Shahid Rajaee
Teacher Training University

Journal of
Electrical and Computer Engineering Innovations (JECEI)

Vol. 10 No. 2, Summer-Fall 2022

Journal of

JECEI

Electrical and Computer Engineering Innovations

Vol. 10 No. 2, Summer-Fall 2022

JECEI

Shahid Rajaee
Teacher Training University

Journal of
Electrical and Computer Engineering Innovations (JECEI)

pISSN 2322-3952
eISSN 2345-3044

Semiannual Publication

Volume 10, Issue 2, Summer-Fall 2022

SRTTU

Journal of Electrical and Computer Engineering Innovations

http://jecei.sru.ac.ir

EISSN: 2345-3044

ISSN: 2322-3952

JECEI

JECEI

License Holder: Shahid Rajaee Teacher Training University (SRTTU)

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

Journal of Electrical and Computer Engineering Innovations

Vol. 10; Issue 2: 2022

Contents

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Review Paper

A Comparative Study on Anonymizing Networks: TOR, I2P, and Riffle Networks Comparison

M. Hosseini Shirvani**, A. Akbarifar

Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran.

Article Info

Article History:
Received 25 July 2021
Reviewed 29 August 2021
Revised 01 November 2021
Accepted 01 November 2021

Keywords:
Tor project
I2P network
Riffle
Dark-web
Anonymity
Security

*Corresponding Author's Email Address:
mirsaeid_hosseini@iausari.ac.ir

Abstract

Background and Objectives: Among miscellaneous networks, onion-based routing network technologies such as The Onion-based Routing (ToR), Invisible Internet Project (I2P), and Riffle networks are used to communicate anonymously by different worldwide users for security, privacy, and safety requirements. Sometimes, these types of networks sacrifice anonymity for the sake of efficient communication or vice-versa. This paper surveys aforementioned networks for investigating their potential and challenges.
Methods: Onion-routing networks encapsulate messages in several layers of encryption similar to layers of an onion. The anonymous communication networks are involved dining cryptographers (DC) problem so-called DC-nets, which need sending anonymous message with unconditional sender and untraceable receipt. So, DC-nets must be resistant against traffic analysis attacks although they will attenuate the network bandwidth. In this line, ToR is a free software that provides anonymous communication, I2P networks are based on hidden internet service project which uses tunnelling for anonymous communications, and Riffle networks include a small set of camouflaging servers that provide anonymity for authorized users. This paper presents a comparative study on anonymizing ToR, I2P, and Riffle networks in terms of associated prominent parameters in this vein.
Results: The comparison is based on similarities, differences, and challenges in network behaviors. This comparison is beneficial for further researches and future improvements.
Conclusion: The review of the current paper reveals that the Riffle networks are more resilient and have great confidentiality and integrity against other onion-based routing networks.

Introduction

Mission critical applications associated to individuals and organizations need meeting security requirements for their business process fulfilment [1]-[4]. In their roles, stakeholders are also interested in privacy and protecting their identity on the web in some cases. In this regards, onion-based routing network technologies such as The Onion-based Routing (ToR), Invisible Internet Project (I2P), and Riffle networks are used to communicate anonymously by different worldwide users for security, privacy, and safety requirements on the web [5].

Sometimes, these types of networks sacrifice anonymity for the sake of efficient communication or vice-versa. Onion-routing networks encapsulate messages in several layers of encryption similar to layers of an onion [6].

Albeit, packet encryption in the network is aimed at confidentiality, integrity, and authentication goals but

Doi: 10.22061/JECEI.2021.8027.466

tracking the packet will be possible because the routers need to know the source and the destination of data [5].

Anonymizing communication networks such as ToR, I2P, and Riffle are attempting to hide the identity of sender's information [5]. ToR networks provide an anonymous communication with hiding the real locations of the senders. Notwithstanding communication profits, hidden web (Dark-web) on ToR's network caused many problems for the government due to its concealment. The Dark-web that is non-registered domains has a specific 16-character address; so, it becomes a place for cyber criminals for some actions like arms trafficking, selling human organs, drug dealing, assassination missions, cyber-attacks and etc. [7]. On the other side, I2P network is a peer to peer and message-oriented anonymizing communication network. Basically, I2P has been developed in order to anonymizing a connection between two regions within the network. The I2P network was firstly introduced in 2003 and its origins could be founded in invisible internet project [7]. This network is built on top of the internet protocol [8].

The Riffle identity anonymizing network is providing safety and security requirements when the attackers in another anonymous system make counterfeit servers to traffic analysis attacks, but the ToR networks are vulnerable to these kinds of attacks. On the other side, a Riffle network utilizes hybrid networks approach as a defensive mechanism against these attack scenarios. Similar to other anonymizing networks, a Riffle network uses of onion routing (OR) protocol that is a method for anonymous information's exchange in computer networks. The packets are encrypted successively and then they will be sent to too many network nodes or ORs. Each OR decrypts a cipher layer to read the routing instruction and then will send it to the next router to do the same process. This method caused the network nodes to have nothing known about the nodes contents and the origin of the packets [9]. A Riffle network is a shuffle network which shuffles data streaming with different keys. So, in the mixed networks, a batch of incoming packets is transmitted between different safe servers, without the use of inefficient public keys. In fact, the traffic will change before and after the logging into the server. Instead of using inefficient public keys, a Riffle network verifies the encryption based on verifiable shuffle during the validation of incoming encrypted packets. However, even unsafe servers on the network could not access the data. To do so, the unsecured servers should shuffle the message correctly, since the input data could be accepted by secure servers which are nearly impossible [10].

Recently, numerous authors have compared low-latency anonymizing communication networks in literature such as ToR, invisible internet project (I2P), and Riffle networks [5], [7], [9]-[11]. In this regard, an overview on utilization of anonymity technologies has been presented [12]. This study discusses user's security perils and describes principal mechanisms to prevent the attacks in anonymizing communication networks. According to this overview, the protection of user's privacy and its violations by government agencies and information security organizations has been investigated. Also, Invisible internet project (I2P) is one of the ToR's capabilities in which many researchers have studied about it in [7], [12], [13]. The contribution of the current paper is to present a comparative study on three anonymity communication networks ToR, I2P, and Riffle networks in terms of advantages and disadvantages along with their commonalities, discrepancies, and possible challenges in their network behavior. This comparison is beneficial for further researches and future scheme improvements.

Related Works

Although the anonymity communication networks are not very novel and stems back to 1997 [14], there is a clear lack in literature to pay on these types of networks. Nevertheless, we bring some of literatures to introduce their behavior.

An efficient communication system with strong anonymity called Riffle has been presented in literature by Albert Know et al. [10]. The proposed Riffle network provides a bandwidth and computation efficient communication network with high anonymity guarantees. To do so, it used hybrid verifiable shuffle and private information retrieval techniques [10]. A novel algorithm called ToRank was proposed that ranks hidden services in ToR networks when the users surf on web; this is because to lessen the harm to ToR network related to suspicious activities. It had successful behavior on famous datasets [15]. A universal serial bus (USB) side-channel attack on ToR has been introduced where a malicious is allowed to reach a public USB charging station [16]. This type of attack depends on power measurements of attacker's device without observing network traffic analysis [16]. In this vein, fingerprint attack is a famous threat [17]. To obviate this problem, an adaptive online website fingerprint attack for ToR networks has been introduced by Attarian et al. [18] To recognize the attack, the authors applied machine learning techniques in dynamic fashion. Also, the round robin queuing process was done to defense against protocol level attack in onion-based routing networks [19]. This utilizes integrity checks and counterfeit traffic auditing in the middle layers of relay nodes to recognize protocol level attacks. The review of related works introduces how the anonymity communication networks perform along with known and contingent attacks and the countermeasures in these fields.

Motivation and Research Plan

As mentioned earlier, users utilize different anonymizing networks to meet their security, privacy, and confidentiality requirements. Although it is clear-cut that these types of networks make fortunities and also challenges, there is a clear lack of a survey study on these kinds of networks. This is the reason of preparing the current review paper in comparative perspective. To do so, our subjective research plan which contains explanation on several sub sections is depicted by Fig. 1.

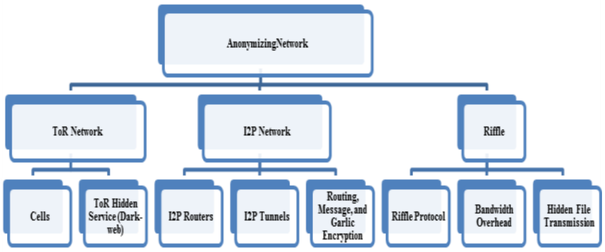


Fig. 1: Subjective research plan.

After introducing the main anonymizing networks along with their attributes, merits, and challenges, the comparative schemes in terms of prominent parameters are presented and discussed in forthcoming sections.

ToR Network

ToR is free network-based software which provides anonymous access to internet servers. There are two important factors in anonymity: firstly, the attacker should not be able to distinguish which IP address interacts with the servers and secondly, the server should not be able to recognize the IP address of data transmitter. Fig. 2 illustrates a typical ToR network structure. Note that the dark grey squares in Fig. 2 represent the core nodes; and the light grey squares represent periphery nodes.

The basic idea for ToR is to transmit traffic by cluster nodes in which they have no information about the source and destination of the transferred packets. This relies on the distributed system principle in which the whole system seems singularly and coherently from the user's point of view [21]-[23].

ToR network utilizes onion routing protocol in order to anonymize user's communication. In onion network, the packets would be encrypted into layers, so to this layered architecture, it is called onion configuration. An onion routing is applied to each node that it is responsible for encrypting each onion layers in order to discover data for the next node within given networks [5], [7], [9], [13], [21], [24]. In a ToR network, the user is required to create an orbital path to communicate with the server. The orbital path is created by using socks via onion proxy on the user's side.

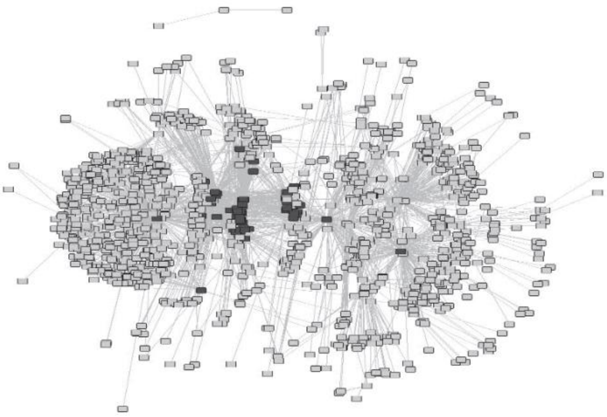


Fig. 2: Core-periphery categorical model of ToR network [20].

A ToR has a directory of verified onion routers associated with their public keys.

The index is examined manually and prevented to make bogus ORs for controlling traffic; hence, it informs onion proxy (OP) from ORs and then onion proxies will fetch the sequence of OR from the intentioned routing list [7].

The first OR is called as a watchdog and the last node which is decrypting the last layer is called exit node. Note that in existing ToR network, there are 3 ORs such as gateway or input, middle, and exit nodes that communicate with OP via transport layer security (TLS) secure connection and disposable keys. This procedure is a development for anonymizing in ToR network.

In order to restore a list of all onion routers, the distributed valid server is used. These servers should be known or being published in particular websites and also be able to track topology changes in the network. Index servers combine network topology and establish subscriber identity form the whole network. These directories automatically will fetch by OPs. Also, user's software contains a default list of directory servers [5], [25]. Directory servers store all ORs information on a list.

The circuits are a virtual duplex communication in ToR network that is established between OPs and a category of ORs. Obviously, a single circuit path in OR's could use several TCP flows simultaneously. In order to prevent flows recognition by the attackers, the lifetime of these circuit paths would be 10 minutes. After spending the circuit path time, the circuit is eliminated and a new one would be used. Note that the new circuit would be made as operational background, so there will be no additional delay in the system [26].

Cells

Circuit path consists of numerous ORs. The client acquired traffic is placed in cells with constant size (512 Byte) to making traffic analysis more difficult. The cells within each OR are rearranged by a symmetric key. The cells within ToR network are considered as a unit which is composed of a header and the only response by a payload. These cells are able to control (build and destroy) or redistribution of the cells (end-to-end data streaming). For instance, if a user intends to create TCP flow at the first point, he should transmit the control cell to first relay station by determining the next OR address; then, the process begins with exchange the symmetric key to second relay station by Diffie-Hellman key exchange model [25]. This process is maintained as a similar model until the circuit configuration was created. Afterwards, the user transferred broadcasted cells, these cells will carry with an end-to-end data stream [5], [27].

ToR Hidden Service (Dark-web)

One of the most important features of ToR network is the Dark-web [7], [13]. In this section, the main focus is on finding pages that the others are not able to see that page, as well as user's identity. Hidden services are included some anonymous websites and servers that could be accessed only by onion routing based networks. The site addresses in ToR are represented by an onion domain that is registered nowhere. On the other hand, the addresses are like “name. Onion” in which the name is a 16-character string.

In computer networks, the Rendezvous protocol is used when two persons have no information for communicating with the second party. This protocol is caused the sources and also counterparts to find each other in peer to peer networks. The Rendezvous protocol, which uses handshake model for communicating, does not send the data before the preparation of the destination. The ToR network uses this protocol in order to build its hidden service.

The hidden service uses three network nodes and calls each of them a “Recommender Node”. Then, the network sends a request to recommender nodes based on the “Recommender Node” placement. If recommender nodes receive a positive response, they will send their public key. Note that the recommender node is not aware of service location. In next step, the hidden service is provided “service descriptors”. A service descriptor contains recommender nodes, their characteristics, and the public key. Service descriptor encrypts the data with the public key and puts it on the hash table. The distributed hash table is a key quantitative database that the value and the key are hidden service descriptor and 16-character address respectively. The 16-character address is generated from the public key of user's service. This address went to the distributed hash table and has no information about IP addresses. The user should have anonymous onion address to be connected. This process is done by using the hash table and users public key for value estimation. Hence, according to the Rendezvous protocol, the user chooses some points named Rendezvous and begins the key exchange. It is worth noting that the user has recommender key and server key in this step. After the intro messages are made by the user, the random point is encrypted by disposable hidden address via server public key and consequently, the only hidden server is able to read that. The 16-character onion address is composed secure hash algorithm 1 (SHA1) and the public key that encrypted with base32. Therefore, the production probability of acquired strings by another person via the same public key is very low and the address would be unique. In making onion address, 3 practical soft wares are used inducing Scallion, shallot, and Escholat [7], [13]. Shallot software is based on the hash structure by using GPU and Escholat which uses a list for a dictionary based search. Although hidden services have a beneficial application they have a wide range of misbehaviour in cyberspace. Creating and using these services as the Dark-web and generating anonymous pages are caused illegal services such as human organ trafficking, drugs, murder, kidnapping, cyber-attacks operations, hiring attackers, and other detractive functions which the government is looking to monitor this network [7], [13].

I2P Network

I2P is a message-oriented and P2P identity anonymizing network. This network is generated due to anonymous communication between two intermediate network sections. Today, various fields of I2P applications are available including unsigned web hosting, browsing web pages, file transfer, and email service.

Utilizing exterior service shows that the service which is not hosted in the I2P network requires external proxies. I2P is a cover network which allows the user to interact with the network anonymously. Technically, I2P has a framework based on java platform which is designed to provide peer to peer anonymous networks.

Each user is responsible for running I2P routers that establish I2P core software. All the packets in this network are transmitted by tunnels via I2P routers, as well as other counterparts. These tunnels would only be performed on route traffic. Therefore, internal and external tunnels are needed for input and output traffic [5]. Fig. 3 depicts a typical I2P network.

In I2P, peer selection is done by an executive row-based algorithm of each I2P router. After creating internal and external tunnels, users will store their data connections in a global database which is called netDB.

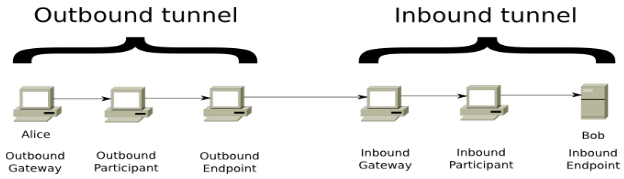


Fig. 3: I2P network with two types of tunnels inbound and outbound [28].

This database contains data connection of each I2P network and other public services within I2P network. The transmitted message is encrypted by Garlic encryption algorithm based on an end to end connection [29], [30].

Garlic encrypting is just same as onion encrypting with the exception that a single Garlic message may be contained within the different messages for various receivers [5], [3].

I2P Routers

I2P network is composed of users, nodes, or executive routers on I2P software which allows application communication to the mentioned network. I2P router is the core of this software. This router is responsible for keeping pairs and counterpart's statistics, which has been used in pair selection including encrypting operation, making the tunnel, providing service and sending the message. The aforementioned applications heavily depend on the tunnels that are created by I2P routers for being anonymous.

The netDB routing information and Lease set: super peers are utilized in creating and managing the database. The netDB is based on a distributed hash table which comprises all discovered information such as I2P pairs and I2P network-based services. Each super peer is only responsible for certain part of network information. Kademlia is a standard distance metric for XOR that determines which super peer is responsible for which part of the network depending on ID Couples [31], [32]. The peers with adequate bandwidth may be promoted to the super peer by decreasing the threshold. The netDB keeps the router infrastructure information as well as a Lease set for all known services.

I2P peers are recognized by a data structure called router information which includes all knowledge about the peers (IP address, port number, I2P stable version number, some of the statistical information, public key, and definitive 256-bit key of hash). In order to restore an available list of I2P peers, lists of routers information are able to be loaded from an identifiable system or a known web server. Restoring primary list from the router information is defined as reseeding operation.

A Lease set is a peer as the input gate to the internal tunnel of the corresponding service that is able to detect the habits. Both router information and Lease set are able to easily accumulated and restored by communicating with the closest super peer. For storing the super peer, the received router information and Lease set are transmitted to seven close super peers. About reseeding, two adjacent super peers communicate with each other. If the requested information did not exist, the super peer will provide a list of the closest super peers. On the other hand, the peers maintain the super peer information until the information restoring.

All routes in the I2P network are identifiable as a 256-bit encrypting key which is composed of a public key with 256-byte, an ID key with a 128 byte and a blank ID. Then, I2P refers to a primary service created by I2P router. Similar to domain name resolver (DNS), to mapping target names to cipher keys, three local host files are applied. In order to merge local and external host files, I2P is intended to make an index directory. Note that, this type of addressing will increase the anonymity.

I2P Tunnels

In I2P network, all the message are transmitted by tunnels. The tunnel is a virtual half-duplex encrypting connection in which two or three I2Ps use. In contrast with ToR, I2P router is going to create a tunnel which itself is contained another tunnel. Primitively, each I2P router makes various tunnels for input or output traffic.

The first I2P peer of a tunnel is called gate tunnel. The last I2P peer is called the end point tunnel. For output tunnel, the I2P router which is responsible for creating a tunnel will always be the gate tunnel and for input tunnel, this will be always the end point tunnel. The default value and tunnel length will be customized by the user in the setting. The tunnel length is always an evaluation of anonymity and functionality. The long tunnels increase the anonymity, whilst decrease the functionality and performance. An application doesn't belong to a particular tunnel and it may need different tunnels for message broadcast. Generally, there are two types of tunnels: exploratory and user tunnels. Exploratory tunnels are the ones with the limited bandwidth that would not be used in sensitive privacy operations. A router uses these tunnels for communicating with super peers and restoring the netDB database. On the other hand, the exploratory tunnels are used for certain, management, and destroying other tunnels. Normally, rebuilding the tunnels prevent data analysis attacks.

Creating new tunnels is performed by the first set of I2P peers. As mentioned in the last section, the peer's selection is based on raw and profile selective algorithm. An exploratory tunnel is used when encrypted tunnel creation requests are transmitted to the first I2P router simultaneously. Each layer includes some information for a single I2P, such as symmetric key and machine address. Similar to OR circuit design, the message is conducted until it reaches to the last I2P peer. The return response is to the successor that each I2P peer will add a layer of encryption. The I2P peer receivers are free to accept or reject the requests.

Routing, Message, and Garlic Encryption

I2P router can send and receive the message by this network while at least one output or input tunnel is created. In order to connect with an I2P service, firstly the router should restore service destination from the super peer. The destination determines a set of input gate tunnels from the counterpart service.

I2P uses the Garlic routing that is a kind of onion routing. This routing uses several Garlic message called Cloves. The Cloves are the data message with traditional routing instruction such as latency. On the other hand, it could be concluded that the Garlic message contains several practical messages. The real data message is encrypted by an end to end connection via a public receiver key. The Garlic message is encrypted multiple times via symmetric encryption by a public key exchange to tunnel peers. When the tunnel is scrolled, each I2P peer eliminates a layer of encryption until the Garlic message reaches the output end-point tunnel. The final point transfers the output message to the input gate. The input gate will transfer the Garlic message to the real receiver until each peer within the tunnel add encrypted layers by symmetric keys of those only the receiver is able to remove all encrypted layers from the Garlic message [5].

Riffle Network

The Riffle network includes small set of identity anonymizing servers that ensure anonymity between authorized users until at least one guaranteed server is presented [5].

The Riffle network is a new approach composed of hash verification, private data recovery for the bandwidth, and computing anonymous communication functionality. The attacker is able to target the Riffle by the ToR node via manipulating some servers and using malicious code.

To prevention of these attacks, the Riffle uses verification and authentication approaches which are located above the ToR stack. This approach presents a verifiable statistical report for the same received or sent the message.

When the secure connection is established between the servers, the system uses encryption and identity verification to confirm the encrypted message by a less computing power, and on the contrary it provides high transmission speed rather than ToR network. By using ID verification mechanism, even the malicious servers are not able to disrupt the network and communications. They need to disrupt message correctly, hence only verified servers are able to receive. Therefore, the network would be safe as long as a single and unique server is presented in the anonymizing network in Riffle [10].

File transfer in Riffle is $\frac{\mathbf{1}}{\mathbf{10}}$ of the time needed for the same operation in ToR and other anonymizing networks [10]. In addition to, the Riffle is able to access 100 kbps bandwidth per user in a set of 200 users and also is able to respond to 100,000 users in microblogs with less than 10-second latency [10]. Similar to described anonymizing networks, the Riffle is expressed to traffic analysis attacks. Two measures caused the traffic analysis attacks to be limited in Riffle network: firstly, DC-nets which technically suggest the information for users and servers safely; secondly, the verifiable mix-nets that are based on hybrid and complex patchworks. In this scheme, the mix sets use disruption for replacing the ciphertext.

Same as DC-nets, the verified mix-nets guarantee anonymity. A DC-net is overloaded by many processes and only scalable for a few thousand users. On the other hand, the verified mix-net allow the user to send a message according to the message size. Therefore, a significant improvement could be observed in bandwidth usage. Although high computation and disruption overhead are guaranteed, the verified mix-nets are prevented from high bandwidth connections. The Riffle considers the problems of mix-nets and DC-nets, while it suggests the same amount of anonymity. The high levels in Riffle are organized as user-server structure. This network has been focused on minimizing bandwidth interface like smartphone and reduces computation overhead of the server and provides support for more users. Obviously, the Riffle users have an appropriate bandwidth given to message size and the members, so the server computation requires symmetric key encryption in common cases. This process allows the user to exchange the message and makes the system suitable for the application of effective file transmission. It is noteworthy that so far, the identity anonymizing systems have not the expected support from anonymity for all users, as well as the servers [9], [10].

The effective bandwidth and efficient computations in Riffle are caused by two factors: first for the verified disruption and a new combination of upstream communication; and second for Private Information Recovery (PIR) for downstream communications.

The current identity anonymizing networks evaluate the computations and bandwidths by the broadcast of all messages to all users via limited computation-high bandwidth or by an expensive PIR computation, high commutation-limited bandwidth. The PIR-based model is able to minimize the download bandwidth by minimizing computation overhead.

Although most communication anonymizing system relies on protecting transmitter anonymity, PIR protects the receiver privacy. In PIR, the user has access to some data via management server or a set of servers which are intended to hide regulatory data. There have been various PIRs for different settings, but some of the PIR strategies have complicated method and rules. In this line, Table 1 show the abbreviations which have been used in Riffle network.

Table 1: Samples of Calibri sizes and styles used for formatting a pes technical work

Terminology	Description
C	Set of Riffle clients
n	The number of clients
C_i	The $i-th$ Riffle client
S	Set of Riffle servers
m	The Number of servers
P_j	Public keys where $j \epsilon [1,m]$
b	Size of a message
π_i	The $i-th$ permutation function
f	The file name
$H_{\vec{f}}$	Hashes of all blocks relevant to file f
$H_j \in H_{\vec{f}}$	Hash blocks of flie f requested by client C_j
$H_{\vec{\pi}}$	Permutated available hash blocks
$\mathrm{M_j}$	$j-th$ plaintext message
r	Number of round
λ	Security parameters

Also, the Fig. 4 illustrates the deployment model of Riffle network. As shown in Fig. 4, the Riffle system is composed of a user-server structure. Users are considered as a set of individuals who are intended to communicate anonymously and the servers are considered as anonymous service. For replacement operations, each server is considered as a separate member or subject. Each user will communicate in a Riffle with a priority service, which is the main server based on the parameters such as host organization and location. The most important and valuable source in Riffle configuration is the bandwidth between user and server, supplying a high bandwidth network between a few servers is a common and feasible process. However, the high bandwidth could not be expected by all users due to their connection mode and the network infrastructure. Therefore, a Riffle has been focused on minimizing the requirements and the bandwidth between users and servers.

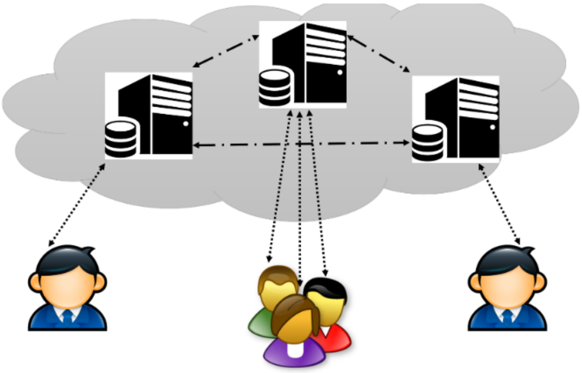


Fig. 4: Deployment model of Riffle [10].

Similar to I2P, a Riffle network tries to prevent traffic analysis attacks. Thus, the Riffle connections are dismissed in a cycle alternately. In both cases, each user receives and sends a message even the non-accession connection existed.

Riffle Protocol

In the following, the Riffle protocol is discussed. During the installation phase, the users establish three sets of ciphers which are coupled with the servers as follows: the $\{K_{ij}\}$ key uses the guaranteed disruption, $\{S_{ij}\}$ and $\{m_{ij}\}$ use a simpler approach such as Diffie-Hellman in PIR [33].

Each server generates π_i permutation for a guaranteed disruption and keeps them for the next use according to connection phase. The $\{K_{ij}\}$ key would be placed in S_i at $\pi_{i-1}\left(\ldots\ \left(\pi_1(j)\right)\right)$ when the installation was done.

In the *r* cycles of connection phase, the protocol uses hybrid shuffle and PIR or distribution for loading and downloads respectively. In loading step, each user C_j encrypts a message by $\left\{K_{ij}\right\}$ key where $i \in [m]$; and loading the encrypted cipher text is done in $S_1$by the main server C_j. At disruption step which is began by S_1, each server S_1 verifies the cipher text and encrypts them by $\left\{K_{ij}\right\}$ key where $i \in [n]$; and they are stored during the installation phase via π_i permutation; then, the results are transmitted to the server. This also means that the ciphertext C_j in S_i, $\left(\ldots\left(AEnc_{kij,r}\left(m_j^r\right)\right)\ldots\right)$ which is confirmed by K_{ij} key. The final server shows the ciphertext for all servers. the final permutation of the message is according to $\pi = \pi_m\left(\pi_{m-1}\ \left(\pi_2(\pi_1)\right)\ldots\right)$.

Bandwidth Overhead

The Riffle has been achieved with the bandwidth optimization between the user and the server during message transmission. The ciphertext from encryption architecture is loaded based on the verified ID layer with $b + m\lambda$ scale. The parameters of this equation are taken from Table 1. If the user is interested in a particular message and the indicator was known; then, the transmission overhead would be included coverage n and the number of users to the main server. The high bandwidth stream $b + m\lambda + m$ as well as low bandwidth stream would be b for each user and cycle.

It should be noted that also the high bandwidth is grown linearly to n; it needs only one bit per user. When a message was anonymous, the download bandwidth like nb for each user. While the loading bandwidth would be decreased by n, the bandwidth requirement between the servers is increased linearly given to the number of users. Each server has to download and a loading n ciphertext (with a crossed off layer) to the next server. The last server has to send an empty text to the others. In addition, although PIR decreases that download bandwidth overhead for the users, it will increase the server to server bandwidth.

Hidden File Transmission

The Riffle makes these systems suitable for the functions of each compressed bandwidth such as file transmission. The file transmission is similar to Bit-Torrent network with a few detailed differences. In a Riffle, the user-server model is presented, but the Bit-Torrent has a peer-to-peer structure. While a user is intended to share a file, he/she will generate torrent file by which is including mixed string off all the file blocks. Then, the user loads the torrent file in the server via Riffle. The servers play the role of the torrent tracker in a Bit-Torrent network and manage all available files on that group. In the simplest design, the file descriptor is propagated to all users and they were able to choose the download directly. Although the torrent files sharing have an expense at once, their distribution would not cost so much the users share the files by distributed torrent files via the Riffle anonymously. Therefore, there are three basic steps:

Firstly, it is Block Request: Each client C_j chooses the file f arbitrarily and disrupts the file block $H_f^{\rightarrow}$ by torrent file. Then, the C_j requests a block by using the Riffle and loading the file $H_j \in H_f^{\rightarrow}$ to S_{pj}. When the user has no block to request, the user sends a dummy non-request message to retain traffic analysis resistant. In this case, all requests in $H_{\pi}^{\rightarrow}$ are dispersed to the users at the end of each round.

Secondly, it is Block Loading: Each client C_j investigates whether the requested block bt disrupted files via $H_{\pi}^{\rightarrow}$ or not. If a matched block such as M_j was found, then, C_j loads the M_j by the Riffle. While the blocks with plaintext were available for the servers, each server will propagate the disrupted file of the existed block $H_{\pi}^{\prime\rightarrow}$.

Thirdly, it is Block Download: Each client C_j uses the H_j to find the $H_{\pi}^{\prime\rightarrow}$, which is the I_j index of the requested C_j block. Then, client C_j downloads the blocks by PIR. Fig. 5 shows the model of anonymous file transmission protocol in Riffle [9].

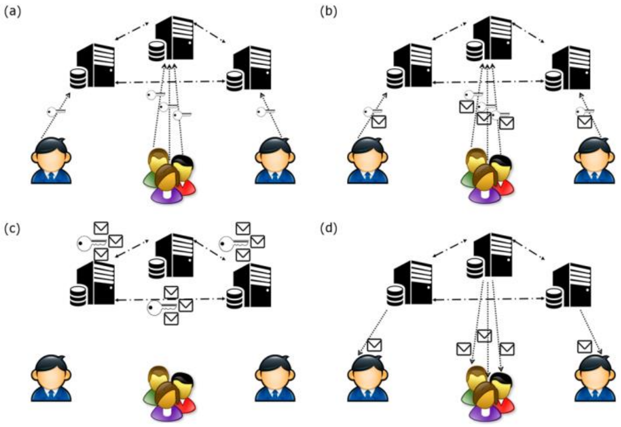


Fig. 5: Hidden File Transfer Protocol [9].

Fig. 5-a, is associated to setup phase in which users share the torrent files anonymously. Fig. 5-b, is related to request phase when a user requests a file upon uploading the hash of the file by utilizing Riffle. Fig. 5-c, is for upload phase when a user uploads an encrypted file according to requests by applying of Riffle. The download phase is depicted in Fig. 5-d, where a user downloads the file that he/she requested via PIR protocol.

Comparison between Identity Anonymizing Networks

There are obvious and various differences between I2P and ToR networks. ToR is based on the services which are provided voluntarily for generating circuit modes while I2P uses the counterparts with a sufficient functionality profile. Also, ToR network has been designed by many routers for optimization in output traffic where the I2P network has been designed for providing intermediate network services and facilities of only one set of output proxies. However, both methods present stable and low latency anonymity. In the following a comparison is given for important aspects of identity anonymizing communication networks:

ToR network uses the safe port interface and socks. In this field, the network is functioned as a proxy server. This expresses that the practical applications are able to use socks without any changes. On the other side, I2P is a middleware that provides the practical applications of the network or communication. Utilizing socks for ToR has two negative aspects: first, the socks interface is able to measure the message over TCP connection while the

I2P has to choose between TCP and UDP. So, I2P has a better performance in particular applications. Second, the message sent by applications may have information which would recognize the transmitter. In order to prevent this process application-level proxies with filtering capability such as Privoxy, have been used.

I2P and ToR have various capabilities. In fact, most I2P applications are designed to access intermediate network services exclusively. On the other hand, ToR network is able to use any program which is configured by proxy socks. Different encryption layers have been used in these three aforementioned networks which begin with transmission layer encryption and require TLS connection maintained by onion routers or I2P counterparts. I2P has additional functions in tunnel encryption. The message which is transmitted by the networks may be encrypted by onions or Garlic encryption. This means that the user connection to tunnel or the circuit would be completely anonymous within every anonymizing network such as the end-to-end encryption in I2P. However, the end-to-end encryption in ToR and other networks has no warranty due to the transmission layer protocol. Therefore, insecure protocols would not be used in these networks whilst a malicious or manipulated output node may store the message in plaintext and restore usernames and even the passwords.

In ToR network, only the first OR in the circuit knows the IP addresses of the real user; all other onion routers only know that before and after routers. User anonymity in ToR is significantly depending on ToR nodes selection algorithm. About I2P, even the first peer has no information about its transmitted data to another peer. In contrast to ToR, I2P does not need an input protector. ToR browser has a better performance compared with I2P except for the HTTP-GET-Request request. Increasing of users in anonymizing networks influence directly on ToR, I2P, and Riffle networks. Conflict and latency effect on the user experience and network usability, but the self-cover traffic for anonymity would be stronger. ToR is not well distributed like the I2P network. The routing in ToR is circuit-based while the routing in I2P drives the stack by implicitly load balancing and prevents any crash or delay in the system. This case is appropriate for high volume file transmission and could highlight I2P network. Two-step identity verification in Riffle makes this network safer and faster than another identity anonymizing network such as I2P and ToR networks. On the other hand, Riffle guarantees the transmitter anonymity by using PIR and protects the receiver's privacy. The most important requirements in these three networks are that the bandwidth supplying between the server is growing linearly given to the number of users. From functionality and performance point of view, Riffle would be better than other identity anonymizing networks for applications with compressed the bandwidth such as file transmission.

Onion Routing (OR), I2P, and Riffle are the anonymizing networks for tunneling issues. Low privacy frameworks can tunnel their information exchange by aforesaid networks. The main difference among the I2P and OR networks is periphery threat specimen and the exterior body of the proxies design. ToR is a directory based approach. So, it has a centralized point to lead the general network with information gathering and report abilities. This case is opposite of the other anonymizing networks that worked based on distributed DBMS. On the Other side, I2P has vulnerabilities such as traffic analysis. The attackers can analysis the traffic when the data came out from the mixed networks. This issue can be done by WATERHOLE and man-in-the-middle (MITM) attacks that made a suitable background to sniff the user's real time communications [34]. In forthcoming subsection, comparison between networks are performed and tabulated.

Comparison of ToR and I2P Terminology

The differences between the ToR and I2P idioms are described in Table 2. The comparison is determined based on the type of network, data transmission policy, kind of routing, etc. [35].

Note that, both I2P and ToR proxy performance have some drawbacks versus especial types of attackers. The used proxies are vulnerable to misuse besides several security penetrates. Although both of them have the same similarity in some cases, their utilizing terminology are rather different in which the deference idioms are tabulated.

Table 2: Differences between the ToR and I2P idioms [35].

Tor	I2P
Cellule	Packet
User	Customer
Circuit	Tunneling
Index	NetDb
Index Server	Router flood fill
Input Supervisor	counterpart
Entrance Point	Entrance Proxy
Egress Point	Egress Proxy
Hidden Service	Lease-Set
Primitive Point	Entrance Gate
Volunteer Users	Router
Onion Proxy	Gate
Point of assignation	Entrance Gateway and Egress point
Onion Service	Conceal the service

Comparison of ToR, I2P, and Riffle Terminology in Terms of Network Requirement

In addition to, a comprehensive comparison of requirement analysis in ToR, I2P, and Riffle networks is proposed on Table 3. The comparison is based on several network and distribution requirements.

As Table 3 shows, the Riffle outperforms against other two networks in terms of reliability, confidentiality, response time, and other network and distribution systems' requirement features.

Table 3: Comparison between ToR, I2P and Riffle in terms of analytical aspects

Network Type	ToR	I2P	Riffle
Requirements			
Reliability	safe	Insecure	more safe
Confidentiality	Confident	Unconfident	more confidently
Availability	accessible	accessible	inaccessible
Usability	Easy	Medium to hard	Normal
Reusability	Yes	Yes	Yes
Cost	free	Non-free	Non-free
Response Time	Medium	Faster	Too Fast
Functionality	Poor	Fast	Faster
Security	Confident	Low Security	Safe
Reputation	Popular	Less popular	Less popular
Performance	efficient	efficient	efficient
Accessibility	available	attainable	attainable
Scalability	unchangeable	changeable	changeable
Adaptability	Inconsistent	Inconsistent	consistent

Comparison of ToR, I2P, and Riffle Terminology in Terms of Analytical Aspects

Also, Table 4 is dedicated to a comparison between anonymizing ToR, I2P, and Riffle networks in terms of analytical aspects and each of literature which paid for.

In terms of analytical aspects, ToR, I2P, and Riffle have competition in some comparison parameters. For instance, in term of upload speed, Riffle is the best but in term of memory usage and utilization, the ToR beats other networks.

Table 4-a: Comparison between ToR, I2P and Riffle in terms of analytical aspects

Network Type	ToR	I2P	Riffle	Ref.
Operational infrastructure	User base	Server base	Server base	[9], [24], [26], [36]
Extra abilities	Hidden web	No extra abilities	Hidden web	[7], [9], [10], [12], [24], [25]
Funding	Considerable funding	Without funding	Limited	[9]-[11], [24], [26]
Developers	More Developers	Limited	Limited	[9]-[11], [24], [26]
transport layer	TLS and Bridge	TLS	TLS	[9]-[11], [24], [26]
Scalability	High	Low	Low	[9]-[11], [24]
Switching	Circuit switched	Packet switched	Packet switched	[9]-[11], [24]
circuits	bidirectional circuits	Unidirectional Tunnels	Unidirectional Tunnels	[9]-[11], [24]
Upload speed	Low	Medium	High	[9]-[11], [24], [26]
Prone to DoS attacks	Larger enough for prevention	Smaller enough to attack	Larger enough for prevention	[5], [6], [9], [10]
Documentation	efficient	inefficient	inefficient	[9]-[11], [3], [13]
Overhead	Low bandwidth	High bandwidth	Low bandwidth	[9]-[11], [24], [26]
Security Focus	Exit Node	Entire network	Sender node	[9], [10], [27], [36]

Table 4-b: Comparison between ToR, I2P and Riffle in terms of analytical aspects

memory usage	Efficient	Inefficient	Inefficient	[9]-[11], [24], [26]
Distributed	Decentralized	centralized	Decentralized	[9]-[11], [24], [26], [21], [36]
Complexity	Reduced	Increased	Increased	[9]-[11], [13], [36]
Latency	Lower	Higher	Lower	[9]-[11], [24], [27]
Throughput	Higher	Lower	Higher	[9]-[12], [24]
Programming platform	C	Java	Probably python	[9]-[10], [27], [36]
Transport protocols	TCP	TCP & UDP	TCP & UDP	[9],[10],[26],[13], [27]
organizing	Server- organizing	Self-organizing	Self-organizing	[9],[10], [24], [36]
Directory servers	Safe	Unsafe	Unsafe	[9],[10], [24], [27]
User friendly	Very satisfying	desirable	Desirable	[9]-[12], [24], [37]

In terms of analytical aspects, ToR, I2P, and Riffle have competition in some comparison parameters. For instance, in term of upload speed, Riffle is the best but in term of memory usage and utilization, the ToR beats other networks.

Table 5: Bilateral comparison between ToR and I2P

Benefits of ToR over I2P	The extremely great user base Answered some obstacle that I2P has not yet to address them Considerable funding More developers TLS transport layer and bridges High scalability and resistance to attacks Circuit switched bidirectional circuits planned and optimized for exit traffic Better documentation, specifications, better website, and translations efficient memory usage low bandwidth overhead Centralized control reduces complexity at each node that can efficiently address Sybil attacks high capacity nodes higher throughput lower latency C language platform
Benefits of I2P over ToR	speedy usage of hidden services than TOR Fully distributed self-organizing peers selection by continuously profiling and ranking performance unvarying and untrustable directory servers Small enough to prone the DOS attacks Peer-to-peer friendly Packet switched implicit transparent load balancing Resilience Unidirectional tunnels Protection against detecting client activity short-lived all peers participate in routing for others The bandwidth overhead of being a full peer is low Integrated automatic update mechanism Both TCP and UDP transports JAVA language platform

Although I2P has great features such as anonymity and speed up in file sharing, there are some problems with its application which cannot be neglected. The I2P is a good framework for companies and organizations' interaction, but this system will drop the packets in two cases: the former is according to technical test with making distances between the received and sent packets; the latter case happens once the length of message exceeds from a certain value. This can be considered an important issue in the above network. Table 6 compares ToR and I2P in implementation details such as in security, scalability, and interfaces.

Table 6: Bilateral comparison between ToR and I2P

Titles	TOR	I2P
Implementation	Easy	Difficult
Scalability	Medium	Suitable enough
Stability	Stable	Unstable
Server Expert (difficulty)	Easy	Medium
Client Expert (difficulty)	Easy	Medium
Security	Medium	Better (with considering attack prevention)
Startup Time	Less	More
GUI features	No	Yes
Speed	As Expected	As Expected
Error rate	Seen	Not Seen
URL Stability	Auto Changed	Not Changed

Investigation over three famous anonymizing networks ToR, I2P, and Riffle reveal that they make fortunities to hide network connections for the sake of security and confidentiality reasons. In some cases, they dissipate network bandwidth and sacrifice network performance to reach security objectives. Nevertheless, some proposed protocols and bandwidth optimization techniques can improve network quality of service (QoS) besides reaching concealing justifications and security objectives. Furthermore, since each anonymizing network has its own merits and demerits, the hybrid network within its protocols design which inherits all of plus points and excludes negative features is favorable.

Results and Discussion

This paper presented a profound comparison between three famous anonymizing ToR, I2P, and Riffle networks which tend to camouflage services from their users. To do so a subjective classification plan for comparison among anonymizing networks has been presented. The comparison has been done based on commonalities, discrepancies, and challenges in this field. This comparison is beneficial for further researches and future improvements to fill the existing gaps. However, by this review, this can be pointed out that the ToR is a popular network whilst other networks like Riffle and I2P are relatively considered novel alternatives. The mentioned systems are updated continuously for performance improvement and also providing more anonymity to protect the users. ToR and I2P differences are in preparation and using virtual communications. ToR network is a set of volunteer servers all over the world which are functioned for anonymous connection to browsing a web page or some particular operations. However, I2P provides anonymous file transmission between two peers. In data transmission, Riffle is faster than other identity anonymizing networks due to using two-step identity verification structure, and the traffic analysis is barely feasible. Also, Riffle has more efficiency because it provides a significant anonymity due to minimizing bandwidth and computation overhead. Totally, anonymizing networks can utilize efficient protocols and bandwidth optimization techniques that can potentially improve network QoS besides reaching concealing justifications and security objectives. Furthermore, as each anonymizing network has its own merits and demerits, the hybrid network within its protocols design which inherits all of plus points and excludes negative features is favorable.

Author Contributions

Dr. Mirsaeid Hosseini Shirvani was the supervisor of the current research plan. He sketched the research framework and the roadmap. Also, he analyzed the results and tabulated the outcome derived from excerpted literatures. In this line, Amir Akbarifar searched in authentic journals to gather all relevant papers. In addition to, he prepared the blueprint of the research plan. He and his supervisor cooperatively summed up the work.

Acknowledgment

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

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.

Abbreviation

I2P	Invisible Internet Project
ToR	the Onion Router

OR	Onion Router
DC	Dining Cryptographers
OP	Onion Proxy
TLS	Transport Layer Security
TCP	Transmission Control Protocol
SHA1	Secure Hash Algorithm 1
P2P	Peer-to-Peer
DNS	Domain Name Resolver
DC-nets	Dining Cryptographers
PIR	Private Information Recovery
UDP	User Datagram protocol
MITIM	Man-in-the-Middle attack
QoS	Quality of Service

References

[1] M.S. Hosseini Shirvani, A.M. Rahmani, A. Sahafi, “An iterative mathematical decision model for cloud migration: a cost and security risk approach,” Software Pract. Ex., 48(3): 449-485, 2018.

[2] M.S. Hosseini Shirvani, “To move or not to move: An iterative four-phase cloud adoption decision model for IT outsourcing based on TCO,” J. Soft Comput. Inf. Technol., 9(1): 7-17, 2020.

[3] M.S. Hosseini Shirvani, “Web Service Composition in multi-cloud environment: A bi-objective genetic optimization algorithm,” in Proc. 2018 Innovations in Intelligent Systems and Applications (INISTA): 1-6, 2018.

[4] The Center for Internet Security (CIS), “The CIS Security Metrics,” v1.0.0, 2010.

[5] B. Conrad, F. Shirazi, “A survey on Tor and I2P,” in Proc. ICIMP 2014: The Ninth International Conference on Internet Monitoring and Protection: 22, 2014.

[6] D. Goldschlag, M. Reed, P. Syverson, “Onion routing for anonymous and private internet connections,” Commun. ACM, 42(2): 39-41, 1999.

[7] G.H. Owenson, N.J. Savage, “The Tor darknet,” Global Commission on Internet Governance,” paper series no. 20, 2015.

[8] http://www.geti2p.net

[9] https://github.com/kwonalbert/riffle

[10] A. Kwon, D. Lazar, S. Devadas, B. Ford, “Riffle: An efficient communication system with strong anonymity,” Proceedings on Privacy Enhancing Technologies, 2: 115-134, 2016.

[11] F. Shirazi, M. Simeonovski, M.R. Asghar, “A survey on routing in anonymous communication protocol,” ACM Comput. Surv. (CSUR), 51(3): 1-39, 2018.

[12] B. Li, E. Erdin, M.H. Gunes, G. Bebis, T. Shipley, “An overview of anonymity technology usage,” Comput. Commun., 36(12): 1269-1283, 2014.

[13] E. Jardine, “The dark web dilemma: Tor, anonymity and online policing,” Global Commission on Internet Governance Paper Series 21, 2015.

