should be assumed, such as $\mathcal{R}_1 = \mathcal{R}_2, \mathcal{R}_3 = \mathcal{R}_4, \dots, \mathcal{R}_{d-1} = \mathcal{R}_d$, where $0 \leq \mathcal{R}_1 \leq \mathcal{R}_3 \leq \dots \leq \mathcal{R}_{d-1}$ and even number of block sizes $d = \beta = 2\ell$ [13]. In Appendix A of [1], it is shown that for any vector $(t_1, \dots, t_\ell) \in \mathbb{R}_{\geq 0}^\ell$, the related polytope is denoted by $\mathcal{P}_\ell(t_1, \dots, t_\ell)$, which is defined as [1]: $\mathcal{P}_\ell(t_1, \dots, t_\ell) = \{(x_1, \dots, x_\ell) \in \mathbb{R}^\ell \mid \forall i \in \{1, \dots, \ell\}, x_i \geq 0 \text{ and } \sum_{j=1}^i x_j \leq t_i\}$. The volume of $\mathcal{P}_\ell(t_1, \dots, t_\ell)$ is computed as follows:

$$\begin{aligned} \text{Vol} \mathcal{P}_\ell(t_1, \dots, t_\ell) &= \\ \int_{x_1=0}^{t_1} \int_{x_2=0}^{t_2-x_1} \dots \int_{x_\ell=0}^{t_\ell-\sum_{i=1}^{\ell-1} x_i} dx_\ell \dots dx_2 dx_1 &\xrightarrow{y_i=\sum_{j=1}^i x_j} \\ \text{Vol} \mathcal{P}_\ell(t_1, \dots, t_\ell) &= \\ \int_{y_1=0}^{t_1} \int_{y_2=y_1}^{t_2} \dots \int_{y_\ell=y_{\ell-1}}^{t_\ell} dy_\ell \dots dy_2 dy_1. \end{aligned} \quad (21)$$

The integral of (21) can be computed numerically as discussed in [1]. For a polytope $\mathcal{P}_\ell(\mathcal{R}_1^2, \mathcal{R}_2^2, \mathcal{R}_3^2, \dots, \mathcal{R}_d^2)$, the coefficient vector $u = (u_1, u_2, \dots, u_\beta)$, which corresponding to the block $\mathcal{L}_{[j,k]}$, can be found in GNR-enumeration by the following probability [1]:

$$\Pr_{u \sim \text{Ball}_d}(\forall j \in [1, d], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2) = \frac{\text{Vol}\mathcal{P}_\ell(\mathcal{R}_2^2, \mathcal{R}_4^2, \dots, \mathcal{R}_d^2)}{\text{Vol}\mathcal{P}_\ell(1, 1, \dots, 1)}. \quad (22)$$

In practice, assuming such this case for bounding function \mathcal{R} does not corrupt the generality of discussion, and just introduces some partial approximations. For bounding function \mathcal{R} which does not satisfy these constraints (i.e., $\mathcal{R}_1 = \mathcal{R}_2, \dots, \mathcal{R}_{d-1} = \mathcal{R}_d$ where $0 \leq \mathcal{R}_1 \leq \mathcal{R}_3 \leq \dots \leq \mathcal{R}_{d-1}$ and $\beta = 2\ell$), the probability of $\Pr_{u \sim \text{Ball}_d}$ can be approximated as follows [1]:

$$\Pr_{u \sim \text{Ball}_d}(\forall j \in [1, d], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2) \approx \left[\frac{d}{2} \right]! \int_{y_1=0}^{\mathcal{R}_2^2} \int_{y_2=y_1}^{\mathcal{R}_4^2} \dots \int_{y_{\lfloor \frac{d}{2} \rfloor} = y_{\lfloor \frac{d}{2} \rfloor - 1}}^{\mathcal{R}_d^2} dy_{\lfloor \frac{d}{2} \rfloor} \dots dy_2 dy_1. \quad (23)$$

Also, to have a better estimation, a partial modification of this approximation is defined as follows:

$$\begin{aligned} \Pr_{u \sim \text{Ball}_d}(\forall j \in [1, d], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2) &\approx \ell! \times \frac{\text{Vol}\mathcal{P}_\ell(\mathcal{R}_2^2, \dots, \mathcal{R}_{2\ell}^2) + \text{Vol}\mathcal{P}_\ell(\mathcal{R}_1^2, \dots, \mathcal{R}_{2\ell-1}^2)}{2} \\ &\approx \ell! \times \frac{\sum_{i=0}^1 \int_{y_1=0}^{\mathcal{R}_{2-i}^2} \dots \int_{y_{\ell/2} = y_{\ell/2-1}}^{\mathcal{R}_{2\ell-i}^2} dy_{\ell/2} \dots dy_1}{2}. \end{aligned} \quad (24)$$

D. Complementary Concepts

The definition of static success probability is the same as the original definition of success probability when enumeration radius R is set to λ_1 as follows [12]:

Static success probability of bounding function: For any lattice block of $\mathcal{L}_{[j,k]} = [b_j, b_{j+1}, \dots, b_k]$, initial enumeration radius $R = \lambda_1$ and bounding function \mathcal{R} , the static success probability of $p_{\text{succ}}(\mathcal{R})$ is defined as the probability of finding solution vector v (with length of λ_1) after GNR pruning by bounding function \mathcal{R} in enumeration tree.

The first version of static success probability is formulated exactly similar to (15) as follows [12]:

$$p_{\text{succ}}^{\text{new0}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) = p_{\text{succ}}(\mathcal{R}) = \Pr_{u \sim \text{Ball}_d}(\forall j \in [1, d], \sum_{i=1}^j u_i^2 \leq \mathcal{R}_j^2). \quad (25)$$

Note: Following expressions are equivalent in this paper: “Success probability”, “Static success probability”, “Success probability of GNR pruned enumeration”, “Success probability of bounding function”.

In other side, by using Rogers’ theorem, dynamic success frequency can be defined as follows [12]:

Dynamic success frequency of bounding function. For any lattice block of $\mathcal{L}_{[j,k]} = [b_j, b_{j+1}, \dots, b_k]$, initial enumeration radius $R = r_{\text{FAC}} \times \text{GH}(\mathcal{L})$ and bounding function \mathcal{R} with static success probability $p_{\text{succ}}(\mathcal{R})$, there are the number of $r_{\text{FAC}}^\beta/2$ solution vectors in n -dimensional ball with radius of R , consequently the frequency of solution vectors v in enumeration tree (where $\|v\| \leq R$) after GNR pruning by \mathcal{R} is estimated by $p_{\text{succ}}(\mathcal{R}) \times \frac{r_{\text{FAC}}^\beta}{2}$;

The dynamic success frequency is formulated as follows [12]:

$$\begin{aligned} f_{\text{succ}}^{\text{new0}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) &= \\ C_{\text{Rogers}} \times \frac{r_{\text{FAC}}^\beta}{2} \times p_{\text{succ}}^{\text{new0}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R). \end{aligned} \quad (26)$$

Note: As suggested in [12], this paper sets C_{Rogers} to 1.

Note: If this is assumed that there is no updating radius in GNR-enumeration, then the dynamic success frequency of bounding function can be assumed as the expected number of solutions visited in enumeration tree, else this dynamic success frequency is more than the expected number of solutions visited in GNR-enumeration.

As discussed in [12], there are different asymptotical/experimental results which verify the convergence of the expected value of the best vectors of lattices with sufficiently big block sizes to $\text{GH}(\mathcal{L}_{[j,k]})$. Based on experimental tests by Chen and Nguyen [1] to compare the final solution norm of enumeration with value of $\text{GH}(\mathcal{L}_{[j,k]})$, depending on the starting index j of a local block for one round of BKZ, following cases are observed:

- For the first lattice blocks in rounds of BKZ, the final solution norm is significantly lower than $\text{GH}(\mathcal{L}_{[j,k]})$. The behaviour of solution norm in running of BKZ is named “head concavity phenomenon” in BKZ, which is discussed in [2].
- For the last lattice blocks in rounds of BKZ (tail of GSO norms), the GSO norms are significantly larger than $\text{GH}(\mathcal{L}_{[j,k]})$. This behaviour of solution norm is named as “tail convexity” in [12].
- For the middle lattice blocks in rounds of BKZ which includes the most of the enumeration calls, the solution norms are mostly bounded as follows [1]:

$$0.95 \text{ GH}(\mathcal{L}_{[j,k]}) \leq \|v\| \leq 1.05 \text{ GH}(\mathcal{L}_{[j,k]}). \quad (27)$$

This third behaviour of BKZ, can be named as “random manner of middle lattice blocks”.

To the best of our knowledge, this test in [1] is performed with some block sizes of $\beta \leq 70$. There are other experimental/asymptotical results on the expected norm of final solution vector which briefly are counted in Section 2.7 from [12].

In fact, the probability distribution of best solution norm for a lattice basis/block is stated in Chen's thesis [16] as following theorem [2]:

Theorem 1. For random lattice \mathcal{L}_1 with rank n and unit volume, the distribution of $V_n(1) \cdot \lambda_1(\mathcal{L}_1)^n$ converges to distribution of $\text{Expo}(1/2)$ as $n \rightarrow \infty$.

The random variable of $\lambda_1(\mathcal{L})$ for lattices with rank d can be sampled by following relation [2]:

$$\lambda_1(\mathcal{L}) \leftarrow \left(\frac{X \text{Vol}(\mathcal{L})}{V_d(1)} \right)^{1/d}, \text{ where } X \leftarrow \text{Expo}\left(\frac{1}{2}\right). \quad (28)$$

Note: Theorem 1 can be considered only for full-enumeration or a GNR-enumeration pruned by a bounding function with success probability $\approx 100\%$, not for any GNR pruned enumeration.

There is a brief, but sufficient survey of the norm of full/pruned enumerations in Section 2.7 from [12].

At this point, some necessary concepts from [12] which are needed in our analysis are counted as follows:

- **Cutting point.** The enumeration cut point index is defined as the last GSO norm index Cut where $\|b_{\text{Cut}}^*\|^2 \leq R^2 \mathcal{R}_{d-\text{Cut}+1}^2$ and $2 \leq \text{Cut} \leq d$.
- **Last non-zero index of \mathcal{g} .** The projected vector $b_{\mathcal{g}}^* \in \{b_1^*, \dots, b_d^*\}$ which is eliminated after inserting the enumeration solution v , has the GSO norm of $\|b_{\mathcal{g}}^*\| \leq \|v\|$; The coefficient $w_{\mathcal{g}}$ is always the last non-zero coefficient in vector of w for lattice block of $\mathcal{L}_{[1,d]}$, as follows (see Theorem 2 in [12]):

$$w_{\mathcal{g}} = y_{\mathcal{g}} = 1. \quad (29)$$

- For a GNR-enumeration with radius $R = r_{\text{FAC}} \times \text{GH}(\mathcal{L}_{[1,d]})$ over lattice block of $\mathcal{L}_{[1,d]}$ with quality q , sufficiently big block size d and cut point index Cut, the probability distribution of \mathcal{g} for the solution vectors v returned by this enumeration, can be estimated by our non-exact approximate formula of (27) in [12] or can be estimated by our exact formula of (44) in Lemma 8 from [12];
- The norm of solution vector v returned by a pruned enumeration with radius factor of r_{FAC} and success probability $p_{\text{succ}}(\mathcal{R}) = \frac{2}{r_{\text{FAC}}^{\text{Cut}}}$ over lattice block \mathcal{L}_{β} can be sampled by (30) (see Lemma 2 from [12]):

$$\|v\| = \sqrt[{\text{Cut}}]{1 + \text{rand}_{[0 \dots 1]}(r_{\text{FAC}}^{\text{Cut}} - 1) \times \text{GH}(\mathcal{L}_{\text{Cut}})}, \quad \text{where } r_{\text{FAC}} = R/\text{GH}(\mathcal{L}_{\text{Cut}}). \quad (30)$$