[14] P.F. Syverson, D.M. Goldschlag, M.G. Reed, "Anonymous connections and onion routing," in Proc. IEEE Symposium on Security and Privacy (Cat. No.97CB36097): 44–54, 1997.

[15] M.W. Al-Nabki, E. Fidalgo, E. Alegre, L. Fernández-Robles, “ToRank: Identifying the most influential suspicious domains in the Tor network,” Expert Syst. Appl., 123: 212–226, 2019.

[16] Q. Yang, P. Gasti, K. Balagani, Y. Li, G. Zhou, “USB side-channel attack on Tor,” Comput. Netw., 141: 57–66, 2018.

[17] A. Kwon, M. AlSabah, D. Lazar, M. Dacier, S. Devadas, “Circuit fingerprinting attacks: Passive deanonymization of tor hidden services,” in Proc. 24th USENIX Security Symposium (USENIX Security 15): 287–302, Washington, D.C., USENIX Association, 2015.

[18] R. Attarian, L. Abdi, S. Hashemi, “AdaWFPA: Adaptive online website fingerprinting attack for Tor anonymous network: A stream-wise paradigm,” Comput. Commun., 148: 74–85, 2019.

[19] K. Sangeetha, K. Ravikumar, “Defense against protocol level attack in Tor network using deficit round robin queuing process,” Egyp. Inf. J., 19 (3) 199–205, 2018.

[20] B. Monk, J. Mitchell, R. Frank, G. Davies, “Uncovering tor: An examination of the network structure,” Secur. Commun. Netw., 2018: 1-13, 2018.

[21] A.S. Tanenbaum, “Distributed operating systems,” Pearson Education India, 1995.

[22] M.S. Hosseini Shirvani, N. Amirsoleimani, S. Salimpour, A. Azab, “Multi-criteria task scheduling in distributed systems based on fuzzy TOPSIS,” in Proc. IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, Canada, 2017.

[23] Y. Ramzanpoor, M. Hosseini Shirvani, M. Golsorkhtabaramiri, “Multi-objective fault-tolerant optimization algorithm for deployment of IoT applications on fog computing infrastructure,” Complex Intell. Syst., 1-32, 2021.

[24] G. Danezis, C. Diaz, “A survey of anonymous communication channels,” Technical Report MSR-TR-2008-35, Microsoft Research, Jan. 2008.

[25] R. Dingledine, N. Mathewson, P., Syverson, “ToR: The second-generation onion router,” in Proc. 13th conference on USENIX Symposium., 13: 1-21, 2004.

[26] D. McCoy, k. Bauer, D. Grunwald, T. Kohno, D. Sicker, “Shining light in dark places: Understanding the Tor network,” in Proc. the 8th International Symposium, PETS 2008 Leuven, Belgium, 2008.

[27] H.Y. Huang, M. Bashir, “Who is behind the Onion? Understanding Tor-Relay operators," CyLab Usable Privacy and Security Laboratory (CUPS), 2016.

[28] https://geti2p.net/en/docs/how/tech-intro.

[29] A. Crenshaw, “Common darknet weaknesses: An Overview of Attack Strategies,” DEFCON 19, Las Vegas, August 6, 2011.

[30] M. Wahal, T. Choudhury ”Anonymous network routing mechanism," in Proc. International Conference on Infocom Technologies and Unmanned Systems, Trends and Future Directions (ICTUS), 2017.

[31] P. Mayamounkov, D.M. Eres, “Kademlia: A Peer-to-peer information system based on the XOR Metric,” International Workshop on Peer-to-Peer Systems. Springer, Berlin, Heidelberg, 2002.

[32] H. Niedermayer, “Architecture and components of secure and anonymous peer-to-peer systems,” Ph.D. dissertation, Dept. Computer Science, Univ. Munich, 2010.

[33] W. Diffie, M. Hellman, "New directions in cryptography," IEEE Trans. Inf. Theory, 22(6): 644-654, 1976.

[34] M. Rouse, "What is watering hole attack?,” Retrieved March. 2015; 6: 2019.

[35] https://geti2p.net/en/comparison/tor

[36] M.G. Reed, P.F. Syverson, D.M. Goldschlag, “Anonymous connections and onion routing,” IEEE J. Sel. Areas Commun., 16(4): 482-494, 1998.

[37] Z.J. Newman “A high-bandwidth, low-latency system for anonymous broadcasting,” Diss. Massachusetts Institute of Technology, 2020.

Biographies

Mirsaeid Hosseini Shirvani received his B.Sc., M.Sc., and Ph.D. all in Computer Software Engineering Systems at Universities in Tehran, IRAN. He has been teaching miscellaneous computer courses in several universities in Mazandaran province of IRAN since 2001. He also published several papers in authentic and worldwide well-reputed journals. Currently, he is an Assistant Professor in Computer Engineering Department at IAU (Sari-Branch). His research interests are in the areas of cloud computing, fog computing, IoT, distributed systems, parallel processing, machine learning, and evolutionary computation.

- Email: mirsaeid_hosseini@iausari.ac.ir
- ORCID: 0000-0001-9396-5765
- Web of Science Researcher ID: AAO-2012-2021
- Scopus Author ID: 55459128300
- Homepage: NA

Amir Akbarifar received his B.Sc. and M.Sc. in Computer Software Engineering Systems in Islamic Azad University in IRAN. Currently, He is a PhD candidate in Computer Engineering Department at IAU (Sari-Branch). He has published numerous articles in prestigious magazines in Iran. Amir is information Security Analyst and his fields of study include: Security, Software Architecture, Artificial Intelligence, Deep Learning, Quantum Computing, and Cryptography.

- Email: msc.akbarifar@gmail.com
- ORCID: 0000-0003-3227-2847
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

M. Hosseini Shirvani, A. Akbarifar, “A comparative study on anonymizing networks: TOR, I2P, and riffle networks comparison,” J. Electr. Comput. Eng. Innovations, 10(2): 259-272, 2022.

DOI: 10.22061/JECEI.2021.8027.466

URL: https://jecei.sru.ac.ir/article_1630.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Polar Formed Histogram of Fast Fourier Transform Feature for Time Series Classification Problems

*K. Kiaei, H. Omranpour**

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

Article Info

Article History:

Received 06 July 2021
Reviewed 15 August 2021
Revised 25 October 2021
Accepted 01 November 2021

Keywords:

Time series
Classification
Fast Fourier transform
Polar form
Random forest classifier

*Corresponding Author's Email Address:

h.omranpour@nit.ac.ir

Abstract

Background and Objectives: Time series classification (TSC) means classifying the data over time and based on their behavior. TSC is one of the main machine learning tasks related to time series. Because the classification accuracy is of particular importance, we have decided to increase it in this research.

Methods: In this paper, we proposed a simple method for TSC problems to achieve higher classification accuracy than other existing methods. Fast Fourier transform is a method that uses in raw time series data preprocess. In this study, we apply the fast Fourier transform (FFT) over the raw datasets. Then we use the polar form of a complex number to create a histogram. The proposed method consists of three steps: preprocessing using FFT, feature extraction by histogram computation, and decision making using a random forest classifier.

Results: The presented method was tested on 12 datasets of the UCR time series classification archive from different domains. Evaluation of our method was performed using k-fold cross-validation and classification accuracy. The experimental results state that our model has been achieved classification accuracy higher or comparable than related methods. Computational complexity has also been significantly reduced.

Conclusion: In the latest years, the TSC problems have been increased. In this work, we proposed a simple method with extracted features from fast Fourier transforms that is efficient to gain more high accuracy.

Introduction

The amount of observations of a variable in many real-world applications depends on its value in earlier times. Such data are referred to as time series, that various researchers have used data mining techniques to predict and analyze such data.

The first step for time series identification and their nature determination is their past study. It can lead to an accurate and reliable prediction about its class [1].

Classification is a task that can be useful in many domains and applications. Any activity that requires thinking and judgment can be a matter of classification [2]. Medical imaging classification is an important classification problem because it plays a critical role in improving the diagnosis of many diseases [3], [4].

On the other hand, every problem using data that is registered taking into account some notion of order can be cast as a TSC problem [5].

In many real applications uses such as recognizing motion imagination by EEG signals, human activity classification based on sound recognition, not only sounds from human activities but also sounds from related objects [6], model prediction with electronic health record data is anticipated to drive personalized medicine and improve healthcare quality. If successful it could provide significant benefits not only for patient

Doi: 10.22061/JECEI.2021.8048.468

safety and quality but also in reducing healthcare costs [7].

Different classification methods have been proposed so far, but what separates TSC problems from usual classification problems is that the order of features in time series is essential. For this reason, must use the methods that retain the best classification features to obtaining acceptable accuracy. The best classification features may be distorted by noise or preprocessing. As a result, classifying time series requires special techniques.

In general, solutions to TSC problems are separated into two separate groups: shape-based and structure-based. In solving shape-based problems, classifiers are classically used in combination with a similarity criterion. But the tendency in solving TSC problems is to discover small patterns to be representation time series classes. Between these patterns, shapelet-based techniques [8]-[10] have gained good classification accuracy. Structure-based methods extract feature vectors from time series, such as feature patterns, or convert a time series into various representations, which often have a high computational cost.

Preliminaries

In this section, we hold forth a few preliminaries and sum up the meaning of frequency notations in Table 1.

Table 1: Notations and their meaning

Notation	Meaning
T	Time series $T = (t_1, t_2, \dots, t_n)$ of the length n
D	Time series T_j without the class label
Z	FFT of $t_{ij}, 1 \le i, j \le n$
F	Fourier coefficient matrix of dataset D
C	the label set of dataset D
R	length of the Z
θ	The angle between Z and the real axis
H	Histogram matrix
n	Radius matrix is divided into n equal parts.
m	Angle matrix is divided into m equal parts.

A time series T is an ordered-value sequence $T = (t_1, t_2, \dots, t_i, \dots, t_n)$, where n is the length of T, t_i is the value observed at timestamp i.

A time series dataset D is a set of time series T_j with C_j = label(T_j); $j \in [1; m]$, where $C_j \in C$ is the class label of the dataset, $C = \{0,1,2, \dots, |C| - 1\}$, and $|C|$ denotes the number of classes.

The discrete Fourier transform is the primary transform used for numerical computation in digital signal processing. It a sequence of N complex numbers $\{x_n\} \coloneqq x_0, x_1, \dots, x_{n-1}$ into another sequence of complex numbers, $\{X_k\} \coloneqq X_0, X_1, \dots, X_{n-1}$ which is defined by:

$$\mathbf{X_k} = \sum_{j=0}^{n-1} x_j e^{\frac{-2\pi i}{n} kj} \qquad (1)$$

These implementations usually employ efficient fast Fourier transform (FFT) algorithms. Applications of Fourier transform include data filtering [11], pattern extraction, and audio processing [12].

Related Work

The critical characteristic that separates TSC issues from the general classification task is that the ordering of the attributes is significant. The best discriminative features for classification may be concealed by the length of the series, embedded in the interaction of observations, or confounded by noise in the phase of the series. Hence, TSC, by and large, need strategies explicit to the nature of the issue.

There are many techniques for the TSC issue. [13] is a great review article that interested people can refer to it. In this section, we mainly survey two significant representative methods, called structure-based methods, and shapelet-based methods. This paper undertakes the structured-based approach with FFT.

A. Shapelet-Based Method

The shapelet-based approach utilizes some similarity criteria to gauge the distance between time series on raw numeric data.

A customary 1NN classifier based upon Euclidean distance [14] or DTW [15] distance is an example of this category. In [16] introduced time series shapelets and proposed the seminal work of it. Algorithms based on the time series shapelet can be interpretable but need high computation time. ST [17] suggests selecting shapelets by extracts the k best shapelets from a dataset in a single pass to represent the original time series. SVM and Neural Networks can be applied to determine shapelets. SD [18] proposed a quick shapelet discovery method that prunes the candidates based on a distance threshold to previously considered other similar candidates. FS [10] proposed an algorithm for shapelet discovery that is fast with transforming the raw time series data into SAX words. The run time is reduced when compared to [16]. However, accuracy is lower than other works [13]. SAX-VFSEQL [19] creates a pool of SAX words as the shapelets and iteratively optimizes the weights of the shapelet. However, this method still cannot produce shapelets, which there isn't in training data. Too, the discovered shapelets are too various and redundant, which makes interpreting the classification decision difficult.

Shape-based methods can achieve high accuracy, but they have a lot of computational costs. So a few approaches are proposed to progress efficiency, such as early abandoning [20]. In addition, shape-based methods are sensitive to noisy data, particularly for long time series.

B. Structured-based method

There are a few Structure-based methods such as Piecewise Aggregate Approximation (PAA) [8], Symbolic Aggregate approximation (SAX) [19], [21], Symbolic Fourier Approximation (SFA) [22], Discrete Fourier Transform (DFT) [23] and other discretization techniques. The dictionary-based method is another significant example. These techniques extract higher-level feature vectors or build a model from the time series before the classification task using classification algorithms such as KNN, SVMs, or random forests. Bag-of-Patterns (BOP) [24] builds a classifier from SAX, calculating the Euclidean distance with the BOP representation of a new instance to specify its class. Another approach, SAXVSM [25], adopts SAX techniques, proposes a vector space model (VSM), and uses TF-IDF to rank time series patterns. In [19], combine various variable-length bag-of-symbolic-words representations and an efficient linear sequence learning approach (SAX-VSEQL [26]) for efficient TSC.

Structure-based methods have resistant to noisy data because of the smooth impact of time series representations. Because of the shortening of representations, these methods are more efficient than shape-based methods.

Some algorithms like BOSS VS[27], which transforms data from the time domain to the frequency domain, give high accuracy, but they can't interpret the classification decision.

Proposed Method

Details of the proposed method are provided in this section. Our method consists of 3 phases. In the first phase, we perform preprocessing on the raw dataset. In the second phase, we prepare a histogram of new data, and in the last phase, we obtain the accuracy of our classifications by using the appropriate classifier. The overview of the method showed in Fig. 1.

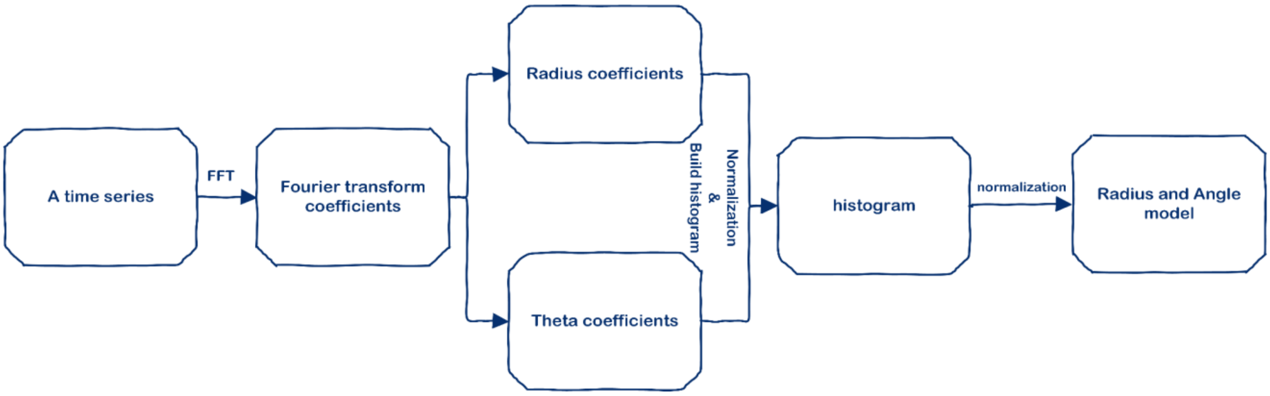


Fig. 1: The overview of our model.

A. Preprocessing

I) Fourier transform

A quick look at a time series shows us that it may be hard to interpret. It's because the underlying pattern behind the jagged peaks and troughs identify hardly in its raw form. Conversions on time series like scaling, shifting, time scaling, time-shifting, and dynamic time warping facilitate the discovery of flexible time series patterns. The high computational efficiency of Fourier transform has made Fourier analysis the preferred method for many applications. Fourier transform-based methods are very well used in all fields of science and engineering [28].

Therefore, in this paper, we use FFT on raw data in the preprocessing step.

In this case, we have a matrix of complex numbers (F) whose rows and columns are the original dataset size.

After applying FFT, we have a complex number $Z = a + bi$ in all elements of matrix F. Fig. 2 shows a complex number representation.

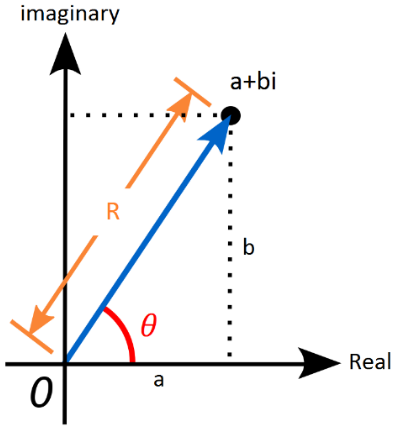


Fig. 2: Complex number representation.

The complex number can also be represented in polar form by (2).

$$Re^{i\theta} \ or \ R(cos\theta + isin\theta) \quad (2)$$

As shown in Fig.2, R is the length of Z and θ is the angle between Z and the real axis. R and θ are obtained from formulas (3) and (4).

$$R = |Z| = \sqrt{(a^2 + b^2)} \quad (3)$$

$$\theta = \begin{cases} \arctan\left(\frac{b}{a}\right) & if a > 0 \\ \arctan\left(\frac{b}{a}\right) + \pi & if a < 0 \ and \ b \geq 0 \\ \arctan\left(\frac{b}{a}\right) - \pi & if a < 0 \ and \ b < 0 \\ \frac{\pi}{2} & if a = 0 \ and \ b > 0 \\ -\left(\frac{\pi}{2}\right) & if a = 0 \ and \ b < 0 \\ indeterminate & if a = 0 \ and \ b = 0 \end{cases} \quad (4)$$

In (3) and (4), a is the real part, and b is the imaginary part of a complex number. Now we derive the radius and angle from each element of F. In this case, we have two matrices, radius matrix and angle matrix, with the dimensions of the original matrix, one of which stores the R values and the other the θ values (Fig. 3).

II) Normalization

Normalization is a good technique used in situations where we do not know how to distribute the data. Normalize data is suitable when the data are at different scales. In addition, the method we use doesn't have any assumptions regarding our data. Change values without eliminating the difference between them are the goal of normalization. Normalization cause converting all the numeric data to the desired scale.

Equation (5) is the general formula for converting data to a range between 0 and 1:

$$x_{normalized} = \frac{x - min(x)}{max(x) - min(x)} \quad (5)$$

Now all elements of radius and angle matrices are between 0 and 1.

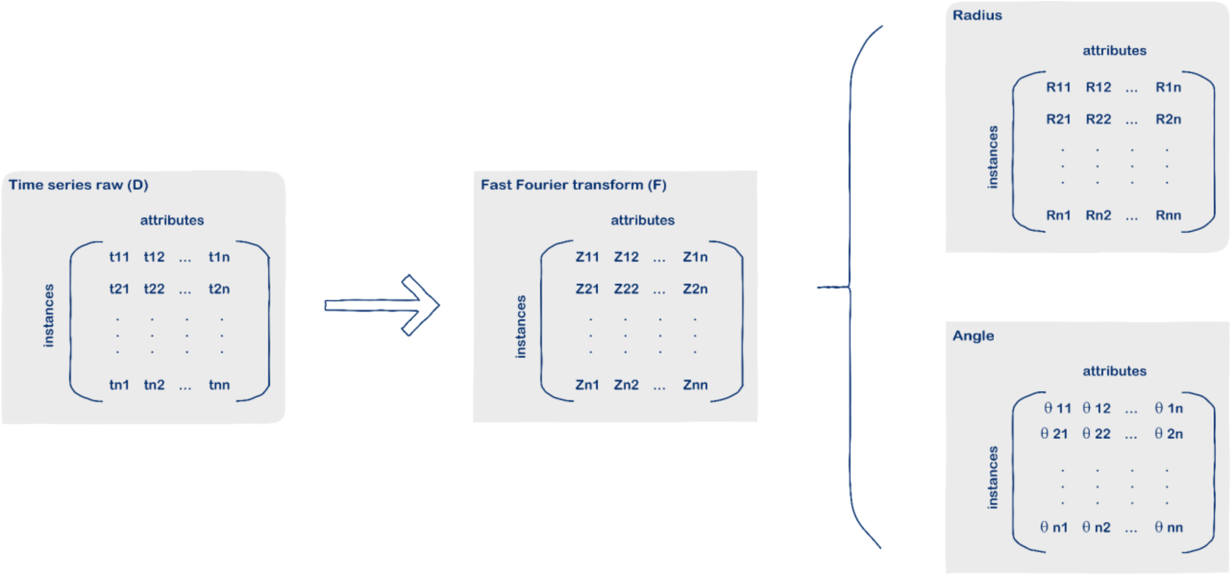


Fig. 3: Derive the radius and angle from F.

B. Histogram computation

A histogram gives us a visual interpretation of numerical data by appearing the count of data points inside a particular range called a bin. To histogram computation, we divided radius and angle matrices into n and m equal parts, respectively. Radius and angle matrices are normal, and the range is between 0 and 1, so the bin size for the radius matrix will be $\frac{1}{n}$, and the bin size for the angle matrix will be $\frac{1}{m}$. This work will usually reduce dimensions.

Equation (6) specifies each Z falls within which bin. We have to do this for all F elements.

$$\mathbf{\frac{j-1}{n} \leq R \leq \frac{j}{n} \quad \& \quad \frac{k-1}{m} \leq \theta \leq \frac{k}{m}} \quad (6)$$
$$\mathbf{for \ 1 \leq j \leq n \quad \& \quad 1 \leq k < m}$$

n and m are variables in (6). It is clear larger n and m, fewer data to be lost, but it also causes the uniformity of the histogram. That means the effect of the proposed feature will be reduced. The best n and m values that lead to high accuracy are listed in Table 3.

For implementation, we consider the matrix H, whose number of rows is the number of original dataset rows and the number of columns is $n * m$. Set the initial value of all H elements to 0. According to Algorithm 1, we have to place each element of row i of the original dataset in one of the rows i columns of the histogram. That means in row i and column j of the original dataset, we add a unit to the value of the element in row i and column $(j-1) * m + k$ from H. The result of this operation is the histogram calculation (Fig. 4).

```
Algorithm 1: Histogram computation
Input: $R_{1,:}, \theta_{1,:}, n, m$
Output: $H_{1,:}$

Initialize H = 0, c=size(R, 2)
For i=1 to c
    For j=1 to n
        For k=1 to m
            If$\left(\frac{j-1}{n} <= R_{1,i} \,\&\, R_{1,i} < \frac{j}{n}\right) \& \left(\frac{k-1}{m} \le \theta_{1,i} \,\&\, \theta_{1,i} < \frac{k}{m}\right)$
                $H_{(j-1)*m+k} ++$
            End If
        End For
    End For
End For
Return $H_{1,i}$
```

C. Classifier

The classifiers used in this article are K-nearest neighbor for simple implementation, fast and efficient [29], support vector machine for the ability to deal with various classification problems like not linearly separable problems and high dimensional [30], and random forest for power, accuracy, can also be used to rate the importance of features in classification or regression problems [31]. Among these classifiers, the results of random forest were better than the others. Only the results of the random forest classification are given in the next section.

```
Algorithm 2: Feature-extraction
Input: time series dataset D
Output: H

Initialize $L = size(D, 2), r = size(normalR, 1)$
For $i = 1$ to $L$
    $A = FFT(D_{i,:})$
    $R_{i,:} = |A|$          #Absolute value
    $\theta_{i,:} = arg(A)$   # argument
End For
$normalR = (R - min(R))/(max(R) - min(R))$
$normal\theta = (\theta - min(\theta))/(max(\theta) - min(\theta))$
For $i = 1$ to $r$
    $H_{i,:} = Histogram-computation(normalR_{i,:}, normal\theta_{i,:}, n, m)$
End For
Return H
```

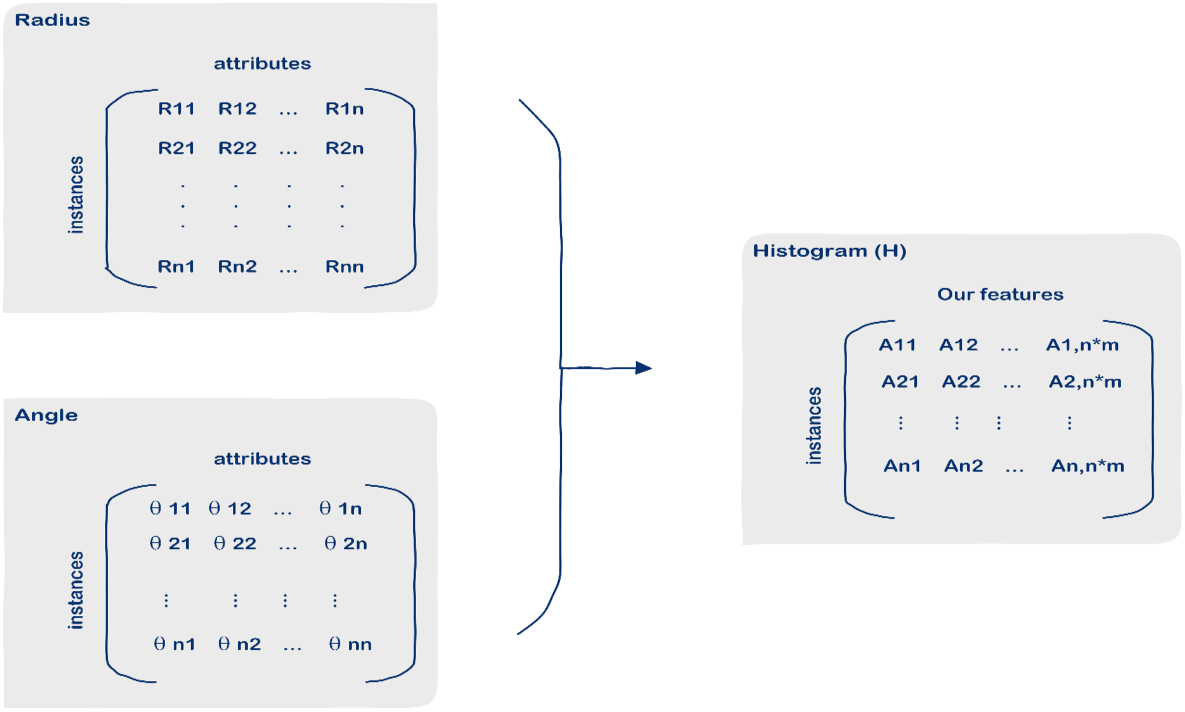


Fig. 4: The histogram matrix.

Results and Discussion

In this section, we show the results of the proposed method on the benchmark dataset used for TSC problems. The algorithm of this article is implemented with MATLAB. All the state-of-the-art methods were conducted on a machine with a Xeon E5-2630 v4 @ 2.2GHz (2S/10C) / 256GB RAM / 128GB SWAP, running on CentOS 7.6 (64-bit). Also, the proposed method was run on a computer with an Intel Core i7-4720HQ 4 x 2.6/ 16GB RAM DDR3-1600/ 256GB SSD, running on windows 8.1.

We can adjust the histogram computation parameters easily since they are efficient at using memory. Cross-validation has been used to evaluate performance.

We compared this method with a few baselines, including BSPCOVER [32] that is both efficient and accurate, RotForest [13] and 1NN-DTW [15] that used as benchmark classifier, and ST [17], ELIS [8], Fast shapelets [10], Scalable Shapelet Discovery [18] that are shapelet-based methods and COTE [33].

A. Datasets

This paper used the UCRARCHIVE [34], the important resource of TSC datasets, to evaluate the model.

We just chose 12 datasets from the various types like Image (4), Motion (2), Spectro (2), Stimulated (2), Sensor (1), ECG (1) because of space limitations. The number of each type is shown in parentheses.

Table 2 shows us information about each dataset, such as the number of instances in the train and test set, number of features, number of classes.

B. Experiments

The classification accuracy results of previous methods and proposed method on 12 datasets are given in Table 4.

I. Analysis of the proposed idea parameters

By repeating the experiments, we tuned the best value for the parameters of the proposed model. In each dataset, we tested numbers smaller than 50 with an interval of 5. Increasing one or two units has little effect on improving accuracy, and this distance allows us to find the best values in the fastest possible time. The values n and m are selected from (7).

$$\boldsymbol{n, m = 5i} \qquad \boldsymbol{i = 1, 2, \ldots, 10} \tag{7}$$

Table 2: Datasets

Dataset	Train set	Test set	Length	Class	Type
BeetleFly	20	20	512	2	IMAGE
ChlorineConcentration	467	3840	166	3	SIMULATED
Coffee	28	28	286	2	SPECTRO
DiatomSizeReduction	16	306	345	4	IMAGE
DistalPhalanxOutlineCorrect	600	276	80	2	IMAGE
Earthquakes	322	139	512	2	SENSOR
ECG5000	500	4500	140	5	ECG
Haptics	155	308	1092	5	MOTION
InlineSkate	100	550	1882	7	MOTION
Mallat	55	2345	1024	8	SIMULATED
Meat	60	60	448	3	SPECTRO
Symbols	25	995	398	6	IMAGE

Table 3: Best values of m and n

dataset	BeetleFly	Chlorine Concentration	Coffee	DiatomSizeReduction	DistalPhalanxOutline Correct	Earthquakes	ECG5000	Haptics	InlineSkate	Mallat	Meat	Symbols
n	10	20	15	35	5	5	10	10	5	20	25	15
m	5	10	10	5	20	20	10	15	10	20	40	10

Table 4: Classification accuracy of the proposed method and state-of-the-art methods on UCRARCHIVE

Dataset	ST – 2012	FS – 2013	COTE – 2015	SD – 2016	DTW Rn 1NN – 2017	RotF – 2017	ELIS – 2018	ResNet – 2019	BSPCOVER – 2020	Proposed method
BeetleFly	90	70	80	75	65	90	85	85	90	90
ChlorineConcentration	69.97	54.64	72.71	55.3	65	84.74	27.39	84.4	61.22	90.08
Coffee	96.43	92.86	100	96.1	100	100	96.43	100	100	98.21
DiatomSizeReduction	92.48	86.6	92.81	89.6	93.46	87.25	89.86	30.1	87.25	96.89
DistalPhalanxOutlineCorrect	77.54	75	76.09	71.7	72.46	75.72	57.83	77.1	83.17	98.58
Earthquakes	74.1	70.5	74.82	63.6	72.66	74.82	77.64	71.2	81.68	80.69
ECG5000	94.38	92.27	94.6	92.4	92.51	94.58	72.69	93.4	94.44	93.6
Haptics	52.27	39.29	52.27	35.6	41.56	43.83	41.56	51.9	45.13	50.11
InlineSkate	37.27	18.91	49.45	38.5	38.73	37.09	35.46	37.3	38.73	51.23
Mallat	96.42	97.61	95.39	92.6	91.43	94.93	81.58	97.2	76.8	96.71
Meat	85	83.33	91.67	93.3	93.33	96.67	55	96.8	75	99.17
Symbols	88.24	93.37	96.38	90.1	93.77	79.3	78.29	90.6	93.37	93.14
Total best acc	2	1	4	0	1	2	0	1	3	6
Rank mean	4.16	6.58	3	6.5	4.91	4.33	7.08	4	4.25	2.08
final rank	4	9	2	8	7	6	10	3	5	1

In the training phase, the best values of n and m are chosen, given in Table 3.

Fig. 5 shows the classification accuracy for different values of n and m parameters in each dataset examined in this paper. According to the results, increasing n and m increases the accuracy, but this does not mean that this always happens. Not only do very high values of n and m not help improve accuracy, but they also reduce the effect of histograms on data classification, resulting in less accuracy.

The highest values are different for each dataset, but on average, $n, m = 5i \quad i = 1,2,3,4$ have the best results.

II. *Analysis of the proposed features*

As described in the previous section, our method is applied in 3 steps: preprocessing, feature extraction, and classifier. To ranking the importance of features in classification issues can use a random forest that rating features based on two criteria called Mean Decrease Impurity (MDI) [35] and Mean Decrease Accuracy (MDA) [35]. MDI is based on the total decrease in node impurity from splitting on the variable, averaged over all trees. Also, MDA is based on the concept that if a feature is not significant, so by changing its value, the accuracy of the prediction should not decrease [31]. To evaluate this effectiveness method, we measured the importance of our features after applying a random forest classifier to the training and test data set.

a) Beetle Fly

Fig. 6 shows how our prepared method discriminates the different classes of the BeetleFly dataset. Fig. 6 (a) a time series plot is presented to visualize data points at consecutive time intervals. To make it clear, we consider five random samples from each class. We extracted ten important features from our model [35] and showed them in Fig. 6 (b), respectively.

The three most important are features 8, 6, and 3, respectively. We show the scatter plot by combining these three properties in Fig. 7 (a)-(b)-(c) that tell us the proposed feature was able to separate the two classes to increase the accuracy of the classification.

b) Chlorine Concentration

Fig. 8 shows how our prepared method discriminates the different classes of the Chlorine Concentration dataset.

Fig. 8 (a) A time series plot is presented to visualize data points at consecutive time intervals. To be clear, we consider five random samples from each class. Fig. 8(b) shows ten features of the model that are more important in predicting data class [35]. The diagram of random forest classification on our model is shown in Fig. 9 that tells us the data of one class have different values than other classes, and they are similar to themselves in the proposed properties. Therefore, it can be concluded that the proposed feature can be a convenient feature for TSC problems. The accuracy we obtained for this dataset is 90.08, significantly higher than all the methods reviewed. In this dataset, after obtaining the importance of each feature [35], we added them to the model in the most important. After adding each feature, we measured the accuracy using the KNN and random forest classifiers. In Fig. 10, we showed the results of the first 100 features.

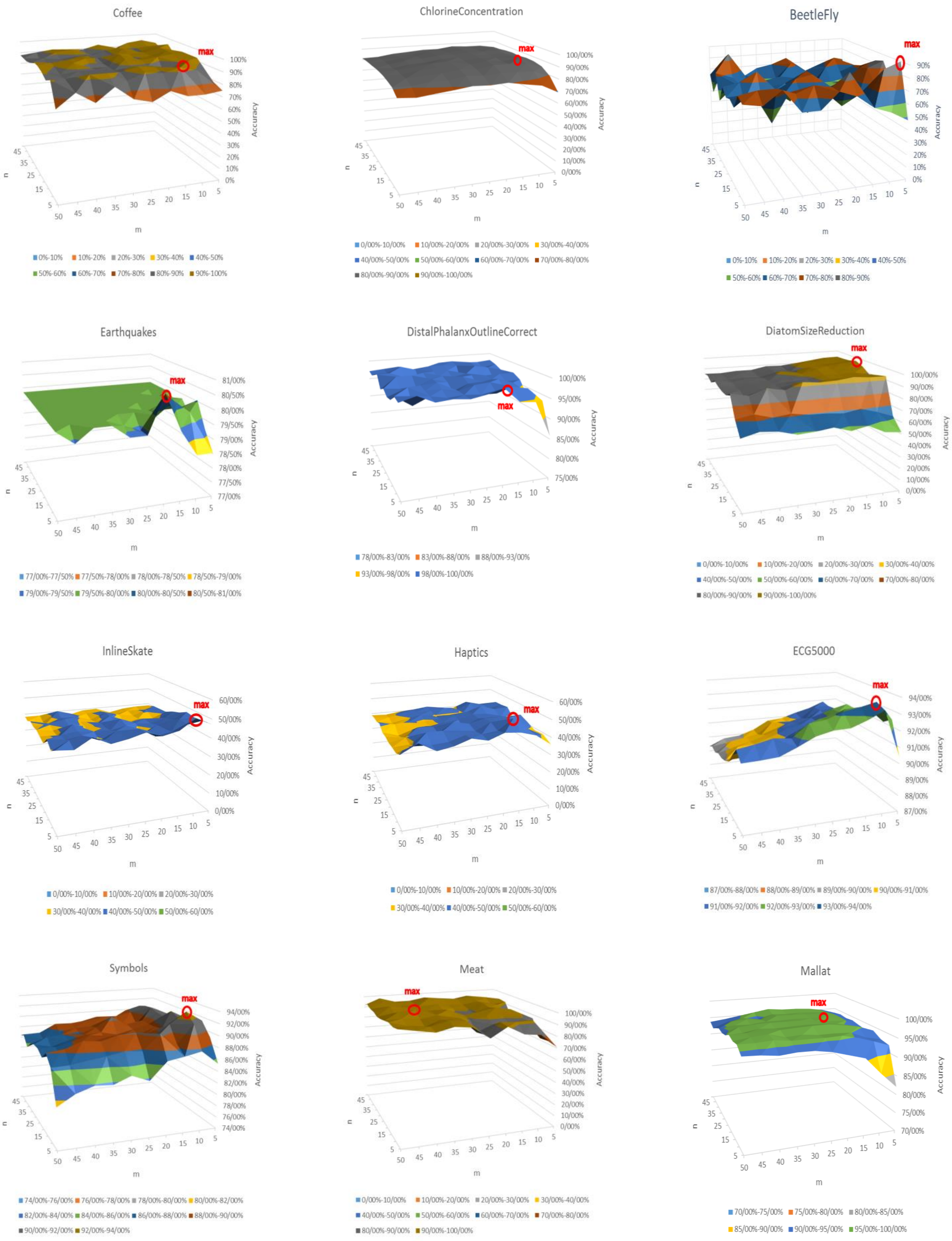


Fig. 5: Accuracy by varying n and m on different datasets.

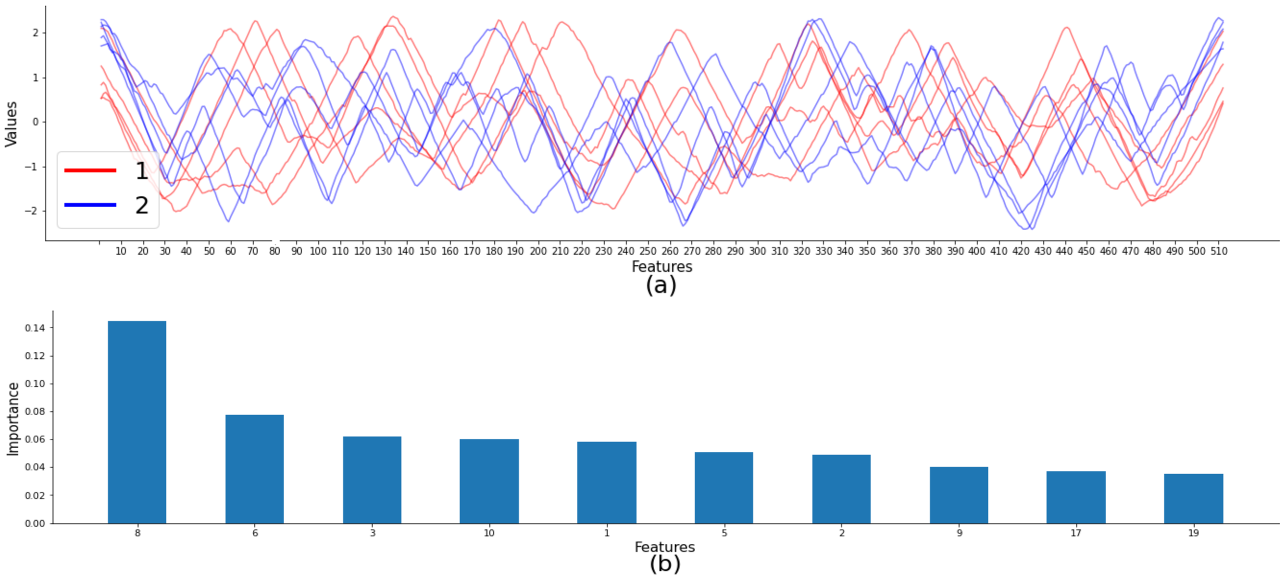


Fig. 6: BeetleFly dataset; (a) Train and test data. To make it clear, five random samples were taken from each class. (b) Feature importance obtained from random forest classifier.

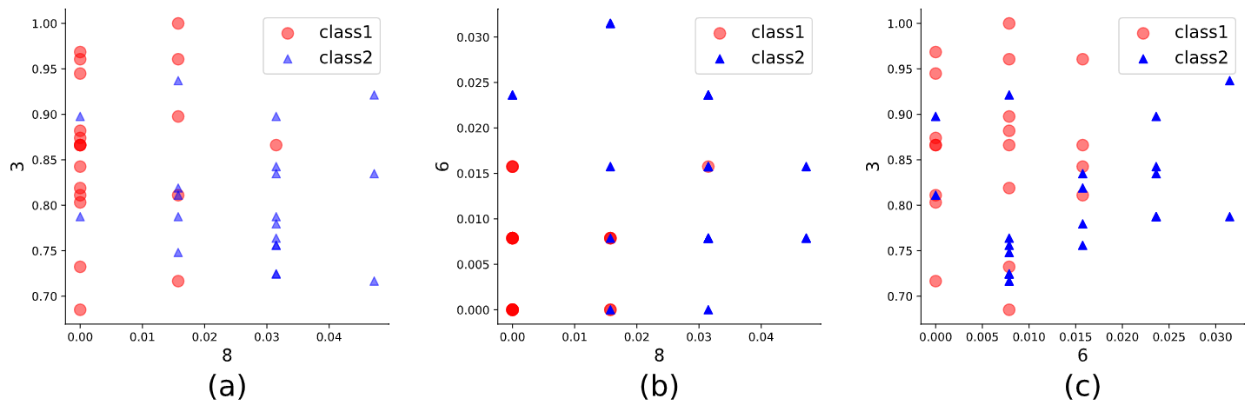


Fig. 7: Scatter plot of BeetleFly dataset; (a) Scatter plot by features 8 and 3. (b) Scatter plot by features 8 and 6. (c) Scatter plot by features 6 and 3.

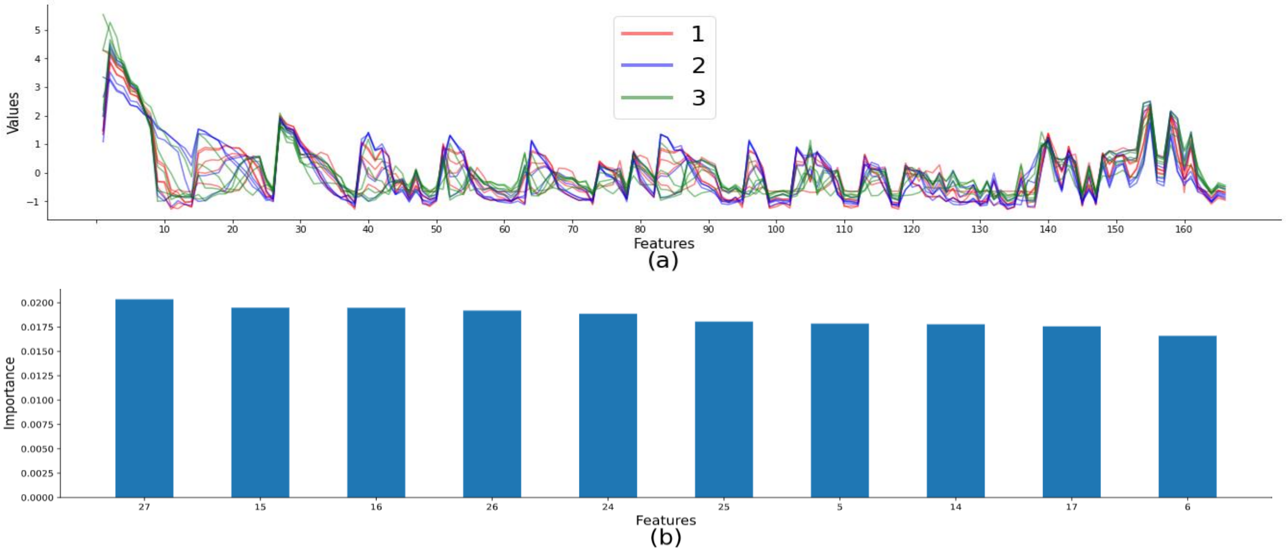


Fig. 8: ChlorineConcentration dataset; (a) Train and test data. To make it clear, five random samples were taken from each class. (b) Feature importance obtained from random forest classifier.

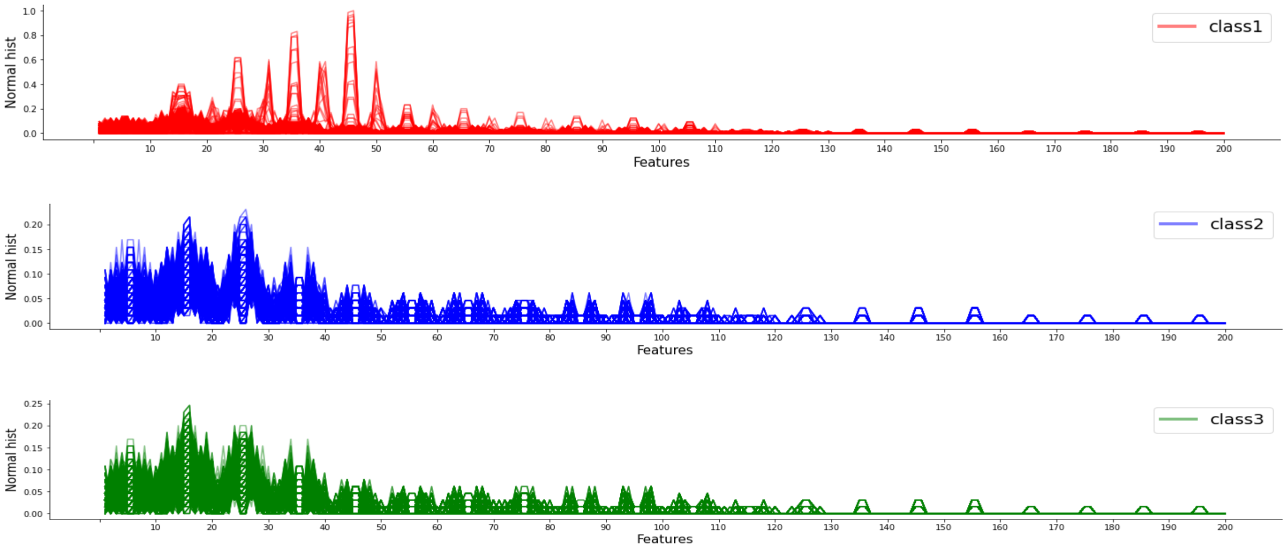


Fig. 9: The diagram of random forest classification in each class.

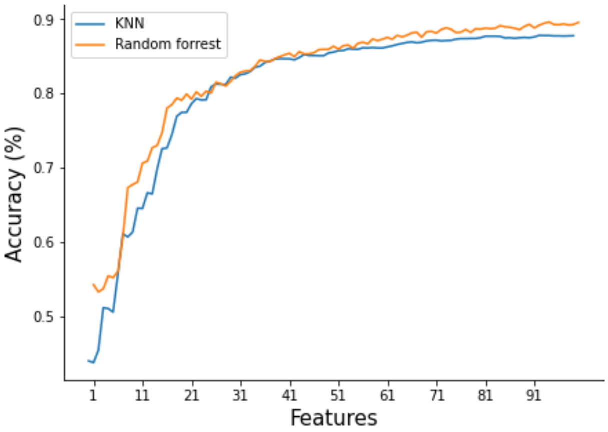


Fig. 10: Accuracy of the 100 most important features on Chlorine Concentration dataset.

As can be seen in Fig. 10, with the addition of features one by one, the accuracy is ascending, which is a good indication of the effect of the features of our proposed model on the final accuracy. With only 40 out of 200 features, we achieved higher accuracy than all the methods studied.

III. Comparison with state-of-the-art

The proposed method is compared with nine different methods that have the best performance. We will only briefly introduce each method here due to space constraints. Interested readers can refer to the main article for more details.

- DTW-Rn-1NN and RotF [13]. The two benchmark classifications are more competitive than many methods, so we added these two to our comparison.

Shapelets are a group of methods that focus on discovering short patterns that define a, and yet can appear anywhere in the series. These independent phase patterns are commonly identified as shapelets. A class is then distinguished by the presence or nonpresence of at least one shapelet someplace in the entire series. [17], [10], [18], and [8] are shapelet-based algorithms.

- Shapelet Transformation (ST) [17]. This algorithm combines a weighted ensemble of standard classifiers that new time series are classified with a weighted vote.
- Fast Shapelets (FS) [10]. FS combines SAX words and random masking techniques to raise efficiency.
- Scalable shapelet Discovery (SD) [18]. SD filters out candidates improving classification with an online clustering/ pruning technique. The dimensionality reduction ratio r is set to 1=2, and pruning distance percentile p is 25 for simplicity if there is no support from.
- ELIS [8]. ELIS is the present efficient shapelet-based algorithm. It utilizes PAA and TF-IDF to improve the efficiency of finding shapelet candidates. The logistic regression classifier is applied to adjust the shapelets. This paper follows ELIS [8] to adjust parameters. Otherwise, the parameters are set as follows: the iteration number is 1000, the regularization λ=0.01, and the learning rate η=0.1.
- COTE [33]. COTE is a combination of 35 TSC algorithms, including EE and ST and actually is the meta ensemble method. COTE works based on two principle ideas. First, the easiest approach to improvement for TSC issues is to convert into a new data space where discriminatory features identify simplify. Second, with a single data representation, improved accuracy could be achieved through simple ensemble schemes.
- ResNet [36]. The primary Property of ResNet is the residual shortcut connection between consecutive

convolutional layers. The network includes three residual blocks followed by a GAP layer and a final softmax classifier whose number of neurons is equal to the number of classes in a dataset. Each residual block comprises three convolutions whose output is added to the residual block's input and then fed to the next layer.

- BSPCOVER [32]. Creates many candidates by Symbolic Aggregate approximation with sliding window, then prunes identical and highly similar candidates by Bloom filters and similarity matching, respectively.

a) Accuracy

The experiment accuracy results for the method proposed in this paper and related methods are measure by cross-validation. The accuracy results for 12 datasets are presented in Table 4. On the dataset tested in this paper, our method is more accurate than other methods and even slightly ahead of COTE, which has been the most accurate classifier ever. Our proposed method has the highest classification accuracy among 12 datasets, so that with an average rank of 2.08, it is in the first final rank. The ranking of the state-of-the-art methods is shown in the last row of Table 4.

b) Efficiency

According to the results of [8], it can be seen that ELIS is about twice as fast as LTS. On the other hand, according to the results of [13], it can be understood that LTS is faster than COTE and ST. Therefore, we compare FS, SD, ELIS, and BSPCOVER with the proposed method for runtime evaluation.

We present the runtime of FS, SD, ELIS, BSPCOVER, and our method, in Table 5. (The unit of the numbers is seconds, and we use h and d to denote for hours and days, respectively.)

According to the results, both FS and SD methods have less execution time compared to the others, but on the other hand, a significant drop in accuracy than the others. The ELIS method has the longest runtime of all the methods and sometimes even takes a few days. The BSPCOVER method has both good runtime and high accuracy.

Table 5: Runtime of the proposed method and state-of-the-art methods on UCRARCHIVE

Dataset	FS	SD	ELIS	BSPCOVER	Proposed method
BeetleFly	13.96	1.62	122.87	42.92	578
ChlorineConcentration	128.56	1.82	31.8h	173.86	2.44h
Coffee	4.1	0.98	53.3	10.96	594
DiatomSizeReduction	4.33	1.02	749.03	30.04	815
DistalPhalanxOutlineCorrect	15.2	4.05	1.1h	52.39	1302
Earthquakes	1152.06	39.75	2.7h	2957.36	1244
ECG5000	36.89	6.95	7.5h	600.37	2.26h
Haptics	2086.17	6.45	3.6d	3.2h	1773
InlineSkate	2.1h	5.42	8.5d	4.2h	2627
Mallat	419.98	4.34	3.9h	2896.15	1.62h
Meat	34.62	1.94	2975.9	44.02	752
Symbols	16.17	1.04	2315.93	90.43	2387

Table 6: Wilcoxon signed rank test results on UCRARCHIVE

Proposed model VS	p-value	Test number	Superior	Inferior	Similar
ST	2.45E-02	12	9	2	1
FS	6.71E-03	12	10	2	0
COTE	8.05E-02	12	8	4	0
SD	5.37E-03	12	12	0	0
DTW Rn 1NN	1.24E-02	12	10	2	0
RotF	1.66E-02	12	9	2	1
ELIS	3.6E-03	12	12	0	0
ResNet	2.08E-02	12	9	3	0
BSPCOVER	5.79E-02	12	7	4	1

Although the method presented in this paper has a longer runtime compared to other related methods, but because the runtime is not very long in any of the datasets, the average runtime of this method is close to BSPCOVER. For example, the average runtime for our method is 2906 seconds and for BSPCOVER is 2794 seconds.

The complexity of this method is directly related to the values of n, m, the instances number and the features number of each dataset. If we call the FFT calculation time T_{FFT}, the instances number r and features number c, the order of complexity according to algorithm1-2 will be $O(T_{FFT} + rcnm)$.

We performed the Wilcoxon signed rank test between the nine state-of-the-art methods and our proposed method to find significant differences between each pair of algorithm behavior [37]. The p-value represents how significant differences the result is: If the value of p is small, the evidence is strong. As Table 6 states, the number of functions on which the proposed method performs better than ST, FS, COTE, SD, DTW Rn 1NN, RotF, ELIS, ResNet, and BSPCOVER are 9, 10, 8, 12, 10, 9, 12, 9 and 7, respectively.

We found out that p-value of all results are smaller than 0.05, except COTE, BSPCOVER (larger than 0.05). Therefore, the overall results show our method is better than related methods.

Conclusion

We present a simple method for time series classification in this paper. Our study of time series processing has led us to find a way to improve the accuracy of TSC problems. The proposed method has three steps. It applies FFT on raw data and then normalized them in preprocessing, histogram computation, extract features, the end using random forest classifier to predict the label of the test set and find accuracy. This method takes a minimum of time to build the model by reducing the dimensions it applies. Using this idea helped us to increase the accuracy of the classification to an acceptable level.

The implementation results of our proposed model show that it has achieved the highest accuracy in half of the datasets, and in total, it is in the first place among ten methods. These results were evaluated using the k-fold cross-validation on the UCR time series classification archive.

In future work, it is suggested that the method presented in this paper be applied to real data, which is usually time-consuming. Also, modifications should be made to the model to use in multivariate time series.

Author Contributions

Kiana Kiaei: Programmer, Software, Validation, Conceptualization, Visualization, Investigation, Writing - Reviewing and Editing, Writing - Original draft preparation. Hesam Omranpour: Supervision, Project administration, Conceptualization, Methodology, Visualization, Investigation, Writing - Reviewing and Editing, Funding acquisition, Programmer, Writing - Original draft preparation.

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

TSC	Time Series Classification
FFT	Fast Fourier Transform
EEG	Electroencephalography
ED	Euclidean Distance
DTW	Dynamic Time Warping
SVM	Support Vector Machine
ST	Shapelet Transformation
SD	Scalable shapelet Discovery
FS	Fast Shapelets
SAX	Symbolic Aggregate approXimation
SFA	Symbolic Fourier Approximation
DFT	Discrete Fourier Transform
KNN	K Nearest Neighbor
BOP	Bag of Patterns
VSM	Vector Space Model
TF-IDF	Term Frequency — Inverse Document Frequency
BOSS	Bag of Symbolic Fourier Approximation Symbols
ELIS	Efficient Learning Interpretable Shapelets
COTE	Collective Of Transformation-based Ensembles
ECG	Electrocardiogram
MDI	Mean Decrease Impurity
MDA	Mean Decrease Accuracy
RotF	rotation forest

References

[1] H. Madsen, Time series analysis. CRC Press, 2007.

[2] M. Langkvist, L. Karlsson, A. Loutfi, “A review of unsupervised feature learning and deep learning for time-series modeling,” Pattern Recognit. Lett., 42: 11–24, 2014.

[3] Z. Turani et al., “Optical radiomic signatures derived from optical coherence tomography images improve identification of melanoma,” Cancer Res., 79(8): 2021–2030, 2019.

[4] S. Adabi et al., “Universal in vivo textural model for human skin based on optical coherence tomograms,” Sci. Rep., 7(1): 1–11, 2017.

[5] J.C.B. Gamboa, “Deep learning for time-series analysis,” arXiv Prepr. arXiv1701.01887, 2017.

[6] M. Jung, S. Chi, “Human activity classification based on sound recognition and residual convolutional neural network,” Autom. Constr., 114: 103177, 2020.

[7] A. Rajkomar et al., “Scalable and accurate deep learning with electronic health records,” NPJ Digit. Med., 1(1): 1–10, 2018.

[8] Z. Fang, P. Wang, W. Wang, “Efficient learning interpretable shapelets for accurate time series classification,” in Proc. 2018 IEEE 34th International Conference on Data Engineering (ICDE): 497–508, 2018.

[9] J. Grabocka, N. Schilling, M. Wistuba, L. Schmidt-Thieme, “Learning time-series shapelets,” in Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 392–401, 2014.

[10] T. Rakthanmanon, E. Keogh, “Fast shapelets: A scalable algorithm for discovering time series shapelets,” in proc. 2013 SIAM International Conference on Data Mining: 668–676, 2013.

[11] B. Wang et al., “Removal of 'strip noise' in radio-echo sounding data using combined wavelet and 2-D DFT filtering,” Ann. Glaciol., 61(81): 124-134, 2020.

[12] G.R. Schleder, A.C. M. Padilha, C.M. Acosta, M. Costa, A. Fazzio, “From DFT to machine learning: recent approaches to materials science--a review,” J. Phys. Mater., 2(3): 32001, 2019.

[13] A. Bagnall, J. Lines, A. Bostrom, J. Large, E. Keogh, “The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances,” Data Min. Knowl. Discov., 31(3): 606–660, 2017.

[14] X. Xi, E. Keogh, C. Shelton, L. Wei, C.A. Ratanamahatana, “Fast time series classification using numerosity reduction,” in Proc. 23rd international conference on Machine learning: 1033–1040, 2006.

[15] H. Sakoe, S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” IEEE Trans. Acoust., 26(1): 43–49, 1978.

[16] L. Ye, E. Keogh, “Time series shapelets: a new primitive for data mining,” in Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 947–956, 2009.

[17] J. Lines, L.M. Davis, J. Hills, A. Bagnall, “A shapelet transform for time series classification,” in Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 289–297, 2012.

[18] J. Grabocka, M. Wistuba, L. Schmidt-Thieme, “Fast classification of univariate and multivariate time series through shapelet discovery,” Knowl. Inf. Syst., 49(2): 429–454, 2016.

[19] T. Le Nguyen, S. Gsponer, G. Ifrim, “Time series classification by sequence learning in all-subsequence space,” in Proc. 2017 IEEE 33rd international Conference on Data Engineering (ICDE): 947–958, 2017.

[20] T. Rakthanmanon et al., “Searching and mining trillions of time series subsequences under dynamic time warping,” in Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 262–270, 2012.

[21] J. Lin, E. Keogh, L. Wei, S. Lonardi, “Experiencing SAX: a novel symbolic representation of time series,” Data Min. Knowl. Discov., 15(2): 107–144, 2007.

[22] P. Schäfer, U. Leser, “Fast and accurate time series classification with weasel,” in Proc. 2017 ACM on Conference on Information and Knowledge Management: 637–646, 2017.

[23] R. Agrawal, C. Faloutsos, A. Swami, “Efficient similarity search in sequence databases,” in Proc. International Conference on Foundations of Data Organization and Algorithms: 69–84, 1993.

[24] J. Lin, R. Khade, Y. Li, “Rotation-invariant similarity in time series using bag-of-patterns representation,” J. Intell. Inf. Syst., 39(2): 287–315, 2012.

[25] P. Senin, S. Malinchik, “Sax-vsm: Interpretable time series classification using sax and vector space model,” in Proc. 2013 IEEE 13th International Conference on Data Mining: 1175–1180, 2013.

[26] G. Ifrim, C. Wiuf, “Bounded coordinate-descent for biological sequence classification in high dimensional predictor space,” in Proc. 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 708–716, 2011.

[27] P. Schäfer, “Scalable time series classification,” Data Min. Knowl. Discov., 30(5): 1273–1298, 2016.

[28] “The Fast Fourier Transform (FFT) as a statistical tool for time-series analysis.”

[29] S. Zhang, X. Li, M. Zong, X. Zhu, D. Cheng, “Learning k for knn classification,” ACM Trans. Intell. Syst. Technol., 8(3): 1–19, 2017.

[30] A. A. Soofi, A. Awan, “Classification techniques in machine learning: applications and issues,” J. Basic Appl. Sci., 13: 459–465, 2017.

[31] G. Biau, E. Scornet, “A random forest guided tour,” Test, 25(2): 197–227, 2016.

[32] G. Li, B.K.K. Choi, J. Xu, S.S. Bhowmick, K.-P. Chun, G.L.H. Wong, “Efficient shapelet discovery for time series classification,” IEEE Trans. Knowl. Data Eng., 2020.

[33] A. Bagnall, J. Lines, J. Hills, A. Bostrom, “Time-series classification with COTE: the collective of transformation-based ensembles,” IEEE Trans. Knowl. Data Eng., 27(9): 2522–2535, 2015.

[34] H. A. Dau et al., “The UCR time series archive,” IEEE/CAA J. Autom. Sin., 6(6): 1293–1305, 2019.

[35] L. Breiman, A. Cutler, “Manual--setting up, using, and understanding random forests V4. 0. 2003,”.

[36] H.I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, “Deep learning for time series classification: a review,” Data Min. Knowl. Discov., 33(4): 917–963, 2019.

[37] J. Derrac, S. García, D. Molina, F. Herrera, “A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms,” Swarm Evol. Comput., 1(1): 3–18, 2011.

Biographies

Kiana Kiaei received her Bachelor's degree in computer engineering from Babol Noshirvani of Technology (BNUT). She is currently a master's student in the field of computer engineering majoring in computer software at (BNUT).

- Email: k.kiaei@stu.nit.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Hesam Omranpour received B.Sc. degree in Computer Engineering-Software Engineering from Iran, University of Science and Technology and M.S. degree and Ph.D. degree in Artificial Intelligence from Amirkabir University of Technology, Tehran, Iran, in 2006, 2009 and 2016 respectively. He is currently an assistant professor with the department of electrical and computer engineering, Babol Noshirvani University of Technology, Iran. His current research interests include optimization, machine learning, pattern recognition.

- Email: h.omranpour@nit.ac.ir
- ORCID: 0000-0003-4253-0811
- Web of Science Researcher ID: NA
- Scopus Author ID: 26423137100
- Homepage: https://ostad.nit.ac.ir/home.php?sp=391011

How to cite this paper:

K. Kiaei, H. Omranpour, “Polar formed histogram of fast fourier transform feature for time series classification problems,” J. Electr. Comput. Eng. Innovations, 10(2): 273-286, 2022.

DOI: 10.22061/JECEI.2021.8048.468

URL: https://jecei.sru.ac.ir/article_1628.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Efficient GAN-based Method for Extractive Summarization

S.V. Moravvej[1,*], *M.J. Maleki Kahaki*[2], *M. Salimi Sartkahti*[3], *M. Joodaki*[1]

[1] *Department of Computer Engineering, Isfahan University of Technology, Isfahan, Iran*

[2] *Department of Computer Engineering, University of Kashan, Kashan, Iran*

[3] *Department of Computer Engineering, Amirkabir University of Technology, Tehran, Iran*

Article Info

Article History:
Received 15 July 2021
Reviewed 27 August 2021
Revised 01 November 2021
Accepted 08 November 2021

Keywords:
Text summarization
Non-greedily
Generative adversarial network
Attention mechanism
Extractive summarization

***Corresponding Author's Email Address:**
sa.moravvej@ec.iut.ac.ir

Abstract

Background and Objectives: Text summarization plays an essential role in reducing time and cost in many domains such as medicine, engineering, etc. On the other hand, manual summarization requires much time. So, we need an automated system for summarizing. How to select sentences is critical in summarizing. Summarization techniques that have been introduced in recent years are usually greedy in the choice of sentences, which leads to a decrease in the quality of the summary. In this paper, a non-greedily method for selecting essential sentences from a text is presented.
Methods: The present paper presents a method based on a generative adversarial network and attention mechanism called GAN-AM for extractive summarization. Generative adversarial networks have two generator and discriminator networks whose parameters are independent of each other. First, the features of the sentences are extracted by two traditional and embedded methods. We extract 12 traditional features. Some of these features are extracted from sentence words and others from the sentence. In addition, we use the well-known Skip-Gram model for embedding. Then, the features are entered into the generator as a condition, and the generator calculates the probability of each sentence in summary. A discriminator is used to check the generated summary of the generator and to strengthen its performance. We introduce a new loss function for discriminator training that includes generator output, real and fake summaries of each document. During training and testing, each document enters the generator with different noises. It allows the generator to see many combinations of sentences that are suitable for quality summaries.
Results: We evaluate our results on CNN/Daily Mail and Medical datasets. Summaries produced by the generator show that our model performs better than other methods compared based on the ROUGE metric. We apply different sizes of noise to the generator to check the effect of noise on our model. The results indicate that the noise-free model has poor results.
Conclusion: Unlike recent works, in our method, the generator selects sentences non-greedily. Experimental results show that the generator with noise can produce summaries that are related to the main subject.

Introduction

Nowadays, the amount of information in different areas of the World Wide Web in various formats such as web pages, articles, emails, log files, and linked data are increasing sharply [1].

However, much of the information is less important.

Doi: 10.22061/JECEI.2021.8051.475

On the other hand, manually identifying important information in a large text is time-consuming and practically impossible.

Therefore, we need a system that extracts vital information from the text in the shortest time and with the highest accuracy. This system is called automatic text summarization (ATS) [2]. Primary hypotheses are presented for summarization. A summary is a text produced from one or more texts and is not more than half of the whole text [3]. According to research in [4], summarization is the extraction of the most important information from a source or sources to produce a shorter version for a specific user or task. In other words, this process is one of the crucial steps in diverse cases such as extracting knowledge from text with huge information about the similarity of diseases in the OMIM dataset [5].

The text summarization methods can be categorized in a different view of points. Output-based summarization can be done as extractive or abstractive. Extractive summarization is done by selecting important sentences from different text parts [6], [7]. In abstractive summarization, the system tries to identify the text key concepts and then convert it to another form that is shorter than the original text [8], [9]. Extractive summarization ensures a semantic and grammatical rule between sentences [10]. Input-based summarization may be single-document or multi-document. In multi-document summarization, several documents with the same subject are considered input, and the final summary is the text produced from all the documents [11], [12]. Based on the content, the summary can be general or query-based. In query-based summarization, a summary is generated according to the user's query [13]. This type of summary focuses on the user query and does not consider the general index of concepts in the text. However, in a general summary, the text is produced regardless of the subject and domain. This article presents an extractive, single-document, and general approach.

Since 1950, many solutions for extractive summarization have been offered. Some methods are based on machine learning concepts such as clustering [14], [15] support vector machines [16]. Recently, optimization algorithms [17], [18] like Cuckoo [19], [20] and fuzzy methods [21] have been used for summarization. Graph-based methods are other techniques that have been selected for this purpose [22]. In these methods, each sentence is usually considered as a graph so that the nodes represent the words, and the graph edges express the interval between the words. With recent advances in deep learning, other previous machine learning methods are less commonly utilized in natural language processing tasks [23], [24], [25]. With its relatively complex structure, deep learning can learn the features of words, sentences, or documents automatically [26].

Despite the deep learning-based methods provided for summarization, many of them have not been able to solve the challenges of extractive summarization. The two main components of summarizing are scoring sentences and selecting them [27]. Most previous works suffer from a greedy choice of sentences [10], [28], [29], [30].

In other words, after choosing a high-ranking sentence, they put it aside and do not consider it in choosing the next sentences, which leads to a decrease in the quality of the summary.

Today, Generative Adversarial Networks have been used in many applications of natural language processing, including text generation [31] and question answering [32].

In this research, an extractive summarizer using generative adversarial networks and attention mechanism is presented.

According to our information, this paper presents the first method based on generative adversarial networks for extractive summarization. These networks are made up of two generator and discriminator components that compete in a process.

In this context, the generator goal is to rate each sentence of the document, while the goal of the discriminator is to distinguish the real from the fake summary, which enhances the performance of the generator.

In a non-greedy way, the generator determines the possibility of the presence of sentences in summary at once.

Other contributions to this article are as follows:

- The generator is trained with the feedback it receives from the discriminator. Therefore, if fake summaries are introduced to the discriminator, it can prevent the generator from producing poor-quality summaries. We extract some fake and real summaries from each document and utilize them for discriminator training, which leading to a new loss function for it.
- Another important characteristic of the proposed model is producing multiple summaries for each document during training and testing. During the training, each document enters the generator with different noises, which the model identifies different and appropriate combinations of sentences that produce quality summaries. During the test, we enter each document into the generator with different noises and use the voting system for the final summary.

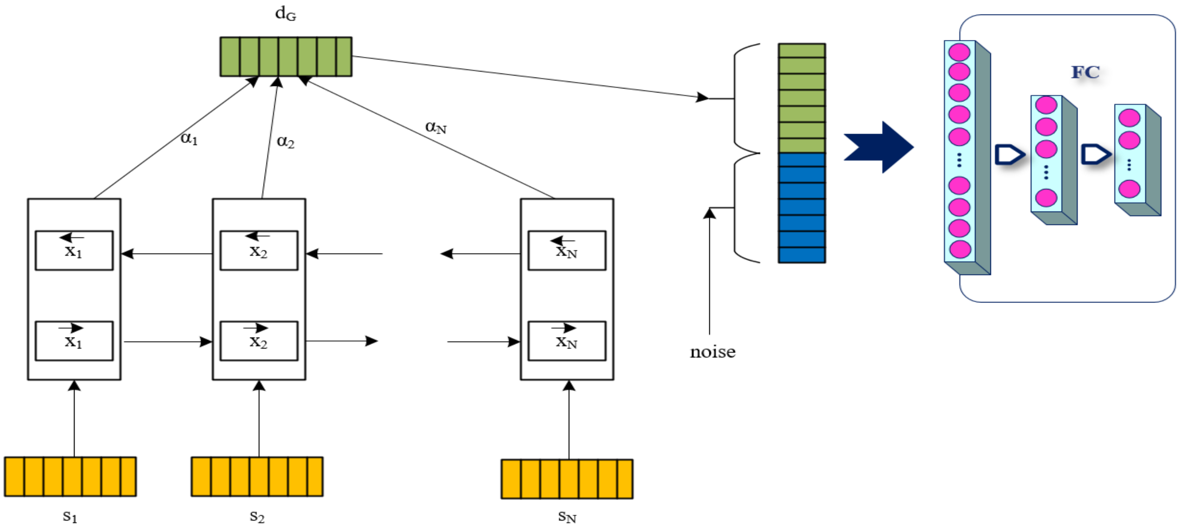


Fig. 1: Our generator architecture.

We evaluate our proposed model on two datasets. In the first evaluation, we use the CNN/Daily Mail dataset[1]. This dataset is a benchmark for evaluating many of the works. In the second application, we utilize the medical dataset available in PubMed Central[2]. The evaluation results show that the proposed model can provide high-quality summaries compared to other compared methods.

Related Work

The two main categories of a summary are abstractive and extractive. In most abstractive methods, an Auto-Encoder is used, in which important features of the text are extracted in the Encoder. In the Decoder section, a summary is generated using the features extracted in the encoder section [33], [34]. So far, many methods for extractive summarization, including graph-based [35], [36], [37], [38] and deep learning, have been proposed. In the following, several extractive summarizations work based on deep learning methods are described.

The works presented in [39], [40] use an auto-encoder to learn sentence features. The authors in [39] make the document word features and then calculate the scores of the sentences using the scores of its words. [40] uses cosine similarity between sentence and subject to score sentences. In [27], the authors used recurrent neural networks to rank sentences. The authors considered each sentence as a tree where the words are on the leaves, and the sentence is at the root. The score of each sentence is obtained from the leaves based on a non-linear process.

Another work in [10] uses reinforcement learning for summarizing. In this work, the coherence between sentences is considered as a reward. The policy is implemented as a multilayer perceptron that assigns a score to each sentence. The authors in [28] introduced a model called Summarunner for extraction summarization based on RNN networks. They used two RNN layers to embed words and sentences. Then, they employed logistic regression to classify sentences.

Recently, the attention mechanism [41] has gained much attention in many areas, including machine translation [42], question answering [43], and text summarization. In [29], the authors proposed a method based on the Siamese neural network (SNN) and the attention mechanism. They utilized the attention mechanism for words and sentences to score. They estimated the features of sentences using the obtained word features. Finally, they calculated the features of the document according to its sentences and used a classifier for the similarity of the summary and the document. [30] uses the hierarchical structure self-attention method to embed sentences and documents. The proposed method model's summarization as a classification problem in which it calculates sentence-summary probability.

Recently, Bidirectional Encoder Representations from Transformers (BERT) [44] has revolutionized the processing of natural languages. BERT is a pre-trained language model [24] for textual data. In [45], a simple version of BERT called BERTSUM is provided for extractive summarization.

In another example in [46], a method is presented for abstractive and extractive summarization. The proposed method can obtain document semantics using the BERT model.

[1] https://cs.nyu.edu/~kcho/DMQA/

[2] https://www.ncbi.nlm.nih.gov/pubmed/

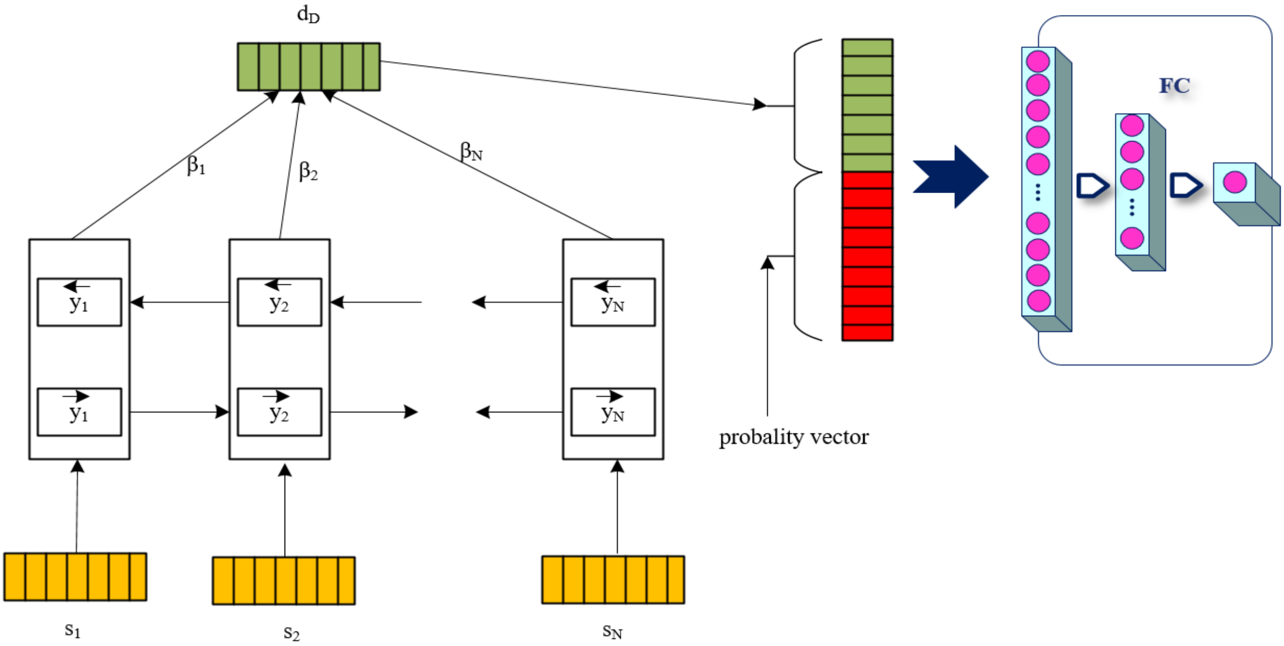


Fig. 2: Our discriminator architecture.

The Proposed Model

In this research, we use adversarial generating networks for extractive summarization. We employ this network to improve the problems of previous methods, including greed. We will first have a description of this network, and then the proposed model is presented.

Generative adversarial networks (GANs) were first proposed by Goodfellow et al. [47]. These networks consist of two separate networks that are similarly trained: the generator and discriminator networks. The purpose of the generator is to produce data such as images, text, etc., which are structurally similar to real data but are fake.On the other hand, the task of the discriminator network is to strengthen the generator.

These two networks play a two-player min-max game with a value function $V(D.G)$ as follows [47]:

$$\min_{G} \max_{D} V\,(D.G) = E_{x \sim p_{data}(x)}[log(D(x))] + E_{z \sim p_z(z)}[log(1 - D(G(z)))] \tag{1}$$

where x and z are input data and noise, respectively. G and D mean the generator and discriminator, respectively. $p_{data}(x)$ and $p_z(z)$ represent the input data distribution and the noise distribution, respectively. E is mathematical expectation. Generative adversarial networks can be extended to a conditional model If condition y is added to the generator and discriminator input. The value function, in this case, changes as follows [48]:

$$\min_{G} \max_{D} V\,(D.G) = E_{x \sim p_{data}(x)}[log(D(x|\mathbf{y}))] + E_{z \sim p_z(z)}[log(1 - D(G(z|\mathbf{y})))] \tag{2}$$

The proposed generator and discriminator model are shown in Fig. 1 and Fig. 2, respectively. We use sentence features as a condition in the generator and discriminator. Let $D = \{s_1 . s_2 \ldots . s_N\}$ represents the document, where the $s_i \in \mathbb{R}^d$ is the extracted features of the $i - th$ sentence. N is the length of document D, which is equal to the number of restricted sentences in each document.

The attention mechanism calculates the representation vector of the document in the generator and the discriminator according to the following equations:

$$d_G = \sum_{i=1}^{N} \alpha_i \, [\overleftarrow{x}_i . \overrightarrow{x}_i] \tag{3}$$

$$d_D = \sum_{i=1}^{N} \beta_i \, [\overleftarrow{y}_i . \overrightarrow{y}_i] \tag{4}$$

where $\overleftarrow{x}_i \in \mathbb{R}^{d_1}$, $\overrightarrow{x}_i \in \mathbb{R}^{d_1}$, $\overleftarrow{y}_i \in \mathbb{R}^{d_2}$, $\overrightarrow{y}_i \in \mathbb{R}^{d_2}$ are the output of step i in BLSTM. α_i and β_i are the coefficients of attention for the i-th sentence in the generator and the discriminator, respectively, which are formulated as follows:

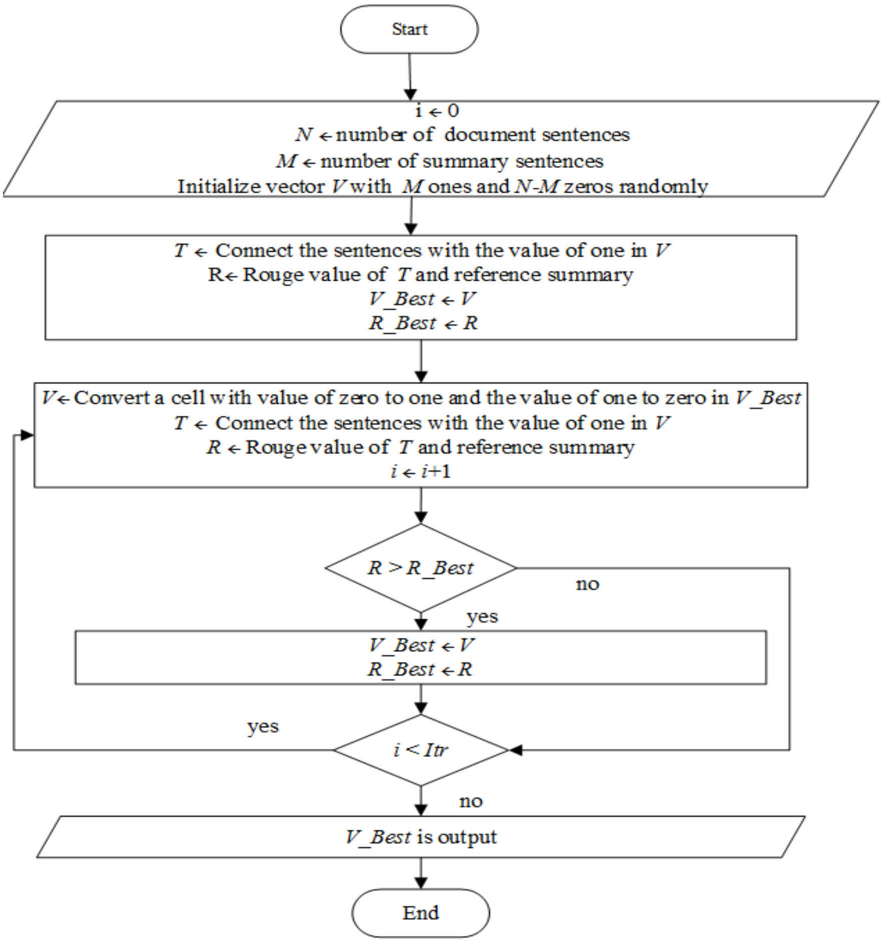

Start
i ← 0
N ← number of document sentences
M ← number of summary sentences
Initialize vector V with M ones and N-M zeros randomly
T ← Connect the sentences with the value of one in V
R← Rouge value of T and reference summary
V_Best ← V
R_Best ← R
V← Convert a cell with value of zero to one and the value of one to zero in V_Best
T ← Connect the sentences with the value of one in V
R ← Rouge value of T and reference summary
i ← i+1
R > R_Best
no
yes
V_Best ← V
R_Best ← R
yes
i < Itr
no
V_Best is output
End

Fig. 3: Generate a real summary for the document.

$$\alpha_i = \frac{e^{u_i}}{\sum_{i=1}^{N} e^{u_i}} \tag{5}$$

$$\beta_i = \frac{e^{v_i}}{\sum_{i=1}^{N} e^{v_i}} \tag{6}$$

$$u_i = tanh(W_u[\overleftarrow{x}_i . \vec{x}_i] + \ b_u) \tag{7}$$

$$v_i = tanh(W_v[\overleftarrow{y}_i . \vec{y}_i] + \ b_v) \tag{8}$$

where $W_u \in \mathbb{R}^{2.d_1}. b_v \in \mathbb{R}$, $W_v \in \mathbb{R}^{2.d_2}.$ and $b_v \in \mathbb{R}$ are the parameters of the attention mechanism for documents.

In Fig. 1, the representation vector of the document is connected to the noise vector and enters a feed-forward neural network. The last layer of this network calculates the probability of the presence of each sentence. Noise causes the generator to produce different outputs. Each document enters the generator to be summarized in different iterations with different noises. The generator tries to produce different summaries of almost the same quality for each document. It allows the generator to identify different combinations of sentences that are appropriate for the summary. Therefore, sentences that may not be useful for the summary alone can lead to a quality summary by being placed next to other sentences.

In the discriminator network, the probability vector of sentences is contacted with the representation vector of the document (see Fig. 2). In this context, the probability vector of sentences is the vector of the number of sentences in a document, and each element is zero or one.

A. The Target Vector

In the general generative adversarial network, the generator output is used as fake data for discriminator training. In addition, a real target is extracted for each sample. In this research, in order to introduce quality summaries to the discriminator, more than one summary is extracted from each document, which is similar in terms of quality. On the other hand, we produce several poor-quality summaries for the document. The real summaries of each document are text and cannot be employed as a target. Therefore, we need a method to express the presence or absence of each sentence in summary as a number. For this purpose, a vector is defined with N element for each document, where N is the number of sentences. Each element of this vector has a value of zero or one. The value of one indicates the presence of a sentence in summary.

We employ a greedy method according to Fig. 3 to produce this vector. First, we get a vector with length N and M number one, where M is the number of sentences in summary. The values of one are randomly arranged in the vector. The sentences corresponding to the value of one are put together in this vector to produce a summary, and the ROUGE metric measures their quality.

After that, a randomly chosen one is converted to zero, and a randomly chosen zero to one, and the Rouge value is recalculated. If its value is better than the previous one, it will be replaced. This process is repeated for Itr times, and the best vector during the process is considered the output. Note that in order to produce any real target, the algorithm must be executed from the beginning. The process of making a fake target is similar to a real target, except that the vector will replace the previous one if the Rogue score is lower. The length of all documents is limited to N sentences. Documents longer than N sentences are cut, and smaller documents are zero-padding.

B. Loss Function

The Loss function is calculated based on the discriminator output for the generator as follows [47]:

$$Loss_G = E_{i\sim Dataset}\left[E_{z\sim p_{z(z)}}\left[log\left(1 - D\left(G(z|y_i)\right)\right)\right]\right] \tag{9}$$

where $Dataset$ is a set of documents, y_i is features of sentences in document i, and E is the mathematical expectation.

The Loss function for the discriminator is computed based on the generator output, real and fake summaries as follows:

$$\begin{aligned} Loss_D &= E_{i\sim Dataset}[\, E_{z\sim p_{z(z)}}[log(1 \\ &- D(G(z|y_i)|\, y_i))] \\ &+ E_{k\sim p_{Fake_i}}\left[log\left(1 - D(k|y_i)\right)\right] \\ &+ \ E_{k\sim p_{Real_i}}[log(D(l|y_i))]] \end{aligned} \tag{10}$$

where p_{Real_i} and p_{Fake_i} show the distribution of real and fake summaries for the document i. Equation (10) forces the discriminator to learn a set of high-quality and low-quality summaries.

On the other hand, be sensitive to the summaries produced by the generator and force the generator to produce a high-quality summary.

C. Summarization

At the time of testing, only the generator is used to generate the summary. By applying different noises in the generator, multiple summaries can be generated for each document, and the quality of these summaries or ROUGE summaries is very close to each other. We consider the voting system to generate a single summary for the document. To do this, the probability of the presence of sentences in summary is calculated for different noises. After that, the sentences are ranked based on their number of selections, and finally, the sentences with the highest rank are selected.

Result

A. Feature Extraction

One of the important components of deep learning is feature extraction. There are many ways to do this. In this research, we use two methods to select features. In the first method, we use 12 traditional features. The list of these features is given in Table 1.

Some of these features are at the word level, and some of them at the sentence level. All features are scaled to [0,1].

An important feature of deep models is the automatic learning of features. Sentence features are used as a condition in the model and cannot be changed during training. We use Skip-Gram [49] to embed words. Finally, by averaging word embedding, we extract sentence embedding.

Table 1: Traditional features

Feature	Description
Common Word	The number of occurrences N common words in the dataset, divided by the sentence length.
Position	The position of the sentence. Supposing there are N sentences in the document, for j the sentence, the position is computed as $1-(j-1)/(N-1)$.
Length	The number of words in the sentence, divided by the length of the largest sentence.
Number Raito	The number of digits, divided by the sentence length.
Named entity ratio	The number of named entities, divided by the sentence length.
Tf/Isf	Term frequency over the sentence, divided by the largest term frequency.
Similarity Sentence	The number of occurrences of words in the sentence with the highest Tf/Isf in the sentence divided by the length of the sentence.
None phrase	The number of None phrases, divided by the sentence length.
Pos Ratio	A four-dimensional vector containing the number of nouns, verbs, adjectives, and adverbs. Each vector cell is divided by the sentence length.

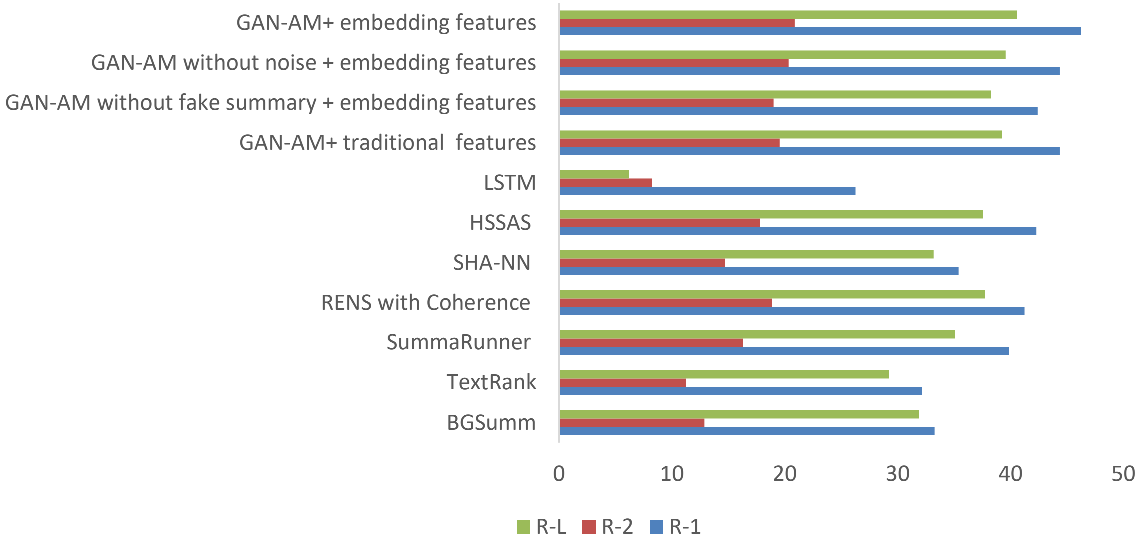


Fig. 4: Graphical comparison of the proposed model and other methods on the CNN/Daily Mail dataset.

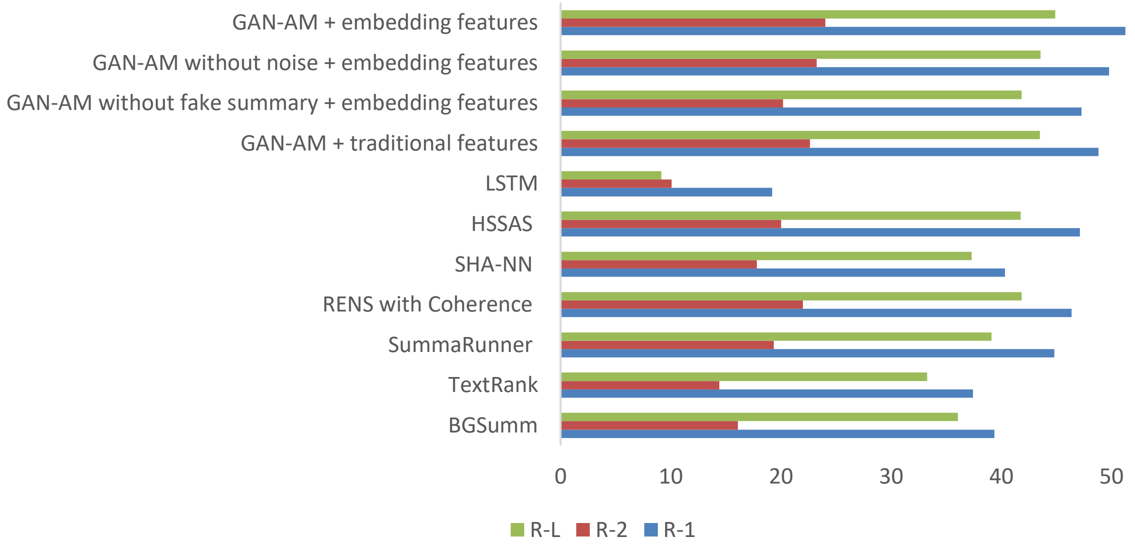


Fig. 5: Graphical comparison of the proposed model and other methods on the Medical dataset.

B. Dataset

We use two known datasets for our evaluations: CNN/Daily Mail and PubMed. The first dataset [50] combines two datasets designed for comprehension, extractive, and abstractive tasks. The datasets have come to the attention of researchers in recent years for automated summarization. This dataset contains 287,226 documents for training, 13,368 for validation and 11,490 for testing. The average number of sentences per document in training data is 28 sentences. The average reference summary of each document is 3-4 sentences, and the average number of words per document in training data is 802 words [28]. More details are available in Table 2.

This dataset consists of two versions. In the first version, all entities are replaced with specific words, while the second version is the original data. We adopt the second version for our model.

Table 2: Statistics of the CNN/Daily Mail dataset

	Train	Validation	Test
Pairs of data	287,113	13,368	11,490
Article length	749	769	778
Summary Length	55	61	58

The second collection is PubMed, which contains many articles in the field of medicine.

The number of these articles is increasing every day. For this purpose, we have randomly downloaded 1000 articles. The dataset was divided into 1334 documents for training, 288 documents for validation, and 378 documents for testing.

C. Detail of Model

In this research, Python language and PyTorch library have been used for implementation. Jupyter has been employed to implement project codes. Another library used in this research is the NLTK library. This library provides classes and methods for processing natural languages in Python. This library can perform a wide range of natural language processing operations.

We employ a two-layer bidirectional LSTM. In generative adversarial networks, the discriminator converges to the optimum point sooner, which causes the generator cannot converge. For this reason, we train the discriminator once for every 15 generator training. In addition, due to the connection of vectors in the two networks, we use batch normalization before the data enters the feed-forward neural network. Table 3 shows the values of the other parameters.

Table 3: The parameters of the model

Parameter	CNN/Daily Mail	Medical
batch size	128	64
embedding dim	60	60
max sentence length	100	50
real summary per document	40	15
fake summary per document	40	15
activation fun(lstm & dense)	relu	relu
dense hidden layer	8	5

D. Metrics

We employ the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) package [51] as an evaluation metric in our experiments.

This metric calculates the similarity between the generated summary and the reference summary by counting the number of common units. Rouge-n recall between an extracted summary and a reference summary is calculated as follows:

$$\text{Rouge-}n = \frac{\sum_{s\,\in\{ref\ sum\}}\sum gram_n \in s\, Count_{match}\,(gram_n))}{\sum_{s\,\in\{ref\ sum\}}\sum gram_n \in s\, Count\,(gram_n))} \quad (11)$$

where n stands for the length of n-gram, $Count_{match}\ (gram_n)$ is the maximum number of n- gram co-occurring in the extracted summary and the reference summary. Rouge-1 and Rouge-2 are special cases of Rouge-n in which $n = 1$ or $n = 2$. R-L calculates the length of the longest common subsequence between the reference summary and the extracted summary. Based on previous works, Rouge-1(R-1), Rouge-2(R-2) and, Rouge-L(R-L) are most widely used in summarization. For this reason, we use these three metrics in all our experiments.