- If the norm of shortest vector in lattice block \mathcal{L}_{β} is less than enumeration radius R , then the norm of solution vector v which is returned by a GNR pruned enumeration with radius factor of r_{FAC} and static success probability P over lattice block \mathcal{L}_{β} , can be sampled by (31):

$$\|v\| =$$

$$\begin{cases} X^{1/\text{Cut}} \text{GH}(\mathcal{L}_{\text{Cut}}), & \text{where } X \leftarrow \text{Expo}\left(\frac{1}{2}\right), \quad \text{if } P \approx 1 \\ \sqrt[{\text{Cut}}]{1 + \text{rand}_{[0 \dots 1]} \left(\frac{2}{P} - 1 \right) \times \text{GH}(\mathcal{L}_{\text{Cut}})}, & \text{if } \frac{2}{r_{\text{FAC}}^{\text{Cut}}} \leq P < 1 \\ \sqrt[{\text{Cut}}]{1 + \text{rand}_{[0 \dots 1]}(r_{\text{FAC}}^{\text{Cut}} - 1) \times \text{GH}(\mathcal{L}_{\text{Cut}})}, & \text{if } P < \frac{2}{r_{\text{FAC}}^{\text{Cut}}} \text{ \& } \text{rand}_{[0 \dots \frac{2}{r_{\text{FAC}}^{\text{Cut}}}] \leq P} \\ \text{Un_Successfull}, & \text{if } P < \frac{2}{r_{\text{FAC}}^{\text{Cut}}} \text{ \& } \text{rand}_{[0 \dots \frac{2}{r_{\text{FAC}}^{\text{Cut}}}] > P} \end{cases}$$

$$\text{where } r_{\text{FAC}} = R/\text{GH}(\mathcal{L}_{\text{Cut}}). \quad (31)$$

Remark 1. For an input lattice block $\mathcal{L}_{[1,d]}$ and enumeration radius R , by using the concept of cutting point "Cut", the formula of (36) in Lemma 2 from [12] and the formula of (37) in Lemma 3 from [12], are revised to formula of (30) and (31) by setting \mathcal{L}_{Cut} with dimension of Cut and GSO basis of $B_{[1,\text{Cut}]}^* = [\|b_1^*\|, \dots, \|b_{\text{Cut}}^*\|]$ instead of \mathcal{L}_{β} with dimension of β and GSO basis of $B_{[1,\beta]}^* = [\|b_1^*\|, \dots, \|b_{\beta}^*\|]$.

Our Contributions

The estimations of GNR-enumeration cost (by relation (19)) and the success probability of GNR-bounding function (by relation (15)) are defined in [1] under Heuristic 2. Unfortunately, paper [1] only considers one type of pruning in these estimations which is defined by condition of (20). In fact, the condition of (20) is used to determine the possibility and probability of laying a partial solution candidate in the corresponding cylinder-intersection by bounding function of \mathcal{R} . Here, three more types of pruning are introduced which are ignored in former estimations of success probability and enumeration cost. These pruning types include following cases:

- **Pruning by concept of full-enumeration success probability.** This type of pruning is discussed and analysed in third section (Part A); Also, we propose the concept of optimal enumeration radius to eliminate this type of pruning while the cost of enumeration is held minimized;
- **Pruning by ignoring the enumeration tree levels of " $l = 1$ to $d - \text{Cut}$ ".** This type of pruning is observed if $\text{Cut} < d$; We discuss massively on this concept in [12]; In third section (Part D), we propose a simple technique to eliminate this type of pruning by introducing a mapping technique which can be included in generating GNR bounding function to force $\text{Cut} = d$;
- **Pruning by finding a final solution in GNR enumeration tree levels of $l = d - \text{Cut} + 2$ to d .** In fact, this item is not a real pruning, but since it prevents from opening the child nodes of an enumeration tree node which includes some final solutions, this is considered as pruning; Moreover, it is impossible to eliminate this type (of pruning) at all, since this is an intrinsic phase in enumeration

function, unless we force the enumeration function to abort the function after finding the first final solution vector, such as the pseudo-code of Algorithm 2 in [13]; In fact, if dynamic success frequency would be small (e.g., $f_0 \approx O(1)$), then aborting enumeration function after first finding of final solution is reasonable, but for big value of dynamic success frequency f_0 , this is expected that enumeration function updates radius after each success in finding solution and then continues to traverse the remain of the enumeration tree (similar to the pseudo-code of Algorithm 9 in [1]).

By introducing these three types of pruning plus the cylinder-intersection pruning by (20), the estimations of success probability and enumeration cost are revised respectively in third section (Part B) and third section (Part C).

A. Definition of Optimal Enumeration Radius

By using the definition of Hermite's constant in second section (Part A), in worst case of the full-enumeration, the optimal enumeration radius can be assumed as $R = \sqrt{\gamma_n} \text{vol}(\mathcal{L})^{1/n}$ [13], while in this section, first definition of optimal enumeration radius is introduced in average-case. The enumeration radius R in [1] is defined as follows (by some partial modification):

$$R = \begin{cases} \min(\sqrt{\gamma} \text{GH}(\mathcal{L}_{[j,k]}), \|b_j^*\|), & \text{if } k - j + 1 \geq 30 \\ \|b_j^*\|, & \text{otherwise} \end{cases}, \quad (32)$$

where $\sqrt{\gamma}$ is the initial radius parameter. For block sizes of $\beta = k - j + 1 \geq 30$, value of r_{FAC} is defined as follows (by using relation (16) and (32)):

$$r_{\text{FAC}} = \frac{\min(\sqrt{\gamma} \text{GH}(\mathcal{L}_{[j,k]}), \|b_j^*\|)}{\text{GH}(\mathcal{L}_{[j,k]})}. \quad (33)$$

The main problem in choosing enumeration radius is to find the smallest radius which is not smaller than the shortest vector in the input lattice block. For this end, Chen and Nguyen claim that, the radius parameter of γ in practice can be selected as $\sqrt{\gamma} = \sqrt{1.1} \approx 1.05$ (see [1]), but to the best of our knowledge, this value is estimated only by some experimental tests over BKZ with block size $\beta < 70$ (see Fig. 3 in [1]). By using Theorem 1, the optimal enumeration radius can be defined by the concept of full-enumeration success probability. A full-enumeration with initial radius R intrinsically prunes enumeration by using enumeration radius (i.e., the use of an enumeration radius is concretely a type of pruning). Following lemma formally defines the success probability of full-enumeration:

Lemma 1. For given lattice block \mathcal{L}_β with block size of β , the success probability of a full-enumeration with initial radius R can be defined by (34):

$$p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = 1 - e^{-\frac{r_{\text{FAC}}^\beta}{2}}. \quad (34)$$

Proof. By using (9), (15), (16) and (28), this success probability of $p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta)$ for lattice block \mathcal{L}_β is estimated as follows (for $X \leftarrow \text{Expo}(1/2)$):

$$p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = \text{prob}(\lambda_1(\mathcal{L}_\beta) < R) = \text{prob}(X < r_{\text{FAC}}^\beta) = 1 - e^{-\frac{r_{\text{FAC}}^\beta}{2}}.$$

Since the success probability of full-enumeration is not noted in former studies, these studies (former studies) always assumed implicitly to use $p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = 1$. For a typical lattice block \mathcal{L}_β , the ideal enumeration radius would be $R = \lambda_1(\mathcal{L}_\beta)$ which defines the radius factor of r_{FAC} by using the tight bound (upper-bound and lower-bound) of $r_{\text{FAC}} = \frac{\lambda_1(\mathcal{L}_\beta)}{\text{GH}(\mathcal{L}_\beta)}$. As mentioned, former estimation of enumeration radius in (27) uses experimental tests to estimate the bound of r_{FAC} in average-case (see Fig. 3 in [1]). Theorem 2 introduces an exact definition of this bound.

Theorem 2. For given number X from random lattice blocks, the effective upper-bound/lower-bound of $r_{\text{FAC}} = \frac{\lambda_1(\mathcal{L}_\beta)}{\text{GH}(\mathcal{L}_\beta)}$ can be estimated in average-case as follows:

$$r_{\text{FACmin}} \leq r_{\text{FAC}} \leq r_{\text{FACopt}}, \quad (35)$$

where

$$r_{\text{FACopt}} = \sqrt[\beta]{-2 \ln(1 - p_{\text{opt}})} \text{ and}$$

$$r_{\text{FACmin}} = \sqrt[\beta]{-2 \ln(1 - p_{\text{min}})} \text{ and}$$

$$p_{\text{min}} = 1/X \text{ and } p_{\text{opt}} = 1 - \varepsilon.$$

Proof. The lower-bound and upper-bound for $r_{\text{FAC}} = \frac{\lambda_1(\mathcal{L}_\beta)}{\text{GH}(\mathcal{L}_\beta)}$ are formally defined based on relation (34), as follows:

Minimum hopeful radius parameter (r_{FACmin}). For given number X of random lattice blocks, the minimum radius parameter leads to success probability of $p_{\text{min}} = p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = \frac{1}{X}$ for full-enumeration over these number of X blocks where $R = r_{\text{FACmin}} \times \text{GH}(\mathcal{L}_\beta)$ (i.e., only one of the full-enumerations over these X blocks probably returns the best solution).

Optimal radius parameter (r_{FACopt}). For given number of X random lattice blocks, the minimum radius parameter leads to success probability of $p_{\text{opt}} = p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = 1 - \varepsilon$ for full enumeration over these X blocks where $R = r_{\text{FACopt}} \times \text{GH}(\mathcal{L}_\beta)$ (i.e., all of full-enumerations over these X blocks return the best solution).

$$r_{\text{FACmin}} \leq r_{\text{FAC}} = \frac{\lambda_1(\mathcal{L}_\beta)}{\text{GH}(\mathcal{L}_\beta)} \leq r_{\text{FACopt}}.$$

By expanding the definitions of r_{FACopt} and r_{FACmin} by relation (34) in Lemma 1:

$$\sqrt{\beta \ln(1 - p_{\min})} \leq r_{\text{FAC}} = \frac{\lambda_1(\mathcal{L}_\beta)}{\text{GH}(\mathcal{L}_\beta)} \leq \sqrt{\beta \ln(1 - p_{\text{opt}})}.$$

Note: The optimal radius parameter $r_{\text{FAC}_{\text{opt}}}$ corresponds with optimal enumeration radius as $R_{\text{opt}} = r_{\text{FAC}_{\text{opt}}} \times \text{GH}(\mathcal{L}_\beta)$.

Remark 2. The random manner of lattice blocks $\mathcal{L}_{[j,k]}$ in BKZ algorithm is observed only for $\text{Hdown} \leq j \leq \text{Tup}$ where “Hdown” represents the maximum index in head concavity and “Tup” represents the minimum index in tail convexity; So for each round of BKZ algorithm (or BKZ-simulation), the number of X random lattice blocks can be assumed as $X = \text{Tup} - \text{Hdown} + 1$;

Our estimation results by formula of (35) for block sizes of $50 \leq \beta \leq 240$ are shown in fourth section (Part A). In actual running of BKZ, simulation of BKZ, also our reasoning and proofs, the value of r_{FAC} is assumed as a variable between 1 to \sqrt{Y} , therefore the success probability of full-enumeration would be mostly $p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) \geq 39\%$ (see our estimations by formula of (35) for block sizes of $50 \leq \beta \leq 240$ in Table 1 and Table 2 from fourth section (Part A)). Also, to have better sense about ignoring full-enumeration success probability in former studies, note to following example:

Chen and Nguyen use the enumeration radius of $R = \text{GH}(\mathcal{L}_{[j,k]})$ in estimation of upper-bound for extreme pruned enumeration cost in Table 5 at [1], while by using our reasoning in this section, all these extreme numerations fail to find best solution with probability $\approx 61\%$, which can be penalized by increasing these estimated costs with factor of at most $\frac{100}{39} \approx 2^{1.36}$. At result, when the value of r_{FAC} is variable between 1 to \sqrt{Y} , the effect of full-enumeration success probability can be ignored in asymptotical analysis of cost estimation.