Table 4: Numerical comparison of the proposed method and other methods on the CNN/Daily Mail dataset

Model	R-1	R-2	R-L
BGSumm [35]	33.27	12.90	31.89
TextRank [38]	32.18	11.26	29.26
SummaRunner [28]	39.60	16.20	35.30
RENS with Coherence [10]	41.25	18.87	37.75
SHA-NN [29]	35.40	14.7	33.2
HSSAS [30]	42.30	17.80	37.60
LSTM	26.28	8.27	6.23
GAN-AM + traditional features	44.36	19.56	39.26
GAN-AM without fake summary + embedding features	42.42	19.02	38.26
GAN-AM without noise + embedding features	44.38	20.34	39.58
GAN-AM + embedding features	**46.26**	**20.89**	**40.56**

E. Experimental Results and Analysis

We consider two techniques for preprocessing in all our experiments: 1- Stop word removal 2- Stemming.

Our project uses a 64-bit Windows operating system with 64 GB of RAM and GPU. The best model was obtained for the CNN/Daily Mail dataset after 50 epochs, while [29] obtained the best model after 70 epochs. The whole process of our training took 5 hours. The best model for the Medical dataset was obtained after 30 epochs. This process took 1.5 hours.

Table 5: Numerical comparison of the proposed method and other methods on the medical dataset

Model	R-1	R-2	R-L
BGSumm [35]	39.37	16.09	36.06
TextRank [38]	37.44	14.40	33.28
SummaRunner [28]	44.82	19.36	39.12
RENS with Coherence [10]	46.39	22.01	41.85
SHA-NN [29]	40.33	17.80	37.32
HSSAS [30]	47.13	20.03	41.76
LSTM	19.21	10.08	9.15
GAN-AM + traditional features	48.84	22.64	43.52
GAN-AM without fake summary + embedding features	47.29	20.19	41.86
GAN-AM without noise + embedding features	49.78	23.24	43.58
GAN-AM + embedding features	**51.26**	**24.04**	**44.91**

Table 6: Execution time of algorithms (in milliseconds)

Length	BGSumm [35]	TextRank [38]	SummaRunner [28]	RENS with Coherence [10]	SHA-NN [29]	HSSAS [30]	GAN-AM
3209	285	250	319	369	332	402	315
3590	319	285	364	413	363	459	346
4006	343	304	401	468	418	509	391
4592	399	369	464	537	482	589	450
5630	496	449	582	751	603	736	563
6972	603	539	710	809	742	860	680
7460	642	572	770	882	793	952	756
8905	801	709	897	1020	942	759	882
9790	873	751	991	1142	1032	1221	958
10196	1006	829	1148	1193	1081	1386	1100

The proposed method is compared with two graph-based methods BGSumm [35], TextRank [38], four deep learning methods SummaRunner [28], RENS with Coherence [10], SHA-NN [29], HSSAS [30] and one basic method LSTM. The LSTM model uses only our generator part. The evaluation results of the proposed system for the two datasets are shown in Table 4 and Table 5. For the CNN/Daily Mail dataset, the results reported for the SummaRunner, RENS with Coherence, SHA-NN methods in [28], [10], and [29] are given in Table 4. BGSumm, TextRank, and HSSAS methods were obtained in the experimentation we did in our laboratory. The total results of Table 5 have been obtained in our laboratory.

As expected, for both datasets, deep learning methods are superior to graph-based methods. Although BGSumm has been tested on a medical dataset, it has failed deep learning methods even in the medical dataset. In general, deep learning-based models are weaker than our model in the two datasets. The RENS with Coherence method is less accurate than our model, although it considers coherence between sentences.

Table 7: Effect of noise on the proposed model for the CNN/Daily Mail dataset

Noise Size	R-1	R-2	R-L
0	44.38	20.34	39.58
10	44.80	20.47	39.62
20	45.34	20.53	39.68
30	45.76	20.6	39.7
40	45.86	20.72	39.92
50	46.13	20.8	40.45
60	46.26	20.89	40.56
70	46.14	20.7	40.42
80	46.07	20.55	40.32
90	45.89	19.92	39.9
100	44.74	19.76	39.72

In addition, the embedding features perform better than the traditional features on our model. By comparing the LSTM model with the GAN-AM method, the importance of the discriminator is seen in our model. As we can see, our method has a relatively strong weakness compared to other models when it does not use a discriminator. In addition, in another experiment, we examined the importance of fake summaries for discriminator training. GAN-AM without fake summary + embedding features shows this model. The difference between the results of this model and our best model for both datasets is noticeable. The superiority of the presented model can be considered in the way the generator scoring of sentences. The generator assigns scores to each sentence, taking into account the rest of the sentences. Another reason that can be mentioned is the voting system used. However, the discriminator has not been ineffective in producing a summary by the generator. To better understanding of results, the evaluation results are shown schematically in Fig. 4 and Fig. 5. To check the time of the algorithms, we selected 10 test documents from the CNN/Daily Mail dataset. The production time of the summary by each method is shown in Table 6. As we can see, graph-based methods take less time than deep learning methods due to less computation. The GAN-AM method consists of an LSTM network and a feed-forward network, so it is expected to take some time to calculate. However, other methods take time to calculate embedding.

Table 8: Effect of noise on the proposed model for medical dataset

Noise Size	R-1	R-2	R-L
0	49.78	23.24	43.58
10	50.69	23.6	43.62
20	50.8	23.72	43.86
30	51.05	23.95	44.78
40	51.26	24.04	44.91
50	51.03	23.86	44.8
60	49.8	23.7	44.62
70	49.52	23.51	44.15
80	49.02	23.42	43.82
90	48.61	22.86	43.42
100	48.2	22.62	42.62

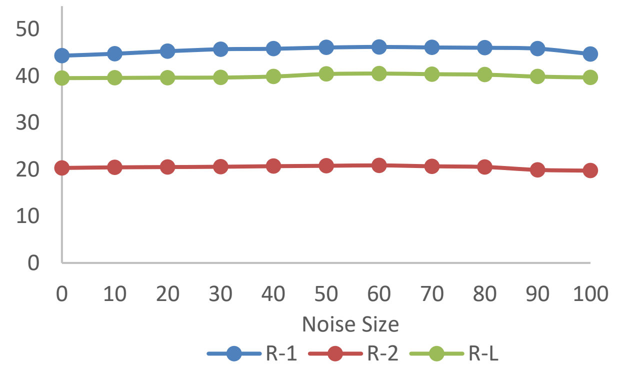


Fig. 6: Results of the proposed model for different noises on the CNN / Daily Mail dataset.

We performed other experiments to determine the effect of noise on the generator. For this purpose, we apply noise of different sizes to the generator. The results R-1, R-2, and R-L for the two datasets are shown in Table 7 and Table 8. For the CNN / Daily Mail dataset, increasing the noise size to 60 raises the R-1, R-2, and R-L criteria, but we have a downtrend from 60 to 100. In this database, the best size for noise is 60. For the Medical dataset, we have an uptrend from 0 to 40 and then a downtrend. For this data set, the noise size is set to 40. As we can see, the proposed model has better performance with noise. For a better understanding, the results for the two datasets are shown in Fig. 6 and Fig. 7.

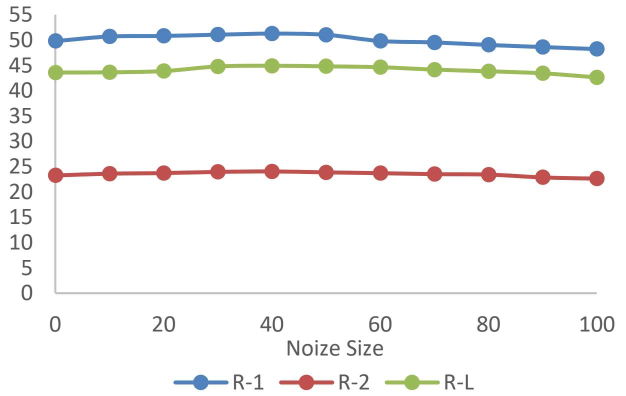


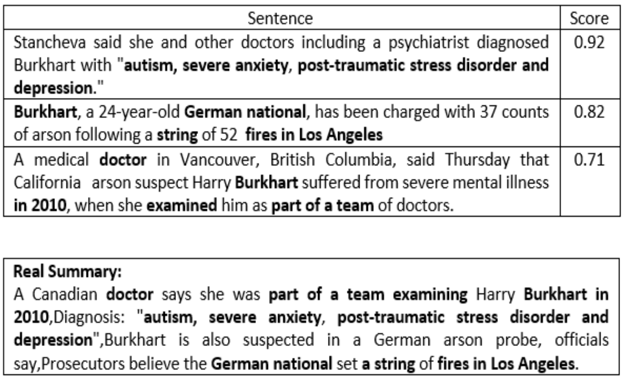
Fig. 8: Results of the proposed model for different noises on the Medical dataset.

As an example of the generator output, the three sentences extracted from a document in the Medical dataset by the generator along with its reference summary are shown in Fig. 8. Common words between each sentence and reference are highlighted. As we can see, the sentences that have more in common with the reference summary are given higher scores.

Conclusion and Future Work

In this study, a method based on the generative adversarial network for extractive summarization was proposed in which sentence features were considered a condition in this network. Sentence features were extracted based on traditional and embedding methods. The generator utilized sentence features to calculate the probabilities of their presence. In addition, due to the use of noise in the generator, several summaries were generated for each document. We set up a new loss function for the discriminator. Experiments have shown that this function can be effective in generating summaries.

In future work, we will consider the coherence between sentences in our model. As a solution, we can consider coherence when constructing the target or as a loss in the generator.



Sentence	Score
Stancheva said she and other doctors including a psychiatrist diagnosed Burkhart with **"autism, severe anxiety, post-traumatic stress disorder and depression.**"	0.92
Burkhart, a 24-year-old **German national**, has been charged with 37 counts of arson following a **string** of 52 **fires in Los Angeles**	0.82
A medical **doctor** in Vancouver, British Columbia, said Thursday that California arson suspect Harry **Burkhart** suffered from severe mental illness **in 2010**, when she **examined** him as **part of a team** of doctors.	0.71

Real Summary:
A Canadian **doctor** says she was **part of a team examining** Harry **Burkhart in 2010**,Diagnosis: **"autism, severe anxiety, post-traumatic stress disorder and depression**",Burkhart is also suspected in a German arson probe, officials say,Prosecutors believe the **German national** set **a string** of **fires in Los Angeles**.

Fig. 7: Three extracted sentences by the generator of the Medical dataset.

Author Contributions

S.V. Moravvej suggested the model and innovation of the problem and wrote the manuscript. M.J. Maleki Kahaki wrote the code of this article. M. Salim designed the experiments. M. Joodaki collected the dataset and edited the text of the article. 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

ATS	Automatic Text Summarization
GANs	Generative Adversarial Networks
ROUGE	Recall-Oriented Understudy for Gisting Evaluation

References

[1] S. Samiei, M. Joodaki, N. Ghadiri, "A scalable pattern mining method using apache spark platform," in Proc. 7th International Conference on Web Research (ICWR): 114-118, 2021.

[2] S.V. Moravvej, A. Mirzaei, M. Safayani, "Biomedical text summarization using Conditional Generative Adversarial Network (CGAN)," arXiv preprint arXiv:2110.11870, 2021.

[3] R. Mitkov, The Oxford handbook of computational linguistics. Oxford University Press, 2004.

[4] I. Mani, M. Maybury, "Automatic summarization John Benjamin's publishing Co," 2001.

[5] M. Joodaki, N. Ghadiri, Z. Maleki, M.L. Shahreza, "A scalable random walk with restart on heterogeneous networks with Apache Spark for ranking disease-related genes through type-II fuzzy data fusion," J. Biomed. Inf., 115: 103688, 2021.

[6] M. Kågebäck, O. Mogren, N. Tahmasebi, D. Dubhashi, "Extractive summarization using continuous vector space models," in Proc. 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC): 31-39, 2014.

[7] M. Jang, P. Kang, "Learning-free unsupervised extractive summarization model," IEEE Access, 9: 14358-14368, 2021.

[8] Y. Liu, P. Liu, "SimCLS: A simple framework for contrastive learning of abstractive summarization," arXiv preprint arXiv:2106.01890, 2021.

[9] Z. Cao, F. Wei, W. Li, S. Li, "Faithful to the original: Fact aware neural abstractive summarization," in Proc. the AAAI Conference on Artificial Intelligence, 32(1), 2018.

[10] Y. Wu, B. Hu, "Learning to extract coherent summary via deep reinforcement learning," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[11] A. Abdi, S. Hasan, S.M. Shamsuddin, N. Idris, J. Piran, "A hybrid deep learning architecture for opinion-oriented multi-document summarization based on multi-feature fusion," Knowledge-Based Syst., 213: 106658, 2021.

[12] S. Lamsiyah, A. El Mahdaouy, B. Espinasse, S.E.A. Ouatik, "An unsupervised method for extractive multi-document summarization based on centroid approach and sentence embeddings," Expert Syst. Appl., 167: 114152, 2021.

[13] S. Murarka, A. Singhal, "Query-based single document summarization using hybrid semantic and graph-based approach," in Proc. 2020 International Conference on Advances in Computing, Communication & Materials (ICACCM): 330-335, 2020.

[14] S. Akter, A.S. Asa, M.P. Uddin, M.D. Hossain, S.K. Roy, M.I. Afjal, "An extractive text summarization technique for Bengali document (s) using K-means clustering algorithm," in Proc. 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR): 1-6, 2017.

[15] M. Joodaki, N. Ghadiri, A.H. Atashkar, "Protein complex detection from PPI networks on Apache Spark," in Proc. 2017 9th International Conference on Information and Knowledge Technology (IKT): 111-115, 2017.

[16] T. Hirao, H. Isozaki, E. Maeda, Y. Matsumoto, "Extracting important sentences with support vector machines," in Proc. COLING 2002: The 19th International Conference on Computational Linguistics, 2002.

[17] S.Y. Ashkoofaraz, S.N.H. Izadi, M. Tajmirriahi, M. Roshanzamir, M.A. Soureshjani, S.V. Moravvej, M. Palhang, AIUT3D 2018 Soccer Simulation 3D League Team Description Paper.

[18] S. Vakilian, S.V. Moravvej, A. Fanian, "Using the Artificial Bee Colony (ABC) algorithm in collaboration with the fog nodes in the internet of things three-layer architecture," in Proc. 29th Iranian Conference on Electrical Engineering (ICEE): 509-513, 2021.

[19] S.H. Mirshojaei, B. Masoomi, "Text summarization using cuckoo search optimization algorithm," J. Comp. Rob., 8(2): 19-24, 2015.

[20] S. Vakilian, S. V. Moravvej, A. Fanian, "Using the cuckoo algorithm to optimizing the response time and energy consumption cost of fog nodes by considering collaboration in the fog layer," in Proc. 2021 5th International Conference on Internet of Things and Applications (IoT): 1-5, 2021.

[21] F.B. Goularte, S.M. Nassar, R. Fileto, H. Saggion, "A text summarization method based on fuzzy rules and applicable to automated assessment," Expert Syst. Appl., 115: 264-275, 2019.

[22] W.S. El-Kassas, C.R. Salama, A.A. Rafea, H.K. Mohamed, "EdgeSumm: Graph-based framework for automatic text summarization," Inf. Proc. Manage., 57(6): 102264, 2020.

[23] S.V. Moravvej, M. Joodaki, M.J.M. Kahaki, M.S. Sartakhti, "A method based on an attention mechanism to measure the similarity of two sentences," in Proc. 2021 7th International Conference on Web Research (ICWR): 238-242, 2021.

[24] M.S. Sartakhti, M.J.M. Kahaki, S.V. Moravvej, M. javadi Joortani, A. Bagheri, "Persian language model based on BiLSTM model on COVID-19 corpus," in Proc. 2021 5th International Conference on Pattern Recognition and Image Analysis (IPRIA): 1-5, 2021.

[25] S.V. Moravvej, M.J.M. Kahaki, M.S. Sartakhti, A. Mirzaei, "A method based on attention mechanism using Bidirectional Long-Short Term Memory (BLSTM) for Question Answering," in 2021 29th Iranian Conference on Electrical Engineering (ICEE): 460-464, 2021.

[26] H. Liang, X. Sun, Y. Sun, Y. Gao, "Text feature extraction based on deep learning: a review," EURASIP J. Wireless Commun. Networking, 2017(1): 1-12, 2017.

[27] Z. Cao, F. Wei, L. Dong, S. Li, M. Zhou, "Ranking with recursive neural networks and its application to multi-document summarization," in Proc. the AAAI Conference on Artificial Intelligence, 29(1): 2015.

[28] R. Nallapati, F. Zhai, B. Zhou, "Summarunner: A recurrent neural network based sequence model for extractive summarization of documents," in Proc. Thirty-First AAAI Conference on Artificial Intelligence: 3075-3081, 2017.

[29] J.Á. González, E. Segarra, F. García-Granada, E. Sanchis, L.ı.F. Hurtado, "Siamese hierarchical attention networks for extractive summarization," J. Intell. Fuzzy Syst., 36(5): 4599-4607, 2019.

[30] K. Al-Sabahi, Z. Zuping, M. Nadher, "A hierarchical structured self-attentive model for extractive document summarization (HSSAS)," IEEE Access, 6: 24205-24212, 2018.

[31] W. Nie, N. Narodytska, A. Patel, "Relgan: Relational generative adversarial networks for text generation," in Proc. International conference on learning representations, 2018.

[32] M. Lewis, A. Fan, "Generative question answering: Learning to answer the whole question," in Proc. International Conference on Learning Representations, 2018.

[33] T. Vo, "SGAN4AbSum: A Semantic-Enhanced Generative Adversarial Network for Abstractive Text Summarization," 2021.

[34] R. Paulus, C. Xiong, R. Socher, "A deep reinforced model for abstractive summarization," arXiv preprint arXiv:1705.04304, 2017.

[35] M. Moradi, "Frequent itemsets as meaningful events in graphs for summarizing biomedical texts," in Proc. 2018 8th International Conference on Computer and Knowledge Engineering (ICCKE): 135-140, 2018.

[36] Z. Wu, R. Koncel-Kedziorski, M. Ostendorf, H. Hajishirzi, "Extracting summary knowledge graphs from long documents," arXiv preprint arXiv:2009.09162, 2020.

[37] T. Safavi, C. Belth, L. Faber, D. Mottin, E. Müller, D. Koutra, "Personalized knowledge graph summarization: From the cloud to your pocket," in Proc. 2019 IEEE International Conference on Data Mining (ICDM): 528-537, 2019.

[38] R. Mihalcea, P. Tarau, "Textrank: Bringing order into text," in Proc. 2004 Conference on Empirical Methods in Natural Language Processing: 404-411, 2004.

[39] S.H. Zhong, Y. Liu, B. Li, J. Long, "Query-oriented unsupervised multi-document summarization via deep learning model," Expert Syst. Appl., 42(1): 8146-8155, 2015.

[40] M. Yousefi-Azar, L. Hamey, "Text summarization using unsupervised deep learning," Expert Syst.Appl., 68: 93-105, 2017.

[41] S.V. Moravvej, S.J. Mousavirad, M.H. Moghadam, M. Saadatmand, "An LSTM-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes," arXiv preprint arXiv:2110.08771, 2021.

[42] Q. Dou, Y. Lu, P. Manakul, X. Wu, M.J. Gales, "Attention forcing for machine translation," arXiv preprint arXiv:2104.01264, 2021.

[43] Y. Zhang, Y. Peng, "Research on answer selection based on LSTM," in Proc. 2018 International Conference on Asian Language Processing (IALP): 357-361, 2018.

[44] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[45] Y. Liu, "Fine-tune BERT for extractive summarization," arXiv preprint arXiv:1903.10318, 2019.

[46] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," arXiv preprint arXiv:1908.08345, 2019.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets," Advances in Neural Information Processing Systems, 27, 2014.

[48] M. Mirza, S. Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv:1411.1784, 2014.

[49] T. Mikolov, K. Chen, G. Corrado, J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[50] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, "Teaching machines to read and comprehend," Advances in Neural Information Processing Systems, 28: 1693-1701, 2015.

[51] C.Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out: 74-81, 2004.

Biographies

Seyed Vahid Moravvej received B.Sc. in software engineering from University of Kashan and MSc in artificial intelligence from Isfahan University of Technology. His research interests are Data Mining, Deep Learning, and Optimization Algorithms. He also has experience in Python.

- Email: sa.moravvej@ec.iut.ac.ir
- ORCID: 0000-0002-8286-9521
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=cMK6FN4AAAAJ&hl=en

Mohammad Javad Maleki Kahaki graduated in B.Sc. of software engineering from University of Kashan. His research interests are Data Mining and deep learning. He also has experience in Python, Matlab.

- Email: mjmaleki@std.kashanu.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Moein Salimi Sartakhti has B.Sc. in software engineering from University of Kashan. He is Master student in software engineering from Amirkabir University of Technology. His research interests are NLP and Graph theory. He also has experience in Python.

- Email: moein.salimi@aut.ac.ir
- ORCID: 0000-0003-3945-6040
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mehdi Joodaki received his B.S. degree in Computer Engineering (software) from Lorestan University in 2016 and his M.Sc. degree in Computer Engineering (software) from Isfahan University of Technology in 2019. His research interests include Bioinformatics, Parallel Computing, Graph mining, Deep Learning and Data mining.

- Email: mehdi.joodaki@ec.iut.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

S.V. Moravvej, M.J. Maleki Kahaki, M. Salimi Sartkahti, M. Joodaki, “Efficient GAN-based method for extractive summarization,” J. Electr. Comput. Eng. Innovations, 10(2): 287-298, 2022.

DOI: 10.22061/JECEI.2021.8051.475

URL: https://jecei.sru.ac.ir/article_1629.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Energy-Efficient Variation-Resilient High-Throughput Processor Design

A. Teymouri[1], *H. Dorosti*[2,*], *M.E. Salehi*[1], *S.M. Fakhraie*[1]

[1]*Nano-Electronics Center of Excellence, School of Electrical and Computer Engineering, University of Tehran, Tehran, Iran.*

[2] *Department of Computer Systems Architecture, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

Article Info

Article History:
Received 14 August 2021
Reviewed 25 September 2021
Revised 20 November 2021
Accepted 21 November 2021

Keywords:
Massive parallel processing
Statistical static timing analysis (SSTA)
Ultra-low-energy
Variation-aware
High-throughput

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

Abstract

Background and Objectives: The future demands of multimedia and signal processing applications forced the IC designers to utilize efficient high performance techniques in more complex SoCs to achieve higher computing throughput besides energy/power efficiency improvement. In recent technologies, variation effects and leakage power highly affect the design specifications and designers need to consider these parameters in design time. Considering both challenges as well as boosting the computation throughput makes the design more difficult.

Methods: In this article, we propose a simple serial core for higher energy/power efficiency and also utilize data level parallel structures to achieve required computation throughput.

Results: Using the proposed core we have 35% (75%) energy (power) improvement and also using parallel structure results in 8x higher throughput. The proposed architecture is able to provide 76 MIPS computation throughput by consuming only 2.7 pj per instruction. The outstanding feature of this processor is its resiliency against the variation effects.

Conclusion: Simple serial architecture reduces the effect of variations on design paths, furthermore, the effect of process variation on throughput loss and energy dissipation is negligible and almost zero. Proposed processor architecture is proper for energy/power constrained applications such as internet of things (IoT) and mobile devices to enable easy energy harvesting for longer lifetime.

Introduction

The future demands of multimedia and signal processing applications forced the IC designers to utilize efficient high performance techniques in more complex System on Chips (SoCs) to achieve higher computing throughput besides energy/power efficiency improvement [1]-[2]. The requirements of target applications are widely various and demanding flexibility makes the custom design method to fail in providing intended performance.

Different parallel structures such as Single Instruction Multiple Data (SIMD), Multiple Instruction Multiple Data (MIMD) and Very Long Instruction Window (VLIW) are developed to reach design goals. These structures use parallel computing in different levels such as instruction-level, task/thread-level, and data-level while each structure has its own features and restrictions [1]. Massive parallel structures such as many-core

Doi: 10.22061/JECEI.2021.8253.499

processors, systolic arrays, and GPUs are another kind of modern parallel processors which provide dramatic computing throughput [3]-[4]. While, increasing the complexity of the design by using parallel structures leads the design to consume more energy/power and causes cooling or supply challenges.

Power budget restriction is resolved using dark silicon technique or near/sub-threshold design [5]-[6]. Near/sub-threshold design dramatically reduces the energy/power consumption of the chip. However, it causes uncertainty in timing specifications of the design due to the variations in transistor parameters within and between dies, known as process variations [7]. In the presence of variations, more development time and cost with higher guard-bands are required to ensure the correct functionality of the chip. These guard-bands directly result in performance loss and lower production yield. Process variations and achieving higher throughput and keeping the power/energy within the bounds are becoming emerging concerns in new technologies and specifically in near/sub-threshold designs [8]. It is important to know that variations in process parameters considerably affect the expected performance of the design. Furthermore, any degradation in performance or reliability requires compensation techniques that result in higher power/energy consumptions. Therefore, design of variation-resilient architectures is of main concerns in new technologies.

In this work, we design a high-throughput and yet ultra-low-energy processor architecture which is resilient against the process variations. The rest of this paper is organized as follows: In the next section, we will introduce related works and will review advantages and disadvantages of the previous architectures. In Section III, we will describe our new energy efficient architecture and its features. Then, we will analyze the implementation results and estimations. At the end of discussions, we present our parallelization structure and its efficiency in terms of performance and energy consumption. Finally, we will summarize the results and conclude the paper in Section V.

Related Works

As mentioned in the previous section, process variations affect the timing specifications of the design, and energy efficient near/sub-threshold design intensifies the effect of these variations. Therefore, variations should be considered to keep demanding performance besides energy efficiency of the design. In this section we introduce recent energy-efficient or high-throughput processors as baseline architectures to be compared with our proposed architecture.

Subliminal [9] is a processor with three different versions of architectures which are designed for energy constrained applications operating at sub-threshold region. Using complex instruction set computer (CISC) architecture with complicated addressing modes, the authors achieved 1.2pj per instruction in 130nm technology providing 0.5 MIPS processing throughput at 200mv. This processor is suitable for low or mid performance applications, while not suitable for high performance applications. Higher power supply will improve the throughput while increasing the energy consumption.

TamaRISC-CS [10]-[11] is another energy efficient processor which uses custom instruction and custom memory architecture to optimize the energy efficiency besides higher performance in a multi-core environment. This processor exploits RISC style architecture with complicated addressing modes to construct a simple energy-efficient core working with high frequency to achieve high computing throughput. The results show that 62x speed-up is achieved through efficient ASIP design. At the other hand, this processor provides limited performance with lower energy efficiency.

In [12] a 10-lane SIMD processor namely Synctium I is designed to work at 530mv and provides high throughput with low timing variation. Timing uncertainty of Synctium I is minimized using three techniques: 1- error detection and correction, 2- decoupling queue between lanes and 3- lane sparing. These techniques have also doubled the performance of the processor. This processor has poor energy efficiency providing dramatic performance.

The authors of [13] have designed a processor for battery-less systems which are able to operate using ambient energy sources. They have explored different architectural configurations such as no-pipeline, n-Stage pipeline and out-of-order execution considering the requirements of 3 different scenarios for target applications. The authors have proposed design guidelines for energy-efficient processor design with respect to the application requirements.

In [14], a novel heterogeneous dual-core processor has been proposed to meet both high throughput and ultra-low-power requirements for continuous operation of target applications. The processor architecture has two cores: one low-power core which operates in near-threshold region and a high-performance one to provide burst-mode requirements. They have proposed an energy-efficient task mapping to employ the underlying cores to achieve both requirements at the same time. The processor consumes 7.7pJ/cycle to have no deadline miss for target benchmarks.

Authors of [15] have proposed an artificial intelligence processor (AIP) with three heterogeneous units to provide performance and power requirements: 1) 8-thread search processor and decision making accelerator, 2) 3-level cache to reduce repetitive computations and 3) on-chip variation monitoring to keep operations stable. The processor is working at 0.5-1.2V consuming 1.1mW to 150mW.Exploiting an embedded variation compensation unit, energy consumption is reduced by 32% while providing 276X search speed-up in comparison to the Cortex-M3 processor.

In [16], the authors proposed a method to use dynamic timing slack of the processor to improve energy consumption. The probability of critical path activation highly depends on the application. The authors present an automated method to detect the dynamic slack based on path activation of the design and achieve 25% average power reduction.

Authors of [17] propose SPARC M7 processor which consists of eight cache clusters, and an on-chip network high speed SerDes to have higher communication rate between underlying units. Their proposed architecture totally has 32 performance improved cores with low latency cache and memory hierarchy structure which provides 3X speed-up in comparison to its previous version SPARC M3.

The authors of [18]-[19] have proposed a bit-serial structure of openMSP430 for energy-constrained sensor applications to achieve higher energy efficiency and provide energy harvesting capability. The serial structure without any ISA modifications results 42% power improvement.

The serial architecture has 1.1X higher clock rate while provides 16X lower computation throughput at nominal supply voltage which is proper for low-performance applications. Their proposed serial architecture operates in near/sub-threshold region and the supply reduction degrades the processor specifications and reduces the improvements.

Authors of [20] proposed a near-threshold processor based on RISC-V architecture for low and mid performance IoT applications. This processor works down to 250mV power supply and consumes 4.5 pJ to 9.6 pJ per cycle. RISC-V architecture is also used in [20] and [22] in a parallel structure to design a processor for ultra-low-power applications such as IoT and wearable sensing. Power consumption of Parallel Ultra-Low Power processor (PULP) proposed in [21] is lower than 12mW to process the sensor data for human machine interaction. The next PULP processor in [22] is able to deliver 2.5 GOPS consuming only 55mW. It is about 38x more energy efficient compared to ARM Cortex-M as a low power processor working with 0.7V power supply.

Authors of [23] proposed a many-core processor for digital signal processing working in sub/near-threshold regime. In this work, a many-core processor is proposed using small processing core to achieve higher throughput while consuming lower power/energy. They claimed that with a negligible hardware overhead, the overall processing time for target benchmark applications (signal processing domain) is improved by 90% and reduced the power consumption by about 30% which results more reduction in energy consumption.

In the next section, we present our proposed simple energy efficient processor core in order to achieve energy efficiency which consumes lower energy in comparison with the literature. In this architecture we have used the basic custom core to build a parallel architecture and to improve the performance of the design. This architecture provides both high throughput and energy efficiency at the same time, while to get the one, the other has sacrificed in the literature.

Processor Architecture

According to our previously published analysis [24], using bit-serial structure for designing computational units in newer technologies, specifically in sub-threshold region leads to less current leakage during computation cycles.

We have also shown that the effect of process variations on working frequency is reduced because of shorter paths, while achieving mid performance. Therefore, we designed a RISC-style serial processor as our basic processing element namely ultra-low-energy (ULE). ULE has higher energy efficiency in comparison to bit-parallel structures and would exploit parallelization to improve the computation throughput.

The proposed architecture is designed to meet the variation resilience besides energy efficiency and higher throughput.

Variation resiliency and energy efficiency are provided using a simple serial core and higher throughput is achieved by massive data level parallel structure. The proposed processor can execute the same program on different data and this feature highly improves the computation throughput.

In this section, we present our energy-efficient serial core including memory organization, controller logic, ISA, and encoding scheme. At the second step we will describe our proposed parallel structure using the basic core.

The processor architecture is based on Harvard architecture with physically separated instruction and data memories and signal pathways. We have divided the data memory into 128-Byte pages to reduce the energy consumption of the memories and also provide concurrent accesses to different pages in our parallel structure.

High-level block diagram of the processor as shown in Fig. 1, includes instruction memory, data memory (at least 128B), and the processing core.

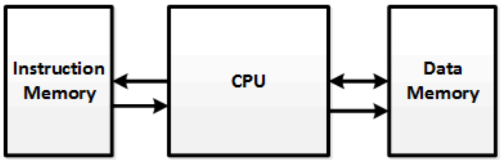


Fig. 1: system-level diagram of processor.

We have chosen a RISC-style bit-serial architecture for ULE processing core to decrease area and logic depth for diminishing process variation effects, and also improving energy efficiency. Fig. 2 shows the bit-serial structure for an adder unit which utilizes a single-bit full adder in combination with a flip-flop to perform addition at different clock cycles (SFA). This structure is more area- and energy-efficient and less vulnerable to variations in comparison to parallel structures [24]. Equations (1), (2) depicts the energy consumption(delay × power) of parallel and serial structures(respectively) to do the same computation considering the variation effects and (3) shows the energy ratio for both structures(nominator is parallel and denominator is serial) which serial structure is 8 times energy efficient.

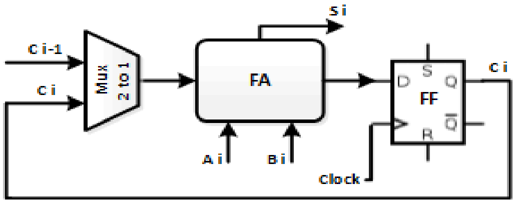


Fig. 2: Single-bit full adder (SFA) in combination with a flip flop to do n-bit addition sequentially at subsequent clock cycles [13].

$$E_{\text{parallel-8bit}} = (8 * \mu + 3*\sqrt{8}\sigma) * 8\, P_{\text{full-adder}} \qquad (1)$$

$$E_{\text{Serial-8bit}} = 8 *(\mu + 3*\sigma) * P_{\text{full-adder}} \qquad (2)$$

$$\frac{E_P}{E_S} = \frac{8\mu+6\sqrt{2}\sigma}{\mu+3\sigma} \sim \frac{8\mu+8\sigma}{\mu+3\sigma} \sim 8 \qquad (3)$$

where μ is the average delay for a full adder unit (FA) and σ defines the standard deviation of delay for the same unit.

Fig. 3 represents the block diagram of the ULE processing core. Data path of the processor includes fetch and decode unit, data memory buffer, constant register, serial shift register file, ALU, and controller units. The fetch unit reads an instruction from instruction memory, then the decode unit decodes the instruction and generates controlling signals. Data memory controller calculates the address of load/store including both page and entity addresses, and then it activates read/write signal serially from/to data memory buffer. Jump shift register is used to save jump address, and then this address is transferred to program counter (PC) for next fetch as the next instruction address. Jump target address would be a direct immediate or register-indirect address.

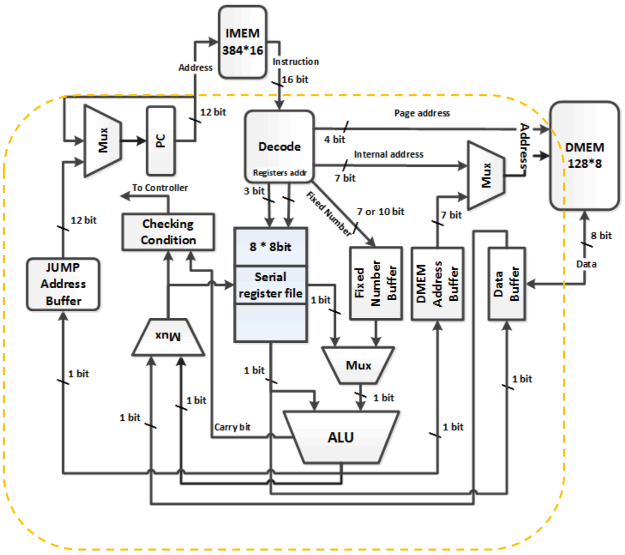


Fig. 3: Structure of ULE core.

ALU is designed in a bit-serial structure and input data is received sequentially while the previous registered carry bit (for add/sub) is used to calculate the next carry and sum. Therefore, the result is generated serially one bit per cycle and the number of clock cycles to complete the operation is equal to the number of bits. Generated results are shifted into register file bit by bit and the operation is performed using a multiplexer. Fig. 4 shows the ALU structure.

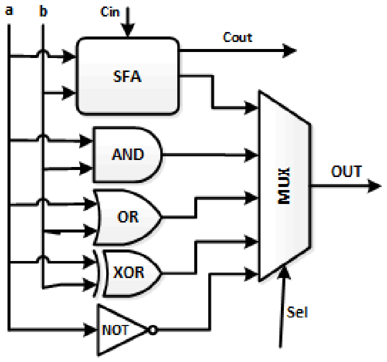


Fig. 4: ALU structure of the CPU.

Serial register file or shift register file of CPU consists of eight 8-bit shift registers, which shifts the data to the destination address at write-back stage in a pipelined manner. The ALU and processor data-path are configurable and designed to operate in various data widths such as 8, 16 or 32 bits. In this case we have used 8-bit version by 8 consequent clock cycles for each instruction.

The RISC-based instruction set architecture (ISA) of the CPU is designed based on a simplified instruction set. An important parameter in ISA design is the number and variety of supported operands. We have chosen2-operand load-store RISC-style ISA for our proposed architecture which is shown in Table 1. The choice of 2-operand instructions is based on [2] which shows that for the representative set of applications, the 3-operand choice generates about 10% larger code in comparison with 2-operand one. The encoding scheme of 16-bit instructions is shown in Fig. 5 and Table 1. The instructions set format have five generic fields: I (Immediate and condition), Op (operation code), Rd (destination register), Rs (source register), and condition field. Encoding scheme is defined in a regular structure: I and OP sections are available for all 7 instruction types and the other parts are specific to the operation and operands of each type. For instruction types listed in rows 1, 3, 4, 5 and 6 in Table 1, different execution modes are provided, which are conditional, unconditional and immediate execution. These modes can be selected by two bits (as I section) at the beginning of instruction encoding.

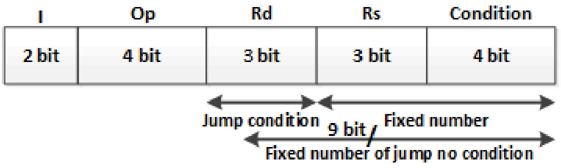


Fig. 5: Generic structure format of instructions.

Table 1: ISA Organization. s: sign bit, c: carry, g: greater than, l: lower than, ge: equal or greater than, le: lower than or equal, e: equal, ne: not equal. Shift input can be 0 (I0), 1 (I1), s, c. Source register (rs) and destination register (rd) are from the eight registers r0 to r7

#	Instructions	Code format	Structure format
1	Add, Sub, And, Or, Xor, Mov, Cmp	(Condition) rd,rs; rd,rs; rd,#Fixed_num;	00 or 01 ; Op 4 bit ; Rd 3 bit ; Rs 3 bit ; Condition 4 bit 10 ; Op 4 bit ; Rd 3 bit ; Fixed number 7 bit
2	Not	rd; (Condition) rd;	00 or 01 ; Op 4 bit ; Rd 3 bit ; xxx ; Condition 4 bit
3	Shl, Shr	rd; (Condition) rd; rd, shift_Input, #num_shift	00 or 01 ; Op 4 bit ; Rd 3 bit ; Shift input 3 bit ; Condition 4 bit 10 ; Op 4 bit ; Rd 3 bit ; Shift input 4 bit ; Fixed number 3 bit

4	Load, Store	rd,[#address_fixed_num];
		rd,[rs];
			10 ; Op 4 bit ; Rd 3 bit ; Fixed number 7 bit
			00 or 01 ; Op 4 bit ; Rd 3 bit ; Rs 3 bit ; Condition 4 bit

5	Jump	(Condition) #jump_address;
		#jump_address;
			11 ; Op 4 bit ; Condition 4 bit ; Fixed number 6 bit
			10 ; Op 4 bit ; Fixed number 10 bit

6	Stc, Ldc	; (Condition);	00 ; Op 4 bit ; xxxxxxxxxx 10 ; Op 4 bit ; Condition 4 bit ; xxxxxx
7	Chp (chpd), Chpi	#Page_address;	10 ; Op 4 bit ; D/M 1bit ; Page number 4 bit ; xxxxxx

Results and Discussion

We have synthesized our proposed processor using 90nm sub-threshold library [24] and estimated the power consumption using VCD files at different supply voltages ranging from 0.2V to 1.0V. Firstly, we present the evaluation results of basic core in comparison to the own implemented subliminal processor in 90nm (named as SSL2 in figures).

Performance, power, and energy estimations is done using FIR and TEA programs [9], [26], [27], [30] as signal processing and encryption algorithms which are used in different applications.

Finally, we will present the parallelization results and throughput improvements. Fig. 6 presents the evaluation flow chart.

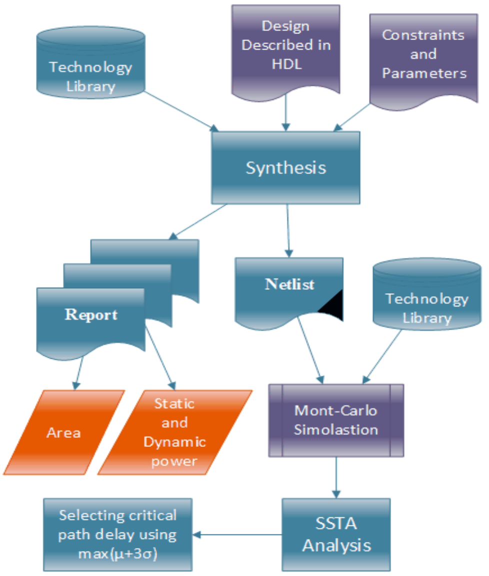


Fig 6: Evaluation flow chart for proposed architecture and re-implemented subliminal processor [9] in 90nm technology.

A. Area

Table 2 summarizes the area results for our proposed architecture in comparison with our implementation of subliminal core. This comparison is done through the same size of program and data memory for fair comparisons. According to this table, with different cores and the same memory area, our proposed architecture has almost 1% smaller area due to bit-serial core structure. The area reduction relaxes the parallelization limits to achieve higher performance/throughput besides lower die cost. Due to bigger memory area in comparison to processing cores in both architectures, total area reduction is not considerable, however, in contrast, excluding the memory area, core area is reduced by 35%.

Table 2: Area results of proposed processor in custom90nm CMOS

Architecture	Subliminal	Proposed	DSP [22]
Total Area(μm2)	242228	240639	248631
Core Area(μm2)	4540	2951	10943

B. Power Consumption

Leakage power in finer technologies and especially in near/sub-threshold region plays an important role and directly relates to the design area. Proposed architecture has lower leakage power due to area reduction. Table 3 and Fig. 7 present the total power consumption of both architectures running benchmark applications at different supply voltages.

According to these results, the proposed architecture dissipates 75% lower power in comparison to the baseline processor.

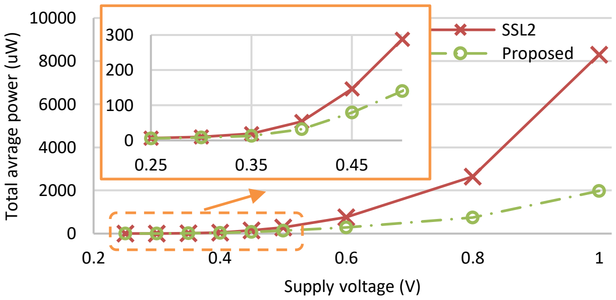


Fig. 7: Total power consumption of the proposed architecture in comparison to subliminal [9].

Table 3: Powerconsumtion results in custom90nm CMOS

Voltage (v)	Subliminal (uW)	Proposed (uW)	DSP [22] (uW)
1.0	8305	1980	-
0.6	754	293	200
0.4	54	31.7	40
0.3	10.1	8	10

C. Performance and Throughput

In the next step, we use different metrics for performance evaluation. The critical path delay indicates the working frequency as a measure of processor performance. Clock per instruction (CPI) is another important factor which is useful in conjunction with working frequency to figure out the other important performance parameters such as instruction latency, execution/computation throughput, and application latency. Million instructions per second (MIPS) is also a well-known parameter to measure the execution throughput of any processor.

Fig. 8 and Table 4 summarize the critical path delay and working frequency at different supply voltages in comparison to the subliminal processor. Also, the required clock cycles to execute each instruction are shown in Table 5. Proposed architecture has about 3x higher clock rate in comparison to the baseline architecture, while more clock cycles are needed to execute each instruction. Combining these parameters (which is the product of critical path delay and CPI) will show that the instruction latency for the proposed architecture is almost 2.6xmore than subliminal. For example, at 1V supply voltage instruction latency of the proposed and baseline architecture is 8.4ns and 3.2ns, respectively.

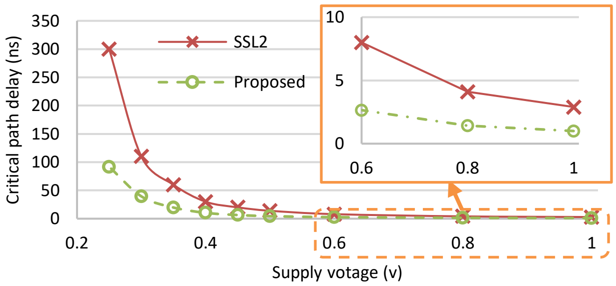


Fig. 8: Critical path delay at different supply voltages.

Table 4: Working frequency in custom90nm CMOS (MHz)

Voltage (v)	Subliminal [9]	Proposed	DSP [22]
1.0	345	1000	350
0.6	125	377	130
0.4	33	97	40
0.3	9	25	15

Table 5: Performance results (CPI)

	Subliminal [9]		Proposed		DSP [22]
CPI	FIR	TEA	FIR	TEA	FIR
	1.2	1	6.5	10.3	1.070
Average	1.1		8.4		1.070

Considering instruction latency, Table 6 and Fig. 9 summarize the MIPS results for both architectures.

According to these results, proposed architecture is capable of providing demanding performance.

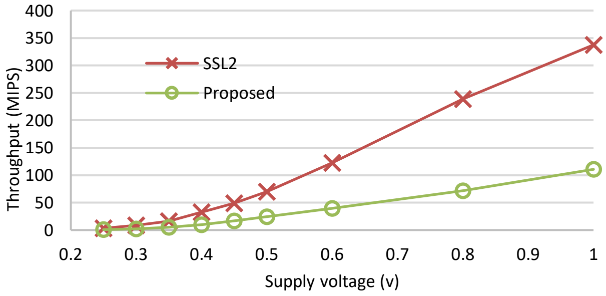


Fig. 9: Throughput vs supply voltages for both architectures.

Table 6: Performance results in90nm CMOS (MIPS)

Voltage (V)	Subliminal	Proposed
1.0	337.5	110.6
0.6	122.3	39.8
0.4	32.6	9.9
0.3	8.9	5.1

D. *Energy Consumption or Power-Delay Product*

Energy consumption is another important metric to be considered in energy-constrained systems. It is important to know that higher performance improves energy efficiency due to dominant effect of leakage power in newer technologies especially in near/sub-threshold regions. Therefore, a system with lower power consumption and higher performance could be more energy efficient in comparison to a system with lower power and lower performance.

Energy per instruction is a common parameter for comparing two different architectures, which is the product of instruction latency and total power consumption. Furthermore, total energy dissipation of application is another important factor which indicates the processor efficiency. Fig. 10 shows the energy per instruction of both architectures in average for FIR and TEA applications at different supply voltages. Table 7 and Table 8 summarize the energy per instruction and total energy consumption of both processors at different supply voltages.

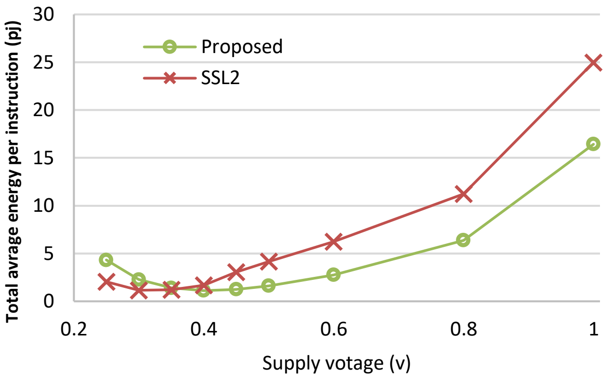


Fig. 10: Total energy consumption of proposed architecture in comparison to subliminal.

Table 7: Energy per instruction (pJ) in custom90nm CMOS

Voltage (V)	Subliminal [9]	Proposed	DSP [22]
1.0	25	16.4	-
0.6	6.2	2.8	3
0.4	1.7	1.1	1.2
0.3	1.2	2.3	0.8

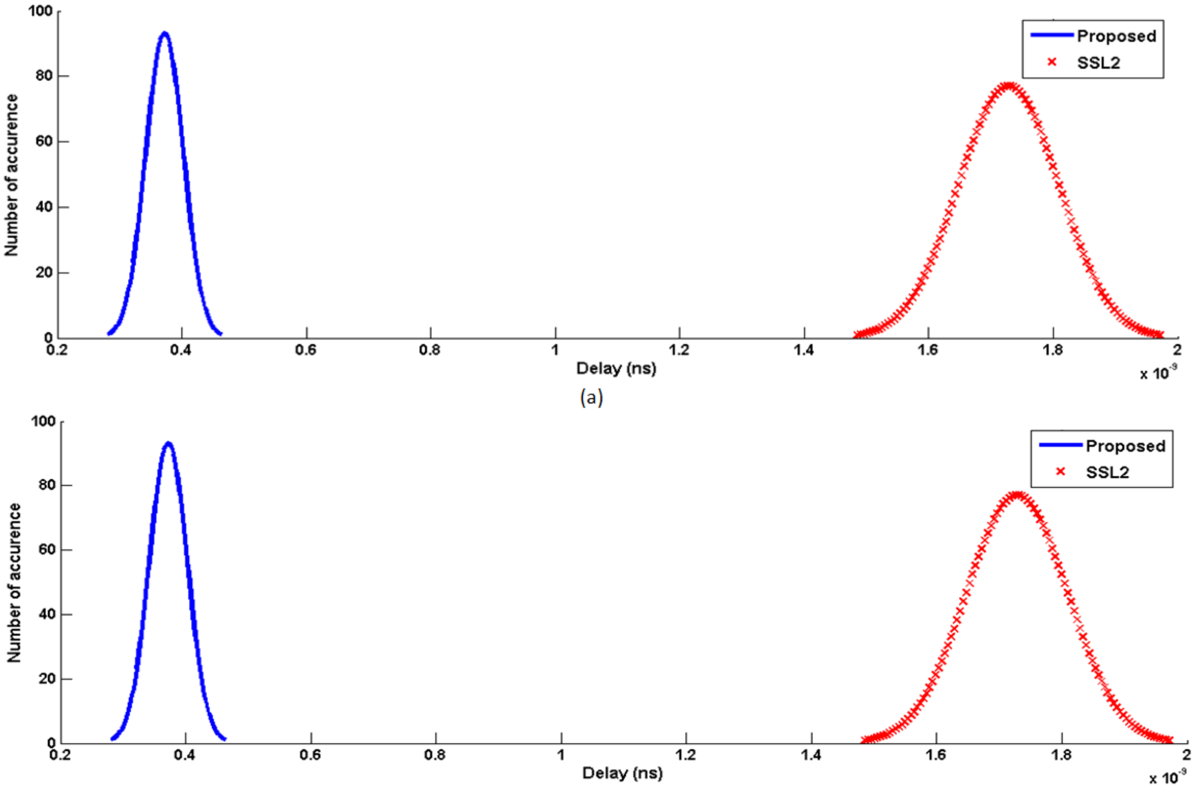
Table 8: Total energy dissipationin custom90nm CMOS (nJ)

Voltage (V)	Subliminal [9]	Proposed	DSP [22]
1.0	2070	1678	-
0.6	493	200	270
0.4	130	85.5	130
0.3	93	43.6	70

According to the results, in average, the proposed architecture has 30%lower energy per instruction and also 20% lower total energy consumption at 1V. The improvement would be even better at lower supply voltages.

E. Variations Analysis

As mentioned previously, the effect of process variations in newer technologies has grown and should be considered in design time to boost demanding performance and energy efficiency. Since shorter paths means higher working frequency and lower performance loss, due to process variation, the proposed architecture suffers almost 60% less than subliminal in terms of path delay. Fig. 11 shows the delay variation of both processors at 1V and 0.4V supply voltages in comparison. This reduction also reduces performance loss and energy overhead. The mean (µ) and standard deviation (σ) of path delay distribution for the proposed architecture (normalized to the delay distribution of the baseline processor) at different supply voltages are presented in Fig. 12 and Fig. 13. According to these figures, the proposed architecture has more compact delay distribution and higher working frequency and these differences are far in higher voltages which is compatible with the results of Fig. 11. Due to lower power consumption, the proposed architecture is capable to operate in higher voltages with the same power consumption to achieve higher working frequency and reduce the performance difference in comparison to the subliminal.



(a)

(a)

Fig. 11: delay distribution for a) 1V and b) 0.4V.

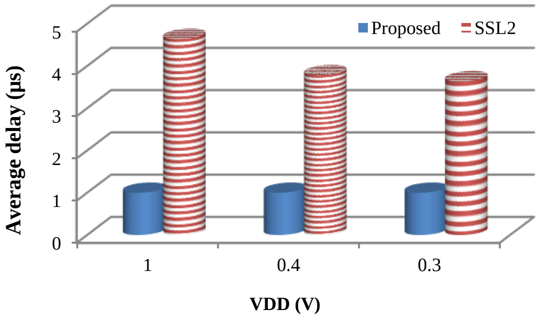


Fig. 12: Average delay (µ) parameter of critical paths normalized to proposed design.

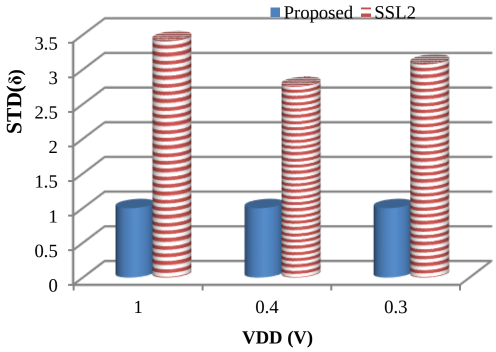


Fig. 13: Standard deviation (σ) parameter of critical paths normalized to proposed design.

F. Parallelization

The results show that the proposed processor has smaller area and lower cost besides higher energy efficiency, lower performance, and lower energy overhead in case of process variation. On the other hand, the proposed core provides similar performance compared to baseline processor. Achieving higher performance/throughput could be provided by parallelizing energy-efficient basic structure. Utilizing serial data-path of basic core in SIMD parallel structure enables the processor to run the same instructions on different data. Using paged data memory provides parallel access to data blocks without any contention and access overhead. Fig. 14 shows the proposed parallel architecture utilizing the energy-efficient basic core. This parallel structure is capable of achieving higher throughputs accordant to application demands. Exploiting data level parallelization reduces the dependency between different processing lanes and reduces the complexity. Therefore, the proposed parallel structure could deliver higher performance besides energy and cost efficiency.

The proposed parallel architecture is organized to use the controlling part of each lane as a common shared unit which is known as shared part in Fig. 14. This unit decodes the instructions and manages the processing lanes with proper controlling signals. Processing unit and other resources such as register file, address and data buffer (for accessing data memory) are isolated for each lane known as private parts (Fig. 15). The parallel architecture is optimized to reduce the parallelization overhead and keep energy efficiency besides higher throughput.

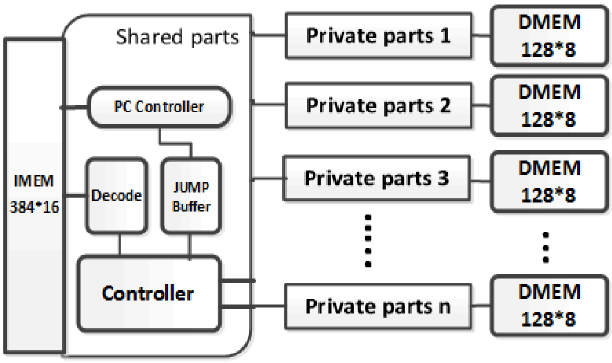


Fig. 14: SIMD processor architecture.

Table 9 summarizes the performance results of parallel SIMD structure using basic data-path with different parallelization degree. In this table, the candidate architectures working at 0.4V and performance results at this voltage are listed for different processors (except for Synctium I which is functional above 0.53V). It is important to know that we have considered the proposed simple core to work with the same frequency as proposed SIMD version and the parallelization overhead is considered in single core structure.

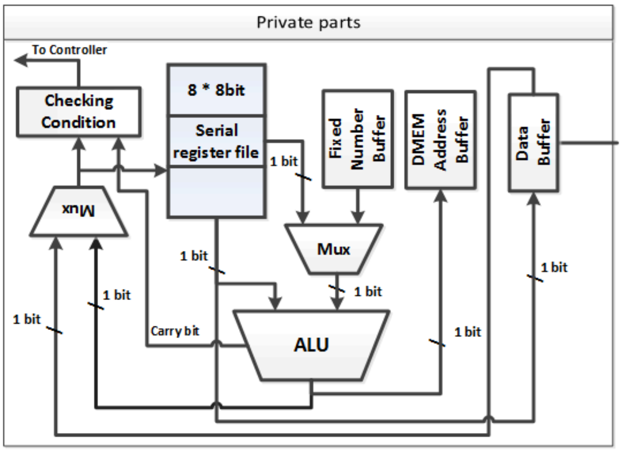


Fig. 15: Private parts of each SIMD lane.

Table 9: MOPS results for candidate processorsat 0.4V.

Processor	Technology (nm)	Energy per instruction (pJ)	Throughput (MOPS)	Freq. (MHz)
Subliminal [9]	90	1.7	19	33
Proposed	90	1.1	9.5	97
8-lane proposed SIMD	90	2.75	76	97
TamaRISC-CS [10]	90	5	0.8	0.11
Synctium-I [12]	45	165	680	85
Many-Core [22]	90	0.8	-	40

Taking an analytical look at performance results may lead to superiority of subliminal [9] running at lower supply voltages in comparison to the proposed core with higher supply voltage, which consumes lower power while providing similar performance with small loss. At the other hand, due to multi-cycle completion of instructions in proposed core, process variations have more effects on the proposed core according to accumulative manner of variation in timing in consecutive cycles. The conclusion is correct for single core design without considering the parallelization. We have designed the underlying energy efficient core to utilize in parallel structure. Then it is important that the goal of basic core is to reduce the energy consumption and the duty of parallel structure is to increase the performance with smaller overhead. Trying to use the subliminal in a parallel structure needs more hardware overhead and the energy requirements will grow in parallel structure. Due to energy efficient basic core, 8-lane proposed SIMD only consumes energy less than 2x of subliminal while providing 4x more performance. in order to achieve the same performance using 4 subliminal cores in parallel structure, the processor will consume energy 2.5x more than proposed SIMD core.

According to the performance results for the proposed and baseline processors (listed in Table 9), the proposed architecture provides higher computation throughput while consuming lower energy. On the other words, performance loss and energy overhead due to process variation is reduced. Massive parallel structures such as GPUs and systolic arrays have bright future for applications with high performance and low energy demands such as multimedia applications and specially hand-held and portable devices. Enabling partial power gating for each lane provides more flexibility for these architectures to support wider applications from low to high performance requirements.

Conclusion

Recent technology advancements help the designer to achieve higher performance realizing complex SoCs. Despite performance improvements, feature size scaling has made the power/energy consumption more challenging and process variation has intensified the crisis.

In this paper, we have proposed area and energy/power efficient bit-serial processor architecture which delivers acceptable performance. On the other words, the performance loss due to process variation is reduced by 60%while consuming 35% less energy. The proposed core exploits RISC ISA with serial structure to reduce power/energy consumption and increases energy efficiency. Furthermore, shorter paths improve performance at the presence of process variation, and reduce performance loss due to fluctuations.

Higher performance/throughput is achieved using parallel structure and banked memory organization. Using the proposed parallel structures dramatically improves energy efficiency besides cost and performance improvements.

It is important to note that when talking about ultra-low-power/energy and near/sub-threshold computation, there are similarity between 90nm and up-to-date technologies (below 10nm) in the ratio between static and dynamic power/energy consumption and sensitivity of the design to any change in path delays. Employed design techniques in sub-threshold regime for 90nm are applicable for newer technologies due to mentioned similarities.

Author Contributions

A. Teymouri, H. Dorosti, and M. E. Salehi designed the experiments with the guidance of S. M. Fakhraie. A. Teymouri collected the data with the help of H. Dorosti through proper simulations, and data analysis is carried out by A. Teymouri and H. Dorosti. Finally, A. Teymouri, H. Dorosti, and M. E. Salehi interpreted the results and wrote the manuscript.

Acknowledgment

The authors would like to thank Alireza. Mazraee for developing sub-threshold library cells.

Conflict of Interest

Authors declare that there is no conflict of interests 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

DSP	Digital Signal Processor
ISA	Instruction Set Architecture
FFT	Fast Fourier Transform
TEA	Tiny Encryption Algorithm
FIR	Finite Impulse Response
RFID	Radio Frequency Identification
ALU	Arithmetic and Logic Unit
MAC	Multiply and Accumulate
RISC	Reduced Instruction Set Computer
CISC	Complex Instruction Set Computer
CPI	Clock Per Instruction
MIPS	Million Instruction Per Second
VCD	Value Change Dump

References

[1] J.L. Hennessy, D.A. Patterson, Computer Architecture: A Quantitative Approach, 4th Edition, San Francisco: Morgan Kaufmann, 2006.

[2] B. Zhai, et al., "Energy-efficient subthreshold processor design," IEEE Trans. Very Large Scale Integr. VLSI Syst., 17: 1127-1137, 2009.

[3] J.D. Owens, et al., "GPU computing," Proc. IEEE, 96(5): 879-899, 2008.

[4] K.T. Johnson, et al., "General-purpose systolic arrays," Computer, 26: 20-31, 1993.

[5] H. Esmaeilzadeh, et al., "Dark silicon and the end of multicore scaling," in Proc. 2011 38th Annual International Symposium on Computer Architecture (ISCA): 365-376, 2011.

[6] M.B. Taylor, "A landscape of the new dark silicon design regime," IEEE Micro, 33(5): 8-19, 2013.

[7] S.R. Sarangi, et al., "VARIUS: A model of process variation and resulting timing errors for microarchitects," IEEE Trans. Semicond. Manuf., 21: 3-13, 2008.

[8] J. Crop, et al., "Design automation methodology for improving the variability of synthesized digital circuits operating in the

sub/near-threshold regime," in Proc. 2011 International Green Computing Conference and Workshops (IGCC): 1-6, 2011.

[9] L. Nazhandali, et al., "Energy optimization of subthreshold-voltage sensor network processors," in Proc. 32nd International Symposium on Computer Architecture (ISCA'05): 197-207, 2005.

[10] J. Constantin, et al., "TamaRISC-CS: An ultra-low-power application-specific processor for compressed sensing," in Proc. 2012 IEEE/IFIP 20th International Conference on VLSI and System-on-Chip (VLSI-SoC): 159-164, 2012.

[11] J.H.F. Constantin, "Application-specific processor design for low-complexity & low-power embedded systems," in Winter School on Design Technologies for Heterogeneous Embedded Systems (FETCH), 2013.

[12] R. Pawlowski, et al., "A 530mV 10-lane SIMD processor with variation resiliency in 45nm SOI," in Proc. 2012 IEEE International Solid-State Circuits Conference (ISSCC), 2012.

[13] K. Ma, et al., "Architecture exploration for ambient energy harvesting nonvolatile processors," in Proc. IEEE 21st International symposium on High Performance Computer Architecture (HPCA), 2015.

[14] Z. Wang, et al., "An energy-efficient heterogeneous dual-core processor for internet of things," in Proc. IEEE International Symposium on Circuits and Systems (ISCAS), 2015.

[15] Y. Kim, D. Shin, J. Lee, Y. Lee, H. J. Yoo, "14.3 A 0.55V 1.1mW artificial intelligence processor with PVT compensation for micro robots," in proc. IEEE International Solid-State Circuits Conference (ISSCC), 2016.

[16] H. Cherupalli, R. Kumar, J. Sartori, "Exploiting dynamic timing slack for energy efficiency in ultra-low-power embedded systems," in Proc. IEEE International Symposium on Computrer Architecture, 2016.

[17] G. K. konstadinidis, et al., "SPARC M7: A 20nm 32-Core 64MB L3 Cache Processor," IEEE J. Solid-State Circuits, 51(1): 79-91, 2016.

[18] M. Tomei, H. Duwe, N.S. Kim, R. Kumar, "Bit serializing a microprocessor for ultra-low-power," in Proc. ISLPED 2016: 200–205, 2016.

[19] C. Kelly, V. Ekanayake, R. Manohar, "SNAP: A sensor-network asynchronous processor," in Proc. Ninth International Symposium on Asynchronous Circuits and Systems: 24-33, 2013.

[20] R. Uytterhoeven, W. Dehaene, "A Sub 10 pJ/Cycle Over a 2 to 200 MHz Performance Range RISC-V Microprocessor in 28nm FDSOI," in Proc. IEEE 44thEuropean Solid-State Circuits Conference (ESSCIRC), 2018.

[21] V. Kartsch, M. Guermandi, S. Benatti, F. Montagna, L. Benini, "An Energy-Efficient IoT node for HMI applications based on an ultra-low power multicore processor," in Proc. IEEE Sensors Applications Symposium (SAS), 2019.

[22] M. Eggimann, S. Mach, M. Magno, L. Benini, "A risc-v based open hardware platform for always-on wearable smart sensing," in Proc. IEEE 8thInternational Workshop Advances in Sensors and Interfaces (IWASI), 2019.

[23] B. Soltani, H. Dorosti, M.E. Salehi, S.M. Fakhraie., "Ultra-low-energy DSP processor design for many-core parallel applications," J. Electr. Comput. Eng. Innovations, 8 (1): 71-84, 2019.

[24] H. Dorosti, et al., "Ultralow-energy variation-aware design: adder architecture study," IEEE Trans. Very Large Scale Integr. VLSI Syst. (TVLSI), 24(3): 1165-1168, 2016.

[25] M. Wang, N. Yu, W. Ma, Q. Sheng, W. Zhang, Z. Huang, " An ultra low-power processor with dynamic regfile configuration," in Proc. 2018 IEEE International Conference on Solid-State and Integrated Circuits Technology (ICSICT): 1-3, 2018.

[26] P. Meinerzhagen, S.M. Sherazi, A. Burg, J.N. Rodrigues, “Benchmarking of standard-cell based memories in the sub-vt domain in 65-nm CMOS technology,” IEEE J. Emerging Sel. Top. Circuits Syst., 1(2): 173-182, 2011.

[27] L. Nazhandali, M. Minuth, T. Austin, "Sensebench: toward an accurate evaluation of sensor network processors," in Proc. 2005 IEEE Workload Characterization Symposium: 197-203, 2005.

[28] S. Yin, P. Ouyang, J. Yang, T. Lu, X. Li, L. Liu, S. Wei, "An ultra-high energy-efficient reconfigurable processor for deep neural networks with binary/ternary weights in 28nm CMOS," in Proc. IEEE Symposium on VLSI Circuits: 37-38, 2018.

[29] A. Wang, B.H. Calhoun, A.P. Chandrakasan, Design for Ultra Low-Power Systems, New York: Springer, 2006.

[30] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, D. Blaauw, “Energy-efficient subthreshold processor design,” IEEE Trans. Very Large Scale Integr. VLSI Syst., 17 (8): 1127-1137, 2009.

Biographies

Ali Teymouri was born in Minoodast City, in 1987. He received the B.S. and degrees in computer (hardware) engineering from the Hamedan University of Technology, Hamedan, Iran, in 2012 and the M.S. degree in Computer Architecture engineering from Tehran University, Tehran, Iran, in 2015. His research interests include embedded systems applications and ultra-low-power processor architecture, multi core and SIMD processor architecture.

- Email: a.teymouri@ut.ac.ir
- ORCID: 0000-0002-5982-8983
- Web of Science Researcher ID: NA
- Scopus Author ID: 57193562636
- Homepage: NA

Hamed Dorosti was born in Khoy, in 1986 and received the B.S. and M.S. degree in computer engineering from University of Tehran, Tehran, Iran, in 2009 and 2011, respectively. He received his Ph.D. in computer engineering (computer architecture) from University of Tehran in 2017. Since 2009, he was member of Silicon Intelligence and VLSI Signal Processing Lab., University of Tehran and co-operated in low-power ASIP project from 2010 to 2012. His research interest includes VLSI design, digital signal processing, adaptive timing error detection and correction and low-power high-throughput/performance processor architecture design considering static and dynamic variations. He is now an assistant professor of Shahid Rajaee University.

- Email: hdorosti@sru.ac.ir
- ORCID: 0000-0001-6554-1607
- Web of Science Researcher ID: L-5928-2019
- Scopus Author ID: 36662029700
- Homepage: https://www.sru.ac.ir/dorosti

Mostafa Ersali Salehi Nasab was born in Kerman, Iran, in 1978. He received the B.Sc. degree in computer engineering from University of Tehran, Tehran, Iran, and the M.Sc. degree in computer architecture from University of Amirkabir, Tehran, Iran, in 2001 and 2003, respectively. He has received his Ph.D. degree in school of Electrical and Computer Engineering, University of Tehran, Tehran, Iran in 2010. From 2004 to 2008, he was a senior digital designer working on ASIC design projects with SINA Microelectronics Inc., Technology Park of University of Tehran, Tehran, Iran. He is now an Assistant Professor in University of Tehran. His research interests include novel techniques for high performance, low-power, and fault-tolerant embedded system design.

- Email: mersali@ut.ac.ir
- ORCID: 0000-0003-1733-6056
- Web of Science Researcher ID: F-1701-2011
- Scopus Author ID: 57194140135
- Homepage: https://profile.ut.ac.ir/en/~mersali

Sied Mehdi Fakhraie received the M.Sc. degree in electronics from the University of Tehran, Tehran, Iran, in 1989, and the Ph.D. degree in electrical and computer engineering from the University of Toronto, Toronto, Canada, in 1995. He was a Professor with the School of Electrical and Computer Engineering, University of Tehran. He was the Director of the Silicon Intelligence and VLSI Signal Processing Laboratory, the Director of Electrical and Electronics Engineering, and the Director of Computer Hardware Engineering with the School of Electrical and Computer Engineering, University of Tehran. He was a Visiting Professor with the University of Toronto, in 1998, 1999, and 2000, where he was involved in efficient implementation of artificial neural networks. He was with Valence Semiconductor Inc., Irvine, CA, USA, from 2000 to 2003. He was in Dubai, United Arab Emirates, and Markham, Canada Offices of Valence as the Director of Application Specified Integrated Circuit (ASIC) and System-on-a-Chip Design, and the Technical Leader of integrated broadband gateway and family radio system baseband processors. He was involved in many industrial integrated circuit design projects, including design of network processors and home gateway access devices, DSL modems, pagers, and digital signal processors for personal and mobile communication devices. He has co-authored a book entitled VLSI-Compatible Implementations for Artificial Neural Networks (Boston, MA, USA: Kluwer, 1997). He has authored or co-authored over 230 reviewed conference and journal papers. His last research interests include system design and ASIC implementation of integrated systems, novel techniques for high-speed digital circuit design, and system-integration and efficient VLSI implementation of intelligent systems. He passed away on December 7, 2014.

- Email: fakhraie@ut.ac.ir
- ORCID: 0000-0003-3507-3261
- Web of Science Researcher ID: F-5138-2011
- Scopus Author ID: 6603636269
- Homepage: https://profile.ut.ac.ir/~fakhraii

How to cite this paper:

A. Teymouri, H. Dorosti, M.E. Salehi, S.M. Fakhraie, “Energy-Efficient Variation-Resilient High-Throughput Processor Design,” J. Electr. Comput. Eng. Innovations, 10(2): 299-310, 2022.

DOI: 10.22061/JECEI.2021.8253.499

URL: https://jecei.sru.ac.ir/article_1637.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

SRTTU

JECEI

Research paper

Semantic Enterprise Architecture Oriented Test case Generation for Business Process

***M. Rahmanian*[1], *R. Nassiri*[2], *M. Mohsenzadeh*[1], *R. Ravanmehr*[2]**

[1]*Department of computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.*

[2]*Department of computer Engineering, Central Tehran Branch, Islamic Azad University, Tehran, Iran.*

Article Info

Article History:

Received 10 August 2021
Reviewed 10 September 2021
Revised 22 November 2021
Accepted 24 November 2021

Keywords:

Enterprise architecture
Business process
Syntactic and semantic modeling
Test case generation

*Corresponding Author's Email Address: *r_nasiri@iauctb.ac.ir*

Abstract

Background and Objectives: The area of enterprise architecture encompasses various domains, the most complicated of which concerns developing enterprise business architecture. Although many state-of-the-art enterprise architecture frameworks describe the architecture by abstract levels, they still fail to provide accurate syntactic and semantic descriptions. Several previous conducted studies were looking for different objectives elaborated on modeling enterprise architectures. However, none of those studies tried to develop a modeling that generates test cases which would later be used for validation and/or verification. Therefore, the main contribution of this study is generating a set of test cases based on the descriptions yielded from enterprise business processes in early steps; then, the amount of later reviews and changes can be significantly lessened.
Methods: Following the objective of accurate validation and/or verification of the enterprise business processes within an enterprise's architecture development, this paper proposes a new method based on the enterprise architecture design. Throughout the iterative cycle of the proposed method, initially, the enterprise goals will be extracted based on the TOGAF framework. Afterwards, it will be subjected to syntactical modeling based on the Archimate language. Then, semantics will be added to the syntactic model of the enterprise business processes based on the WSMO framework and formalize manually to B language by using defined transition rule. Therefore, in order to discover test cases, a set of test coverage will be tested on the formal model.
Results: The proposed method has been implemented in the marketing and sales department of a petrochemical corporation, where the results show the validity and also the effectiveness of the method. Based on the implementation of our method on the selected case study, the details of the business process have been defined based on an enterprise level, the level of abstraction is decreased by syntactic and semantic modeling of enterprise architecture description, the formal descriptions created using the proposed transition rules for sampling.
Conclusion: The proposed method starts from the goals of enterprises; therefore, the output samples are efficiently precise. By adding semantics to the syntactic models of enterprise architecture, the degree of abstraction has been decreased. By creating a formal model, the model can be subjected to sampling. For future work, it is suggested to use the proposed method for the automatic generation of codes.

Introduction

Enterprise architecture is a comprehensive integrated approach that separates and analyzes an enterprise in various aspects and objects from an engineering, but IT-based point of view, to acquire a better understanding of

Doi: 10.22061/JECEI.2021.8218.496

the entire structures and elements of an enterprise, as well as the forms in which they are connected. Currently, there are multiple enterprise architecture frameworks available worldwide, all of which provide a highly abstract description of an enterprise architecture [1]. In this context, one of the most deployed among those frameworks is the TOGAF, which is based on an iterative development method known as ADM [1]. Based on the TOGAF framework, every enterprise architecture is to be shaped of four domain layers known as the business, data, application, and technology architecture layers [2].

Software testing is a critical and costly process in the Software Development Lifecycle. In fact, a considerable portion of the cost of producing reliable software is associated with this process phase [3], [4]. Nevertheless, if one can generate a set of test cases based on the descriptions yielded from enterprise business processes in early steps; then, the amount of later reviews and changes can be significantly lessened. This is the main idea behind of later steps in this paper.

Several previous studies following different objectives tried to model enterprise architecture [1], [5]-[9], [11]-[24]. However, none of those studies elaborated on modeling to generate test cases for the purpose of verification and/or validation. The main issues addressed by this paper is how to start from the enterprise level, get benefited from the enriched descriptions yielded for enterprise business processes, move to generate proper syntactic and semantic models of the descriptions, and generate prioritized test cases from the well-established models to come up with samples to be used for verification and/or validation.

Based on what aforementioned, we may cope with two main challenges:

A) Syntactic and semantic modeling of the enterprise business processes.

B) Generating proper test cases using the syntactic and semantic models.

Compared to previous studies, innovation in this paper is as follows:

- Starting from the enterprise level (vision, mission and goal) to get descriptions of the enterprise business process.
- Reducing the level of abstraction in architectural descriptions using syntactic and semantic modeling.
- Enabling model sampling using formal transition rules.

Later on, in the next section the related studies would be reviewed, and once the proposed method is explained, it would be implemented in the form of a case study for further evaluation purposes. Finally, this paper ends up with the proposed method conclusions and further suggestions for future works.

Review of Literature

A. Syntactic and Semantic Description of Enterprise Architecture

Several studies have tried to provide syntactic and semantic descriptions of enterprise architecture, while each of them used the descriptions for a specific purpose.

In [1], Zhou et al. to identify, classify, analyze, and evaluate existing methods for EA visualization, reviewed the research papers on EA visualization systematically. They selected and analyzed 112 research papers, and then they categorized them according to their purposes. In none of the studies reviewed in this study, the issue of modeling with the aim of generating test cases has been addressed. In [5], Bouafia and Molnar defined the basics concepts of EA and the purpose and utility of an EA and its place in the IS environment are discussed. The approach presented provides a formal way to use the mathematical analytic methods for exploring misalignment based on different concepts and relation between them. In [6], Hinkelmann et al. believe that modeling for humans is different from modeling for machines. They proposed a combined approach that would be suitable for both humans and machines. In order to create a graphical modeling language, a graphical symbol was designed for each ontology. However, the perspective proposed in this research is to be known as general and not pertaining to any specific domain, particularly while someone is seeking a test case generation solution at this level. In [7], Babkin proposed a method for detecting any logical paradoxes in enterprise architecture models that works under the approach of model checking (verification) in the business process model. Babkin's study uses the ArchiMate language and the MIT Alloy Analyzer tool to describe enterprise architecture and to analyze model limitations, respectively. Furthermore, the study has also developed an editor module that translates enterprise architecture models to the MIT Alloy Analyzer system's language. The main drawback of this model may be its failure to support semantics, whilst the major effort was mainly on model consistency check and syntax matters among models. In [8], Caetano attempted to address three major issues. The first one was about how to use ontology to present enterprise models; the second one was how to use ontology to integrate enterprise models; and the third one was how can use semantic computing techniques to analyze integrated enterprise models. He stated that the main challenge should be on the determination of mapping function among different schemas. Sometimes based on the extent of semantic difference among various schemas, it may become virtually impossible to select a mapping function. This study stated that a conceptual model could be determined via three components, namely as the subject, the interface, and the object. Through this method, the concepts underlying each of the model components would be described in an ontological fashion. The main drawback of this research is its focus on semantic modeling where sampling is ignored. In [9], Hinkelmann et al. proposed a combined modeling approach for convergence between business and technology. The authors used enterprise ontology proportional to BPaaS concepts. It is noteworthy that OMiLAB LifeCycle backs the development of the BPaaS design environment. The authors believed that the BPaaS ontology is the format of ArchiMEO ontology [10]. The study has expanded various models with different algorithms and mechanisms of semantic transformation to connect graphical models to the BPaaS ontology. The main pitfall here is deployment area if the proposed method for other cases is not considered by authors. In [11], Gokalo believes that the complexity of enterprise architectures is the reason that manual analysis seems to be impossible, and so he proposed using descriptive logic along with ontology. This study categorized inferential activities into five main groups: subscription, sample inspector, relation inspector, compatibility of concepts, and compatibility of the database. Each of the mentioned activities was being described by a different descriptive logic also with varying descriptive capabilities. The study used OWL to provide a high-level description of the meta-model. Using the stated descriptive logic, concepts relating to the elements and the relationships among them had been described in ArchiMate. However, still no sampling capability is available in this method. In [12], Chen et al. suggested that semantic technology should allow different datasets extracted from different data sources in enterprise, to be later integrated into an EA repository, and would prepare the basic information required for decision making corresponding to the outlook of the information systems at hand. In addition, enterprise architecture frameworks such as the TOGAF produce meta-models to be used as guides for generating EA repositories. Considering this content, the authors defined a process for generating SEAM repositories. Furthermore, the data have been subjected to ontology and related to the enterprise architecture. SEAM focuses on modeling the dependencies among the business, information systems, and IT infrastructures. However, still no sampling capability is available in this method. In [13], Hinkelmann et al. developed a framework intended to make a balance between technology and business. Model-based engineering is presented either as a graphical or as a formal model. Enterprise architecture frameworks solely display a general schema of the enterprise architecture and its structures and elements, while in some cases such as the Zackman framework, there is a lack of any specific modeling instrument. As a result, such frameworks cannot be used as a tool for decision making. The study assumed that no language can provide a formal description of an enterprise architecture definition since a perfect modeling language is the one that encapsulates the three components of syntax, semantics, notations, and symbols.

B. Generating Test Cases Based on Enterprise Architecture Description

An important debate pertaining to the domain of software testing includes generating test cases which are normally generated in different forms from various software models. In the following, previous studies related to this domain are to be discussed. Since the present study focuses on enterprise business processes, we will only discuss the studies that fall into this category.

In [3], Sharma et al. identified various factors affecting related aspects of software testing process and therefore the impact of ontology has been observed in the testing and analyzed. They believe that such an elucidation is significant for having knowledge-oriented verification and validation and the wide adoption of ontology helps the domain in manifolds. In [14], Yazdani et al. present a model-based approach to automatically generate test cases from business process models. They first model business processes and convert them to state graphs. Then, the graphs are traversed and transformed to the input format of the "Spec explorer" tool that generates the test cases. The limitations of the proposed algorithm is that it works only for well-structured processes and if the input process is not well-structured, it cannot define states correctly and it so captures invalid paths. In [15], Zhang et al. developed a tool for the automated generation of test cases based on descriptions. Their study used preconditions and post-conditions to produce formal descriptions. Their method was thoroughly based on programming which was unusable for sets containing infinite elements. In [16], Ajay et al. used the activity diagram for the automated generation of test cases. In their method, an activity flow table was established based on the activity diagram, and then based on the former table, an activity flow graph will be yielded. In addition, their study used the Genetic Algorithm to generate a set of optimal test cases. It is noteworthy that their method generates the set of test cases according to the structure of the activity diagram via selecting different paths within it. In [24], Bures et al. created a tool named PCTgen which automatically generated a set of test cases to check a workflow. Unlike previously presented methods, this method assumed that there were no UML documents available. To this end, a directed graph will be used to signify the workflow. The test cases generated through this method are resulted solely from the sequence of activities existing in the work flow, irrespective of semantics.

C. *Motivation of the paper*

Investigating the previous studies led us to the conclusion that although many studies have tried to generate semantic models of enterprise architecture, but none of them, have adopted the approach of test case generation with the aim of verification and/or validation using syntactic and semantic models. Therefore, considering enterprises objectives and missions, we intend to generate proper test cases for further use in the system through a new method known as "Semantic Enterprise Architecture Oriented Test Case Generation for Business process ", which is based on the views relating to the TOGAF and model checking of a formal model of syntactic and semantic modeling.

The Proposed Method

The overall framework of the proposed method is based on an iterative cycle, which is derived from the Architecture Development Method of the TOGAF, which itself is a stepped iterative process. The proposed method doesn't elaborate on the transition from the existing status of the enterprise to the desired status; rather than, in this method, the only input is the architecture of the desired status which will be further developed via an iterative cycle. The overall framework of the proposed method is presented in Fig. 1.

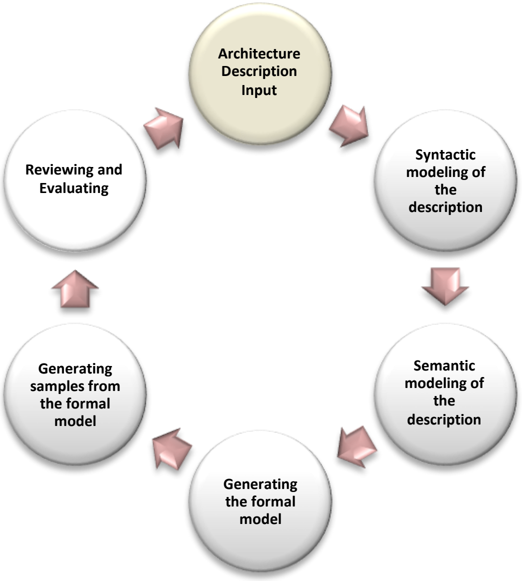

Architecture Description Input
Syntactic modeling of the description
Semantic modeling of the description
Generating the formal model
Generating samples from the formal model
Reviewing and Evaluating

Fig. 1: The overall framework of the proposed method.

A. *Receiving the Enterprise Architecture's Description*

Based on the TOGAF's view, the entire business processes of an enterprise are rooted and validated in its objectives and missions. At this phase, we will start from the enterprise-level objectives and missions to reach the enterprise business processes. In the meantime, it is assumed that the descriptions have been received from the domain experts. It is worthy of mentioning that the mentioned descriptions have been provided in an oral and/or semi-documented unofficial manner, and hence could not be directly subjected to sampling.

B. *Syntactic Modeling*

Each model depicts a part of the reality, but a single model alone cannot express all the realities by itself. Since the received (input) descriptions are oral and/or semi-documented and unofficial, at this phase, the products of TOGAF will be generated using the ArchiMate [17]-[20] language and according to the received data. Since the present paper focuses on enterprise business processes, we will only elaborate on the required products corresponding to the TOGAF and the business architecture layer.

The steps involved in syntactic modeling are described in Table 1, as displayed in Fig. 2.

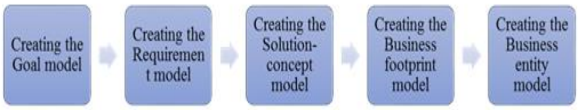

Creating the Goal model
Creating the Requiremen t model
Creating the Solution-concept model
Creating the Business footprint model
Creating the Business entity model

Fig. 2: Steps involved in syntactic modeling.

At the first step, we determine the goals and their respective hierarchies of details based on the extracted descriptions. For each goal, an enterprise face certain requirement. At the second step, we define respected requirements of goals. In order to meet the expressed requirements, one or more proper solutions will be proposed. At the third step, solutions proportional to different requirements will be determined. At the fourth step, to provide an overview that traces essential elements to be built or revised from goals through to components, we create a business footprint model. The Footprint is a complete collection of process, data, application, business unit, and business objective that validates a capability as in TOGAF and finally at the fifth step, to model the entities identified in a business process and the relationship among them, the business entity model is created.

C. *Semantic Modeling*

Syntactic models are unable to cover the entire knowledge pertaining to an enterprise alone. Additional data required to describe a model and should be expressed in the form of business rules. These rules are either defined by the process, or by the entities existing in a process. To this end, it would be necessary to provide a semantic description of the enterprise business processes. The algorithm used for semantic modeling has been illustrated in Fig. 3. Nevertheless, we use the semantic modeling process by concepts involved in WSMO [21]-[27]. This is because of the capability of WSMO as a semantic modeling language compliant with the aims of this paper.

For semantic modeling of an enterprise business process, at the first step, the business entities are semantically modeled to determine the ontology of business entities by specifying the name, attributes, types, and constraints on each entity. At the second step, we use axioms to express the rules governing a business process, and finally, at the third step, for modeling the goal of a business process, we use the Goal concept in WSMO to express the goal of an enterprise business process.

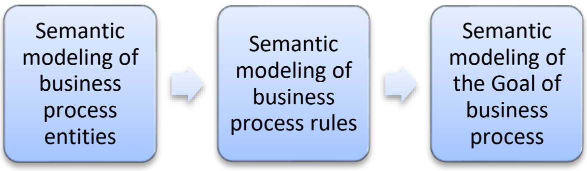

Semantic modeling of business process entities
Semantic modeling of business process rules
Semantic modeling of the Goal of business process

Fig. 3: Steps in semantic modeling.

D. *Creating the Formal Model*

Since semantic models developed with the WSML language [21] cannot be subjected to sampling, in order to provide the required conditions for sampling, at this step, a formal model that can be subjected to sampling will be generated from the description. In order to express the description in a formal sense, the B language [28], [29] will be used. The reason by which the B language is selected for use in the proposed method is that the target description is based on the Abstract State Machine model. Similarly, description B is also based on the abstract state machine. Language B is equipped with suitable supports for validating and checking the model; therefore, the model created using the B language will be adequately suitable for sampling at the next phase. We have defined a set of rules to transform semantic descriptions from WSML language to the B formal language. The description generated in WSML language contains two major parts, the first being the ontology of concepts, and the second being goal description.

1) Rules of Ontology Transformation

Ontology is the main part of WSMO and comprises three parts: the header, concept, and rules. The following rules are abided, while transforming into the B language.

Rule 1: Transforming the Ontology Header.

The header is comprised of several parts of the name, imports-ontology, mediator, and non-functional properties. However, since the sections of mediator and non-functional properties are ineffective, they won't be used in transformation. To this end, according to the relation (1) in Table 1, the transformation of machine B will be executed under the same name given to the defined ontology.

Rule 2: Transforming the Imports-ontology in the ontology.

If a specific ontology is added to the header, it will be added to the machine in the INCLUDES section according to relation (2) in Table 1.

Rule 3: Transforming Concept names in the Ontology.

The concepts in a WSMO ontology are transformed to sets in a B machine in the SETS section. The deferred sets in B usually declare the sets. A deferred set is one that is not initialized at the time of the set declaration. However, the explicit initialization of a set is represented by the initialized set. The set initialization can also be used to map the inheritance of the concepts. A concept with multiple sub-concepts can be transformed as the initialized set with elements representing its sub-concepts.

A concept comprises three sections, namely the name, attribute, and (attribute) type. Concepts are transformed into language B using relation (3) in Table 1. In this sense, the concept's name will be transformed into a set's name (in capital letters) in the SETS section. It is noteworthy that these sets will not be primarily quantified during defining.

Rule 4: Transforming the Attributes of Ontology Concepts.

The attributes of a concept in a WSMO ontology are transformed using the B relations. An attribute of a concept is transformed as a relation over the set representing the concept and the set representing the type of the attribute. Such relations are defined in the INVARIANT section of the B machine. The attribute of a concept is defined as a variable for defining this relationship.

The attributes will be transformed into the B language in the form of relations and based on relation (4) in Table 1. It is worthy of mentioning that the attributes will be defined in the VARIABLES section of the machine.

Rule 5: Transforming the Attribute Types of Ontology Concepts.

Attribute types are defined as a relation from a set representing the concept to a set representing the attribute type. These relations are defined in the INVARIANTS section of the machine based on relation (5) in Table 1. The type of a variable can be one of the main defined types or the types added to the machine in the imports-ontology section.

Rule 6: Transforming the Rules of Concepts.

Rules are constraints expressed in a logical form. In WSML language, rules are added in different sections under the name of axioms in order to show a restriction. For mapping, it is necessary to transform the mappings between the operators from WSML to B language. Using the mappings between different operators that expressed in the Table 2, rules can be transformed into the B language. It is noteworthy to mention that axioms will be written in the INVARIANT section of the machine.

Table 1: Rules of transforming the ontology from WSML to B language

Relation	Relation Transformation Rule
Header name Transformation	**Ontology**(ontology-name)- > **MACHINE**(ontology-name)Machine (1)
Imports-Ontology Transformation	**Imports-Ontology**(ontology-name)- > **INCLUDES**(ontology-name-Machine) (2)
Concept Name Transformation	**Concept**(name)-> cap(name) in **SETS** (3)
Concept Attributes Transformation	**Concept**(Attribute)- > var in **VARIABLES** (4)
Attribute Type Transformation	Attribute(Type)- > **VARIABLES**(var) : **SETS**(concept) < - > Type (5)

Table 2: Transforming the operators from WSML to B language

Description	WSMO operator	B operator
Conjunction	And	&
Disjunction	Or	or
Negation	Neg, naf	Not
Universal quantifier	For all	!x
Exist quantifier	exists	#=
Equality	=, :=:	=
Inequality	!=	/=
Implication	Implies, ImpliedBy	=>
Reverse implication	ImpliedBy	=>
Membership	memberOf	:
Typing	ofType, impliesType	:
Inheritance	subConceptOf	:>

II) Rules of Goal Transformation

The goal shows the system's behavior and performance in the user's view. A goal's description includes three parts, namely as header, capability, and interface.

A goal specification G is defined as a 3-tuple G = (H, I, C), where H is a goal header, I is a goal interface specification, and C is a goal capability. Below we describe the mapping in the same order.

Rule 1: Transforming Goal Header.

The header of a WSMO goal Specification consists of names, imports-ontology, mediator, and non-functional properties. Since they do not affect the non-functional and the mediator, we do not use them in transformation. As shown in relation (6) in Table 3, the goal declaration is transformed to a B machine declaration by the MACHINE statement. Note that the symbol "->" denotes the "is transformed to" statement. This means that the goal declaration is transformed to the machine declaration in the translated B machine. The naming convention is to use the name of goal Specification, with the suffix "Machine".

Rule 2: Transforming Goal Imports-ontology.

An ontology imported in a goal Specification using the imports-ontology makes all the ontology concepts and instances visible to the goal Specification as if they were included. Therefore, the imports-ontology statement in goal specification is transformed using the INCLUDES statement in the B machine that also makes the included machine visible and accessible in the including machine. The B machine representing the ontology is imported using the INCLUDES statement in the B machine representing the goal Specification. This is shown in relation (7) in Table 3.

Rule 3: Transforming Goal Capability.

Capability is determined with four parameters namely the precondition, assumptions, post-condition, and effects. Each of these parameters is a set of axioms, therefore using the previously mentioned rules for transforming axioms; they will be transformed into the B language via the rules stated in Table 2.

Rule 4: Transforming Goal Interface.

In describing the interface, the parts "signature" and "transition rule" are of importance. States are sets of concepts used in describing the interface. In WSMO, states are based on ASM with the variables in the B machine functioning in a similar way.