B. Revised Estimation of Enumeration Success Prob.

This is worthy of mentioning that the GSO partial solution candidates in level l from GNR pruned-enumeration tree are only limited to those enumeration tree nodes which satisfy “bounding condition” at level l , which is defined in (20) and the probability of this condition is referred in this paper as $\text{Pr}_{u \sim \text{Ball}_l}$ (also see this condition in line 10 from Algorithm 2 in [13] or line 16 from Algorithm 9 in [1]). Moreover, the final solutions are GSO partial candidates in level $l = d$, and the probability of this condition is referred generally as success probability p_{succ} . In fact, this section tries to revise the probability of this condition as $\text{Pr}_{u \sim \text{Ball}_l}$ (or p_{succ}) to be more exact. By assuming Heuristic 2, this section introduces an exact estimation of success probability in following lemma:

Lemma 2. Under Gaussian Heuristic and Heuristic 2, for an input lattice block of $\mathcal{L}_{[1,d]} = [b_1, b_2, \dots, b_d]$ with

shortest vector of v with norm of $\|v\| = \lambda_1(\mathcal{L}_{[1,d]})$, the success probability of finding this solution vector v by GNR-enumeration with enumeration radius $R \geq \|v\|$ can be estimated by (36):

$$p_{\text{succ}}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) = \sum_{j=2}^{\text{Cut}} \left[\text{Prob}(\mathcal{G} = j) \times p_{\text{succ}}(1^j, R, \mathcal{L}_j) \times \text{Pr}_{u \sim \text{Ball}_{j-1}} \left(\forall t \in [d - j + 2, d], \sum_{i=d-j+2}^t u_i^2 \leq \frac{R^2 \mathcal{R}_t^2 - \|b_j^*\|^2}{R^2 - \|b_j^*\|^2} \right) \right]. \quad (36)$$

Proof. Under assumption of Heuristic 2, the success probability can be estimated by the idea proposed in relation of (15). Also, since $w_{\mathcal{G}} = 1$ (by using Theorem 2 in [12]), for $\mathcal{G} = 1$, this is only needed to determine whether the first vector of block $\mathcal{L}_{[1,d]}$ as b_1^* has the norm of $w_{\mathcal{G}} \|b_{\mathcal{G}}^*\| = \|b_1^*\| \leq R$ or not? The probability of this case as $v = b_1^*$, with respect to all other linear combinations of v by using vectors of $\{b_1^*, b_2^*, \dots, b_{\text{Cut}}^*\}$ is zero, so $\mathcal{G} = 1$ is ignored in (36). By using our definition of “Cutting Point”, the probability of visiting GSO partial solution candidates in level l from GNR pruned-enumeration tree for given bounding function \mathcal{R} and lattice block $\mathcal{L}_{[1,d]}$ with cut point of Cut, can be estimated as follows:

$$\begin{aligned} & \text{Pr}_{u \sim \text{Ball}_l}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, \mathcal{G} = \text{Cut}) \approx \\ & \text{Pr}_{u \sim \text{Ball}_l} \left(\forall t \in [d - \text{Cut} + 2, l], \frac{w_{\text{Cut}}^2 \|b_{\text{Cut}}^*\|^2}{R^2} + \sum_{i=d-\text{Cut}+2}^t u_i^2 \leq \mathcal{R}_t^2 \right), \end{aligned} \quad (37)$$

where $d - \text{Cut} + 2 \leq l \leq d$.

The relation (37) assumes that last non-zero index for all partial (and final) solutions is $\mathcal{G} = \text{Cut}$. Let’s try to estimate the probability of finding the partial solution vectors which are limited to the ones with any possible last non-zero index of $\mathcal{G} = j \leq \text{Cut}$. For this end, $\text{Pr}_{u \sim \text{Ball}_l}^{\text{new1}}$ in (37) can be modified into (38):

$$\begin{aligned} & \text{Pr}_{u \sim \text{Ball}_l}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, \mathcal{G} = j \leq \text{Cut}) \approx \\ & \text{Pr}_{u \sim \text{Ball}_l} \left(\forall t \in [d - \mathcal{G} + 2, l], \frac{w_{\mathcal{G}}^2 \|b_{\mathcal{G}}^*\|^2}{R^2} + \sum_{i=d-\mathcal{G}+2}^t u_i^2 \leq \mathcal{R}_t^2 \right), \end{aligned} \quad (38)$$

Remark 3. By our definition of cutting point of Cut and last non-zero index of $\mathcal{G} \leq \text{Cut}$ (see Section 3.2.1 and Section 3.2.2 from [12]), this is clear that the condition of “ $\frac{w_{\mathcal{G}}^2 \|b_{\mathcal{G}}^*\|^2}{R^2} = \frac{\|b_{\mathcal{G}}^*\|^2}{R^2} \leq \mathcal{R}_{d-\mathcal{G}+1}^2$ ” is always expected to be “True”, therefore the probability of visiting GSO partial solution candidates in level $l = d - \mathcal{G} + 1$ can be defined as follows:

$$\begin{aligned} & \text{Pr}_{u \sim \text{Ball}_{l=d-\mathcal{G}+1}}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, \mathcal{G}) = \\ & \text{Pr}_{u \sim \text{Ball}_{l=d-\mathcal{G}+1}} \left(\frac{w_{\mathcal{G}}^2 \|b_{\mathcal{G}}^*\|^2}{R^2} \leq \mathcal{R}_{d-\mathcal{G}+1}^2 \right) = 1. \end{aligned} \quad (39)$$

By using $w_g = 1$ (which is in Theorem 2 from [12]):

$$\Pr_{u \sim \text{Ball}_l}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, l, g = j) \approx$$

$$\Pr_{u \sim \text{Ball}_l} \left(\forall t \in [d - g + 2, l], \sum_{i=d-g+2}^t u_i^2 \leq \mathcal{R}_t^2 - \frac{\|b_g^*\|^2}{R^2} \right),$$

where $d - g + 2 \leq l \leq d$. (40)

At this point, we use the definition of last non-zero index of g (in Section 3.2.1 from [12]) in sampling of random vector u from Ball_l with radius of unit-length. By only focusing on the GSO partial solution candidates in level l with any possible last non-zero index of $g = j$, GNR-enumeration opens the child nodes of these partial solutions (unless, at last level $l = d$ which returns these final solutions, and comes back to previous level of enumeration tree to find the other solutions). The direction of visiting nodes in a GNR-enumeration tree is from the last index of GSO block to the first one. Accordingly, by using Lemma A.1 in [1] and considering this fact that the effective radius of surrounding unit ball of dimension $\mathcal{D} = d$ is reduced into a ball of dimension $\mathcal{D} = l - d + g - 1$ with radius of $1 - \frac{\|b_g^*\|^2}{R^2}$, the estimation of $\Pr_{u \sim \text{Ball}_l}^{\text{new1}}$ in (40) can be revised into $\Pr_{u \sim \text{Ball}_l}^{\text{new2}}$ as follows:

$$\Pr_{u \sim \text{Ball}_l}^{\text{new2}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = j) \approx \frac{\text{VolP}_\ell \left(\mathcal{R}_{d-g+2}^2 - \frac{\|b_g^*\|^2}{R^2}, \dots, \mathcal{R}_l^2 - \frac{\|b_g^*\|^2}{R^2} \right)}{\text{VolP}_\ell \left(1 - \frac{\|b_g^*\|^2}{R^2}, \dots, 1 - \frac{\|b_g^*\|^2}{R^2} \right)} \approx$$

$$\frac{\text{VolP}_\ell \left(\frac{R^2 \mathcal{R}_{d-g+2}^2 - \|b_g^*\|^2}{R^2 - \|b_g^*\|^2}, \dots, \frac{R^2 \mathcal{R}_l^2 - \|b_g^*\|^2}{R^2 - \|b_g^*\|^2} \right)}{\text{VolP}_\ell(1, 1, \dots, 1)} \approx$$

$$\Pr_{u \sim \text{Ball}_\mathcal{D}} \left(\forall t \in [d - g + 2, l], \sum_{i=d-g+2}^t u_i^2 \leq \frac{R^2 \mathcal{R}_t^2 - \|b_g^*\|^2}{R^2 - \|b_g^*\|^2} \right) \approx$$

(41)

$$\left[\frac{\mathcal{D}}{2} \right]! \times \text{VolP}_\ell(T_1, \dots, T_{\lfloor \mathcal{D}/2 \rfloor}) \approx$$

$$\left[\frac{\mathcal{D}}{2} \right]! \times \int_{y_1=0}^{T_1} \dots \int_{y_{\lfloor \mathcal{D}/2 \rfloor} = y_{\lfloor \mathcal{D}/2 \rfloor - 1}}^{T_{\lfloor \mathcal{D}/2 \rfloor}} dy_{\lfloor \mathcal{D}/2 \rfloor} \dots dy_1, \quad (42)$$

$$\text{where } T_i = \frac{R^2 \mathcal{R}_{2[(d-g+2)/2+i]}^2 - \|b_g^*\|^2}{R^2 - \|b_g^*\|^2} \text{ and}$$

$$1 \leq \mathcal{D} = l - d + g - 1 \leq g - 1 \text{ and}$$

$$d - g + 2 \leq l \leq d.$$

Note: All the notations with formats of $\Pr_{u \sim \text{Ball}_l}$, $\Pr_{u \sim \text{Ball}_l}^{\text{new...}}$ and p_{succ} in this paper show the probability value and obviously are upper-bounded by 1.

The pseudo-code of estimator for $\Pr_{u \sim \text{Ball}_l}^{\text{new2}}$ in relations of (41) and (42) as “Our estimator of success probability” is proposed in Algorithm 1:

Algorithm 1: Estimation of probability of $\Pr_{u \sim \text{Ball}_l}^{\text{new2}}$ in relation (41)

Input: Bounding func. \mathcal{R} , enum radius R , GSO norms $\{\|b_1^*\|, \dots\}$ level l , total block size d , last non zero index of g .

```

1:  for( $t = d - g + 2, \dots, l$ )  $\mathcal{R}''_{t-d+g-1} \leftarrow$ 
     $\min\left(\frac{R^2 \mathcal{R}_t^2 - \|b_g^*\|^2}{R^2 - \|b_g^*\|^2}, 1\right); /* \text{see (41)} */$ 
2:   $\mathcal{D} = l - d + g - 1;$ 
3:  for( $k = 0, 1$ )  $\{ // \text{begin for1}$ 
4:     $C \leftarrow 1; // C \in \mathbb{R}[X] \text{ is a polynomial}$ 
5:    for( $j = \mathcal{D}, \mathcal{D} - 2, \dots, 2$ )  $\{ // \text{begin for2}$ 
       $C \leftarrow \int_{t=0}^x C(t) dt; C \leftarrow C(\mathcal{R}''_{j-k}) - C(x); \}$ 
    end for2
6:     $p_k \leftarrow C(0) \times \left[ \frac{\mathcal{D}}{2} \right]!; /* \text{see (42)} */ \}$  end for1

```

Output: $(p_1 + p_2)/2$ as the success probability

By applying the probability of last non-zero index as $\text{Prob}(g = j)$ by using Lemma 8 in [12], and our proposed concept of full-enumeration success probability (see (34) in Lemma 1 at third section (Part A)), our revised estimation of the probability of finding the GSO partial solution candidates in level l , with any possible last non-zero index of $g = j$, can be defined as follows:

$$\Pr_{u \sim \text{Ball}_l}^{\text{new3}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = j) \approx \text{Prob}(g = j) \times p_{\text{succ}}(1^j, R, \mathcal{L}_j) \times \Pr_{u \sim \text{Ball}_l}^{\text{new2}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = j) \approx \quad (43)$$

$$\text{Prob}(g = j) \times p_{\text{succ}}(1^j, R, \mathcal{L}_j) \times \Pr_{u \sim \text{Ball}_\mathcal{D}} \left(\forall t \in [d - j + 2, l], \sum_{i=d-j+2}^t u_i^2 \leq \frac{R^2 \mathcal{R}_t^2 - \|b_j^*\|^2}{R^2 - \|b_j^*\|^2} \right), \quad (44)$$

$$\text{where } 1 \leq \mathcal{D} = l - d + g - 1 \leq g - 1 \text{ and}$$

$$d - g + 2 \leq l \leq d.$$

Now, the expected value of the probability of finding the GSO partial solution candidates in level l (by considering whole indices of $2 \leq g \leq \text{Cut}$) can be estimated as follows:

$$\mathbb{E}[\Pr_{u \sim \text{Ball}_l}^{\text{new3}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, 2 \leq g \leq \text{Cut})] \approx$$

$$\left\{ \sum_{j=d-l+2}^{\text{Cut}} \Pr_{u \sim \text{Ball}_l}^{\text{new3}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = j), \text{ for } d - \text{Cut} + 2 \leq l \leq d \right.$$

$$\left. \Pr_{u \sim \text{Ball}_{l_0=d-\text{Cut}+2}}^{\text{new3}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = \text{Cut}), \text{ for } l = d - \text{Cut} + 1. \right. \quad (45)$$

Note: For level of $l = d - \text{Cut} + 1$ in (45), the probability in this level is equal to the probability of $\Pr_{u \sim \text{Ball}_{l=d-\text{Cut}+2}}^{\text{new3}}$ at level of $l = d - \text{Cut} + 2$;

Note: Since $\mathcal{D} = l - d + j - 1 \geq 1$ in (44), therefore the index of j in (45) starts from $j = d - l + 2$, instead of index of $j = 2$;

Finally by using (45), our revised estimation of success probability of bounding function \mathcal{R} , as the expected value of the probability of finding final solutions in level $l = d$

(by considering whole indices of $2 \leq g \leq \text{Cut}$) can be estimated as follows:

$$p_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) \approx E[\text{Pr}_{u \sim \text{Ball}_{l=d}}^{new3}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, 2 \leq g \leq \text{Cut})].$$

Finally this lemma (Lemma 2) is proved.

Since our estimation of success probability by relation (36) in Lemma 2, is not easy to work and analyze, Remark 4 introduces a suitable formula which approximates the success probability.

Remark 4. Our estimation of the success probability in Lemma 2 can be approximated by (46):

$$p_{succ}^{\text{Approx1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) = (\sum_{j=1}^{\text{Cut}} \text{Prob}(g = j)) \times p_{succ}(1^{\text{Cut}}, R, \mathcal{L}_{\text{Cut}}) \times \text{Pr}_{u \sim \text{Ball}_{l=d}}^{new2}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, g = \text{Cut}). \quad (46)$$

As mentioned in second section (Part D), dynamic success frequency shows the expected number of solutions in enumeration tree (by assumption of no updating radius). By using formula (21) in paper [12], the success probability of p_{succ}^{new1} in (36) can be changed into dynamic success frequency of f_{succ}^{new1} as follows:

$$f_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) \approx C_{Rogier} \frac{r_{\text{FAC}}^{\text{Cut}}}{2} p_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R),$$

$$\text{where } r_{\text{FAC}} = \frac{R}{\text{GH}(\mathcal{L}_{[1,\text{Cut}]})}. \quad (47)$$

Note: As suggested in [12], this paper sets C_{Rogier} to 1.

Accordingly, by using (47), the sampling method for computing the number of solutions (showing by the notation of K) in a typical GNR-enumeration with dynamic success frequency of $f_{succ}^{new1} = f_0$ can be defined as follows:

$$K = \begin{cases} \lfloor f_0 \rfloor, & \text{if } \text{rand}_{[0 \dots 1]} \leq f_0 - \lfloor f_0 \rfloor \\ \lfloor f_0 \rfloor, & \text{otherwise.} \end{cases} \quad (48)$$

C. Revised Estimation of Enumeration Cost

This section proposes following algorithm to estimate the total nodes of GNR-enumeration tree. This algorithm includes the concepts of all four pruning types, along with the process of updating radius. Lemma 3 formally introduces our revised estimation of GNR-enumeration cost by using Algorithm 2. Besides the better estimation of enumeration cost, Algorithm 2 can be used as a sampling method of solution norm too, similar to Lemma 3 in [12].

Note: The array of “Solution” defined in line 9 of Algorithm 2, includes the number of K entries, in the way that each of these entries has two fields: “index” (as the index of that leaf node in enumeration tree) and “norm” (as the norm of final solution in that leaf node). The notation of “Solution[i]#index” and “Solution[i]#norm” in Algorithm 2 respectively represent the index and norm of final solution in leaf node i .

Algorithm 2: Enumeration cost with updating radius (enum_cost_UpdateR)

Input: GSO norms $B_{[1,d]}^* = [\|b_1^*\|, \dots, \|b_d^*\|]$ of $\mathcal{L}_{[1,d]}$,

Bounding function \mathcal{R} , enum radius R , parameter “abort”.

```

1:  Cut = GETCUT( $B_{[1,d]}^*$ ,  $\mathcal{R}$ ,  $R$ );
2:  gh = GH( $\mathcal{L}_{[1,\text{Cut}]}$ );  $r_{\text{FAC}} = \frac{R}{\text{GH}(\mathcal{L}_{[1,\text{Cut}]})}$ ;
3:   $f_0 = f_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$ ; //by formula (47)
4:   $K = \begin{cases} \lfloor f_0 \rfloor, & \text{if } \text{rand}_{[0 \dots 1]} \leq f_0 - \lfloor f_0 \rfloor \\ \lfloor f_0 \rfloor, & \text{otherwise} \end{cases}$ ; //
    by formula (48)
5:   $N_{\text{new1}} = 0$ ;
6:  for( $l = d - \text{Cut} + 1, \dots, d$ ) { //begin for1*/
7:     $H_l^{\text{new}} = E[\text{Pr}_{u \sim \text{Ball}_l}^{new3}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, 2 \leq g \leq \text{Cut})]$   $\times$ 
       $H_l$ ;
      //by using (45) where  $H_l$  is defined in (11).*/
8:     $N_{\text{new1}} += H_l^{\text{new}}$ ; } //end for1*/
9:  Solution := array[1 ...  $K$ ] of Struct {index, norm};
10: for( $t = 1, \dots, K$ ) { //begin for2
11:   loop{ $j \leftarrow \text{randINT}_{[1 \dots N_{\text{new1}}]}$ ; }
12:   until( $\forall 1 \leq i < t$ : Solution[ $i$ ]#index  $\neq j$ );
      //uniform random selection without substitution
13:   Solution[ $t$ ]#index  $\leftarrow j$ ;
14:   Solution[ $t$ ]#norm  $\leftarrow$ 
       $\sqrt[{\text{Cut}}]{1 + \text{rand}_{[0 \dots 1]}(r_{\text{FAC}}^{\text{Cut}} - 1) \times \text{gh}}$ ;
      //see (30) by Remark 1*/ //end for2
15:  Sort(array = "Solution", key = "index");
    /* Sorting of array of "Solution" based on
    "key = index" in an increase order */
16:   $R_{\text{new}} \leftarrow R$ ; lastidx = 0;  $N_{\text{new2}} \leftarrow 0$ ;
17:  for( $t = 1, \dots, K$ ) { //begin for3
18:    if(Solution[ $t$ ]#norm <  $R_{\text{new}}$ ) { //begin if2
19:      for( $l = d - \text{Cut} + 1, \dots, d$ ) { //begin for4
20:         $N_{\text{new2}} += \frac{\text{Solution}[t]\#index - \text{last}_{\text{idx}}}{N_{\text{new1}}} H_l^{\text{new}} \left(\frac{R_{\text{new}}}{R}\right)^l$ ;
          //end for4
21:         $R_{\text{new}} \leftarrow \text{Solution}[t]\#norm$ ; Lastidx =  $t$ ;
22:        if(abort = true) return [ $N_{\text{new2}}$ ,  $R_{\text{new}}$ ];
23:      } //end if2 //end for3
24:    } //begin for5
25:     $N_{\text{new2}} += \frac{N_{\text{new1}} - \text{last}_{\text{idx}}}{N_{\text{new1}}} H_l^{\text{new}} \left(\frac{R_{\text{new}}}{R}\right)^l$ ;
      //for last update of radius up to end*/ //end for5

```

Output: [N_{new2} , R_{new}]. /* N_{new2} is returned
as the total enumeration cost and R_{new} is returned
as the sampled norm of enumeration solution*/

Note: For better speedup in running-time of Algorithm 2, “Insertion Sort” can be used instead of line 12, so that

the repeated indices can be checked and eliminated easily by “Insertion Sort”.

Lemma 3. For an input lattice block $\mathcal{L}_{[1,d]}$, bounding function \mathcal{R} and enumeration radius R , under [Gaussian Heuristic](#) and [Heuristic 2](#), by assuming that each node at the same level of GNR-enumeration tree includes same number of child nodes, [Algorithm 2](#) samples the norm of final solution, also it estimates the total nodes of GNR enumeration after being pruned by four proposed types of pruning and updating the enumeration radius after each success of finding solution.

Proof. To prove this lemma, this is needed to show that the concept of four types of pruning (which are proposed at the beginning of [third section](#)) and also updating radius are considered collectively by [Algorithm 2](#), in estimation of the total nodes of GNR-enumeration (also sampling the norm of final solution of GNR-enumeration);

The function of GET_{CUT} in line 1 from [Algorithm 2](#) returns the cut point of bounding function \mathcal{R} with enumeration radius R , for an input lattice block (i.e., the same operations as lines 11 to 16 from [Algorithm 3](#) in [\[12\]](#)). The line 2 from [Algorithm 2](#) works based on [Remark 1](#). Dynamic success frequency and number of final solutions in GNR-enumeration tree with no updating radius, are respectively defined in lines of 3 and 4 in [Algorithm 2](#).

Lines of 5 to 8 estimate the total nodes of GNR-enumeration tree after four types of pruning (which are proposed at the beginning of [third section](#)) as follows:

$$N_{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) = \sum_{l=d-\text{Cut}+1}^d E[\text{Pr}_{u \sim \text{Ball}_l}^{\text{new3}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R, 2 \leq \varphi \leq \text{Cut})] \times H_l, \quad (49)$$

where H_l is defined in [\(11\)](#).

To complete this proof, updating radius should be considered in revising our estimation of N_{new1} to be more exact. For this end, we should describe the assumption that each node at the same level of GNR-enumeration tree includes same number of child nodes. This is clear that GNR-enumeration is a pre-order tree search. The root of this tree which corresponds with b_{Cut}^* , against the ordinary trees, has two nodes with coefficient of $w_{\text{Cut}} = 0$ or $w_{\text{Cut}} = 1$. This assumption is illustrated with a simple example as follows:

Lets assume that for an input lattice block of $\mathcal{L}_{[1,6]} = \{b_1^, b_2^*, b_3^*, b_4^*, b_5^*, b_6^*\}$ with block size of $d = 6$, this enumeration tree has the depth of 5 (i.e., $\text{Cut} = 5$). Also assume the following number of nodes at each level after four types of pruning (which is computed in line 7 from [Algorithm 2](#)): $H_{l=2}^{\text{new}} = 2$, $H_{l=3}^{\text{new}} = 4$, $H_{l=4}^{\text{new}} = 8$, $H_{l=5}^{\text{new}} = 5$, $H_{l=6}^{\text{new}} = 3$; The total number of nodes in pre-order search (with no update of radius) is $N_{\text{new1}} = \sum_{l=d-\text{Cut}+1}^d H_l^{\text{new}} = 22$. Now by the assumption that each node at the same level of GNR-*

enumeration tree includes same number of child nodes, each nodes in root (corresponding with $l = 2$ and GSO vector of b_5^) has $\frac{H_{l=3}^{\text{new}}}{H_{l=2}^{\text{new}}} = 2$ child nodes, each nodes in level $l = 3$ has $\frac{H_{l=4}^{\text{new}}}{H_{l=3}^{\text{new}}} = 2$ child nodes, each nodes in level $l = 4$ has $\frac{H_{l=5}^{\text{new}}}{H_{l=4}^{\text{new}}} = 0.625$ child nodes, each nodes in level $l = 5$ has $\frac{H_{l=6}^{\text{new}}}{H_{l=5}^{\text{new}}} = 0.6$ child nodes, and the nodes in level $l = 6$ are leaf nodes;*

By this example, we introduce an outline of our main assumption in [Lemma 3](#). In fact, we use this assumption to determine the approximate number of nodes at each level which should be visited between two specific nodes in pre-order search of GNR-enumeration tree. For this case, again the previous example can be used, and it is asked to determine the approximate number of nodes at each level which are visited after 5th node up to 14th node in pre-order search. There are 9 nodes which should be visited after node of $i = 5$ to reach the node of $j = 14$, so by using our main assumption, the number of nodes at each level l , which should be visited between these two specific nodes of i and j , can be estimated by $\frac{j-i}{N_{\text{new1}}} H_l^{\text{new}}$ (e.g., for level $l = 4$, this number of node is ≈ 3.27);

Three states are considered for output of this algorithm:

- If parameter of “abort” would be “true”, then [Algorithm 2](#) (at line 22) returns the total nodes of enumeration as “Solution[1]#index”, and samples the solution norm as “Solution[1]#norm”:

$$N_{\text{new2}} = \sum_{l=d-\text{Cut}+1}^d \frac{\text{Solution}[t=1]\#\text{index}-\text{last_idx}}{N_{\text{new1}}} H_l^{\text{new}} \left(\frac{R_{\text{new}}}{R}\right)^l = \text{Solution}[t=1]\#\text{index},$$

where $R_{\text{new}} = R$ and $\text{last_idx} = 0$.