Suppose SSIG is the set of states used in the description of the interface, and VAR is the set of variables defined in machine B. in this sense, while mapping the states, according to the relation (8) in Table 3, the set of states will be defined in the variables section. In contrast, their types will be signified in the INVARIANTS section.

Rule 5: Transforming the Rules of Goal Interface Transition.

Suppose that T (G) is the transition rules defined in the target description, and OP(M) is the set of operations defined in machine M. In this case, all transition rules will be mapped into an operation in the machine using the relation (9) in Table 3.

Rule 6: Transforming the Inputs and Outputs of Goal Interface Transition Rule.

An operation is specified with a name; input values, and return values. According to relation (10) in Table 3, the name of the transition rule will be mapped into the operation name while the input parameters and return values of the transition rule will be mapped into operation input parameters and operation return values, respectively.

Table 3: Rules of transforming the goal from WSML to B

Relation	Transformation Rule
Transforming Goal header name	**Goal**(Goal-name)-> **MACHINE**(Goal-name)Machine (6)
Transforming Goal imports-ontology	**Imports-Ontology**(ontology-name)-> **INCLUDES**(ontology-name-Machine) (7)
Transforming Interface States	**Interface**(SSIG)-> **Machine(**VAR) Type(SSIG)-> VAR : **SETS**(concept) < - > Type (8)
Transforming Transition Rules	T(G)- > OP(M) (9)
Transforming the Inputs and Outputs of Transition Rule	Tri- > Opi Tri(in-concept)= Opi (inArg) ^ Tri(out-concept)= Opi (retype) ^ Tri-name = Opi-Name (10)

E. *Generating Samples from the Formal Model*

Once the formal model is generated in B language, it will be subjected to sampling to generate test cases, to which end the method of Model Checking will be used. The mentioned method is usually used to study the validity and reliability of state-based formal models. By this method, firstly, a set of traps for formal descriptions are set and subsequently added to the assertion section of formal description; then, the model will be checked. A negative trap is a test predicate obtained through various criteria of test coverage. The model checker searches through different system states for a state in which the assertion is contravened.

We use ProB [30] for model checking and generating test cases. The reason for using ProB is that: first of all, it is fast and automatic; second, it applies a perfect mechanism on the formal description to find contravention instances, and also checks the entire states space; and third, it is suitable for state-based descriptions. The steps involved in this section are shown in Fig. 4.

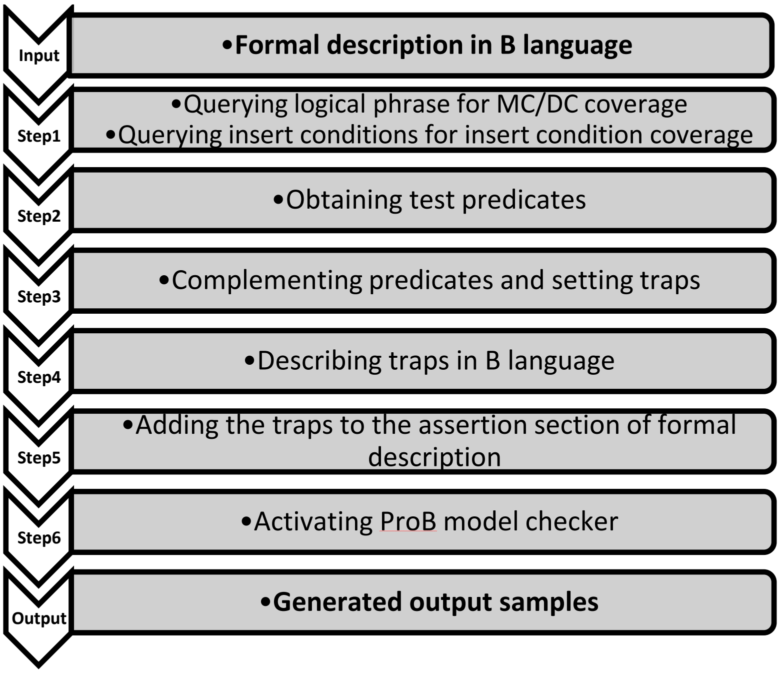

Input
•Formal description in B language
Step1
•Querying logical phrase for MC/DC coverage
•Querying insert conditions for insert condition coverage
Step2
•Obtaining test predicates
Step3
•Complementing predicates and setting traps
Step4
•Describing traps in B language
Step5
•Adding the traps to the assertion section of formal description
Step6
•Activating ProB model checker
Output
•Generated output samples

Fig. 4: The flowchart of generating samples from formal model.

I) *Obtaining a Trap from Formal B Descriptions*

The present paper uses the two criteria of boundary condition coverage and modified condition decision coverage to cover the formal descriptive model. Boundary condition coverage efficiently tests relational phrases, while modified condition decision coverage elaborates on testing predicates and logical phrases.

The traps are obtained based on formal descriptions and criteria of test coverage which are added to the assertion section of the description.

The Boundary Value Testing method tests the system’s behavior on the boundary of a variable.

In order to obtain the traps from the MCDC coverage criterion, a logical phrase containing atomic parts will be tested by n+1 test cases. The steps involved in the process of extracting traps from this coverage criterion are described as follows:

1- Determining logical conditions from various parts of the description.

2- Obtaining test predicates for logical phrases through the application of a table-based method.

3- Complementing the entire test predicates and obtaining the traps.

Afterwards, the obtained traps will be added to the assertion section of the description.

Once created, the traps will be consecutively added; then, the model checker will be activated. Since the traps are added to the ASSERTIONS section, we will request the checker to check the model for the defined assertions. If the descriptive model is valid, the checker will run into an error. The steps involved in trap calculation using MC/DC coverage are displayed in Fig. 5.

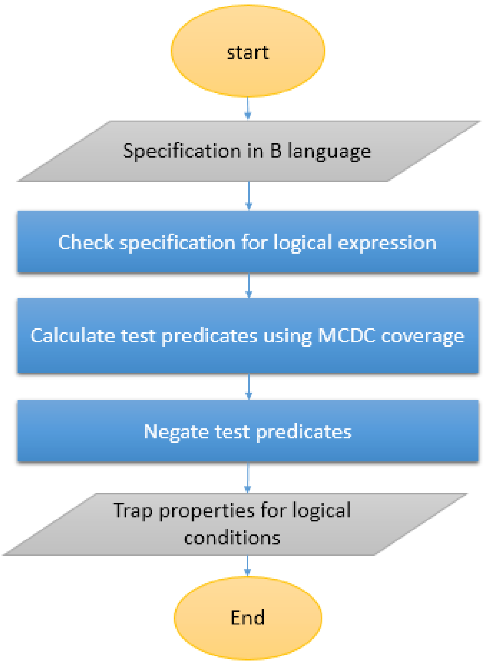

start
Specification in B language
Check specification for logical expression
Calculate test predicates using MCDC coverage
Negate test predicates
Trap properties for logical conditions
End

Fig. 5: Steps for trap calculation by using MCDC coverage.

F. Reviewing and Evaluating

Based on the views expressed in the TOGAF, developing an enterprise architecture is a gradually iterative process. At this phase, in case of a need for a change, the evaluation will be made, and the corresponding cycle of the proposed method will be iterated.

Implementing and Evaluating

In order to evaluate the efficiency of the proposed method, we deployed it in the marketing and sales department of a petrochemical corporation as a good case study.

In the following, firstly, the case study will be described, and afterwards, the details and results of the implementation of each section of the proposed method will be studied.

A. Description of the Case Study: Petrochemical Corporation

The vision of this corporation is to become the most well-known producer and distributor of isocyanate in the entire Middle-east.

The studied corporation uses nitric acid and gases such as chlorines, carbon monoxide, hydrogen and toluene to produce high-quality basic petrochemical products such as various isocyanates which are of a higher added value and also to distribute these products in domestic and foreign markets.

The goals of the marketing and sales department of the studied petrochemical corporation have been shown in Table 4, using well-known format of the Balance Score Cards technique.

Table 4: The marketing and sales department goals

Aspect	Goal1	Goal2	Goal3	Goal4	Goal5	Goal6
Financial Aspect	Development of income opportunities	Increased domestic sales	Increased income	Increased exports		
Customer Aspect	Being satisfied by sales staff	Customer loyalty	Being satisfied by the sales mechanism			
Processes Aspect	Development of marketing for new grades	Variability of customers	Development of the communications process	Improving the process of development of major and regional market studies	Development of mechanisms for improving customers' loyalty	Development of relationships with the global pricing centers
Growth and Learning Aspect	Promoting the personnel's knowledge and skills	Improving the sales mechanism,	Development of knowledge management	Increasing the contacts between the sales and other Departments of the enterprise	Creating a database and promoting integrated Documentation and data bank.	

B. Syntactic Modeling of the Case Study

Syntactic modeling of the descriptions has been carried out using the ArchiMate language and the software of Modelio V.7.0. In the following, the steps involved in syntactic modeling, as well as the outputs of each section, will be described. Afterwards, implementing the details of the proposed method will be discussed in terms of a process aspect.

I) Modeling the Vision Layer

To provide a full vision that can be used to scope all the work area, the vision phase uses initial schemas of an essentially informal nature. These artifacts are very high level and do not yet involve detailed modeling activities. They will be developed free hand, in the form of images or matrices, in order to prepare later phases. TOGAF defines an enterprise as being a collection of business units with a common set of goals. This shows just how important goals are within an enterprise; they are its reason for existence. Goals are constructed hierarchically. Goals constitute the roots of the goal/objective tree.

According to Table 5, increasing sales and income is a strategic goal for the realization of which there is a need for other objectives, including developing marketing for new grades, leveraging customers, developing the process of efficient communications, developing mechanisms for improving customer loyalty, and developing contacts with global sales centers must be already realized. In order to realize the mentioned requirements, proper and adequate solutions must be found. For instance, in order to realize the determined goals, develop income opportunities, increasing domestic sales, and improving exports, the corporation under study has been suggested to develop its sales process.

Based on the proposed algorithm, in this step, we first draw the Goal model, and then, based on that we draw the requirement model, the solution concept model, and the business footprint model. A solution-concept diagram has been shown in Fig. 6. This model uses preliminary information to share a preliminary vision with all stakeholders by providing general information on the changes that are going to be implemented.

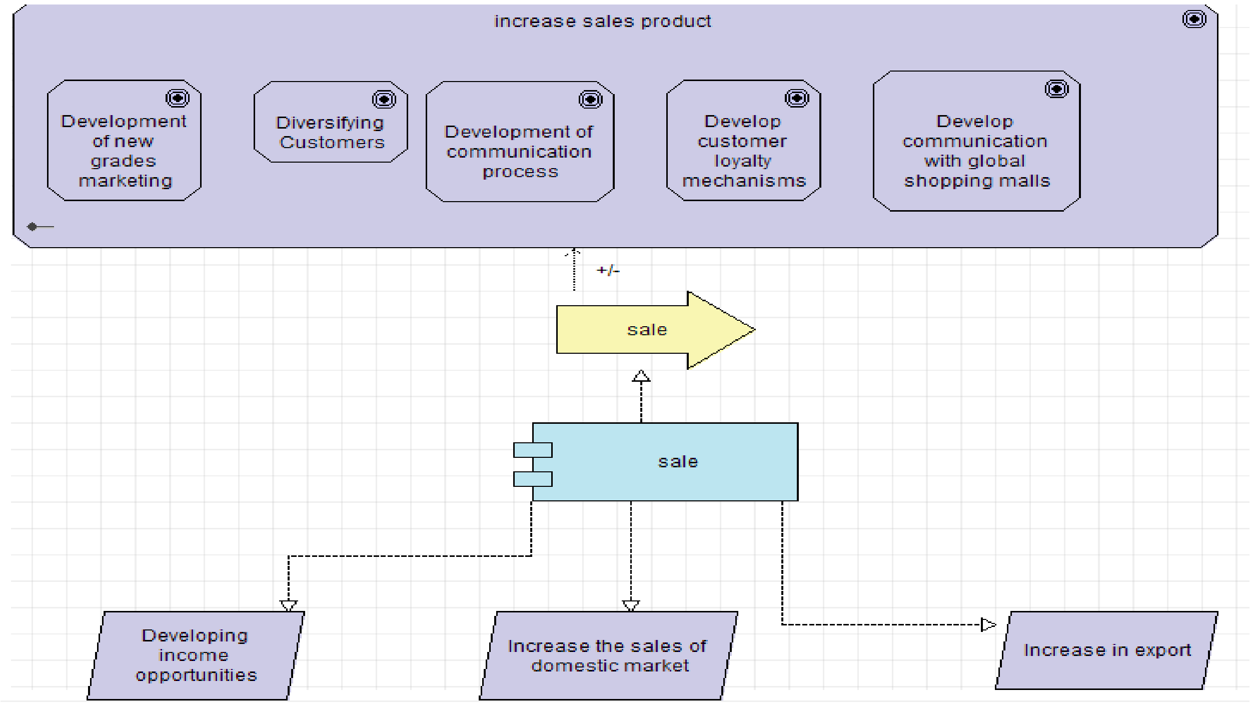

increase sales product
Development of new grades marketing
Diversifying Customers
Development of communication process
Develop customer loyalty mechanisms
Develop communication with global shopping malls
+/-
sale
sale
Developing income opportunities
Increase the sales of domestic market
Increase in export

Fig. 6: Solution concept diagram.

II) Modeling the Business Architecture Layer

Enterprise architecture puts a very strong emphasis on business architecture. Business architecture endeavors to identify the key business processes to fulfill Business strategies and goals. Based on the obtained information, the sales process group analysis is reported in Table 5. At this step, based on the collected information, models pertaining to the business process are created. A business footprint diagram describes the links between business goals, enterprise departments, business functions, and business services. These functions and services are also traced with technical components producing the required capabilities. A business footprint diagram is only interested in essential elements that show the connection between organization units and functions in order to produce services. A business footprint diagram has been drawn in Fig. 7 based on the sales department information. It is used to communicate with the management of the enterprise. Business footprint diagrams focus on the current concerns of the business.

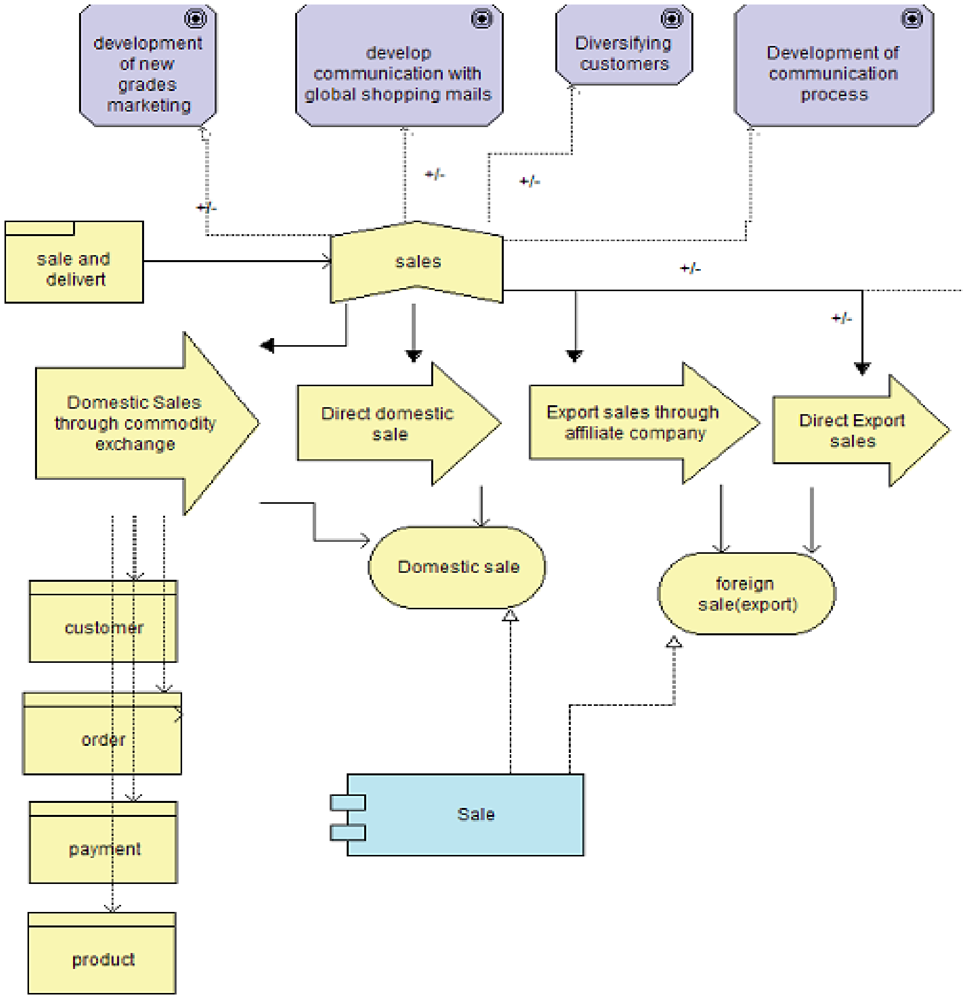

development of new grades marketing
develop communication with global shopping mails
Diversifying customers
Development of communication process
Develop customer loyality mechanisms
+/-
+/-
+/-
+/-
+/-
sale and delivert
sales
Domestic Sales through commodity exchange
Direct domestic sale
Export sales through affiliate company
Direct Export sales
Domestic sale
foreign sale(export)
customer
order
payment
product
Sale

Fig. 7: Business footprint diagram.

The retrieved information shows that in the domain of business knowledge, it is necessary to pay close attention to terminology.

For instance, it must be clear what is meant by customers, purchase request, product, payment, sale bill, freight bill, freighter, and invoice.

Each of the above-mentioned entities is defined by a set of properties and rules governing them. For example, a customer is specified by a name, an ID, an address, and a credit card number. In fact, every customer has their unique IDs, and there could not be any two customers with the same ID. Furthermore, every customer has a unique account number and so on.

C. Semantic Modeling of Business Processes

Among the processes pertaining to the case of the study, we have selected the process of direct domestic sales for the purpose of semantic description. Semantic modeling and the rules governing the client entity are written as follows:

```
ontology petro_EA_ontology
  nonFunctionalProperties
    wsmostudio#version hasValue "0.7.3"
  endNonFunctionalProperties
concept client
  name impliesType _string
  surname impliesType _string
  Identifier impliesType _integer
  address impliesType _string
  crediet_card_type impliesType _string
  credit_card_balance impliesType _integer
  credit_card_number impliesType _integer
```

Table 5: Sales process group analysis

The ID of the Sales and Delivery Process Group			
Existing process group	Sales and delivery	The code of existing process group	EA-company-SAL
The range of existing process group			
Includes the entire products of the company including the final products, middle products, side products, and waste products			
The goals of the existing process group			
Planning and executing the entire activities relating to sales including receiving orders, reviewing orders, sealing contracts, and delivery of products			
Business services			
• Domestic/foreign sales • Product Delivery report • Managing the contracts			
The indices of the existing process group			
• Monetary realization of domestic sales goals • Monetary realization of exported sales goals • Weight realization of domestic sales goals • Weight realization of exported sales			
Main inputs			
Data		**From process**	
Customer needs		Customer services	
Cash received approval		Cash received approval	
Main outputs			
Data		**From process**	
Customer data		CRM	
Sales plan		Production planning and controlling	
Sales invoices		Management of accounts receivable	
Providing product services		Customer services	
Order basket		Supplying the feed	
The owner of the existing process group			
• The chief of sales department			
The beneficiaries of the existing process group			
• petrochemical company as well as the customers			
The existing processes			
1. Domestic Sales through commodity exchange 2. Direct domestic sales 3. Export sales through another company 4. Direct Export sales			

For the client entity, a certain axiom is that account numbers are unique per person (client) and this axiom is described in the following fashion:

axiom uniq_credit_card
  definedBy
    ?x[creditcardnumber **hasValue** ?ccn1] **memberOf** client **and** ?y[creditcardnumber **hasValue** ?ccn2] **memberOf** client:-?ccn1 != ?ccn2.

In order to describe a process, we will try to describe its interface and capability parts. While describing the capability, we use axioms to write the preconditions, post-conditions, assumptions, and effects. For example, one important assumption in this process is that a valid credit card is the one that is either the MCARD or a SCARD. For this purpose, we use an axiom for this assumption named as a valid card. While describing the interface, we will describe the set of states along with the rules of transition among them.

D. Creating the Formal Model

In this section, we use the predefined mapping rules to transform the WSML language semantic description into formal B language. In the first step, the concepts and their attributes will be transformed. The following presents a partial transformation for the concept of client:

```
MACHINE petroEAontologymachine
SETS
   CLIENT
VARIABLES
         address,crediet_card_type,crediet_card_balanc
         e,creadit_card_number,Identifier,name,surnam
         e
INVARIANT
address:CLIENT<->STRING&
credit_card_balance:CLIENT<->INT&
credit_card_number:CLIENT<->INT&
credit_card_type:CLIENT<->STRING&
Identifier:CLIENT<->INT&
name:CLIENT<->STRING&
surname:CLIENT<-> STRING
```

Describing the goal is comprised of three parts, being the header, capability, and interface, respectively. Since in describing the goal, we have used "input_saleontology". Then, by implementing the expressed rules for transforming axioms, the preconditions, post-conditions, assumption, and effects in the description of capability will be transformed into B language. For example, one assumption maintained in the description is that a credit card is only valid if it is either the MCARD or a SCARD. The following presents the description in both WSML and formal B languages.

Presentation in WSML language:

assumption valid_card
  definedBy
    ?x[credit_card_type **hasValue** Mcard] **memberOf** client **or** ?x[credit_card_type **hasValue** Scard] **memberOf** client.

Presentation in B language:

```
#(x,credit_card_type).(x:CLIENT                    &
credit_card_type:CLIENT<->STRING                  =>
credit_card_type(x)="Mcard"                        or
credit_card_type(x)="Scard”)
```

In the interface part of the goal, we have a set of states and transitions among them. The set of states is written in the VARIABLES section, whereas their types are written in the INVARIANTS section.

E. Obtaining Samples from the Formal Model

In this section, to obtain the traps based on the expressed algorithm in Fig. 6, we used the MC/DC coverage criterion for a logical phrase in the INVARIANT section. Valid_credit_card is a logical phrase in the formal description. The logical phrase is as follows:

A=credit_card_type(x)="Mcard"
B=credit_card_type(x)="Scard”)

These two parts are connected by an OR connection operator. In the following, test phrases will be obtained using the table-based method.

Table 6: Sample from logical expression

No.	A	B	A or B	Effect A	Effect B
1	F	F	F	×	×
2	F	T	T		×
3	T	F	T	×	
4	T	T	T		

In Table 6, columns 5 and 6 denote effects A and B, respectively. These two columns signify what parts of a phrase are responsible for the sum of that phrase. In every phrase, the part that results in the occurrence of overall result is referred to as the main part, and the rest are called subsidiary parts.

According to the above table, the test cases of rows 1 and 2 test the effect B, while test cases 1 and 3 test the effect A. as a result, the test cases that test A and B effects are rows 1, 2, and 3. Hence, using these three test cases, one can test the above-mentioned phrase in terms of MCDC coverage.

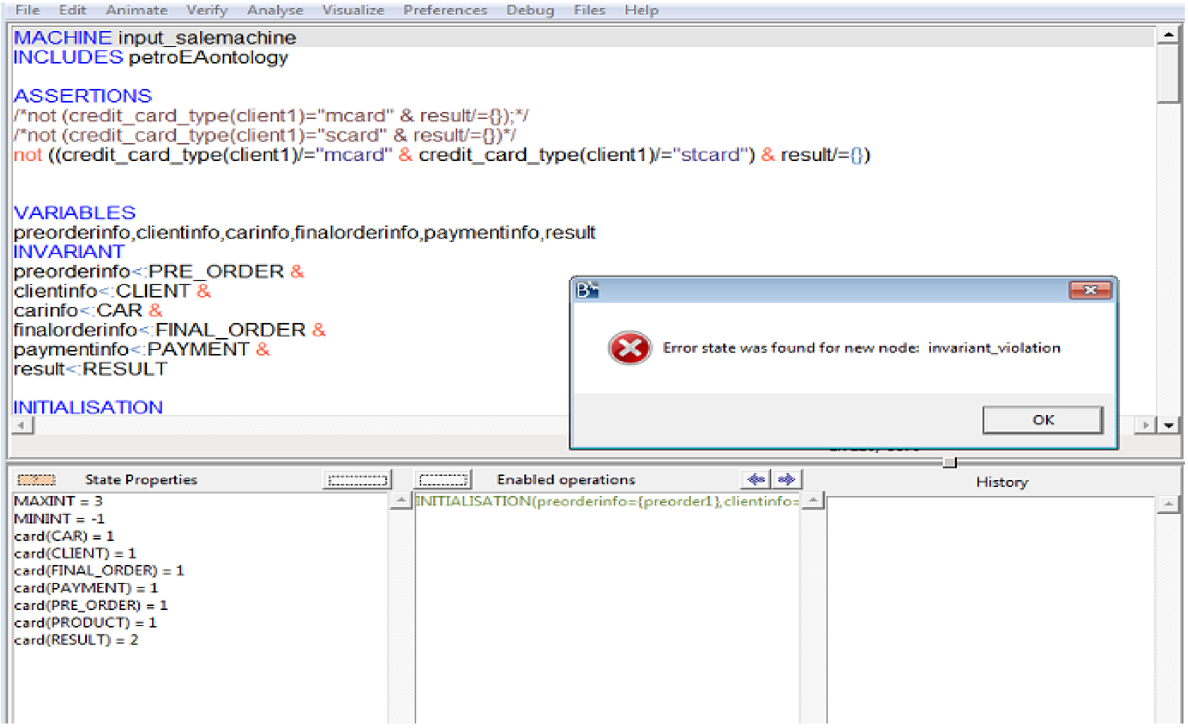
Afterwards, the obtained test cases will be complemented and then added to the ASSERTION section of the description B.

ASSERTIONS
Not (credit_card_type(x)="Mcard")
Not (credit_card_type(x)="Scard")
Not (credit_card_type(x)/="Mcard" and credit_card_type(x)/=”Scard”)

The third assertion in the last section has been checked using the model checker and displays the output as Fig. 8.

The distance of the fault location from the beginning is the machine mode which is a test case. The result illustrates suitable input and expected values under posed semantic limitations.

A piece of the output graph of the test case has shown in Fig.9.



```
MACHINE input_salemachine
INCLUDES petroEAontology

ASSERTIONS
/*not (credit_card_type(client1)="mcard" & result/={});*/
/*not (credit_card_type(client1)="scard" & result/={})*/
not ((credit_card_type(client1)/="mcard" & credit_card_type(client1)/="stcard") & result/={})


VARIABLES
preorderinfo,clientinfo,carinfo,finalorderinfo,paymentinfo,result
INVARIANT
preorderinfo<:PRE_ORDER &
clientinfo<:CLIENT &
carinfo<:CAR &
finalorderinfo<:FINAL_ORDER &
paymentinfo<:PAYMENT &
result<:RESULT

INITIALISATION
```

Error state was found for new node: invariant_violation

```
MAXINT = 3
MININT = -1
card(CAR) = 1
card(CLIENT) = 1
card(FINAL_ORDER) = 1
card(PAYMENT) = 1
card(PRE_ORDER) = 1
card(PRODUCT) = 1
card(RESULT) = 2
```

```
INITIALISATION(preorderinfo={preorder1},clientinfo=
```

Fig. 8: The output of activating model checker.

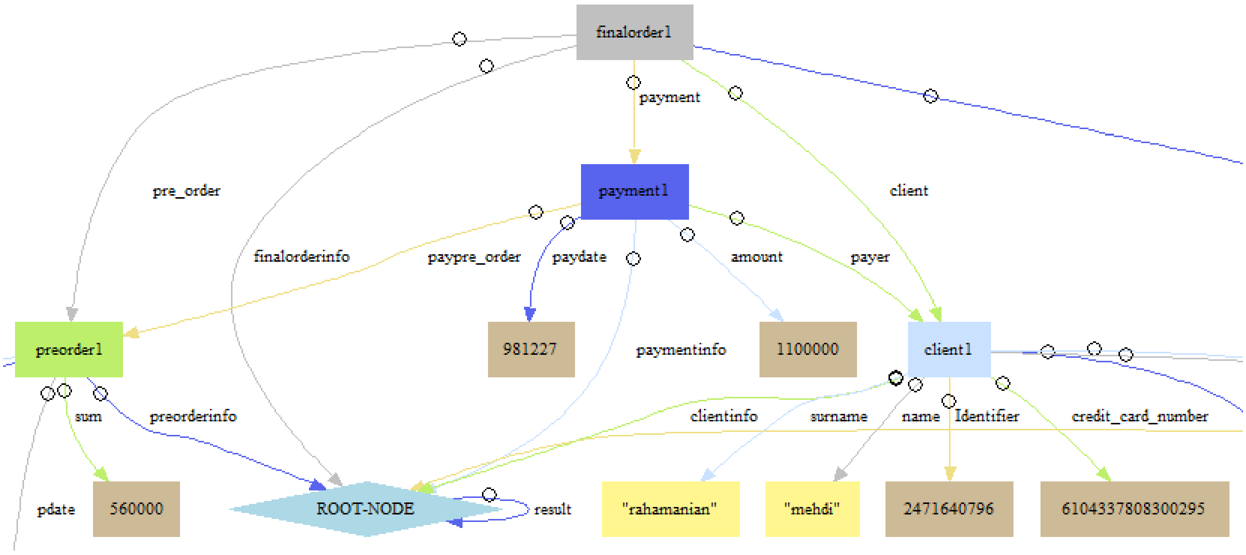


Fig. 9: A piece of output graph of the test case .

F. *Evaluation of Implementation*

The recent studies are categorized by two indicators:

1- Studies that have studied syntactic and semantic modeling in the field of enterprise architecture. None of which was intended to generate test cases.

2- Studies that have produced test cases without considering the discussion of enterprise architecture.

In comparison to recent studies, as shown in Table 7, the proposed method starts from the enterprise level, get benefited from the enriched descriptions yielded for enterprise business processes, move to generate proper syntactic and semantic models of the descriptions, and generate prioritized test cases from the well-established models to come up with test cases to be used for verification and/or validation. While in [1], [5]-[9], [11]-[13], syntactic and semantic modeling has been used for other objectives and a formal model has rarely been created.

These studies have in no way generated test cases based on enterprise architecture design. In [15], test cases have been generated, regardless of enterprise architecture design.

Table 7: Comparison of propose method by recent studies

Reference	Main contribution	Enterprise architecture design	Syntactic modeling	Semantic modeling	Formal modeling	Test case generation
[1]	Identify, classify, analyze, and evaluate existing methods for EA visualization.	yes	yes	yes	no	no
[5]	Providing a formal way for exploring misalignment of concepts.	yes	yes	no	yes	no
[6]	Proposing a combined approach that enterprise modeling would be suitable for both humans and machines.	yes	yes	no	no	no
[7]	Proposing a method for detection of any logical paradoxes in enterprise architecture models.	yes	yes	no	no	no
[8]	Using ontology to present, integrate and analysis enterprise models.	yes	no	yes	no	no
[9]	Proposing a combined modeling approach for convergence between business and technology.	yes	no	yes	no	no
[11]	Using descriptive logic along with ontology for manual analysis of enterprise design.	yes	no	yes	no	no
[12]	Modeling the dependencies between the business, information systems and IT infrastructures.	yes	no	yes	no	no
[13]	Develop a framework intended to make a balance between technology and business.	yes	yes	no	no	no
[14]	Presenting a model-based approach to automatically generate test cases from business process models	no	yes	no	no	yes
[15]	Developing a tool for the automated generation of test cases based on descriptions	no	no	no	yes	yes
[16]	Using the activity diagram for the automated generation of test cases	no	yes	no	no	yes
[17]	Creating a tool named PCTgen which automatically generated a set of test cases to check a workflow	no	yes	no	no	yes
Proposed method	Generating semantic test case from the enterprise level	yes	yes	yes	yes	yes

Results and Discussion

As mentioned earlier, no studies have been conducted to generate test cases based on descriptions received from enterprise architecture. However, the proposed method has the following advantages and limitations based on implementation of our method on the selected case study.

1- The details of the business process have been defined and started based on the enterprise level.

2- The level of abstraction is decreased by syntactic and semantic modeling of the enterprise architecture description.

3- In order to create a description that would be sampled, the semantic descriptions created using the proposed transition rules.

4- The generated test cases can be used in the validation and/or validation of business software, and because their descriptions are started at the enterprise level, they have high validity.

The limitations of the proposed method:

1- Due to the generality of TOGAF, it has been used for describing an enterprise architecture in this research. It is suggested that other enterprise architectures can also be used.

2- The focus of the proposed method was on enterprise business process. It is suggested that other business elements such as enterprise business services can be examined.

3- We transform semantic description into formal form by predefined rules, manually. It is suggested that, this can be done automatically by writing a parser.

Conclusions

The frameworks of enterprise architecture describe the elements of enterprise architecture at an abstract level and thus fail to elaborate on the details. However, they do provide architecture developers with a general primitive perspective. The main core in every enterprise's architecture consists its business processes. Enterprise processes are resulted by enterprise's goals and missions.

In order to test the verification and/or validation of a software product, it must be evaluated against the expressed descriptions and business rules. Previously, several studies following different objectives have tried to model enterprise architectures.

However, no previously conducted study has elaborated on modeling following the objective of creating test cases for further verification and/or validation purpose.

Therefore, the subject of testing for business software practically starts from the enterprise level (goals, missions, etc.). Therefore, the main contribution of this study is generating a set of test cases based on the descriptions yielded from enterprise business processes in early steps; then, the amount of later reviews and changes can be significantly lessened.

The overall framework of the proposed method is an iterative cycle adopted from the TOGAF Architecture Development Method. Following this cycle, once the primary descriptions of goals, missions, and strategies of the enterprise are retrieved, we will syntactically model the architecture.

Afterwards, the business processes will be modeled semantically, and the description will be transformed into formal B language for the purpose of sampling from the syntactic-semantic modeling.

In order to evaluate the proposed method, it has been implemented on the sales and marketing department of a petrochemical corporation, and the yielded results confirmed the validity of the proposed method. Based on the proposed method, the following values have been created:

1- By adding semantics to the syntactic models of enterprise architecture, more precise and exhaustive descriptions of the processes have been yielded, and in fact, the degree of abstraction has been decreased.

2- Creating a formal model of the syntactic and semantic descriptions can be subjected to sampling. And the resulting samples will be covering both syntax and semantics.

3- The proposed method starts from the missions and strategic goals of enterprises; therefore, the output samples are efficiently precise and complete.

One of the most applied fields in the domain of software engineering is automatic code generation. For future work, it is suggested to use the proposed method for automatic generation of codes according to the descriptions retrieved relating to enterprise architecture, and according to the TOGAF framework, which is a general framework.

Semantic descriptions are suitable tools for providing precise and yet understandable descriptions for machines.

The focus of the present study was centered on business processes in the architecture layer of enterprises. For future work, it is suggested to elaborate on providing semantic descriptions for other layers of architecture as well.

In the present study, the authors have used the TOGAF standard due to its publicity and high applicability; however, it is suggested to use also other standards for analyzing enterprise architecture and providing high-level descriptions.

Author Contributions

M. Rahmanian designed and implemented the proposed method. R. Nassiri collected data. M. Mohsenzade interpreted the results and carried out the data analysis. R. Ravanmehr wrote the manuscript and carried out the data analysis.

Acknowledgment

The authors received no funding from any organization in the course of carrying out the current study. It is certified that the Islamic Azad University is a private research and academic institute.

Conflict of Interest

No potential conflict of interest regarding the publication of this work. Besides, the authors have been completely witnessed the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy.

Abbreviations

TOGAF	The Open Group Architecture Framework
ADM	Architecture Development Method
WSMO	Web Service Modeling Ontology
WSML	Web Service Modeling Language
OWL	Ontology Word Language
EA	Enterprise Architecture
SEAM	Semantic Enterprise Architecture Modeling
SBVR	Semantic of Business Vocabulary and Rule
IS	Information System
BPaaS	Business Process as a Service
OWL	Ontology World Language
SEAM	Semantic Enterprise Architecture Modeling
MC/DC	Modified Condition Decision Coverage
BSC	Balance Score Cards

References

[1] Z. Zhou, Q. Zhi, S. Morisaki, S. Yamamoto, "A systematic literature review on enterprise architecture visualization methodologies," IEEE Access, 8(1): 96404-96427, 2020.

[2] "The TOGAF® Standard, Version 9.2." accessed 1 October 2021.

[3] S. Sharma, L. Raja, D. Pallavi Bhatt, "Role of ontology in software testing," J. Inf. Optim. Sci., 41(2): 641-649, 2020.

[4] A. Mili, F. Tcheir, Software Testing Concepts and Operations, John Wiley & Sons, 2015.

[5] K. Bouafia, B. Molnár, "Analysis approach for enterprise information systems architecture based on hypergraph to aligned business process requirements," Procedia Comput. Sci., 164: 19-24, 2019.

[6] K. Hinkelmann, E. Laurenzi, A. Martin, B. Thönssen, Ontology-Based Metamodeling" Business Information Systems and Technology 4.0. Studies in Systems, Decision and Control, 141: 177-194, 2018.

[7] E. Babkin, A. Ponomarev, "Analysis of the consistency of enterprise architecture models using formal verification methods," Bus. Inf., 3 (41): 30–40, 2017.

[8] A. Caetano, G. Antunes, J. Pombinho, M. Bakhshandeh, J. Granjo, "Representation and analysis of enterprise models with semantic techniques: an application to ArchiMate, e3value and business model canvas," Knowledge Inf. Syst., 50: 315–346, 2016.

[9] K. Hinkelmann, E. Laurenzi, B. Lammel, S. Kurjakovic, "A semantically-enhanced modelling environment for business process as a service," in Proc. 4th International Conference on Enterprise Systems, 4: 143-152, 2016.

[10] K. Hinkelmann, E. Laurenzi, A. Martin, D. Montecchiari, M. Spahic, B. Thönssen, "ArchiMEO: A standardized enterprise ontology based on the ArchiMate conceptual model," in Proc. the 8th International Conference on Model-Driven Engineering and Software Development – MODELSWARD: 417-424, 2020.

[11] G. Antunes, M. Bakhshandeh, R. Mayer, J. Borbinha, A. Caetano, "Using ontologies for enterprise architecture analysis," in Proc. 17th IEEE International Enterprise Distributed Object Computing Conference Workshops, 17: 361-368, 2013.

[12] W. Chen, C. Hess, M. Langermeier, "Semantic enterprise architecture management," in Proc. the 15th International Conference on Enterprise Information Systems (ICEIS-2013): 318-325, 2013.

[13] K. Hinkelmann, D. Karagiannis, B. Thoenssen, R. Woitsch, A. Gerber, "A new paradigm for continuous alignment of business and IT: combining enterprise architecture," Model. Enterpr. Ontol., 79: 77-86, 2015.

[14] A. Yazdani Seqerloo, M.J. Amiri, S. Parsa, "Automatic test cases generation from business process models," Requirements Eng. 24: 119-132, 2019.

[15] W. Zhang, S. Liu, "Supporting tool for automatic specification-based test case generation," in Proc. International Workshop on Structured Object-Oriented Formal Language and Method, 7787: 12-25, 2013.

[16] A.K. Jena, S.K. Swain, D.P. Mohapatra, "A novel approach for test case generation from UML activity diagram," in Proc. International Conference on Issues and Challenges in Intelligent Computing Techniques: 621-629, 2014.

[17] "The ArchiMate® Standards" accessed 1 October 2021.

[18] G. Wierda, Mastering ArchiMate Edition 3.1. : R & A, 2021.

[19] P. Desfray, G. Raymond, Modeling Enterprise Architecture With TOGAF. : Elsevier, 2014.

[20] "Archi-Open Source ArchiMate Modelling" accessed 1 Octobr 2021.

[21] "Web Service Modeling Ontology" accessed 1 October 2021.

[22] D. Fensel, H. Lausen, J. Bruijn, Enabling Semantic Web Services: The Web Service Modeling Ontology. : Springer-Verlag Berlin, 2007.

[23] D. Fensel, F.M. Facca, E. Simperl, I. Toma Web Service Modeling Ontology. In: Semantic Web Services. Springer, Berlin, Heidelberg, 2011.

[24] M. Bures, T. Cerny, M. Klima, "Prioritized process test: More efficiency in testing of business processes and workflows," in Proc. International Conference on Information Science and Applications, 424: 585-593, 2017.

[25] J. Bruijn, H. Lausen, A. Polleres, D. Fensel, “The web service modeling language WSML: An overview,” in Proc. European Semantic Web Conference: 590-604, 2006.

[26] J. Bruijn, D. Fensel, U. Keller, "Using the web service modeling ontology to enable semantic e-Business," Commun. ACM, 8(12): 43-47, 2005.

[27] F. Christina, P. Axel, D. Roman, D. John, " Towards intelligent web services: the web service modeling ontology (WSMO)," in Proc. International Conference on Intelligent Computing (ICIC'05): 23-26, 2005.

[28] "The Programming Language B" accessed 1 October 2021.

[29] K. Lano, The B Language and Method: A Guide to practical Formal development. : Springer Verlog, 1996.

[30] M. Leuschel, M. Butler, "ProB: a model checker for B," in Proc. International Symposium of Formal Methods Europe Springer, Berlin-Hei- delberg: 855–874, 2003.

Biographies

Mehdi Rahmanian received his B.S. degree in software Engineering from Shahid Chamran University, Ahwaz, Iran in 2007. He graduated in M.Sc. degree in software Engineering from IAU University, Ahwaz, Iran in 2011. Currently he is a Ph.D. student in IAU University, Tehran, Iran. His interests include software engineering, Enterprise Architecture and Software testing.

- Email: mehdi.rahmanian@srbiau.ac.ir
- ORCID: 0000-0002-3575-5230
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Ramin Nassiri received his B.S. in Computer software engineering from Tehran University, Tehran, in 1989, the M.S. in Computer software engineering, in 1995 and the Ph.D. degree in Computer software engineering from IAU University, Tehran, in 2003. Currently, he is a faculty in the Department of Computer engineering at the IAU University. He is the author/coauthor of more than 100 publications in professional and/or academic journals and conferences. He is co-founder of three IT companies in Iran and UAE since 2000. Also he has been managing a few national IT projects since past decade to promote public welfare by deploying ICT technologies and enablers. His research focus is basically on Software engineering, Enterprise architecture, Big data, Software testing, IoT and Etc.

- Email: r_nasiri@iauctb.ac.ir
- ORCID: 0000-0002-9488-9044
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mehran Mohsenzadeh received his B.E degree (Software Engineering) in 1997 from Shahid Beheshti University and M.E (in 1999) and Ph.D. (Software Engineering) in 2004 from IAU University, Tehran. His major interests are Cloud Computing, Software Engineering and Big Data and has published more than 85 papers (author/co-author) in International Conferences and journals. He is Assistant Professor in the Department of Computer Engineering, Science and Research Branch, IAU University of Iran.

- Email: r.mohsenzadeh@srbiau.ac.ir
- ORCID: 0000-0001-6835-409x
- Web of Science Researcher ID: NA
- Scopus Author ID: 26435355100
- Homepage: NA

Reza Ravanmehr graduated in computer engineering from Shahid Beheshti University, Tehran, in 1996. After that, he gained his M.Sc. and Ph.D. degrees, both in computer engineering, from Islamic Azad University, Science and Research Branch, Tehran, in 1999 and 2004, respectively. His main research interests are distributed/parallel systems, large-scale data management systems, and social network analysis. He has been a faculty member of the Computer Engineering Department at Central Tehran Branch, Islamic Azad University, since 2001.

- Email: r.ravanmehr@iauctb.ac.ir
- ORCID: 0000-0001-9605-5839
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:
M. Rahmanian, R. Nassiri, M. Mohsenzadeh, R. Ravanmehr, "Semantic enterprise architecture oriented test case generation for business process," J. Electr. Comput. Eng. Innovations, 10(2): 311-328, 2022.

DOI: 10.22061/JECEI.2021.8218.496

URL: https://jecei.sru.ac.ir/article_1636.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Resource Allocation for Full-Duplex Wireless Information and Power Transfer in Wireless Body Area Network

N. Khatami, M. Majidi*

Department of Electrical and Computer Engineering, University of Kashan, Kashan, Iran.

Article Info

Article History:
Received 10 August 2021
Reviewed 26 September 2021
Revised 23 November 2021
Accepted 28 November 2021

Keywords:
Wireless body area network
Simultaneous wireless
Information and power transfer
Full-duplex
Resource allocation
Body energy harvesting
Optimization

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

Abstract

Background and Objectives: The purpose of a wireless body area network (WBAN) is to collect and send vital body signals to the physician to make timely decisions, improve the efficiency of medical informatics systems, and save costs. The sensors of the WBAN network have limited size and energy, and hence, to extend the lifetime of these sensors, they can be powered wirelessly. Our focus in this paper is on a two-tier full-duplex (FD) cooperative WBAN in which sensors, in addition to transmitting physiological information, harvest energy from radio frequency (RF) coordinator signals and body sources. Our goal is to maximize average weighted sum throughput (AWST) under the constraints of each sensor, including meeting the minimum data rate, delay limitation, energy and transmission power constraints.

Methods: The resources allocated to solve this optimization problem are the time slots, the transmission rates of the sensors and coordinator, and the transmission powers of sensors in each time slot. The time scheduling problem in the first step is modeled in the form of a mixed-integer linear programming (MILP) problem and the second step problem is convex. Also, Karush–Kuhn–Tucker (KKT) conditions are presented for power and rate allocation.

Results: In the optimal allocation (OA) mode, contrary to the equal time allocation (ETA) one, with increasing the relay power, the AWST increases despite increasing self-interference (SI). Energy harvesting from the body, nevertheless the power consumption for transmission, makes positive the slope of the instantaneous energy curve for the motion sensor and reduces the corresponding slope for the electrocardiogram (ECG) one. Comparison of the proposed method with previous methods shows that the proposed method has better control over the information flow of sensors, and also in allocating rate to users, fairness is satisfied.

Conclusion: According to the simulation results in our method, the system showed better performance than the equal time allocation mode. We also used the FD technique and with the help of the optimal time scheduling index, we were able to control the SI.

Introduction

Prevention, early detection, and treatment of diseases through the wireless body area networks (WBANs) create a dynamic healthcare system. These real-time networks, which are a new generation of wireless personal area networks (WPANs), include several small devices on the human body or implanted in it.

These nodes convert the body's physiological parameters into electrical signals, and finally, a coordinator collects these signals and sends them to the access point (AP). The collected data is then presented to the hospital, physician, emergency department, or relatives for tracking the patient's vital signs, emergency actions, awareness of the patient's condition, and updating medical records [1]-[6].

There are two significant challenges associated with WBANs. The first challenge is to provide sustainable power to the body sensors and extend the network's lifetime. The limited capacity of the battery limits the network lifetime. On the other hand, it is difficult to replace and recharge the battery [7]. The authors in [8] have noted that the second challenge is to guarantee the quality of service (QoS) for the delay issue and deliver the data streams of sensors with different priorities. Therefore, in this paper, concerning the importance, the average weighted throughput is considered. By limiting the delay, the loss of critical information should be prevented. The energy needed for the sensors can be provided with the help of simultaneous wireless information and power transfer (SWIPT) technique, and harvesting wireless energy from radio frequency (RF) sources or body energy sources (such as biochemical and biomechanical energy) [9]-[11].

The combined use of SWIPT and relays results in maintaining the connectivity of WBANs, reducing interference, increasing reliability and spectrum and energy efficiency in these networks.

A. *Related Works*

In [12], a half-duplex (HD) cooperative system that has SWIPT capability with one sensor is investigated. This system aims to find the best relay location to maximize the throughput from the source to the destination. The system model studied in [13] as in [12] is single-sensor and cooperative, where the relay can harvest RF energy from an AP and does not have a buffer. The relay transmits the harvested energy to the sensor and forwards the received information to the AP.

The purpose in [13] is to maximize the information transmission rate under the constraint of the balance between the consumed and harvested energies. Finally, the optimal power splitting ratio and optimal time switching ratio are obtained, and the effect of relay location on system performance is discussed. In [14], a full-duplex (FD) cooperative system with one sensor and one buffer is explained. The continuous transmitted information rate maximization problem in adaptive power allocation mode and constant power allocation mode is formulated. In [15], a cooperative multi-sensor system model is considered, which includes links from sensors to the handset and from the handset to the AP. Also, several sources of RF signal propagation are employed to power WBAN sensors wirelessly and the goal is to maximize the throughput. However, the transmission delay of each user is not taken into account.

The authors are encouraged in [16] and [17] to investigate a cooperative multi-sensor WBAN in which both sensors and relay have energy harvesting, and the goal is to maximize throughput. In both system models, the relay is HD.

In [1], a cooperative multi-sensor WBAN that destination nodes (DNs) and relay node (RN) receive RF energy simultaneously from the source is considered. Power splitting ratio at the relay and each DN are optimization variables. The purpose of the optimization problem is to maximize the information sum-throughput. In [18], a multi-sensor WBAN system model without buffer is expressed in both normal and abnormal states. In the normal mode, the combination of time switching and power splitting is used, but in the abnormal mode, only time switching is used. In the system model of this paper, the coordinator is not considered, and the sensors are in direct contact with the AP, so due to the lack of a coordinator and buffer, it is not possible to control the delay, and sensors must consume more energy to send their information directly to AP.

In [19], a FD cooperative multi-sensor system with buffers is expressed. The purpose is to solve the problem of maximizing average delay limited weighted throughput under the stability of the average queue length in the context of a two-level WBAN architecture. The ability to harvest energy from the body by sensors, constraints on having limited instantaneous buffer length, and having positive instantaneous energy for the sensors are not considered. It should be noted that in all the above papers that have the WBAN system model, harvesting energy from the body and in [1], [8], [15], [16], [20] buffer is not considered.

B. *Contribution*

In this paper, a two-level multi-sensor WBAN is considered with several sensors, a FD coordinator, and an AP. Sensors send vital information extracted from the body, such as electrocardiograph information, to the coordinator. They can harvest RF wireless energy and body energy, and also, for each sensor, there is a buffer in the coordinator to control the delay of the information of that sensor. The time scheduling of sending different sensors, the transmission rates and powers of the sensors, and the transmission rates of the coordinator are determined Optimally. Resources are allocated to maximize average weighted sum throughput (AWST). Constraints such as the limitation of sensors' transmission powers, delays, and initial sensor energy are considered.

Innovations added to the problem model are as follows:

- In addition to battery-powered sensors, they are also equipped with RF and body energy harvesting (BEH) system (like mechanical energy).

- To achieve fairness, the average rate of each sensor over all time slots must be greater than a minimum threshold.
- For instantaneous buffer length, a constraint is considered to limit the data transfer delay of each user.
- The time scheduling optimization problem is modeled in the form of mixed-integer linear programming (MILP).
- Allocation of power and rate of each sensor and relay is performed by modeling the problem as a convex one, and to analyze it, Karush–Kuhn–Tucker (KKT) equations are presented.

C. Structure of the Paper

The rest of this paper is as follows: In the next section, we describe the signal and intended system model. Then, we formulate the problem and explain the optimal allocation (OA) and equal time allocation (ETA) modes. After that, the numerical solution and simulation results are presented. The last section concludes the paper.

Signal and System Model

The proposed scenario in Fig. 1, which is a two-level cooperative WBAN, consists of the RN, the source nodes (SNs), and the DN. The FD relay node, called full-duplex relay node (FD-RN), is a decode and forward relay equipped with N first-in-first-out (FIFO) buffers and is responsible for relaying messages to the destination, where *i*th buffer is shown by *Qi*. SNs have limited energy, sense the body's physiological signals, and send them to the coordinator. The FD-RN relays the information received from the sensors to the destination, and at the same time, the sensors receive RF energy from that signal. All sensors wirelessly receive RF power from the coordinator, and some sensors can receive energy from the body.

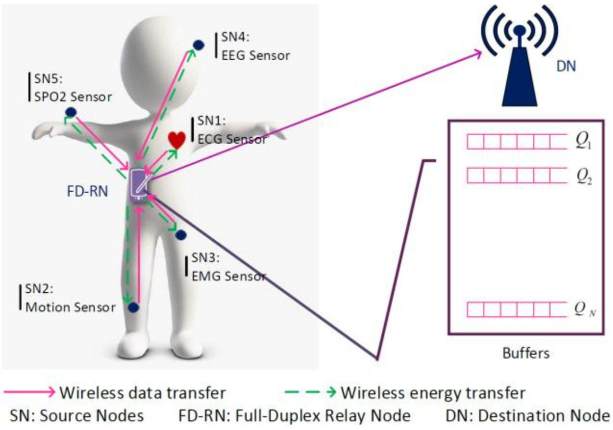


Fig. 1: System model.

The additive white Gaussian noise (AWGN) channel variances in DN and RN are σ_{nD}^2 and σ_{nR}^2, respectively. $h_{sr}^q(j) = \frac{|g_{sr}^q(j)|^2}{\sigma_{nR}^2}$, $h_{rd}(j) = \frac{|g_{rd}(j)|^2}{\sigma_{nD}^2}$, $h_{rr}(j) = \frac{|g_{rr}(j)|^2}{\sigma_{nR}^2}$ are normalized channel gain between *q*th sensor and relay, normalized channel gain between relay and destination, and normalized channel gain of SI at the relay, respectively.

Problem Formulation

According to Shannon's Theorem, the capacity of the channel from the *q*th sensor to the relay and from the relay to the destination in the *j*th time slot is:

$$C_{sr}^{q,l}(j) = W\, log_2\left(1 + \frac{P_s^{q,l}(j) h_{sr}^q(j)}{P_r^{q,l}(j) h_{rr}(j)+1}\right), \forall q, l, j \tag{1}$$

$$C_{rd}^{q,l}(j) = W\, log_2\left(1 + P_r^{q,l}(j) h_{rr}(j)\right), \forall q, l, j, \tag{2}$$

where the channel capacity of *q*th-RN and RN-DN are $C_{sr}^{q,l}(j)$ and $C_{rd}^{q,l}(j)$, respectively. Also, W and *q* and *l* are bandwidth, and sensor number and buffer number, and $P_s^{q,l}(j)$ is the power of the sensor in the *j*th time slot as *q*th sensor is sending data to the relay, and at the same time, *l*th buffer is emptying from the relay to its destination. $P_r^{q,l}(j)$ is the power of the relay in the *j*th time slot. We assume $m_{q,l}(j)$ as a binary time scheduling variable where if it is one, it indicates that in *j*th time slot data from *q*th sensor is being sent to the relay and from *l*th buffer is leaving towards AP [21], [22]. In any time slot *j*, only one sensor is allowed to send, and only the data of one buffer can be read and sent from the relay.

$$\sum_{q=1}^{N} \sum_{l=1}^{N} m_{q,l}(j) = 1, \forall j \tag{3}$$

$$m_{q,l}(j) \in \{0,1\}, \forall q, l, j \tag{4}$$

$$E_q(j) = E_q(j-1) - \sum_{l=1}^{N} m_{q,l}(j)\, P_s^{q,l}(j)\, \tau + E_{BH}(q)$$

$$+ \sum_{k=1,k\neq q}^{N} \sum_{l=1}^{N} m_{k,l}(j)\, P_r^{k,l}(j)\, \tau\, \eta\, \left|g_{sr}^q(j)\right|^2, \forall q, j, \tag{5}$$

where $E_q(j)$ is the energy of the *q*th sensor in the time slot *j*, $E_{BH}(q)$ is the amount of energy harvested from the body by the *q*th sensor, τ is the length of each time slot, and η is the energy conversion efficiency of the sensor.

The energy limitation of the *q*th sensor means that the total energy consumption of that sensor must be less than the total energy that is harvested from the RF and the body, plus its primary energy as follows:

$$\sum_{j=1}^{J}\sum_{l=1}^{N} m_{q,l}(j)\, P_s^{q,l}(j)\,\tau \leq$$

$$\sum_{j=1}^{J}\sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j)P_r^{k,l}(j)\,\eta\,\left|g_{sr}^{q}(j)\right|^2\tau + E_0$$

$$+JE_{BH}(q), \forall q, \tag{6}$$

where E_0 is the initial energy of each sensor. The instantaneous energy of each sensor is positive:

$$0 \leq E_q(j), \forall q, j. \tag{7}$$

$B_q(j)$ is qth buffer length in the time slot *j*. Also, the length of the buffer cannot be negative, and BF^{max} is the maximum value of the buffer length, hence

$$B_q(j) = B_q(j-1) - \sum_{k=1}^{N}\tau R_{rd}^{k,q}(j) + \sum_{l=1}^{N}\tau R_{sr}^{q,l}(j), \forall q, j \tag{8}$$

$$0 \leq B_q(j), \forall q, j \tag{9}$$

$$B_q(j) \leq BF^{max}, \forall q, j \tag{10}$$

$$\tfrac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j) \leq \tfrac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{N} R_{rd}^{k,q}(j)\,, \forall q, \tag{11}$$

$R_{sr}^{q,l}(j)$ indicates transmission rate from the qth sensor to the relay in the *j* time slot when the *l*th buffer at the relay gets empty to send its information towards the destination. Also, $R_{rd}^{k,q}(j)$ is the transmission rate when the *k*th sensor is active and qth buffer sends its information from the relay to the destination in the time slot *j*. Equation (11) is written because buffers do not waste data, and data queues are stable. Transmission rates cannot exceed channel capacity, therefore:

$$R_{sr}^{q,l}(j) \leq W m_{q,l}(j)\, log_2\left(1 + \frac{P_s^{q,l}(j)h_{sr}^{q}(j)}{P_r^{q,l}(j)h_{rr}(j)+1}\right), \forall q, l, j \tag{12}$$

$$R_{rd}^{q,l}(j) \leq W m_{q,l}(j)\, log_2\left(1 + P_r^{q,l}(j)h_{rd}(j)\right), \forall q, l, j. \tag{13}$$

To ensure fairness among the sensors, the average rates must be higher than a minimum.

$$R_{min}^{q} \leq \tfrac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j), \forall q. \tag{14}$$

The following is a short-term constraint for power $P_s^{q,l}(j)$:

$$0 \leq P_s^{q,l}(j) \leq P_s^{max}, \forall q, l, j, \tag{15}$$

where P_s^{max} is the maximum transmit power. It should be noted that $P_r^{q,l}(j)$ is assumed to be equal to P_r^{max}.

The difference between OA mode and ETA is that in OA mode, we determine $m_{q,l}(j)$ optimally by solving the optimization problem according to the constraints in our first step. However, in ETA mode, $m_{q,l}(j)$ is predefined.

A. OA mode

In OA mode, the allocation of scheduling power, sensor transmission rate, and relay transmission rate are dynamic. Our purpose is to maximize AWST. $m_{q,l}(j)$, $P_s^{q,l}(j)$, $R_{sr}^{q,l}(j)$ and $R_{rd}^{q,l}(j)$ are the optimization variables. Since $m_{q,l}(j)$ is a binary optimization variable, and the other variables are real, according to constraints such as (12), the problem is mixed-integer nonlinear programming (MINLP) and non-convex. Hence, we consider the following two steps to solve it.

In the first step, we assume that $P_s^{q,l}(j)$ is equal to the constant value P_s^{max}. Then we get $m_{q,l}(j)$. Therefore, $m_{q,l}(j)$, $R_{sr}^{q,l}(j)$ and $R_{rd}^{q,l}(j)$ are the optimization variables of this step and the optimization problem is as follows:

P1: $\max \frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N} W_q\, R_{sr}^{q,l}(j)$

s.t. a) $\sum_{q=1}^{N}\sum_{l=1}^{N} m_{q,l}(j) = 1, \forall q$

b) $m_{q,l}(j) \in \{0,1\}\,, \forall q, l, j$

c) $E_q(j) = E_q(j-1) - \sum_{l=1}^{N} m_{q,l}(j)P_s^{max}\,\tau +$
$\sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j)P_r^{max}\,\tau\,\eta\,\left|g_{sr}^{q}(j)\right|^2 +$
$E_{BH}(q), \forall q, j$

d) $\sum_{j=1}^{J}\sum_{l=1}^{N} m_{q,l}(j)\, P_s^{max}\tau \leq$
$\sum_{j=1}^{J}\sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j)P_r^{max}\,\tau\,\eta\,\left|g_{sr}^{q}(j)\right|^2$
$+E_0 + JE_{BH}(q), \forall q$

e) $0 \leq E_q(j), \forall q, j$

f) $B_q(j) = B_q(j-1) - \sum_{k=1}^{N}\tau R_{rd}^{k,q}(j) +$
$\sum_{l=1}^{N}\tau\, R_{sr}^{q,l}(j), \forall q, j$

g) $0 \leq B_q(j), \forall q, j$

h) $B_q(j) \leq BF^{max}, \forall q, j$

i) $\frac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j) \leq \frac{1}{J}\sum_{j=1}^{J}\sum_{k=1}^{N} R_{rd}^{k,q}(j)\,, \forall q$

j) $R_{sr}^{q,l}(j) \leq W m_{q,l}(j).$
$log_2\left(1 + \frac{P_s^{max}h_{sr}^{q}(j)}{P_r^{max}h_{rr}(j)+1}\right), \forall q, l, j$

k) $R_{rd}^{q,l}(j) \leq W m_{q,l}(j).$
$log_2\left(1 + P_r^{max}\, h_{rr}(j)\right), \forall q, l, j$

l) $R_{min}^{q} \leq \frac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j), \forall q.$

Because this problem contains integer and continuous variables and also functions are affine, it is a MILP.

In the second step, we consider $m_{q,l}(j)$ obtained from the first step to be known. $R_{sr}^{q,l}(j)$, $P_s^{q,l}(j)$, $R_{rd}^{q,l}(j)$

are optimization variables and the resulting problem, according to the objective function and constraints, is convex as follows:

$$\text{P2: } \max \frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N} W_q\, R_{sr}^{q,l}(j)$$

s.t : c,d,e,f,g,h,i,j,k,l

m) $P_s^{q,l}(j) \le P_s^{max}, \forall q,l,j$

n) $0 \le P_s^{q,l}(j), \forall q,l,j$

To solve problem P2, CVX software can be used. The whole problem-solving process is presented in Algorithm 1 in two steps.

Algorithm 1. Solution of OA mode problem

```
1.  Initialization: $P_s^{max}$,$P_r^{max}$, $R_{min}^{q}$ , and $BF^{max}$
        Number of sensors and buffers.
        Number of time slots.
        Duration of each time slots.
        Energy conversion efficiency.
        Weight of each SN (User Priorities).
        Initial energy of each SN.
2.  First optimization step :
        Solve optimization problem P1.
        Output: $m_{q,l}(j)$.
3.  Second optimization step :
        $m_{q,l}(j)$ is determined from first step.
        Solve optimization problem P2.
            Output: $P_s^{q,l}(j), R_{sr}^{q,l}(j)$ and $R_{rd}^{q,l}(j)$.
```

For the analytical solution of the second step, we can write KKT conditions [23], [24]. To do so, first, consider the Lagrangian function of problem P_2 as follows.

$$\sum_{j=1}^{J} L_j\left(P_s^{q,l}(j), R_{sr}^{q,l}(j), R_{rd}^{q,l}(j), \varphi, \mu, \zeta, \nu, o, \rho, \lambda, \alpha, \beta, \delta, \psi, \varepsilon\right)$$

$$= -\frac{1}{J}\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N} W_q\, R_{sr}^{q,l}(j) +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N} \varphi_q(j)[E_q(j) - E_q(j-1) + \sum_{l=1}^{N} m_{q,l}(j) P_s^{q,l}(j)\tau$$

$$-E_{BH}(q) - \sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j) P_r^{max}\eta\left|g_{sr}^{q}(j)\right|^2\tau] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\mu_q[\sum_{l=1}^{N} m_{q,l}(j)\, P_s^{q,l}(j)\tau - \frac{E_0}{J} -$$

$$\sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j) P_r^{max}\eta\left|g_{sr}^{q}(j)\right|^2\tau - E_{BH}(q)] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\zeta_q(j)[-E_q(j)] + \sum_{j=1}^{J}\sum_{q=1}^{N}\nu_q(j)[B_q(j) -$$

$$B_q(j-1) + \tau\sum_{k=1}^{N} R_{rd}^{k,q}(j) - \tau\sum_{l=1}^{N} R_{sr}^{q,l}(j)] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N} o_q(j)[-B_q(j)] + \sum_{j=1}^{J}\sum_{q=1}^{N}\rho_q(j)[B_q(j) -$$

$$BF^{max}] + \sum_{j=1}^{J}\sum_{q=1}^{N}\lambda_q[\sum_{l=1}^{N} R_{sr}^{q,l}(j) -$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N}\alpha_{q,l}(j)[R_{sr}^{q,l}(j) - W m_{q,l}(j) log_2(1 +$$

$$\frac{P_s^{q,l}(j) h_{sr}^{q}(j)}{P_r^{max} h_{rr}(j)+1})] + \sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N}\beta_{q,l}(j)\left[R_{rd}^{k,q}(j) -\right.$$

$$\left.W m_{q,l}(j)\, log_2\left(1 + P_r^{max} h_{rd}(j)\right)\right] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\delta_q\left[R_{min}^{q} - \sum_{l=1}^{N} R_{sr}^{q,l}(j)\right] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N}\psi_{q,l}(j)\left[P_s^{q,l}(j) - P_s^{max}\right] +$$

$$\sum_{j=1}^{J}\sum_{q=1}^{N}\sum_{l=1}^{N}\varepsilon_{q,l}(j)\left[-P_s^{q,l}(j)\right],$$

where φ, μ, ζ, ν, o, ρ, λ, α, β, $\delta, \psi, \varepsilon$ are Lagrange multipliers. To achieve the KKT conditions, we follow below four items:

1) Conditions C_1 to C_{12} must be met.

2) The Lagrangian multipliers of unequal constraints must all be positive.

$\mu \ge 0, \zeta \ge 0, o \ge 0, \rho \ge 0, \lambda \ge 0, \alpha \ge 0, \beta \ge 0, \delta \ge 0,$
$\psi \ge 0, \varepsilon \ge 0$

3) For each of the optimization variables, a partial derivative is taken. The gradient of the Lagrange function must be equal to zero.

$$G_1: \frac{\partial L}{\partial P_s^{q,l}(j)} = \sum_{j=1}^{J}\mu_q\, m_{q,l}(j)\,\tau +$$

$$\sum_{j=1}^{J}\varphi_q(j) m_{q,l}(j)\,\tau - \sum_{j=1}^{J}\alpha_{q,l}(j) W m_{q,l}(j)$$

$$\left[\frac{h_{sr}^{q}(j)}{(P_r^{max} h_{rr}(j)+1)+\left(P_s^{q,l}(j) h_{sr}^{q}(j)\right) ln2}\right] +$$

$$\sum_{j=1}^{J}\psi_{q,l}(j) - \sum_{j=1}^{J}\varepsilon_{q,l}(j) = 0$$

$$G_2: \frac{\partial L}{\partial R_{sr}^{q,l}(j)} = -\frac{1}{J}\sum_{j=1}^{J} W_q - \sum_{j=1}^{J}\nu_q(j)\ \tau +$$

$$\sum_{j=1}^{J}\lambda_q + \sum_{j=1}^{J}\alpha_{q,l}(j) - \sum_{j=1}^{J}\delta_q = 0$$

$$G_3: \frac{\partial L}{\partial R_{rd}^{q,l}(j)} = \sum_{j=1}^{J}\nu_l(j)\ \tau - \sum_{j=1}^{J}\lambda_l +$$

$$\sum_{j=1}^{J}\beta_{q,l}(j) = 0$$

4) First, we have to rewrite the inequality constraint functions to the standard form fi(x) ≤ 0, where fi (x) is the ith inequality constraint function [23]. The Complementary slackness conditions are written as follows:

$$CS1: \mu_q[\sum_{j=1}^{J}\sum_{l=1}^{N} m_{q,l}(j) P_s^{q,l}(j)\,\tau -$$

$$\sum_{j=1}^{J}\sum_{k=1,k\neq q}^{N}\sum_{l=1}^{N} m_{k,l}(j) P_r^{max}\ \eta\left|g_{sr}^{q}(j)\right|^2\tau$$

$$-E_0 - J E_{BH}(q)] = 0$$

$$CS2: \lambda_q[\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j) - \sum_{j=1}^{J}\sum_{k=1}^{N} R_{rd}^{k,q}(j)] = 0$$

$$CS3: \alpha_{q,l}(j)\left[R_{sr}^{q,l}(j) - W m_{q,l}(j) log_2\left(1 + \frac{P_s^{q,l}(j) h_{sr}^{q}(j)}{P_r^{max} h_{rr}(j)+1}\right)\right] = 0$$

$$CS4: \beta_{q,l}(j)\left[R_{rd}^{k,q}(j) - W m_{q,l}(j)\, log_2(1 + P_r^{max} h_{rd}(j))\right] = 0$$

$$CS5: \psi_{q,l}(j)\left[P_s^{q,l}(j) - P_s^{max}\right] = 0$$

$$CS6: \varepsilon_{q,l}(j)\left[-P_s^{q,l}(j)\right] = 0$$

$$CS7: \delta_q\left[R_{min}^{q} - \frac{1}{J}\sum_{j=1}^{J}\sum_{l=1}^{N} R_{sr}^{q,l}(j)\right] = 0$$

$$CS8: \rho_q(j)\left[B_q(j) - BF^{max}\right] = 0$$

$$CS9: \zeta_q(j)\left[-E_q(j)\right] = 0$$

$$CS10: o_q(j)[-B_q(j)] = 0$$

With the help of Newton's method, the set of KKT equations can be solved, and the optimal point can be obtained.

B. ETA Mode

In ETA mode, we assign equal access times to the sensors, and the sensors change their turn rotationally. Given that $m_{q,l}(j)$ is definite and static in this case, and the ETA optimization problem is similar to the P2 problem. This case was proposed for comparison with the performance of the optimal OA mode.

Simulation Results and Discussion

Using MATLAB software and CVX software, we solve the problems of OA and ETA modes and analyze the results. In the simulations, we have considered one electrocardiogram (ECG) and one motion sensor. The parameters used in the simulations are described in Table 1. To model the channel, the Lognormal distribution is most compatible with dynamic scenarios [25], [26]. Therefore, here too, we use the Lognormal distribution to implement the channel model between the sensors and the coordinator.

$$h_{ij}[\mathrm{dB}] \sim N(\mu_{ij}, \sigma_{ij})$$

We take the distribution parameters, i.e., μ_{ij} (mean) and σ_{ij} (variance), from the information in Table 2 [20]. In addition to the random Lognormal distribution, $P_{ij}[dB]$, which is the relative power loss of the link between transmitter *j* and receiver *i*, is also considered. d_{ij} is the distance between the transmitter *j*, and the receiver *i*. *j* indicates coordinator, and *i* is the body's sensor. $h_{ij}[dB]$ and $P_{ij}[dB]$ are added together.

For the channel gain model between the coordinator node and the base station, we use $h = {}^{200}\!/_{d^4}$,where *d* is the distance between those two nodes with a uniform distribution between 50 to 200 meters.

Table 1: Simulation parameters

Parameter	Symbol	Value
Number of sensors and buffers	*N*	2
Number of time slots	*J*	20
Channel bandwidth [19]	*B*	10 kHz
Duration of each time slot	τ	20 ms
Noise power at the RN [19]	σ_{nR}^2	-124 dBm
Noise power at the DN [19]	σ_{nD}^2	-124 dBm
Energy conversion efficiency [19]	η	0.8
Maximum power of sensors [19]	P_s^{max}	0.1mW
Maximum power of relay [19]	P_r^{max}	1 mW
Array of weight of each Sensor (User Priorities) [27]	*W*	$(2/11) \times [6{,}5]$
Carrier frequency [28]	f_c	4.2 GHz
Energy harvesting by ECG sensor [29]	$E_{BH}(1)$	4 μJ
Energy harvesting by Motion sensor [30]	$E_{BH}(2)$	0.4 μJ
Maximum length of the buffer	BF^{max}	20 Kbit

Table 2: Channel parameters between sensors and synchronizer [20]

Location of sensor	$P_{ij}[dB]\backslash d_{ij}[cm]$	$\mu_{ij}\backslash\sigma_{ij}$
chest	-43.29\26	-0.72\2.67
right foot	-51.00\92	-3.25\6.21

In Fig. 2, the instantaneous rates of the first and second sensors are plotted according to the time slot number. In this case, the initial energy value E_0 varies from 10 microjoules to 400 microjoules. Fig. 2 shows that in the OA mode, the sensor allocates more time slots to itself due to its higher weight and better channel condition. The time scheduling index is such that one of the sensors must be sent in a time slot, so the diagrams in Fig. 2 are also based on complementary time slots.

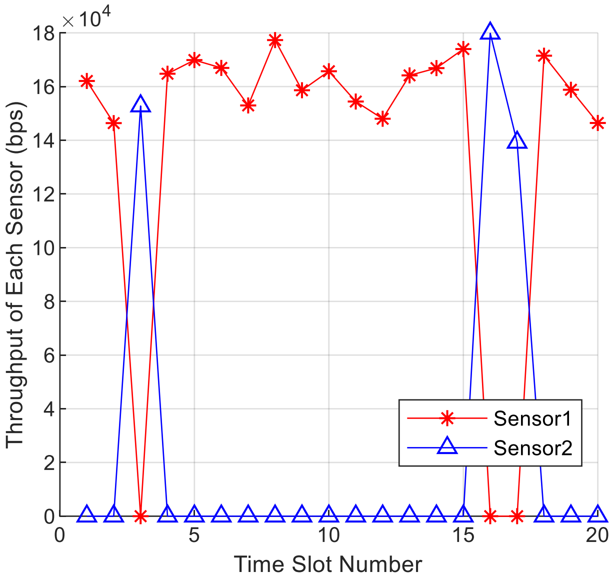
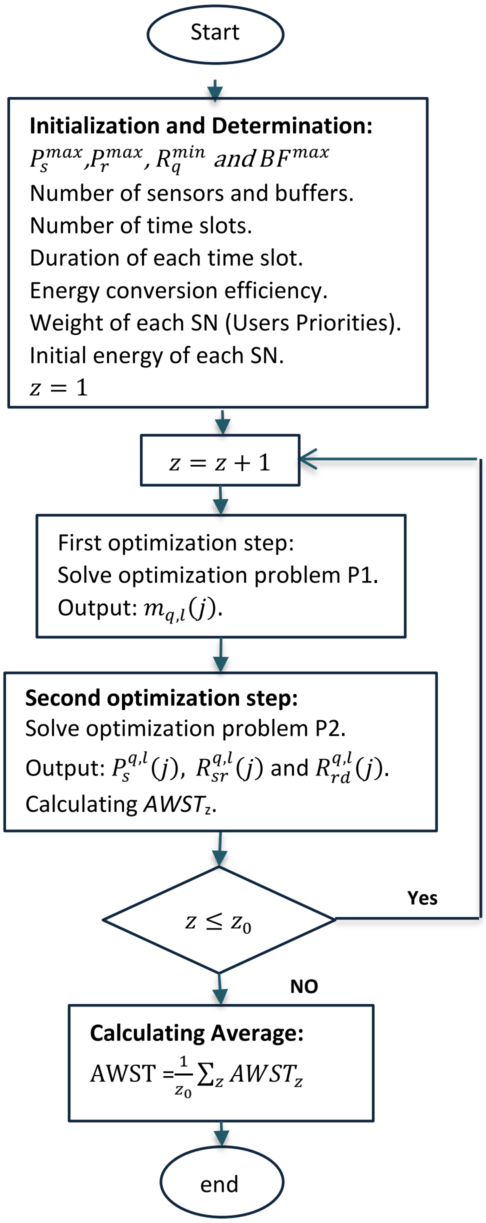


Fig. 2: Instantaneous rate of the first and second sensors according to the time slot number.

Flowchart 1 is based on Algorithm 1. Only a loop is added to perform z_0 Monte Carlo simulations. We use Flowchart 1 for calculating AWST.

Fig. 3 shows the mean weighted throughputs of the first and second sensors in terms of initial energy E_0. When the initial energy of the sensors is low, they can send with less power. Hence, they have lower mean weighted throughputs. By increasing the initial energy, the sensors can send at their maximum power, so their mean weighted throughputs go up and reach saturation.

As a result, the mean weighted throughput of the first sensor (ECG) reaches a maximum of 140 Kbps due to higher weight and better channel condition, and the mean weighted throughputs of the second sensor (Motion) reaches a maximum of 20 Kbps due to lower rate weight and worse channel condition.


Start
Initialization and Determination:
P_s^{max},P_r^{max}, R_q^{min} and BF^{max}
Number of sensors and buffers.
Number of time slots.
Duration of each time slot.
Energy conversion efficiency.
Weight of each SN (Users Priorities).
Initial energy of each SN.
$z = 1$
$z = z + 1$
First optimization step:
Solve optimization problem P1.
Output: $m_{q,l}(j)$.
Second optimization step:
Solve optimization problem P2.
Output: $P_s^{q,l}(j)$, $R_{sr}^{q,l}(j)$ and $R_{rd}^{q,l}(j)$.
Calculating $AWST_z$.
$z \leq z_0$
Yes
NO
Calculating Average:
AWST $= \frac{1}{z_0} \sum_z AWST_z$
end

Flowchart 1: Monte Carlo simulation for calculating AWST.

Fig. 4, Fig. 5, and Fig. 6 show the AWST comparisons for OA and ETA modes. In Fig. 4, the AWST is plotted in terms of the initial energy E_0. When the initial energy of the sensors is small, they cannot send a large percentage of their maximum power, so the throughput is low, and the AWST is also low. As the initial energy increases, the sensors can transmit at a higher percentage of their maximum power; as a result, the diagram goes up. By increasing E_0, the sensors can send P_s^{max} at their maximum power; consequently, AWST reaches its maximum value, and the graph goes to saturation. By comparing the two diagrams for the OA and ETA modes, it can be seen that at the initial energy of 100 micro joules, the AWST value in OA mode is 6.85% higher than in ETA mode.

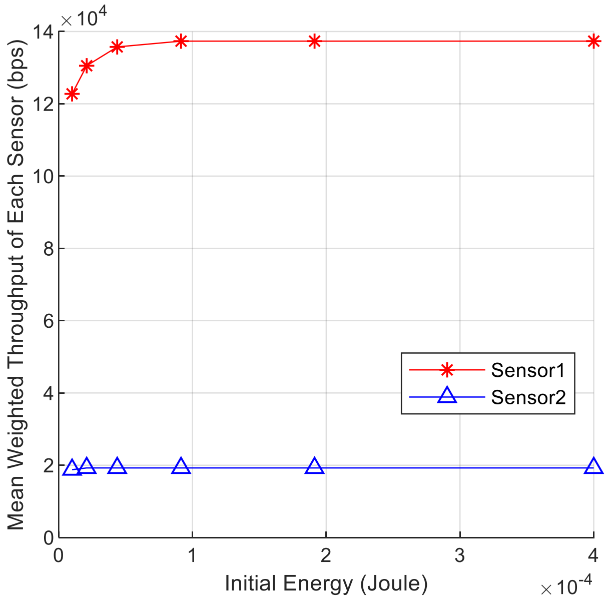


Fig. 3: Comparison of the mean weighted throughput of the first and second sensors in terms of initial energy.

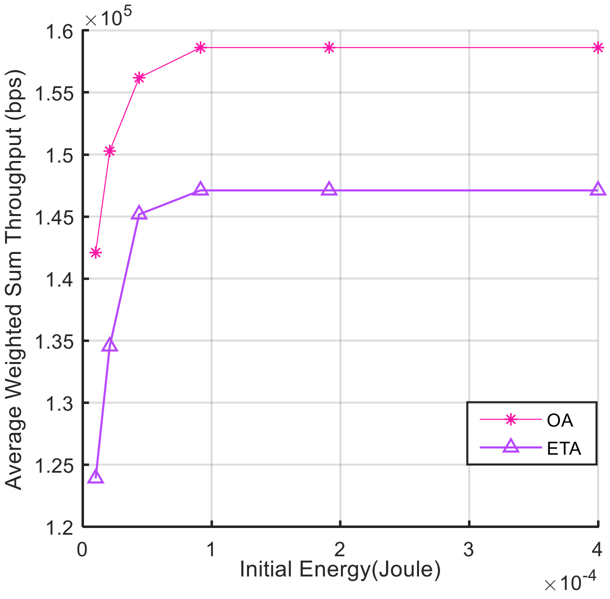


Fig. 4: Comparison of AWST diagrams in E_0 for both OA and ETA modes.

Fig. 5 shows the AWST in terms of different h_{rr} for the OA and ETA modes. In this case, the initial energy value is 200 microjoules. Initially, h_{rr} is very small, and its power can be ignored compared with the power of noise at the denominator of the Shannon relationship. By increasing the h_{rr} at the denominator of the Shannon relationship, the amount of self-interference (SI) power of the FD relay becomes significant, resulting in a decrease in throughput and AWST. It can be seen that in h_{rr}=-140 dB, AWST in OA mode is 10 Kbps higher than ETA mode.

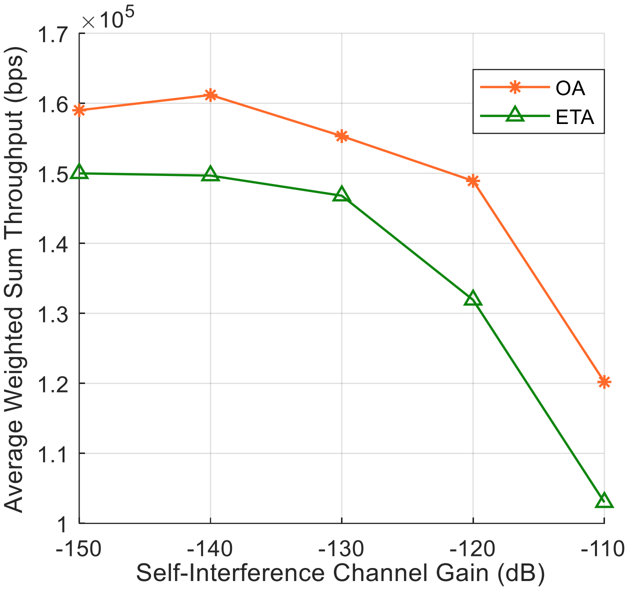


Fig. 5: Comparison of AWST diagrams in h_{rr} for OA and ETA modes.

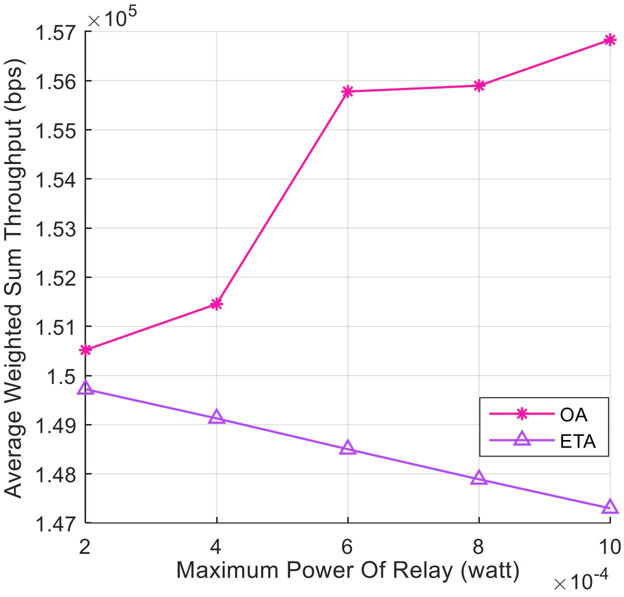


Fig. 6: Comparison of AWST diagrams in terms of P_r^{max} for OA and ETA modes.

In Fig. 6 AWST for different P_r^{max} the diagrams of the OA and ETA modes are compared. According to Fig. 6, for the OA mode, when P_r^{max} increases, the capacity of the relay-destination channel increases; hence the buffers in the relay may have a higher output rate. It causes the total rate of the sensors to relay to increase, but with increasing P_r^{max}, the diagram slope decreases because increasing P_r^{max} also increases SI. For the ETA mode, we see that the AWST is decreasing because only half of the time slots are allocated to the first sensor (ECG), which has a higher weight rate, and the number of its time slots cannot be increased. On the other hand, with the increase in P_r^{max}, the SI is also increasing, which leads to a decrease in AWST.

Therefore, in contrast to the OA mode, which can allocate more time slots to the ECG sensor and increase the total rate, in the ETA mode, it is impossible to increase the rate.

Fig. 7 shows AWST in terms of E_0 for sitting position, walking mode, and no BEH. In the small E_0s, the AWST of the walking mode is higher than the sitting position and the non-harvested mode.

At larger E_0s, the initial energy is saturated, and the increase in E_0 no longer has an effect on the increase in throughput, so all three curves are getting closer to each other. For validation, we compared the method of [19] with our method. In Fig. 8, Fig. 9, and Fig. 10, BF^{max}= 5 Kbit and R_{min}=40 Kbps.

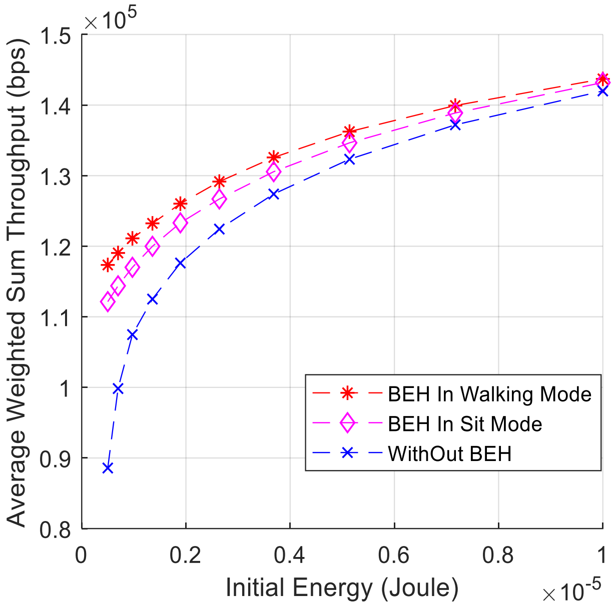


Fig. 7: Comparison of AWST diagrams for walking, sitting, and without energy harvesting modes.

Fig. 8 shows filled buffer length of the first and second sensors in terms of time slot numbers, and the initial energy value E_0 is 100 microjoules. In our method, we defined the threshold (BF^{max}=5 Kbit), which controls the filled length of the buffers, but in the method of [19], there is no threshold, so their buffers have overflowed.

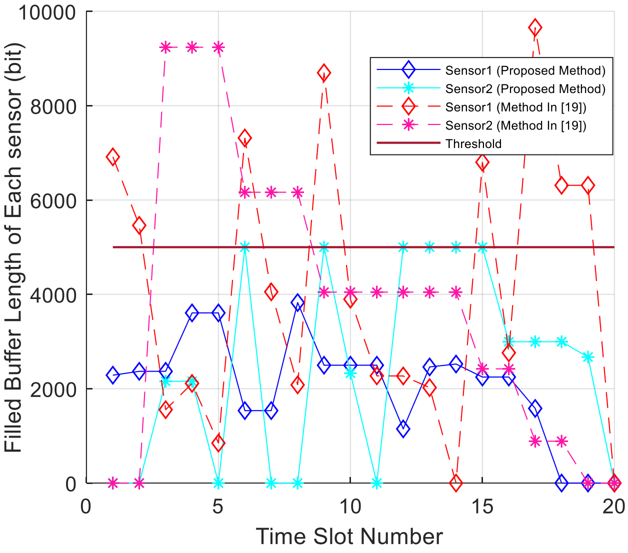


Fig. 8: Filled buffer lengths of the sensors in terms of time slot numbers to compare the proposed method with [19].

As can be seen in Fig. 9, the fairness is not satisfied in [19], because the rate of sensor 2 is zero and lower than threshold (R_{min}=40 Kbps).

In our method, the fairness causes the minimum rate be provided for each sensor.

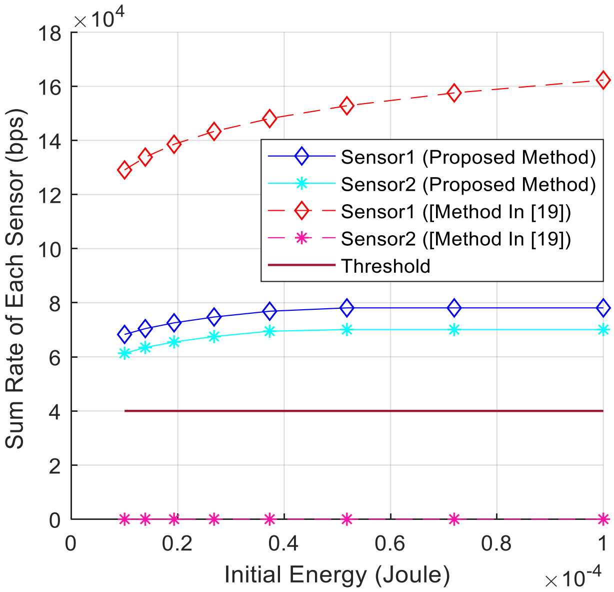


Fig. 9: Comparison of sum rate diagrams in terms of E_0 for proposed method and method of [19].

In Fig. 10, the AWST is plotted in terms of the initial energy E_0. In our method, compared to the method of [19], the value of obtained AWST is less because we have two more constraints, h, and l. Although AWST of [19] is higher than our method, it has two problems: First, fairness is not satisfied in [19], and second, buffers may overflow. Therefore, the obtained AWST theoretically in [19] may not be achieved in practice.

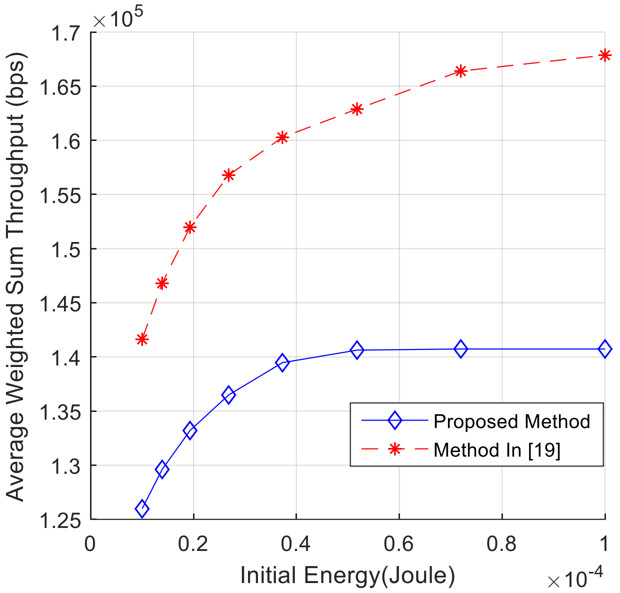


Fig. 10: Comparison of AWST diagrams in terms of E_0 for Proposed Method and Method of [19].

In Algorithm 1, the minimum rate, R_{min}, is a limiting condition that may make the problems infeasible in some channel conditions.

Therefore, we obtained the feasibility of the problems using Monte Carlo simulation by changing R_{min}, which can be seen in Fig. 11. This figure shows the effect of the minimum rate of sensors on the feasibility of the resource allocation problem. In Fig. 11, BF^{max}= 5 Kbit and the initial energy value E_0 is 10 microjoules. Resource allocation in the proposed method is more flexible than ETA mode. The first sensor has a higher rate due to a better channel, and a higher rate requires more power.

The buffer of both sensors is not negative as expected, and in addition, it remains smaller than its specified maximum value, which means that vital information is not lost. When the initial energy of the sensors is low at first, so they can send with less power, as a result, AWST is low. Different h_{rr} affects the channel capacity.

Increasing the h_{rr} reduces the capacity of the sensor channel to the relay, so the rate and consequently the AWST decreases.

In future work, the objective function can be the maximization of the energy efficiency or minimization of energy consumption. The relay can harvest energy from the body or RF signal. A system model can be considered, when the energy of one of the sensors is finished It sends an alert message in an uplink channel.

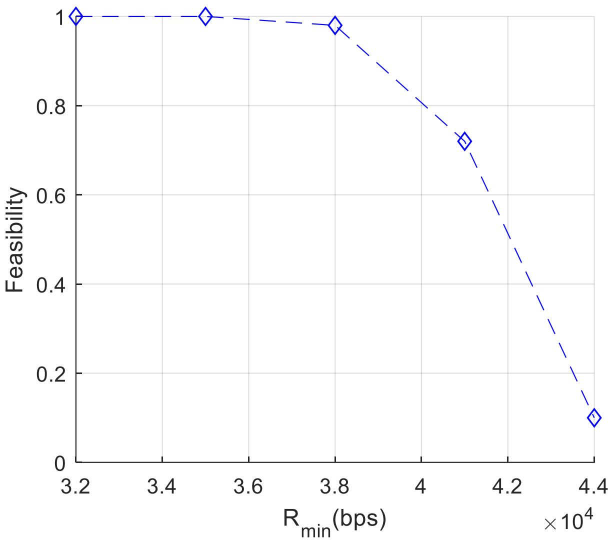


Fig. 11: The feasibility versus the minimum rate.

Conclusion

In this paper, a two-tier cooperative WBAN architecture including a FD relay, several sensors, and a destination were introduced was presented. In this system model, a coordinator based on the OA of time slots between sensors received the body's data from the sensors and simultaneously sent energy to them; it also sent sensors information to a destination, which is the AP here. In this system model, to face the problem of an energy shortage, the ability to harvest energy from the body was considered for the sensors. The purpose of this work is to maximize AWST with limited delay in the form of an MINLP and non-convex problem which included minimum data rate limit, energy limit, latency limit, and limited transmission power satisfaction. The goal of optimization can be to maximize energy efficiency or minimize energy consumption. In addition, the ability to command the sensors and having some actuators can be considered.

Author Contributions

All stages of research have been achieved with the participation, consensus, and equal efforts of both authors.