- Also, if expected number of final solutions would be $K = 0$, then lines of 17 to 23 are not performed, and lines of 24 and 25 are only performed and finally this algorithm returns the total nodes of enumeration as “ N_{new1} ” and samples the solution norm as “ R ”;
- After sampling the solution indices in pre-order search of GNR-enumeration in lines of 10 to 15 from [Algorithm 2](#), by using our proposed idea, the number of enumeration nodes, after finding a solution and before finding next solution (lines 19 to 20 from [Algorithm 2](#)) or finishing the search of enumeration tree (lines 24 to 25 from [Algorithm 2](#)), are estimated.

Also, by using the factor of $\left(\frac{R_{\text{new}}}{R}\right)^l$ in lines 20 and 25 from [Algorithm 2](#), we apply the process of updating radius in estimation of total nodes of enumeration (in the way that it is discussed for our example in this proof). Moreover, this is clear that the final solution norm which is returned by GNR-enumeration is equal

to the last update of enumeration radius (as last setting of R_{new} in Algorithm 2).

D. Revised Generation of Bounding Function

To find better solution vector, it is reasonable to run enumeration function over bigger block sizes (i.e., cutting point of $\text{Cut} < d$ is not pleasant). Also, by forcing $\text{Cut} = d$, it is easier to generate of a bounding function with an intended success probability by relation (36) in Lemma 2. Moreover, by forcing $\text{Cut} = d$, some other functions, relations, propositions and formulations in BKZ-simulation can be simplified too. Following lemma formally defines our technique to force $\text{Cut} = d$:

Lemma 4. The bounding function \mathcal{R} with dimension d with our revised success probability defined by (36) can be generated as follows:

$$\mathcal{R}_{i+1}^2 \leftarrow \left(1 - \frac{\|b_d^*\|^2}{R^2}\right) \mathcal{R}_i'^2 + \frac{\|b_d^*\|^2}{R^2}, \quad (50)$$

where $1 \leq i \leq d-1$ and $\mathcal{R}_1^2 \leftarrow \frac{\|b_d^*\|^2}{R^2}$ and

$p_{\text{succ}}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) \approx p_{\text{succ}}(\mathcal{R}') \times p_{\text{succ}}(1^d, R, \mathcal{L}_d)$ and bounding function \mathcal{R}' with dimension $d-1$ and $p_{\text{succ}}(\mathcal{R}')$ defined by (15).

Proof. Assume that the bounding function of \mathcal{R}' with dimension $d-1$ in this lemma is defined as follows:

$$\mathcal{R}_i'^2 = \frac{R^2 \mathcal{R}_{i+1}^2 - \|b_d^*\|^2}{R^2 - \|b_d^*\|^2}, \quad 1 \leq i \leq d-1.$$

The corresponding success probability of \mathcal{R}' is defined by (15), as follows:

$$p_{\text{succ}}(\mathcal{R}') = \Pr_{u \sim \text{Ball}_{d-1}} \left(\forall i \in [1, d-1], \sum_{l=1}^i u_l^2 \leq \mathcal{R}_i'^2 \right) =$$

$$\Pr_{u \sim \text{Ball}_{d-1}} \left(\forall i \in [1, d-1], \sum_{l=1}^i u_l^2 \leq \frac{R^2 \mathcal{R}_{i+1}^2 - \|b_d^*\|^2}{R^2 - \|b_d^*\|^2} \right) =$$

$$\Pr_{u \sim \text{Ball}_{d-1}} \left(\forall t \in [2, d], \sum_{l=1}^t u_l^2 \leq \frac{R^2 \mathcal{R}_t^2 - \|b_d^*\|^2}{R^2 - \|b_d^*\|^2} \right) \Rightarrow$$

By using (41):

$$p_{\text{succ}}(\mathcal{R}') = \Pr_{u \sim \text{Ball}_{d-1}}^{\text{new2}} (\mathcal{L}_{[1,d]}, \mathcal{R}, R, \mathcal{G} = \text{Cut}) =$$

$$\Pr_{u \sim \text{Ball}_{d-1}} \left(\forall t \in [d - \text{Cut} + 2, l], \sum_{i=d-\text{Cut}+2}^t u_i^2 \leq \frac{R^2 \mathcal{R}_t^2 - \|b_{\text{Cut}}^*\|^2}{R^2 - \|b_{\text{Cut}}^*\|^2} \right).$$

Because of $\mathcal{R}_1^2 = \frac{\|b_d^*\|^2}{R^2}$ (see relation (50)), the cutting point is $\text{Cut} = d$, and summation of $\sum_{j=1}^{\text{Cut}=d} \text{Prob}(\mathcal{G} = j)$ equals to 1. Also, since enumeration radius is not changed, and Gaussian Heuristic of $\mathcal{L}_{[1,d]}$ is close to Gaussian Heuristic of $\mathcal{L}_{[1,d-1]}$ (by relation (5)) as $\text{GH}(\mathcal{L}_{[1,d-1]}) \approx \text{GH}(\mathcal{L}_{[1,d]})$, so the enumeration radius factor r_{FAC} is not changed nearly for $\mathcal{L}_{[1,d-1]}$ and $\mathcal{L}_{[1,d]}$. Accordingly, by using (34) and Remark 4:

$$p_{\text{succ}}(\mathcal{R}') \times p_{\text{succ}}(1^d, R, \mathcal{L}_d) = p_{\text{succ}}^{\text{Approx1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R) \approx p_{\text{succ}}^{\text{new1}}(\mathcal{L}_{[1,d]}, \mathcal{R}, R).$$

This proof is completed.

By using Lemma 4, Remark 5 generates the extreme/non-extreme bounding function with given dynamic success frequency f_0 which is estimated by (47). Remark 5. The bounding function \mathcal{R} with dimension d and given dynamic success frequency f_0 estimated by (47), can be generated by proposed three steps in Fig. 1:

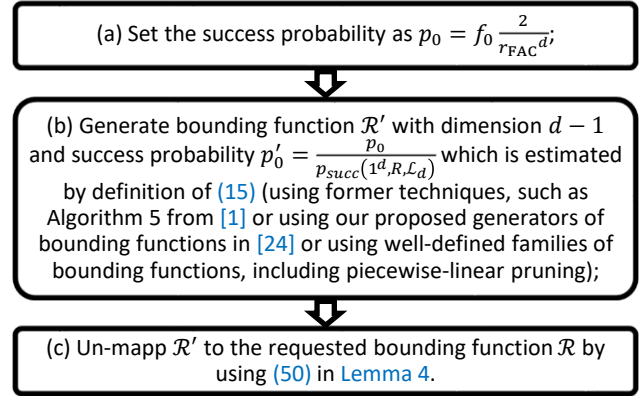


Fig. 1: Steps of Generating bounding function \mathcal{R} with dimension d and success frequency f_0 estimated by (47).

Results and Discussion

In this section, our test results show the simulation/experimental outcomes of our contributions in this paper. The tests in this paper use the random instances of SVP lattice challenges in the sense of Goldstein and Mayer [17], [18]. Also, two libraries of fpLLL library [19] and NTL library [20] are used for these tests. All the implementations and simulations are compiled with MSVC x64 bit C++. These tests use the following hardware platform: ASUS motherboard series Z97-K, Intel® Core™ i7-4790K processor with the base frequency of 4 GHz, 16 GB RAM; also, the running times are provided only for a single real-core.

A. Our Estimations for Parameter of \sqrt{Y}

As mentioned, former studies usually use the initial enumeration radius parameter of $\sqrt{Y} = 1.05$, but Theorem 2 defines the optimal initial radius parameter \sqrt{Y} (as optimal initial radius factor r_{FAC}) and a bound for the norm of shortest vector of lattice blocks in average-case. Our definition in Theorem 2 can be used dynamically to compute optimal enumeration radius in actual running of BKZ-algorithm (or BKZ-simulation). By using relation (34) in Lemma 1, the success probability of full-enumeration for block sizes of $50 \leq \beta \leq 240$ in different values of initial radius parameter \sqrt{Y} (as initial radius factor of r_{FAC}) is shown in Table 1 and Table 2.

Note: The success probability of full-enumeration in former studies is set to 1.

Table 1: Success probability of full-enumeration for $50 \leq \beta \leq 90$ in different values of r_{FAC}

radFac	0.91	0.92	0.93	0.94	0.95	0.96	0.97	0.98	0.99	1	1.01	1.02	1.03	1.04	1.05
$\beta=50$	0.00	0.01	0.01	0.02	0.04	0.06	0.10	0.17	0.26	0.39	0.56	0.74	0.89	0.97	1
$\beta=60$	0.00	0.00	0.01	0.01	0.02	0.04	0.08	0.14	0.24	0.39	0.60	0.81	0.95	0.99	1
$\beta=70$	0.00	0.00	0.00	0.01	0.01	0.03	0.06	0.11	0.22	0.39	0.63	0.86	0.98	1	1
$\beta=80$	0.00	0.00	0.00	0.00	0.01	0.02	0.04	0.09	0.20	0.39	0.67	0.91	1	1	1
$\beta=90$	0.00	0.00	0.00	0.00	0.00	0.01	0.03	0.08	0.18	0.39	0.71	0.95	1	1	1

Table 2: Success probability of full-enumeration for $100 \leq \beta \leq 240$ in different values of r_{FAC}

radFac	0.95	0.96	0.97	0.98	0.99	1	1.002	1.004	1.006	1.008	1.01	1.012	1.014	1.016	1.018	1.02	1.03
$\beta=100$	0.00	0.01	0.02	0.06	0.17	0.39	0.46	0.53	0.60	0.67	0.74	0.81	0.87	0.91	0.95	0.97	1
$\beta=120$	0.00	0.00	0.01	0.04	0.14	0.39	0.47	0.55	0.64	0.73	0.81	0.88	0.93	0.97	0.99	1	1
$\beta=140$	0.00	0.00	0.01	0.03	0.12	0.39	0.48	0.58	0.69	0.78	0.87	0.93	0.97	0.99	1	1	1
$\beta=160$	0.00	0.00	0.00	0.02	0.10	0.39	0.50	0.61	0.73	0.83	0.91	0.97	0.99	1	1	1	1
$\beta=180$	0.00	0.00	0.00	0.01	0.08	0.39	0.51	0.64	0.77	0.88	0.95	0.99	1	1	1	1	1
$\beta=200$	0.00	0.00	0.00	0.01	0.06	0.39	0.53	0.67	0.81	0.91	0.97	1	1	1	1	1	1
$\beta=220$	0.00	0.00	0.00	0.01	0.05	0.39	0.54	0.70	0.85	0.94	0.99	1	1	1	1	1	1
$\beta=240$	0.00	0.00	0.00	0.00	0.04	0.39	0.55	0.73	0.88	0.97	1	1	1	1	1	1	1

Also for block sizes of $50 \leq \beta \leq 240$, our proposed bound of radius factor \sqrt{Y} , which is defined by formula of (35), is shown in Table 3.

Table 3: Our proposed lower-bound/upper-bound for initial radius factor \sqrt{Y} for $50 \leq \beta \leq 240$ with assumption of 100 middle random lattice blocks

Block Size	radFac _{min}	radFac _{opt}
$\beta = 50$	0.925	1.045
$\beta = 60$	0.937	1.038
$\beta = 70$	0.946	1.032
$\beta = 80$	0.952	1.028
$\beta = 90$	0.958	1.025
$\beta = 100$	0.962	1.022
$\beta = 120$	0.968	1.019
$\beta = 140$	0.972	1.016
$\beta = 160$	0.976	1.014
$\beta = 180$	0.979	1.012
$\beta = 200$	0.981	1.011
$\beta = 220$	0.982	1.01
$\beta = 240$	0.984	1.009

By assuming the number of 100 middle random lattice blocks in BKZ running, Table 3 introduces the optimal initial radius parameter of r_{FACopt} for this number of random lattice blocks in middle of BKZ with full-enumeration success probability of $p_{\text{opt}} = p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = 0.99$. Also, Table 3 introduces the minimum radius parameter r_{FACmin} for this number of random lattice blocks in middle of BKZ with full-enumeration success probability of $p_{\text{min}} = p_{\text{succ}}(1^\beta, R, \mathcal{L}_\beta) = 0.01$ (see our discussions in third section (Part A)). By these estimations, the values of r_{FACopt} in Table 3 can be used instead of initial radius parameter of $\sqrt{Y} = 1.05$.

B. Test Results for Our Revision of Success Probability

By definition of cutting point in [12], this is found that if this point (cutting point) would be less than the block size, then in some specific cases (e.g., the demanded success probability is extremely small or the input block is much reduced), former studies (such as [1], [13]) may generate bounding functions with some success probabilities which unintentionally becomes less than intended one! To show the importance of this case, a test is introduced to show that GNR-enumeration with extremely small success probability generated by [1], [13] over strong-reduced bases (nearly HKZ-reduced bases)

can be actually smaller than estimated one by relation (15). This test uses following studies for comparison:

- The estimation of success probability by Chen-Nguyen [1] in (25),
- The estimation of dynamic success frequency by Aono et al. [1], [3] in (26),
- Monte-Carlo estimation of success probability [13] by condition of (20).

This test uses following bounding functions: full-enumeration, some bounding functions with no known families (for success probabilities of 0.25, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95 which are estimated by Monte-Carlo), linear-pruning (with success probability of 0.01) and five piecewise-linear bounding function with parameters of “a = 0.4”, “a = 0.3”, “a = 0.2”, “a = 0.1”, “a = 0.05”. The entries of these bounding function plotted on Fig. 2; The success probability of these bounding functions is defined by (15) which uses Monte-Carlo estimation.

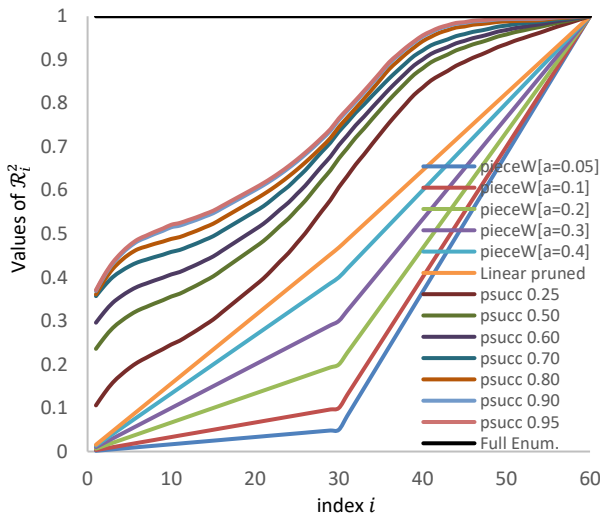


Fig. 2: Some bounding functions with different success probabilities.

The quality of randomization, LLL-reduction, and nearly HKZ-reduction for 20 random lattice bases in the sense of Goldstein and Mayer [17], [18] in dimension of $n = 60$ is illustrated in Fig. 3; In this test, for randomization of lattice blocks, the re-randomization strategy of fplll library [19] is used, which works by permuting basis vectors and the triangular transformation matrix with coefficients of $\{-1, 0, 1\}$, also for LLL reduction the parameter of $\delta = 0.99$ is set, finally for nearly HKZ reduction, this paper uses $BKZ_{\beta=60}$ from NTL library [20].

The quality of these three types of reduction are shown by GSO norms of $\|b_i^*\|^2$ which are plotted in \log_2 form in Fig. 3.

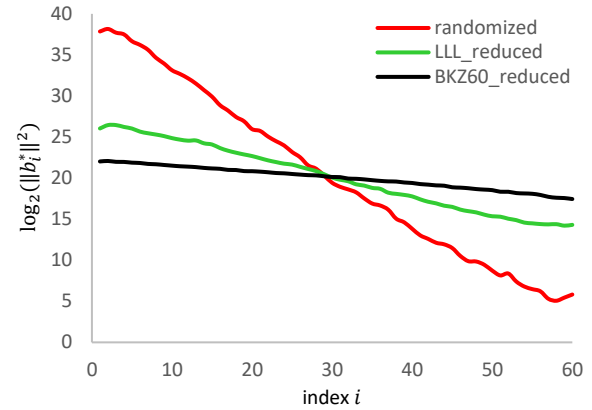


Fig. 3: Quality of randomized/LLL-reduced/nearly HKZ-reduced basis with dimension 60.

For determining the cutting point in Fig. 4, the entries of bounding function (i.e., \mathcal{R}_i^2 for $1 \leq i \leq 60$) are scaled by multiplying with squared value of enumeration radius (i.e., in the form of $R^2 \mathcal{R}_i^2$). The initial radius parameter in this test is set to $\gamma = 1.13$, so the squared value of enumeration radius would be $R^2 = 1.13 \times \text{GH}^2(\mathcal{L}_{[1,60]})$. This is worthy of noting that the indices of entries in bounding function \mathcal{R}_i^2 correspond with the inverse of indices in squared GSO norm of $\|b_i^*\|^2$ (i.e., in Fig. 4, the value of $R^2 \mathcal{R}_{\beta-i+1}$ corresponds with $\|b_i^*\|^2$). The values of squared GSO norm of $\|b_i^*\|^2$ and values of $R^2 \mathcal{R}_{\beta-i+1}$ in Fig. 4 are plotted in form of \log_2 (the parameter of “GH^2” in Fig. 4 represents the squared value of the shortest vector norm estimation in (5) by Gaussian Heuristic).

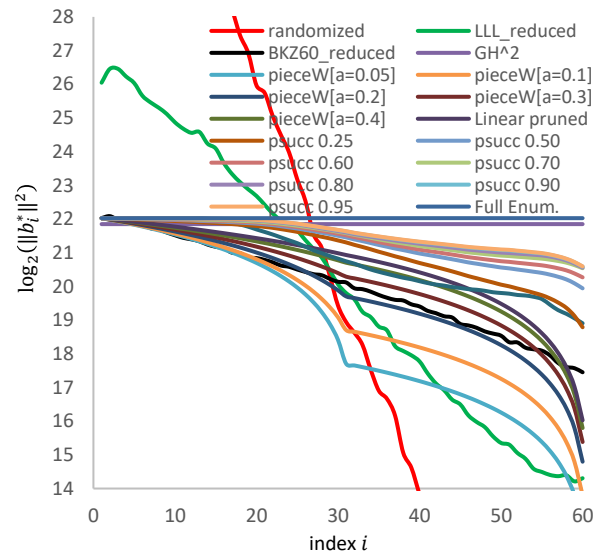


Fig. 4: Different qualities of basis and different bounding functions with scaled entries of $R^2 \times \mathcal{R}_{\beta-i+1}$ in dimension 60 to find the cutting point.

Fig. 4 shows that the BKZ_{60} -reduced bases are nearly cut with linear pruning at index of Cut =58, cut with piecewise-linear with parameter $a = 0.4$ at index of Cut =57, cut with piecewise-linear by parameter of $a = 0.3$ at index of Cut =54, cut with piecewise-linear by parameter of $a = 0.2$ at index of Cut =26, cut with piecewise-linear by parameter of $a = 0.1$ at index of Cut =20, and cut with piecewise-linear by parameter of $a = 0.05$ at index of Cut =17; Other bounding functions in Fig. 4 have cutting point of Cut =60 for GSO norms of three types of reduced basis. In Table 4, our test results show the comparison of our revised estimation of success

probability in (36) and our revised estimation of dynamic success frequency in (47) with some former estimations including: success probability by Chen-Nguyen technique [1] (see relation (25)), dynamic success frequency by Aono et al. [3] (see relation (26)), and static success probability by Monte-Carlo estimator with condition of (20) [13]. By using the initial radius parameter of $\gamma = 1.13$, the count of solutions in full-enumeration tree would be estimated as $\approx \frac{r_{FAC}^\beta}{2} \approx \frac{(\sqrt{\gamma})^\beta}{2} \approx \frac{(\sqrt{1.13})^{60}}{2} \approx 19.6$;

Table 4: Comparison of our revised estimation of success probability and dynamic success frequency with former estimations in [1], [3], [13] over nearly HKZ-reduced bases in dimension 60

	Cut Point	p_{succ} by Monte Carlo [13]	p_{succ} by Chen-Nguyen [1]	f_{succ} by Aono et al. [3]	p_{succ} by our estimator of (36)	f_{succ} by our estimator of (47)
PieceWise[a=0.05]	17	-	$2^{-44.6}$	$2^{-40.3}$	0	0
PieceWise[a=0.1]	20	-	$2^{-30.5}$	$2^{-26.2}$	0	0
PieceWise[a=0.2]	26	$2^{-19.8}$	$2^{-17.4}$	$2^{-13.1}$	0	0
PieceWise[a=0.3]	54	$2^{-12.9}$	$2^{-10.7}$	0.012	$2^{-19.8}$	2^{-16}
PieceWise[a=0.4]	57	0.0024	0.01	0.195	$2^{-11.8}$	0.005
Linear-pruning	58	0.01	0.036	0.7	0.003	0.046
BF[$p_{succ}=0.25$]	60	0.25	0.48	9.4	0.4	7.8
BF[$p_{succ}=0.5$]	60	0.5	0.82	16	0.77	15.1
BF[$p_{succ}=0.6$]	60	0.6	0.9	17.6	0.87	17
BF[$p_{succ}=0.7$]	60	0.7	0.95	18.6	0.93	18.3
BF[$p_{succ}=0.8$]	60	0.8	0.98	19.1	0.96	18.9
BF[$p_{succ}=0.9$]	60	0.9	0.993	19.4	0.98	19.2
BF[$p_{succ}=0.95$]	60	0.95	0.995	19.5	0.99	19.3
Full-Enum.	60	1	1	19.6	1	19.6

However, using $r_{FAC} \leq 1$ is not a common practice, but in final rounds of BKZ -reduction with high block sizes, this may be observed! Therefore by using the concepts of full-enumeration success probability in third section (Part A), if the enumeration radius factor in this test reaches to $r_{FAC} \approx 0.98$, this is expected that all the numerical results in column of “ p_{succ} by our estimator of (36)” and “ f_{succ} by our estimator of (47)” are decreased by factor of 0.01 (see Table 1 and Table 2 in fourth section (Part A)). As shown in Table 4, when the input basis (or lattice block) is

strongly reduced (near to HKZ-reduced) and the success probability of bounding function is extremely small (near to extreme pruning), the actual value of success probability and dynamic success frequency can be decreased asymptotically. This test focuses on moderate block sizes ($\beta = 60$), while for bigger block sizes, this problem is relaxed automatically! As the block sizes are increased, even for extreme-pruned bounding functions over HKZ-reduced lattice blocks, the cutting point stays around the size of β , so this special case (which makes the

value of success probability dropped asymptotically) cannot be observed for high block size! However for some special setting, this can be seen even for high block sizes yet; For example, Fig. 5 shows the average shape of 7 HKZ-reduced bases with dimension 200 and different piecewise-linear bounding functions. As shown in Fig. 5, the cutting points of all the bounding functions nearly are equal to 200, but piecewise-linear bounding function by parameter of “ $a=0.01$ ”, with extreme pruning and estimated success probability $\approx 2^{-246}$ by relation (25), has the cutting point of $\text{Cut} = 84$, and consequently the estimated success probability and dynamic success frequency of it would be zero by our formulas in (36) and (47)!

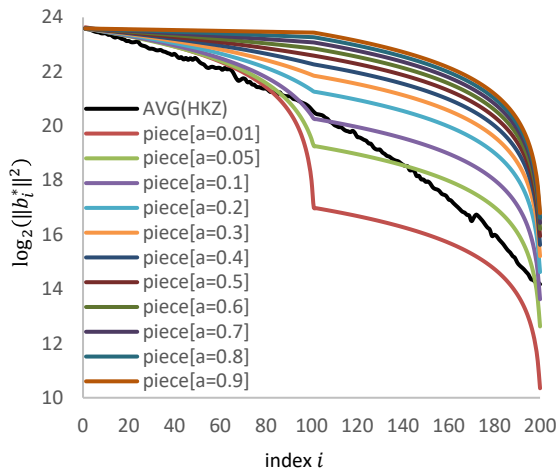


Fig. 5: Different shapes of quality of basis and different bounding functions with scaled entries of $R^2\mathcal{R}_{\beta-i+1}$ in dimension 200 to find the cutting point.

C. Test Results for Our Revision of Enumeration Cost

In this section, the exactness of our revised estimation of GNR-enumeration cost by Algorithm 2 is compared with Chen-Nguyen estimation of GNR-enumeration cost in Algorithm 8 of Appendix A from [1] (see relation (23) and (19)). For actual implementation of GNR-pruned enumeration function, the pseudo-code in Appendix B from [13] is used, but after each success of the enumeration function, the enumeration is not aborted, rather, the best solution and enumeration radius are updated (similar to the pseudo-code in Appendix B from [1]). The cost of experimental running of GNR-enumeration is determined by using a counter in actual enumeration function which counts the number of enumeration tree nodes. This test uses piecewise-linear bounding function with different success probabilities. The bounding functions which are used in this test, use mapping technique (see Lemma 4 and Remark 5 in third section (Part D)). The initial enumeration radius factor in this test is set to $\Upsilon = r_{\text{FAC}} = 1.2$; Table 5 shows our numerical results for this comparison. In Table 5, parameter of “ f_{succ} ” represents the Dynamic Success Frequency and parameter of “ p_{succ} ” represents the success probability. This test uses some random lattice basis of dimension 70 in the sense of Goldstein and Mayer [17], [18].

Note: Although floating point arithmetic is known to cause some stability problems during LLL reduction, based on the experiences in [13], such problems during enumeration (even up to the dimension of 110) are not seen.

Table 5: Comparison of our revised estimation of GNR-enumeration cost (in Algorithm 2) with the cost estimation proposed by [1] and the cost computed in experimental running of enumeration

Enum. Cost by compared cases Success Frequency & Success Probability	Enumeration cost by experimental running	Enumeration cost by our estimator in Algorithm 2	Enumeration cost by Chen- Nguyen in [1]
$f_{\text{succ}}=0.01$ & $p_{\text{succ}}=2^{-24}$	$2^{9.5}$	$2^{8.4}$	$2^{12.2}$
$f_{\text{succ}}=0.1$ & $p_{\text{succ}}=2^{-21}$	$2^{10.2}$	$2^{11.35}$	$2^{14.21}$
$f_{\text{succ}}=1$ & $p_{\text{succ}}=2^{-17}$	2^{11}	$2^{13.5}$	$2^{16.82}$
$f_{\text{succ}}=10^1$ & $p_{\text{succ}}=2^{-14}$	$2^{12.4}$	$2^{15.7}$	$2^{20.04}$
$f_{\text{succ}}=10^2$ & $p_{\text{succ}}=2^{-11}$	$2^{14.6}$	$2^{18.68}$	$2^{23.9}$
$f_{\text{succ}}=10^3$ & $p_{\text{succ}}=2^{-7}$	2^{18}	$2^{21.28}$	$2^{28.52}$
$f_{\text{succ}}=10^4$ & $p_{\text{succ}}=0.06$	$2^{23.3}$	$2^{23.78}$	$2^{34.63}$
$f_{\text{succ}}=10^5$ & $p_{\text{succ}}=0.57$	$2^{36.47}$	$2^{34.45}$	$2^{47.81}$
$f_{\text{succ}}=130964$ & $p_{\text{succ}}=0.75$	$2^{39.67}$	$2^{37.12}$	$2^{50.2}$

As shown in Table 5, this is clear that, the cost results by our proposed estimator of GNR-enumeration cost in Algorithm 2 are closer to the cost determined in experimental running of enumeration, than the closeness of enumeration cost by Chen-Nguyen estimator (in Algorithm 8 from Appendix A in [1]) to this experimental running cost.

Conclusions

BKZ algorithm has a determinative role in security analysis of lattice-based cryptographic primitives, therefore the total cost of BKZ and quality of output basis should be computed exactly to be used in parameter selection of these primitives. Although the exact manner of BKZ algorithm with small block sizes can be studied by practical running of BKZ, this manner for higher block sizes (e.g., $\beta \geq 100$) should be simulated. Designing a BKZ-simulation with GNR-pruned enumeration needs to some necessary building-blocks which includes definition of enumeration radius, generation of bounding function, estimation of success probability, LLL simulation, estimation of GNR enumeration cost, sampling method for enumeration solution, simulation of updating GSO. This paper tries to introduce some exact definition of optimal enumeration radius, generation of bounding function, estimation of success probability and GNR enumeration cost. Our contributions and results in this paper are described as follows:

- **Formal definition of optimal enumeration radius.** By definition of full-enumeration success probability in this paper, the optimal value for radius parameter \sqrt{Y} (as initial radius factor r_{FAC}) and corresponding bound for solution norm of full-enumeration are defined exactly in Theorem 2 in average case (see our estimations in fourth section (Part A)). This definition can be used dynamically to compute optimal enumeration radius in BKZ simulation and even actual running of BKZ algorithm. In other sides, former studies on BKZ-simulation don't use optimal version of the radius parameter of r_{FAC} . Paper [1] uses as an invariant factor just based on some limited experimental observations (see Figure 3 in [1]), paper [3] uses non-exact assumption of GSA to determine r_{FAC} dynamically for each block size (see relation (9) in [3]), and paper [2] uses no new idea to make the exactness of radius factor better (to the best of our knowledge).

Test Results. Against the success probability of full-enumeration in former studies which is set to 1, our exact estimation of success probability in full-enumeration, for some practical range of block sizes of $50 \leq \beta \leq 240$, is shown in this paper for different values of r_{FAC} based on our proposed theorem (Theorem 2). Also for block sizes of $50 \leq \beta \leq 240$,

our better bound of radius factors of \sqrt{Y} defined by Theorem 2, is introduced in this paper.

- **Revised estimation of success probability for GNR bounding function.** The former studies [1]-[3] use the efficient idea by [1] to estimate the success probability of GNR-enumerations (see formulas (15), (23) and (24)); In fact, the estimation by [1] only considers the pruning type by condition of (20) for cylinder-intersection of bounding function; This paper proposes to consider three more types of pruning in estimation of success probability (see our discussions at the beginning of third section); All of these four types of pruning are applied collectively in our estimation of success probability in relation (36); Our results in fourth section (Part B) shows non-negligible gap of our exact estimation of (36) from former estimations in some cases.

Test Results. Our revised estimation of success probability (for GNR bounding function) in our test results on (nearly) HKZ-reduced bases in dimension 60, shows non-negligible gap from former estimations by some main former studies of [1], [3], [13]. Also to have better sense, this paper shows the shape of bounding functions with different success probabilities and the shapes of quality of randomized/LLL-reduced/nearly-HKZ-reduced bases with dimension 60 (also 200) and the corresponding cutting points.

- **Revised cost estimation of GNR-enumeration.** The former studies [1]-[3] use the efficient idea by [1] to estimate the cost of GNR-enumerations (see formulas (19), (23) and (24)); Similar to success probability, the cost estimations in [1] only consider the pruning type by condition of (20); Our paper considers all of four proposed types of pruning in estimation of GNR-enumeration cost along with the process of updating enumeration radius in Algorithm 2; Our results in fourth section (Part C) shows the exactness of the estimation by Algorithm 2 against the former studies.

Test Results. Our results show that the cost results by our proposed estimator of GNR-enumeration cost in Algorithm 2 are closer to the cost determined in experimental running of enumeration, than the difference of enumeration cost results by Chen-Nguyen estimator in Algorithm 8 from Appendix A in [1] against the experimental cost results.

- **A novel technique in generation of bounding function.** By using Lemma 4, this is possible to generate a bounding function including cutting point of $\text{Cut} = d$ (see Remark 5); In former studies [1]-[3], if the simulation tries to generate bounding functions with much small success probability, this is possible that the success probability of this bounding functions unintentionally becomes much less than intended one

or even zero (because of ignoring the cutting points which are less than the block size, i.e., $\text{Cut} < d$; see our results and discussions in [fourth section \(Part B\)](#)).

This is worthy of noting that if we use another SVP-solver instead of GNR-enumeration (e.g., sieving algorithm in [22], enumeration by integrating sparse orthogonalized integer representations in [23], etc.), none of our contributions can be used in BKZ algorithm or BKZ-simulation!

Future Works. Three of our proposed components in this paper (include optimal enumeration radius, generation of bounding function and estimation of success probability) can be used in actual running of BKZ-algorithm, such as our technique of “BKZ with Progressive Success Probabilities” [21], [25] which massively generates bounding functions with different success probabilities, and consequently introduce new lattice-reduction security estimates to fix the problem of non-exactness in bit-security estimations of current cryptography schemes (e.g., [26]-[28]), so this is worthy of re-estimating their bit-securities by our revised components. Also [Algorithm 2](#) in this paper samples the norm of final solution which can be used in our revised method for sampling coefficient vector of GNR-enumeration solution in [29]. Moreover, the authors suggest the formal verification and proof of [Algorithm 2](#) corresponding with each claims in [Lemma 3](#), by some theorem provers such as Isabelle/HOL (see our similar works in [30]). At the end, nearly all components and concepts introduced in this paper can be used in design of new BKZ-simulation with better exactness, and consequently it is expected to use this new BKZ-simulation in introducing new lattice-reduction security estimates (as bit-security level by reduction).

Author Contributions

Gholam Reza Moghissi suggested the innovations and wrote the manuscript with the guidance of Dr. Ali Payandeh.

Acknowledgment

The authors gratefully thank the anonymous reviewers and the editor of JECEI for their useful comments and suggestions.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Abbreviations

$\mathcal{L}(b_1 \dots b_n)$ A lattice by basis vectors of $b_1 \dots b_n$

$\mathcal{L}(B)$	A lattice by basis matrix of B
$\mathcal{L}_{[j,k]}$	A lattice by GSO-projected basis vectors of $\pi_j(b_j) \dots \pi_k(b_k)$
$\det B$	Determinant of basis matrix of B
$\lambda_1(\mathcal{L})$	First Successive-Minima of lattice \mathcal{L}
$\ v\ $	Euclidean norm of lattice vector v
SVP	Shortest Vector Problem
LLL	Lenstra-Lenstra-Lovász algorithm
BKZ	Block Korkin-Zolotarev algorithm
β	Input parameter of Lattice block size
d	Lattice block size in running BKZ which is varied from β to 2
BKZ ₆₀	BKZ algorithm with block size $\beta = 60$
HKZ	Hermite-Korkine-Zolotarev algorithm
GSA	Geometric Series Assumption
Enum	Enumeration
$\text{Vol}(\mathcal{L}(B))$	Volume of lattice $\mathcal{L}(B)$
$\text{Ball}_n(R)$	n -dimensional sphere with radius R
$\text{Vol}(\text{Ball}_n(R))$	Volume of $\text{Ball}_n(R)$
$V_n(R)$	Volume of $\text{Ball}_n(R)$
$\Gamma(x)$	Gamma function with parameter x
$\text{GH}(\mathcal{L})$	The estimation of value of $\lambda_1(\mathcal{L})$ by Gaussian Heuristic of lattice \mathcal{L}
GSO	Gram-Schmidt Orthogonal
$B_{[1,d]}^*$	GSO basis of lattice $\mathcal{L}_{[1,d]}$ as $[b_1^*, b_2^*, \dots, b_d^*]$
GSO norms $B_{[1,d]}^*$	The norms of $\ b_1^*\ , \dots, \ b_d^*\ $
$\pi_i(b_i)$	i -th vector of GSO basis of B^*
b_i^*	Another notation for $\pi_i(b_i)$
$\mu_{i,j}$	GSO coefficient of i, j as $\mu_{i,j} = \frac{b_i b_j^*}{\ b_j^*\ ^2}$
$\text{Gamma}(x; k, \theta)$	Gamma distribution function
$\text{Expo}(x; \lambda)$	Exponential distribution function

1^x	A vector with length of x whose all entries are 1	$\text{Vol}(C_{R_1, \dots, R_l})$	Volume of cylinder-intersection of C_{R_1, \dots, R_l}
GNR	Gamma-Nguyen-Regev (pruning)	V_{R_1, \dots, R_l}	Another notation for $\text{Vol}(C_{R_1, \dots, R_l})$
$\pi_j(b_j, \dots, b_k)$	The projected form of the lattice block of $[b_j, \dots, b_k]$ whose vectors are projected on the vectors of (b_1, \dots, b_{j-1})	$\mathcal{P}_\ell(t_1, \dots, t_\ell)$	A polytope with radii of t_1, \dots, t_ℓ
$\mathcal{L}(b_j, \dots, b_k)$	Another notation for $\pi_j(b_j, \dots, b_k)$	$\text{Vol}\mathcal{P}_\ell(t_1, \dots, t_\ell)$	Volume of polytope $\mathcal{P}_\ell(t_1, \dots, t_\ell)$
N	Number of total nodes of full-enumeration tree	$p_{succ}^{new0}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	Original version of success probability (the equivalent notation for $p_{succ}(\mathcal{R})$)
N'	Total number of nodes in GNR pruned enumeration tree	$p_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	Our version of success probability
H_l	Gaussian Heuristic prediction of number of nodes at level l in full-enumeration tree	$f_{succ}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	Original dynamic success frequency
H'_l	Gaussian Heuristic prediction of number of nodes at the level l in GNR pruned enumeration tree	$f_{succ}^{new0}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	The equivalent notation for f_{succ}
		$f_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	Our revised version of dynamic success frequency
		C_{Roger}	An abstract parameter (not a real parameter) in f_{succ}^{new} , and is set to 1
R	Radius of n -dimensional ball in enumeration tree	\sqrt{Y}	Initial radius parameter
$C_{R_1 \dots R_l}$	l -dimensional cylinder-intersection with radii of $[R_1, \dots, R_l]$	\mathbf{p}_{min}	Success probability of full-enum. corresponding with r_{FACmin}
\mathcal{R}	Vector of $\mathcal{R} = [\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_\beta]$ as the bounding function	\mathbf{p}_{opt}	Success probability of full-enum. corresponding with r_{FACopt}
$p_{succ}(\mathcal{R})$	Success probability of bounding function \mathcal{R}	r_{FACmin}	Minimum hopeful radius parameter
γ_n	Hermite's constant	r_{FACopt}	Optimal radius parameter
$\mathbf{Pr}_{u \sim Ball_d}$	Probability of visiting GSO partial solution candidates in level l from GNR enumeration tree by assuming vector u is chosen uniformly distributed from d -dimensional ball of the radius 1	R_{opt}	Optimal enumeration radius
$\mathbf{Pr}_{u \sim Ball_d}^{new1}$	Our revised version of $\mathbf{Pr}_{u \sim Ball_d}$ including cut point of Cut	Hdown	Maximum index in head concavity
$\mathbf{Pr}_{u \sim Ball_d}^{new2}$	A minor revision of $\mathbf{Pr}_{u \sim Ball_d}^{new1}$	Tup	Minimum index in tail convexity
$\mathbf{Pr}_{u \sim Ball_d}^{new3}$	Our revised version of $\mathbf{Pr}_{u \sim Ball_d}^{new2}$ with any last non-zero index of $\mathcal{G} = j$	$\text{randINT}_{[x \dots y]}$	Return a uniformly random integer number between x to y
$p_{succ}^{Approx1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	Our approximation of success probability $p_{succ}^{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, R)$	$\text{rand}_{[x \dots y]}$	Return a uniformly random real number between x to y
r_{FAC}	Radius parameter, defined as $\frac{R}{GH(\mathcal{L})}$	\mathcal{G}	Last non-zero coefficient index in coefficient vector w (see [12])
		w	A Coefficient Vector defined for GNR enumeration solution [12]
		Cut	GNR enum cut point index (see [12])
		$\text{Prob}(\mathcal{G} = j)$	Probability distribution of \mathcal{G} for solution vectors of v (see [12])
		$\mathbf{E}[X]$	Expected value of X

K	Sampled number of solutions in GNR-enumeration function
$N_{new1}(\mathcal{L}_{[1,d]}, \mathcal{R}, \mathbf{R})$	Total nodes of GNR-enumeration tree after four types of pruning
H_l^{new}	Gaussian Heuristic prediction of nodes count at level l of GNR enumeration tree (line 7 in Algorithm 2).
$N_{new2}(\mathcal{L}_{[1,d]}, \mathcal{R}, \mathbf{R})$	Total nodes of GNR-enumeration tree after four types of pruning and aborting after finding first solution
a	The parameter of piecewise-linear bounding function

References

- [1] Y. Chen, P. Q. Nguyen, "BKZ 2.0: Better lattice security estimates," in Proc. International Conference on the Theory and Application of Cryptology and Information Security: 1-20, Berlin Heidelberg, 2011.
- [2] S. Bai, D. Stehlé, W. Wen, "Measuring, Simulating and Exploiting the Head Concavity Phenomenon in BKZ," in Proc. Advances in Cryptology – ASIACRYPT 2018: 369-404, 2018.
- [3] Y. Aono, Y. Wang, T. Hayashi, T. Takagi, "Improved progressive BKZ algorithms and their precise cost estimation by sharp simulator," in Proc. Annual International Conference on the Theory and Applications of Cryptographic Techniques: 789-819, Berlin, Heidelberg, 2016.
- [4] J. Hoffstein, J. Pipher, J. M. Schanck, J. H. Silverman, W. Whyte, Zhenfei Zhang, "Choosing parameters for NTRUEncrypt," Cryptology ePrint Archive, Report 2015/708, 2015.
- [5] M. R. Albrecht, B. R. Curtis, A. Deo, A. Davidson, R. Player, E. W. Postlethwaite, F. Virdia, T. Wunderer, "Estimate all the {LWE, NTRU} schemes!," in Proc. International Conference on Security and Cryptography for Networks, 2018.
- [6] M. R. Albrecht, et al., "Estimate all the {LWE, NTRU} schemes!," [Online]. Available at: <https://estimate-all-the-lwe-ntru-schemes.github.io/docs/>.
- [7] "Post-Quantum Cryptography Standardization Project", [Online]. Available at: <https://csrc.nist.gov/Projects/post-quantum-cryptography>.
- [8] J. Sharafi, H. Daghighi, "A Ring-LWE-based digital signature inspired by Lindner–Peikert scheme," J. Math. Cryptology, 16(1): 205-214, 2022.
- [9] N. Samardzic, A. Feldmann, A. Krastev et al., "CraterLake: a hardware accelerator for efficient unbounded computation on encrypted data," in Proc. ISCA: 173-187, 2022.
- [10] K. Cong, D. Cozzo, V. Maram, N. P. Smart, "Gladius: LWR based efficient hybrid public key encryption with distributed decryption," in Proc. International Conference on the Theory and Application of Cryptology and Information Security: 125-155, Cham, 2021.
- [11] T. Espitau, A. Joux, N. Kharchenko, "On a dual/hybrid approach to small secret LWE," in Proc. International Conference on Cryptology in India: 440-462, Cham, 2020.
- [12] G. Moghissi, A. Payandeh, "Better sampling method of enumeration solution for BKZ-Simulation," ISC Int. J. Inf. Secur., 13(2): 177-208, 2021.
- [13] N. Gama, P. Q. Nguyen, O. Regev, "Lattice enumeration using extreme pruning," in Proc. EUROCRYPT '10, volume 6110 of LNCS. Springer, 2010.
- [14] G. R. Moghissi, A. Payandeh, "Rejecting claimed speedup of $2^{B/2}$ in extreme pruning and revising BKZ 2.0 for better speedup," J. Comput. Secur., 8(1): 65-91, 2021.
- [15] L. Devroye, "Sample-based non-uniform random variate generation," in Proc. the 18th conference on Winter simulation: 260-265, 1986.
- [16] Y. Chen, "Reduction de reseau et securite concrete du chiffrement completement homomorphe," PhD thesis, Paris 7, 2013.
- [17] "SVP Challenge," [Online]. Available at: <https://www.latticechallenge.org/svp-challenge/index.php>.
- [18] D. Goldstein, A. Mayer, "On the equidistribution of Hecke points," Forum Math., 15(2): 165-190, Berlin, 2003.
- [19] GitHub hosting service, "fpLLL library project," [Online]. Available at: <https://github.com/fplll/>.
- [20] V. Shoup, "NTL: a library for doing number theory". [Online]. Available at: <http://www.shoup.net/ntl/>.
- [21] G. R. Moghissi, A. Payandeh, "Using progressive success probabilities for sound-pruned enumerations in BKZ algorithm," Int. J. Comput. Network Inf. Secur., 10(9): 10-24, 2018.
- [22] L. Ducas, "Shortest vector from lattice sieving: A few dimensions for free," in Proc. EUROCRYPT: 125-145, 2018.
- [23] Z. Zheng, X. Wang, Y. Yu, "Orthogonalized lattice enumeration for solving SVP," Sci. China Inf. Sci. 61: 032115, 2018.
- [24] G. R. Moghissi, A. Payandeh, "Optimal bounding function for GNR-enumeration," Int. J. Math. Sci. Comput. (IJMSC), 8(1): 1-17, 2022.
- [25] G. R. Moghissi, A. Payandeh, "Design of optimal progressive BKZ with increasing success-probabilities and increasing block-sizes," J. Comput. Secur., 9(2): 65-93, 2022.
- [26] D. J. Bernstein et al., NTRU Prime. Technical report, National Institute of Standards and Technology, 2020.
- [27] J. Bos et al., "CRYSTALS - kyber: A CCA-secure module-lattice-based KEM," in Proc. 2018 IEEE European Symposium on Security and Privacy (EuroS&P): 353-367, London, UK, 2018.
- [28] J. P. D'Anvers et al., SABER. Technical report, National Institute of Standards and Technology (2020).
- [29] G. R. Moghissi, A. Payandeh, "Revised method for sampling coefficient vector of GNR-enumeration solution," Int. J. Math. Sci. Comput. (IJMSC), 8(3): 1-20, 2022.
- [30] G. R. Moghissi, A. Payandeh, "Formal verification of NTRUEncrypt scheme," Int. J. Comput. Network Inf. Secur., 8(4): 44, 2016.

Biographies



Gholam Reza Moghissi received the M.S. degree in department of ICT at Malek-e-Ashtar University of Technology, Tehran, Iran, in 2016. His researches focus on information security.

- Email: fumoghissi@chmail.ir
- ORCID: [0000-0001-9189-6786](https://orcid.org/0000-0001-9189-6786)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA



Ali Payandeh received the M.S. degree in Electrical Engineering from Tarbiat Modares University in 1994, and the Ph.D. degree in Electrical Engineering from K.N. Toosi University of Technology (Tehran, Iran) in 2006. He is now an assistant professor in the Department of Information and Communications Technology at the Malek-e-Ashtar University of Technology, Iran. He has published many papers in international journals

and conferences. His research interests include information theory, coding theory, cryptography, security protocols, secure communications, and satellite communications.

- Email: payandeh@mut.ac.ir
- ORCID: [9246-9953-0002-0000](https://orcid.org/9246-9953-0002-0000)
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

G. R. Moghissi, A. Payandeh, "Revised estimations for cost and success probability of GNR-enumeration," J. Electr. Comput. Eng. Innovations, 11(2): 459-480, 2023.

DOI: [10.22061/jecei.2023.9228.588](https://doi.org/10.22061/jecei.2023.9228.588)

URL: https://jecei.sru.ac.ir/article_1880.html