Acknowledgment

We thank reviewers for their so-called insights. We are also immensely grateful for their comments on an earlier version of the manuscript.

Conflict of Interest

The authors state that there is no obstacle and conflict of interest to publish 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 entirely and carefully witnessed by the authors.

Abbreviations

AP	*Access Point*
AWGN	*Additive White Gaussian Noise*
AWST	*Average Weighted Sum Throughput*
BEH	*Body Energy Harvesting*
ECG	*ElectroCardioGram*
DN	*Destination Node*
ETA	*Equal Time Allocation*
FD	*Full-Duplex*
FD-RN	*Full-Duplex Relay Node*
FIFO	*First-In-First-Out*
HD	*Half-Duplex*
KKT	*Karush–Kuhn–Tucker*
MILP	*Mixed-Integer Linear Programming*
MINLP	*Mixed-Integer Non Linear Programming*
OA	*Optimal Allocation*
QoS	*Quality of Service*
RF	*Radio Frequency*
RN	*Relay Node*
SI	*Self-Interference*
SN	*Sensor Node*
SWIPT	*Simultaneous Wireless Information and Power Transfer*
WBAN	*Wireless Body Area Network*
WPAN	*Wireless Personal Area Network*

References

[1] S. Li, F. Hu, Z. Mao, Z. Ling, Y. Zou, “Sum-throughput maximization by power allocation in WBAN with relay cooperation,” IEEE Access, 7: 124727–124736, 2019.

[2] S. Movassaghi, M. Abolhasan, J. Lipman, D. Smith, A. Jamalipour, “Wireless body area networks : A survey,” IEEE Commun. Surv. Tutorials, 16(3): 1658–1686, 2014.

[3] S. Dehghanpour, M. Majidi, "Simultaneous wireless information and power transfer in a network of on-body and implantable sensors with temperature constraint and intelligent channel prediction," Comput. Intell. Electr. Eng., 2021.

[4] A. Razavi, M. Jahed, “Capacity-outage joint analysis and optimal power allocation for wireless body area networks,” IEEE Syst. J., 13(1): 635–646, 2019.

[5] X. Liu, F. Hu, M. Shao, D. Sui, G. He, “Power allocation for energy harvesting in wireless body area networks,” China Commun., 14(6): 22–31, 2017.

[6] A. Vyas, S. Pal, B.K. Saha, “Relay-based communications in WBANs : A comprehensive survey,” ACM Comput. Surv. (CSUR), 54(1): 1-34, 2020.

[7] M. Boumaiz, M. El Ghazi, S. Mazer, M. Fattah, A. Bouayad, M.E.l Bekkali, Y. Balboul, ‘‘Energy harvesting based WBANs: EH optimization methods,’’ Procedia Comput. Sci., 151: 1040–1045, 2019.

[8] Z. Liu, B. Liu, C. Chen, C.W. Chen, “Energy-efficient resource allocation with QoS support in wireless body area networks,” presented at the 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, USA, 2015.

[9] H. Yektamoghadam, A. Nikoofard, “Fault detection in thermoelectric energy harvesting of human body,” J. Electr. Comput. Eng. Innov., 9(1): 57–66, 2021.

[10] F. Akhtar, M.H. Rehmani, P. Design, “Energy harvesting for self-sustainable wireless body area networks,” IT Prof., 19(2): 32–40, 2017.

[11] J.C. Kwan, A.O. Fapojuwo, “Radio frequency energy harvesting and data rate optimization in wireless information and power transfer sensor networks,” IEEE Sens. J., 17(15): 4862–4874, 2017.

[12] L. Wang, F. Hu, Z. Ling, B. Wang, “Wireless information and power transfer to maximize information throughput in WBAN,” IEEE Internet Things J., 4(5): 1663–1670, 2017.

[13] H. Gu, Z. Li, L. Wang, Z. Ling, “Resource allocation for wireless information and power transfer based on WBAN,” Phys. Commun., 37, 2019.

[14] M. Mohammadkhani Razlighi, N. Zlatanov, “Buffer-aided relaying for the two-hop full-duplex relay channel with self-interference,” IEEE Trans. Wirel. Commun., 17(1): 477–491, 2018.

[15] S. Shen, J. Qian, D. Cheng, K. Yang, G. Zhang, “A Sum-utility maximization approach for fairness resource allocation in wireless powered body area networks,” IEEE Access, 7: 20014–20022, 2019.

[16] S. Li, F. Hu, J. Yu, Z. Huang, “Optimal power allocation with a cooperative relay in multi-point WBAN,” presented at the IEEE/CIC International Conference on Communications in China (ICCC), Changchun, China, 2019.

[17] S. Li, F. Hu, Z. Xu, Z. Mao, Z. Ling, H. Liu, “Joint power allocation in classified WBANs with wireless information and power transfer,” IEEE Internet Things J., 8(2): 989–1000, 2021.

[18] H. Liu, F. Hu, S. Qu, Z. Li, D. Li, “Multipoint wireless information and power transfer to maximize sum-throughput in WBAN with energy harvesting,” IEEE Internet Things J., 6(4): 7069–7078, 2019.

[19] X. Zhang, K. Liu, L. Tao, “A cooperative communication scheme for full-duplex simultaneous wireless information and power transfer wireless body area networks,” IEEE Sensors Lett., 2(4): 1–4, 2018.

[20] Z. Huang, Y. Cong, Z. Ling, Z. Mao, F. Hu, “Optimal dynamic resource allocation for multi-point communication in WBAN,” IEEE Access, 8: 114153–114161, 2020.

[21] Y.H. Xu, G. Yu, Y.T. Yong, “Deep reinforcement learning-based resource scheduling strategy for reliability-oriented wireless body area networks,” IEEE Sensors Lett., 5(1): 3–6, 2021

[22] Z. Liu, B. Liu, C. W. Chen, “Buffer-aware resource allocation scheme with energy efficiency and QoS effectiveness in wireless body area networks,” IEEE Access, 5: 20763–20776, 2017.

[23] S. Boyd, L. Vandenberghe, Convex Optimization, United Kingdom: Cambridge, 2004.

[24] H. Zhang, S. Huang, C. Jiang, K. Long, V.C.M. Leung, H.V. Poor, “Energy-efficient user association and power allocation in millimeter-wave-based ultra-dense networks with energy harvesting base stations,” IEEE J. Sel. Areas Commun., 35(9): 1936–1947, 2017.

[25] A.K. Shukla, P.K. Upadhyay, A. Srivastava, J.M. Moualeu, “Enabling co-existence of cognitive sensor nodes with energy harvesting in body,” IEEE Sens. J., 21(9): 11213–11223, 2021.

[26] D. Sui, F. Hu, W. Zhou, M. Shao, M. Chen, “Relay selection for radio frequency energy-harvesting wireless body area network with buffer,” IEEE Internet Things J., 5(2): 1100–1107, 2018.

[27] IEEE standard for local and metropolitan area networks part 15.6: Wireless body area networks, 2012.

[28] S. van Roy, F. Quitin, L. Liu, C. Oestges, F.Horlin, J.M. Dricot, P. De Doncker, “Dynamic channel modeling for multi-sensor body area networks,” IEEE Trans. Antennas Propag., 61(4): 2200–2208, 2013.

[29] D. Cavalheiro, A.C. Silva, S. Valtchev, J.P. Teixeira, V. Vassilenko, “Energy harvested from the respiratory effort,” presented at the 5th International Joint Conference on Biomedical Engineering Systems and Technology- BIOSTEC, Vilamoura, Portugal, 2012.

[30] E. Romero Ramirez, “Energy harvesting from body motion using rotational micro-generation,” Ph.D. dissertation, Dept. Mech. Eng., Univ. Michigan Tech, 2010.

Biographies

Negarsadat Khatami received the B.SC. degree in biomedical engineering from the Islamic Azad University Khomeinishahr Branch, Isfahan, Iran, in 2014. She is graduated in the field of telecommunication engineering at the master's level of Kashan University, Isfahan, Iran, in 2020. Her research interests include wireless body area network (WBAN), optimization, and wireless powered communications.

- Email: negarsadat.khatami@grad.kashanu.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Mahdi Majidi received the B.Sc. degree in electrical engineering from Isfahan University of Technology, Isfahan, Iran, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2007 and 2014, respectively. In 2012, he joined the Communications and Networks Laboratory, Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore, as a Visiting Ph.D. Student. In 2015, he worked as a researcher at the Iran Telecommunication Research Center (ITRC). Since 2016, he is an assistant professor of the Department of Electrical and Computer Engineering, University of Kashan, Iran. His research interests include numerical and analytical optimization methods, Intelligent reflecting surfaces, machine learning, wireless transceiver design, wireless powered communications, and multi-antenna communications.

- Email: m.majidi@kashanu.ac.ir
- ORCID: 0000-0003-0245-4738
- Web of Science Researcher ID: NA
- Scopus Author ID: 55512805400
- Homepage: https://faculty.kashanu.ac.ir/mmajidi/en

How to cite this paper:

N. Khatami, M. Majidi, "Resource allocation for full-duplex wireless information and power transfer in wireless body area network," J. Electr. Comput. Eng. Innovations, 10(2): 329-340, 2022.

DOI: 10.22061/JECEI.2021.8112.485

URL: https://jecei.sru.ac.ir/article_1638.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Fast DC Offset Removal for Accurate Phasor Estimation using Half-Cycle Data Window

H. Sardari[1], B. Mozafari[1,*], H.A. Shayanfar[2]

[1]Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran.

[2]Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran.

Article Info

Article History:
Received: 03 September 2021
Reviewed 15 October 2021
Revised 11 November 2021
Accepted 20 December 2021

Keywords:
Phasor estimation
Digital filter
HCFA
Mimic filter
Exponential decaying-DC component
Synchrophasor measurements

*Corresponding Author's E-mail Address: *mozafari@srbiau.ac.ir*

Abstract

Background and Objectives: Current and voltage signals' distortion caused by the fault in the power system has negative effects upon the operation of the protective devices. One of the influencing factors is the existence of the exponential DC which can significantly distort the signals and lead to a possible malfunction of the protective devices, especially distance and over-current relays. The main problem is the lack of clarity about this component due to the dependence of its time constant and initial amplitude to the configuration of the electrical grid, location and resistance of faulty point. This makes it hard to extract the main frequency phasors of the voltage and current.

Methods: Considering the importance of a fast clearance of the fault, this paper offers a method for an effective and fast removal of the decaying-DC that employs a data window with a length that is equal to the half cycle of the main frequency, while the conventional methods mostly use data from one cycle or even more. The proposed method is based upon the extraction of the decaying-DC component's parameters.

Results: The efficiency of this method is compared to the conventional Fourier algorithm of Half-Cycle (HCFA) and the mimic filter plus the HCFA.

Conclusion: The outcomes display that the proposed method presents a better efficiency from the point of view of the speed and the accuracy of convergence to the final results.

Introduction

Fast fault clearance in the power system is a crucial requirement for the system operation. Its main purpose is to separate the grid faulty areas and to prevent the instability. This is performed by the operation of protective relays installed in the power system and has to happen in a fraction of a power frequency cycle. Input signals of different relays are filtered according to the protective logic and their operation by removing the unwanted quantities and only preserving the desired ones [1].

Since the most of the protective relays such as distance and over-current relays operate based on the main phasors of the voltages and currents, the employed digital protective algorithms should be designed so that they eliminate the DC component and harmonics. Otherwise, the proper function of the protective relay may be disrupted due to any these quantities. For instance, presence of the decaying-DC in the current signal will lead to reduction of the impedance obtained in the distance relay and the overreach phenomenon. Consequently, the relay reacts for a fault which has not happened in its operational zone.

Doi: 10.22061/JECEI.2021.8205.492

Algorithms used in digital filters, known as phasor estimation algorithms, can structurally be classified as follows:

a) Algorithms based on the small window data, such as: i) The Sample and Its First Derivative method by Mann-Morrison [2], ii) The First and the Second Derivative method by Gilchrist-Rockfeller-Udern [3], and iii) Two Samples method by Mokino-Miki [4].
b) Algorithms based on the orthogonal such as: Fourier Filter algorithm [5], [6] and its products i.e. Cosine and Sinusoidal Filters as well as the Walsh Filter algorithm [7] and its products i.e. CAL and SAL.
c) The Least Error Squares algorithms (LES), such as: i) Integral LSQ Fit [8], ii) Power Series LSQ Fit [9], and iii) Multi-Variable Series LSQ technique [10].
d) Algorithms based on the Kalman Filter [11].

The conventional Discrete Fourier Transform, DFT i.e. the group (b) algorithm, is the most sought-after algorithm used in the digital protection because of its proper operation and the ease of implementation. DFT algorithms are classified into Half-Cycle and Full-Cycle algorithms.

The DFT cannot eliminate the DC component because of its non-periodic nature and large frequency spectrum. In the recent years, some algorithms are offered in order to eliminate or to weaken the adverse aspects of the exponential DC component in the output of the full-cycle algorithms [12]-[40]. In [12], a mimic filter with the Fourier algorithm is proposed to remove the DC component. In this method, if the time constants (τ) of the decaying-DC component and the mimic filter are the same, the impact of the DC component can be completely removed.

In [12], the decaying-DC parameters are calculated by two Full-Cycle successive outputs Discrete Fourier Transform (FCDFT). In the modified version of the method in [12] by the same authors [14], the effect of analog anti-aliasing filter i.e. production of additional decaying-DC has been overcome. The method proposed in [15] uses two parallel DFT filters, one of them is set to the main frequency and the other to the m^{th} harmonic. The latter is used for calculating the decaying-DC component's parameters.

In [16], two partial sums are employed for complete removing the DC component's effects. One of the partial sums is the sum of odd samples and the other is the sum of even samples during a full cycle of the power frequency. The amplitude and τ of the DC component in [17] are obtained by two mathematical expressions which directly use the values from four samples. This method, which can be used in both full-cycle and half-cycle data windows, requires two extra samples.

In [18], the phasor is computed from three consecutive DFT estimates by using a recursive computing. So, it requires two extra samples. The method in [19] eliminates the DC impact by means of the difference between the outputs of the FCDFT for even and odd samples. The method proposed in [20] calculates the value of the actual DC offset by integrating the input signal. And then, the DC component is subtracted from the main signal for each sample. In [21], the DC component impact is removed by combining the outputs of FCDFT for even and odd samples extracted by decimation of the full cycle data window by two and by four.

The FCDFT output in [22] is corrected by integrating the input signal in a full cycle data window. To consider the changing frequency scenario of the electric network, [23] proposes LES method iteratively which fulfills the steady state and dynamic performance criteria of the IEEE standard for Synchrophasor Measurements for Power Systems [24]. The proposed method in [23] requires extra memory for storing LES filter coefficients of various frequencies.

The method in [25] computes the amplitude of the main frequency component by combining the FCDFT outputs filters for odd and even samples. In [26], the decaying-DC parameters are calculated by integrating the fault current signal in a full cycle. Then, the DC is subtracted from the main fault current.

The method in [27] uses MATLAB's fsolve function to estimate the fundamental frequency fault signal component which is developed for two cases including i) decaying-DC with known time constant, and ii) unknown time constant. For improving the fault location estimates, [28] removes the effect of the DC component by curve fitting by means of Non-Linear Least Squares method. Algorithms based upon wavelet transform [29] and neural network [30] have been utilized for the protection and phasor estimation applications.

Recently, phasor estimation under dynamic conditions has been under investigation. The methods in [31]-[33] propose dynamic phasor estimation which consider the off-nominal frequency condition. These methods may produce more accurate results for phasor estimation. However, they entail higher computational burden. In one of the most recent algorithms in this category, the DC amplitude and time constant are calculated by applying Hilbert transform and integrating the fault current signals within one cycle [33]. Hilbert transform has been utilized due to its effectiveness in the analysis of time-varying signals. Over the past few years some studies are conducted to forecast phenomena with uncertainties [34]-[37]. In [34] Gaussian model, in [35] ensemble learning based method and in [36] deep learning-based approach are used for forecasting.

All of the above methods are proposed for the full cycle algorithms and there are only few methods proposed for the half cycle algorithms. Half-cycle algorithms have a higher convergence speed, in the order of two times faster than full cycle methods. Among the most important half-cycle algorithms, the Half-Cycle DFT algorithm (HCDFT) and the combination of digital mimic filter and the HCDFT algorithm can be nominated. These methods are unable to completely remove the effects of the DC component [38].

One of the recently proposed methods to extract the phasor by means of the half-cycle data window is presented in [39] in which three offline look-up tables have to be created prior to processing the input signal for determining the decaying-DC component's parameters and removing its effects from the main signal. The look-up tables should be referred to during the online process which in turn increases the computational burden.

The method in [40] proposes a general modified DFT algorithm, so that it is possible to employ the method in both HCDFT and FCDFT algorithms. In this method, two successive outputs of the imaginary and real part filters are combined to eliminate the DC impact. In the method proposed in [40], three parallel filters are used. In addition, the data window length for HCDFT will be $n/2+1$, where n is the number of samples per cycle. A hybrid algorithm based upon integration and half-cycle DFT is proposed in [41]. This method computes the DC component parameters and the unwanted share of DC in the phasor estimation. However, it requires two movements in the sampling window.

In this paper, a method is presented to improve the efficiency of the Half-Cycle algorithm against the DC component. In the proposed method, the influence of the decaying-DC is entirely eliminated by means of its parameters' estimation. The proposed method can be used for a wide range of decaying-DC time constants and it is not dependent on the amount of the time constant.

This paper is structured as follows: the first section introduces the problem description, the second section formulates the proposed method, the third section evaluates the performance of the proposed methods, and the final section concludes this work.

Problem Description

The unpredictable nature of the fault signals in the power grid makes the main component phasor estimation a challenging process. Under the usual operating conditions, the voltage and current signals are almost clear sinusoidal with the main frequency of the grid. However, after failures or disturbances in the grid, these waveforms are distorted containing decaying-DC, harmonics, and the non-main frequency components [15].

The reactive-resistive feature of the network results in the generation of decaying-DC signal. The DC component considerably impacts the current signal where it has an insignificant influence of the voltage signal. There have been reports on up to 15% error in the phasor estimation by the deteriorative effect of DC component on the calculations [12]. Besides, DC component parameters cannot be determined with a high level of certainty. For instance, its time constant can depend on the configuration of the grid, the resistance and the location of fault and is specified by means of the X/R ratio seen from the fault point in general. For highly resistive earth faults, decaying rate will be so high that the decaying-DC would decay in less than half a cycle in some cases.

Generally, decaying-DC time constant range of variation is from 0.5 a cycle up to 5 cycles. It is not an alternating signal and thus, contains a wide frequency spectrum. Therefore, convergence speed and accuracy of the digital filtering methods are affected which leads to errors in the estimated phasors. Fig. 1 shows the frequency spectrum of the DC component with different time constants.

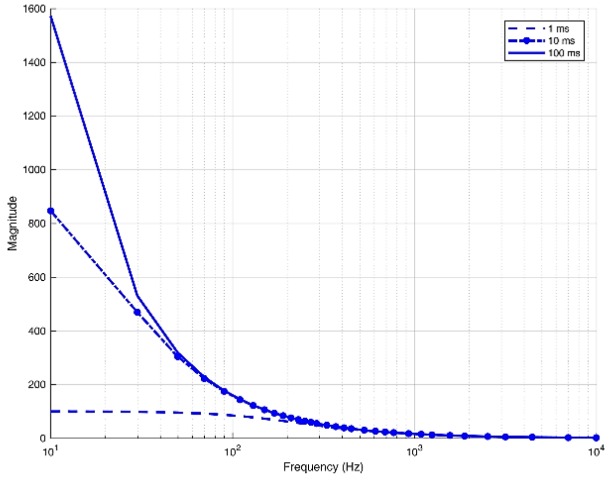


Fig. 1: Frequency spectrum of the DC component with various time constants.

As it can be observed, the ratio of low frequency component to high frequency one changes with the time constant. In other words, a fast decaying-DC contains less low frequency components compared to a slow decaying one.

The Proposed Method

In this part, the structure of the proposed method is introduced. First, influence of the DC component on the HCFA will be examined and then, for removing this effect a method will be presented.

A. Effect of the Decaying-DC Component

Let the input fault current signal contain: i) fundamental component, ii) first harmonic to p^{th} harmonics, iii) decaying-DC. It can be presented by the formula below:

$$i(t)=I_0e^{-t/\tau}+\sum_{k=1}^{p}I_k.\sin(k\omega_1t+\theta_k) \tag{1}$$

where I_0 is the DC amplitude and τ is its time constant. I_k is k^{th} harmonic amplitude, ω_1 is main angular frequency, ϑ_k is of k^{th} harmonic phase angle, and *p* is the largest order of harmonic that exists in the waveform.

It is assumed that the harmonic components that have higher orders than *p* have been eliminated in the input using the anti-aliasing low-pass filter. The analog to digital conversion is performed by an A/D converter as:

$$i(n)=I_0e^{-nT/\tau}+\sum_{k=1}^{p}I_k.\sin(k\omega_1nT+\theta_k) \tag{2}$$

where *T* represents the sampling time period and *n* points to the n^{th} sample.

The main frequency HCFA generates its output using the following equation:

$$\begin{aligned} HCdft_1^t &= \frac{4}{N}\sum_{n=0}^{\frac{N}{2}-1}i(n)\times(\sin\omega_1nT+j\cos\omega_1nT) \\ &= \frac{4}{N}\sum_{n=0}^{\frac{N}{2}-1}i(n)\times j\times e^{-j\omega_1nT} \end{aligned} \tag{3}$$

where $HCdft_1^t$ is output of the main frequency HCFA for the total input signal, i.e., the signal that includes main frequency, harmonics, and the decaying-DC, and *N* is the quantity of samples per each cycle.

The harmonic components with odd order are eliminated by the HCFA and the input signal does not include even harmonics [42], the output will only contain the main frequency and the DC. The main frequency phasor will be found by removing the DC from the output of this algorithm. The output of main frequency HCFA for the exponential DC input can be calculated as follows:

$$HCdft_1^{dc}=\frac{4}{N}\sum_{n=0}^{\frac{N}{2}-1}I_0e^{-nT/\tau}\times je^{-j\omega_1nT}=\frac{4}{N}\times jI_0\frac{1+e^{-NT/2\tau}}{1-e^{-T/\tau}e^{-j\omega_1T}} \tag{4}$$

where $HCdft_1^{dc}$ is the output of main frequency HCFA; resulted from the DC component.

Once $HCdft_1^{dc}$ is determined, the output of main frequency HCFA for the main frequency component can be calculated using:

$$HCdft_1^{1f}=HCdft_1^t-HCdft_1^{dc} \tag{5}$$

where $HCdft_1^{1f}$ is the output of main frequency HCFA for the main frequency component which is the main frequency phasor.

According to (4), $HCdft_1^{dc}$ is a function of time constant and amplitude of the decaying-DC. Therefore, to obtain the output of the HCFA for the decaying-DC, these parameters have to be determined first.

B. Determining Decaying-DC Component's Parameters

As it was mentioned in the previous subsection, to obtain the main frequency phasors, the main frequency HCFA's output for the decaying-DC is required. According to (4), $HCdft_1^{dc}$ is a function of τ and amplitude of the DC component. Therefore, the mentioned parameters must be calculated first.

The current and voltage signals of the fault may consist main frequency, decaying-DC, high-frequency harmonics, and noise. Protective equipment use a filter with anti-aliasing low-pass features in each analog channel input to remove the high-frequency components. As a result, the components with the frequencies higher than the filter cut-off frequency of the anti-aliasing filter do not show up in the channel output.

Correspondingly, a Fourier filter of half-cycle set to a harmonic frequency higher than the low-pass filter cut-off frequency can be designed so that the main frequency and the other harmonics will not emerge in its output. Consequently, the output will only be influenced by the DC component. Time constant and Amplitude of the DC can be calculated by the output of the m^{th} harmonic frequency HCFA.

The Fourier filter of Half-Cycle is set to the m^{th} harmonic frequency. This frequency has to be higher than the low-pass filter cut-off frequency and lower than the half of the sampling frequency. Subsequently, output of the Fourier filter of Half-Cycle will only contain the effect of decaying DC and it goes as follows:

$$HCdft_m^{dc}=\frac{4}{N}\sum_{n=0}^{\frac{N}{2}-1}I_0e^{-nT/\tau}\times je^{-j\omega_1nmT} \tag{6}$$

With the assumption that the m^{th} harmonic is odd, one can rewrite the above equation as:

$$HCdft_m^{dc}=\frac{4}{N}\times jI_0\frac{1+e^{-NT/2\tau}}{1-e^{-T/\tau}e^{-j\omega_1mT}} \tag{7}$$

where $HCdft_m^{dc}$ is the outcome of the m^{th} harmonic frequency Half-Cycle Fourier filter.

Dividing (7) into imaginary and real parts results in the equations below, where $e^{-T/\tau}$ is substituted for *E*. The real part *R* is:

$$R=\frac{4}{N}\frac{I_0(1+E^{N/2})E\sin(\omega_1mT)}{1+E^2-2E\cos(\omega_1mT)} \tag{8}$$

and the imaginary part *I* is:

$$I=\frac{4}{N}\frac{I_0(1+E^{N/2})(1-E\cos(\omega_1mT))}{1+E^2-2E\cos(\omega_1mT)} \tag{9}$$

By using (8) and (9), the values for *E* and $(4/N)I_0(1+E^{N/2})$ can be calculated as:

$$E=\frac{R}{R\cos(\omega_1mT)+I\sin(\omega_1mT)} \tag{10}$$

$$\frac{4}{N}I_0\left(1+E^{N/2}\right)=\frac{R\left(1+E^2-2E\cos(\omega_1 mT)\right)}{E\sin(\omega_1 mT)} \quad (11)$$

The above equations use imaginary and real parts of the m^{th} harmonic frequency Half-Cycle Fourier algorithm's output and the specified values of $\sin(\omega_1 mT)$ and $\cos(\omega_1 mT)$. By placing (10) and (11) in (4), the main frequency Half-Cycle Fourier algorithm's output for the DC component is resulted. Finally, the main frequency phasor of the input signal, $HCdft_1^{1f}$, is achieved via (5).

In line with the above explanations, it can be observed that the proposed method requires two Half-Cycle Fourier filters; one set to the fundamental frequency and the other set to the m^{th} harmonic, where m is odd. The main purpose of using the m^{th} harmonic Fourier filtering is to acquire the parameters of decaying-DC. The needed calculations of the proposed method are: i) the implementation of two Fourier filters of Half-Cycle and ii) the calculations pertaining to (10), (11), (4), and (5). The proposed method flowchart is illustrated in Fig. 2.

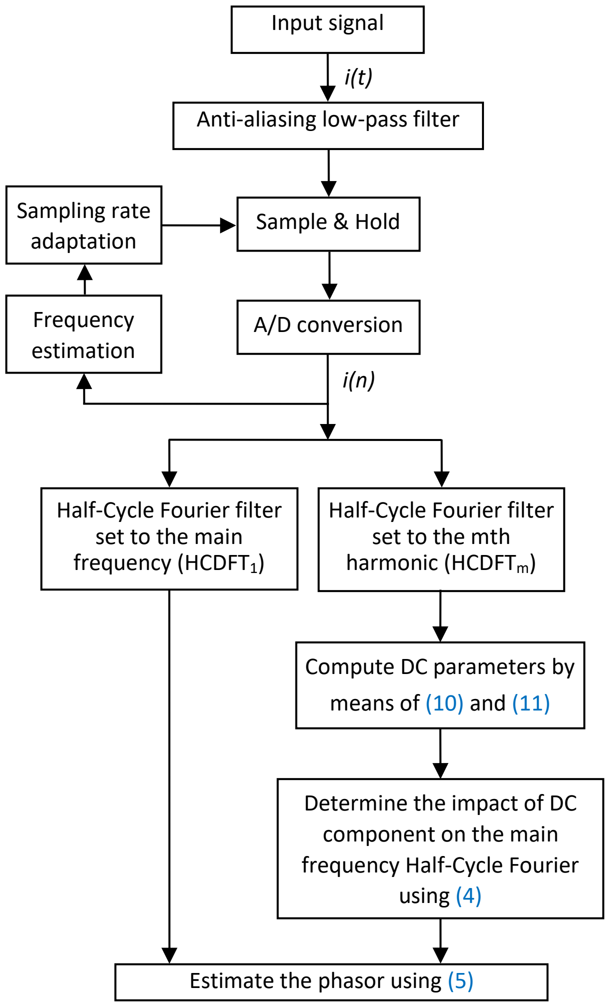

Input signal
i(t)
Anti-aliasing low-pass filter
Sampling rate adaptation
Sample & Hold
Frequency estimation
A/D conversion
i(n)
Half-Cycle Fourier filter set to the main frequency (HCDFT₁)
Half-Cycle Fourier filter set to the mth harmonic (HCDFTₘ)
Compute DC parameters by means of (10) and (11)
Determine the impact of DC component on the main frequency Half-Cycle Fourier using (4)
Estimate the phasor using (5)

Fig. 2: The proposed method flowchart for the phasor estimation.

Results and Discussion

Algorithms efficiency is being assessed by the application of the following input signal:

$$i(t)=I_1\cos(\omega_0 t+\theta)-I_0 e^{-\frac{t}{\tau}} \quad (12)$$

in which I_0, amplitude of the DC component, and I_1, amplitude of the main frequency component are selected as 1 per-unit. *i(t)* is applied to the various algorithms with a variable time constant of the decaying-DC component (τ) and their sensitivity versus τ variation is evaluated.

To make a comparison between different methods, the performance indices (PI_1 and PI_2) are utilized [12]. The performance indices are defined based upon the output of the digital phasor extraction filters for the input signal *i(t)*. *y(t)* is the waveform of the filter's output for the applied input signal. *y(t)* oscillates around 1 per-unit before permanently settling in this value. The first performance index PI_1 is calculated using the following equation:

$$PI_1(\tau)=\int_{T_0}^{NT}\left[1-y(t)\right]^2 dt \quad (13)$$

As soon as *y(t)*'s amplitude exceeds 1 per-unit, the integration starts (T_0) and proceeds until *NT*, which represents an integer number of the main frequency cycles. In the simulations, let *N* be 3. PI_1 represents the extent of the amplitude oscillations around the steady-state final value in the filter's output in the presence of the DC component in the input.

The second performance index PI_2 is equal to the highest overshoot percentage in *y(t)*'s amplitude. There is a straight relevance between this index and the protective devices' overreach potential.

$$PI_2(\tau)=\left(Max\left[y(t)-1\right]\right)\times 100 \quad (14)$$

As much as these indices get closer to zero, the higher quality of the tested algorithm is inferred. The input signal's sampling rate is 36 samples per cycle and the value for *m* is selected as 13 for the proposed method. The sampling window used in the simulations is the half of the main frequency cycle that means 18 samples.

The frequency response of the Half-Cycle Fourier filter set to the main frequency is presented in Fig. 3. As it can be observed, this filter cannot remove the decaying-DC component when used standalone. The time response generated by applying the input signal to the HCFA is illustrated in Fig. 4.

The values for the performance indices of the HCFA versus τ variation in the range of 0.5 cycle to 5 cycles are presented in Table 1.

If the current waveform passes a mimic circuit including a series resistor and inductor, the exponential decaying component will be removed or deteriorated in the circuit's output. The transfer function for the mimic circuit in the Laplace domain would be:

$$H_{mimic}(S) = K(1 + S\tau_1) \qquad (15)$$

where τ_1 is the time constant which mimic filter is set to.

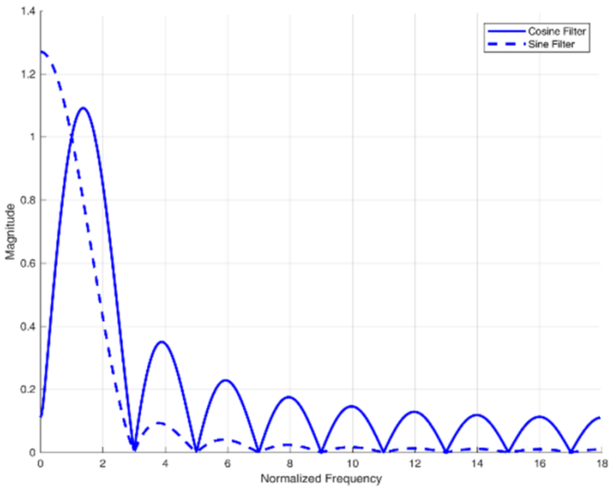


Fig. 3: Frequency response of the Half-Cycle Fourier filter for the main frequency.

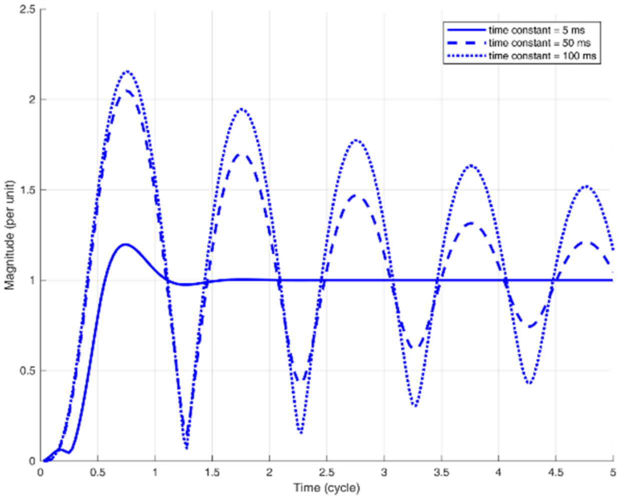


Fig. 4: Time response of the HCFA.

Table 1: Performance Indices for the HCFA

Time constant (mSec)	*PI₁*	*PI₂ (%)*
10	2.8692	49.1603
20	9.9800	78.5331
40	22.6705	99.7476
60	31.7330	108.1275
80	38.0549	112.6007
100	42.5512	115.3807

If the decaying component's time constant is equal to τ_1, its effect will be eliminated in the output of the mimic filter and if the time constant has a different value, its effect will be significantly reduced. The mimic circuit including a resistor and an inductor can also be digitally modeled. In the case *S* is replaced using the following equation, the *Z* domain representation of the mimic circuit's transfer function can be obtained:

$$S = \frac{1 - Z^{-1}}{\Delta T} \qquad (16)$$

where *ΔT* is the sampling period.

The time constant is set to 50 ms in the mimic filter's design which is approximately located in the middle of its variation range. The digital mimic filter's frequency response is shown in Fig. 5. It is clear that the mimic filter is a high-pass filter that means boosting the high frequency components. Therefore, it is prone to high frequency noise.

By combining the digital mimic filter and the HCFA, the performance of the HCFA in confronting with the decaying-DC can be improved to some extent. The frequency response of the combination of digital mimic filter and the HCFA is presented in Fig. 6. The time response obtained by applying the input signal to the combination of the mimic filter and the HCFA is illustrated in Fig. 7.

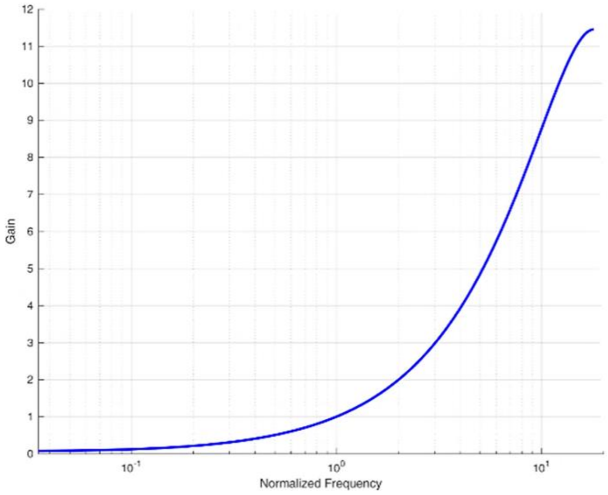


Fig. 5: Digital mimic filter's frequency response.

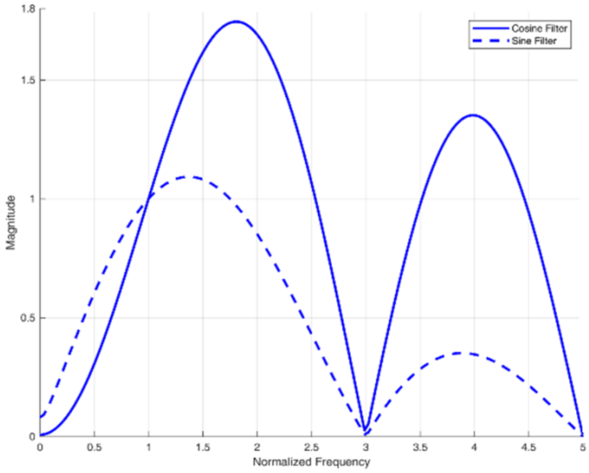


Fig. 6: Frequency response of the digital mimic plus the HCFA.

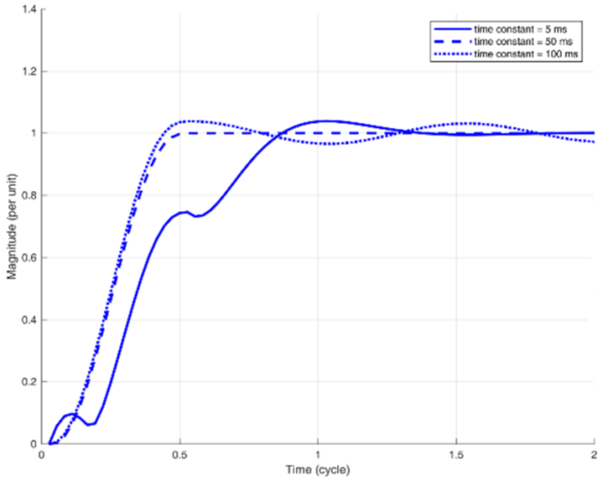


Fig. 7: Time response of the combination of digital mimic filter and the HCFA.

Performance indices for the combination of digital mimic filter and the HCFA are presented in Table 2.

Table 2: Performance Indices for the Combination of Digital Mimic Filter and the HCFA

Time constant (mSec)	PI_1	PI_2 (%)
10	0.054969	7.2968
20	0.046537	5.7402
40	0.004078	1.4166
60	0.003376	1.1969
80	0.021745	2.8038
100	0.044052	3.8010

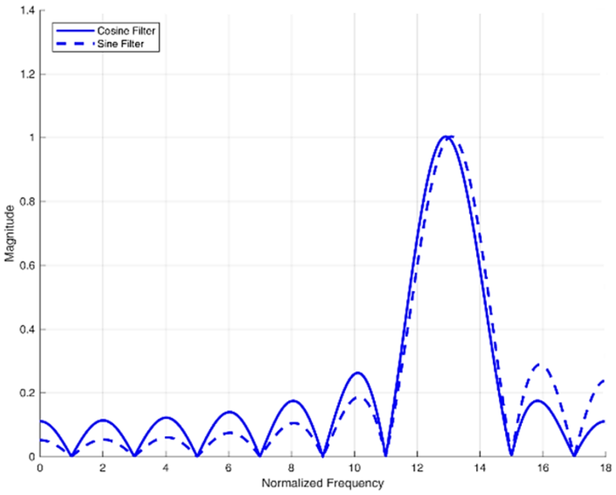


Fig. 8: Frequency response of the Fourier filter set to the 13th harmonic.

By using two parallel Half-Cycle Fourier filters, impact of the DC component upon the extracted phasor can be totally eliminated. As it was mentioned before, one of these Half-Cycle Fourier filters is set to the *m*th harmonic (*m*=13) and the other is set to the main frequency. Fig. 8 demonstrates the frequency response of the Fourier filter set to the 13th harmonic.

The time response obtained by applying the input signal to the proposed algorithm is shown in Fig. 9.

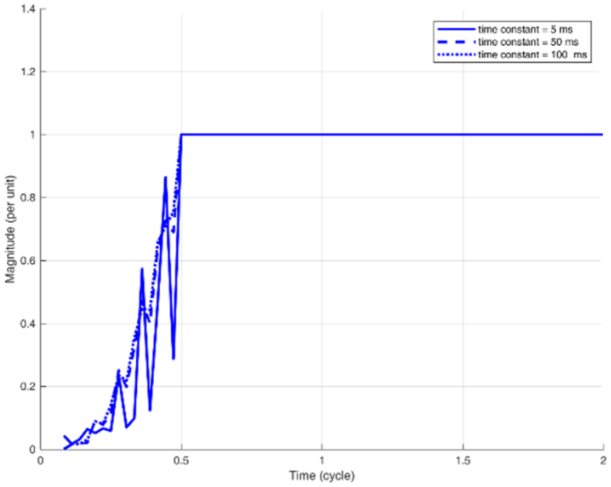


Fig. 9: Time response of the proposed algorithm.

For the proposed algorithm the values of the performance indices for τ variation in the range of 0.5 cycle to 5 cycles are presented in Table 3.

Table 3: Performance Indices for the Proposed Algorithm

Time constant (mSec)	PI_1	PI_2 (%)
10	0.00	0.00
20	0.00	0.00
40	0.00	0.00
60	0.00	0.00
80	0.00	0.00
100	0.00	0.00

By a careful examination of the time responses obtained from different methods, it can be observed that the Half-Cycle Fourier filter and the combination of digital mimic filter and the Half-Cycle Fourier both have overshoots in their outputs. Whereas, the proposed method does not have such overshoots and as soon as the data window fills with the valid fault data, its output reaches the desired value. In addition, the proposed method generates favorable responses for different time constants and it is not dependent on the value of the *τ*.

More simulations are performed to have a more vivid representation of different algorithms' performance for a wider range of *τ* variations of the decaying-DC, where the *τ* varies from 1 to 120 ms. Outputs after filling their data windows with the fault data are shown in Fig. 10.

The highest deviation of the HCFA from the desired output is 49.18% which happens in 120 ms time constant. The highest deviation from the desired output for the combination of digital mimic filter and the HCFA is 25.40% happening in 5 ms time constant. The proposed method's output comes to the favorite value as soon as the data window fills with the first half cycle data.

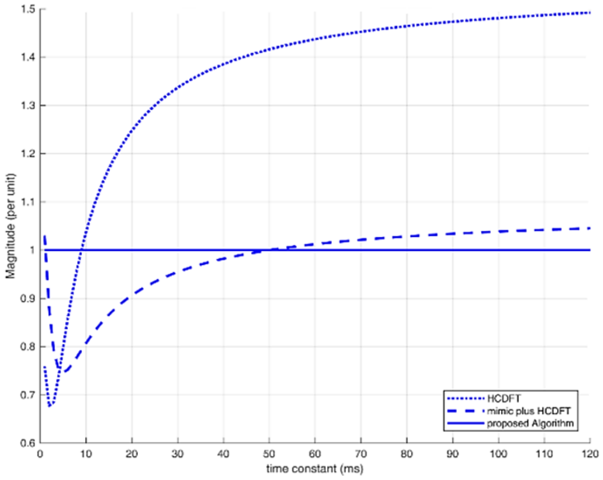


Fig. 10: The extracted phasor at the end of the fault's first half cycle.

Fig. 11 demonstrates the variations of the highest overshoot in the algorithms output as a function of the decaying-DC's time constant. The highest overshoot in the HCFA is 117.27% happening in 120 ms time constant. The highest overshoot in the combination of digital mimic filter and the HCFA is 7.39% happening in 11 ms time constant, whereas the highest overshoot in the proposed method is 2.59% happening in 1 ms time constant. As it can be observed, the proposed method does not generate a large overshoot for a wide range of the time constant variation.

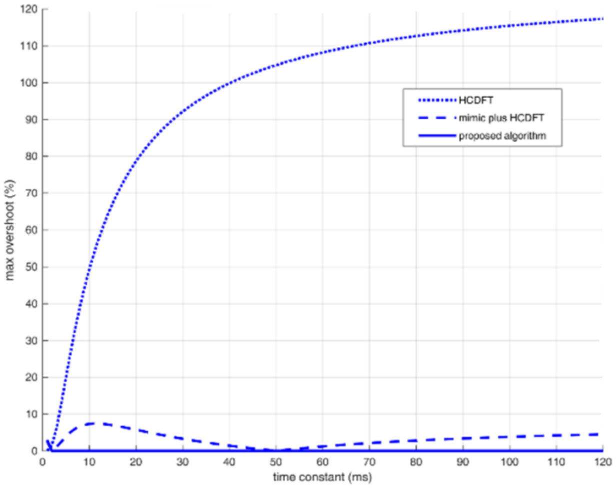


Fig. 11: The highest overshoot in the extracted phasor.

Conclusion

In this paper, a method for extracting the main frequency phasor was proposed which is favorably robust against the impact of the DC component. The proposed method estimates the phasors using a data window equal to the half cycle of the power grid's main frequency.

The proposed method utilizes two parallel filters set to different frequencies, so that after filling the data window with the fault data, precise and stable outputs are generated. In the proposed method, once the data window is filled with half-cycle data ($n/2$ of samples), the main phasor component is a computed, while in the presented method in reference [39] three look-up tables are referred to during online processing which causes an increase in computational work. The offered data window length for HCDFT method is $n/2+1$ in reference [40] which is one sample longer than that of our presented method.

Finally, in the proposed method of reference [41] it is necessary to move the data window two samples. As a result, the main phasor component will be calculated with a two-sample delay. Moreover, the Efficiency of the proposed method was compared to the HCFA and the combination of digital mimic filter and the HCFA which showed a higher speed and accuracy of the proposed method. The performance indices (PI_1, PI_2) are calculated for various algorithms and the indices are almost zero for the proposed method. The more these indices get closer to zero, the higher quality of the tested algorithm is inferred and therefore the desired performance of the proposed method is confirmed.

Author Contributions

Authors have had an equal contribution in the problem and data analysis, interpreting the results and writing the manuscript.

Acknowledgement

Authors want to warmly acknowledge the kind helps provided by Dr. Saeed RamezanJamaat for the writing assistance and by Mr. Mehdi Bakhshandeh for proof reading of the 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

DC	Direct Current
DFT	Discrete Fourier Transform
FCDFT	Full-Cycle Discrete Fourier Transform

HCDFT	Half-Cycle Discrete Fourier Transform
HCFA	Half-Cycle Fourier Algorithm

References

[1] E.O. Schweitzer, D. Hou, "Filtering for protective relays," in Proc. IEEE WESCANEX 93 Communications, Computers and Power in the Modern Environment., 1993.

[2] B.J. Mann, I.F. Morrison, "Digital calculation of impedance for transmission line protection," IEEE Trans. Power Appar., PAS-90(1): 270-279, 1971.

[3] G.B. Gilchrist, G.D. Rockefeller, E.A. Udren, "High-speed distance relaying using a digital computer, Part I: System description," IEEE Trans. Power Appar., PAS-91(3): 1235-1243, 1972.

[4] J. Makino, Y. Miki, "Study of operating principles and digital filters for protective relays with a digital computer," in Proc. Conf. Pap. IEEE Power. Eng. Soc., 1975: 661-668, 1975.

[5] M. Ramamoorty, "Application of digital computers to power system protection," J. Inst. Eng., 52(10): 235-238, 1972.

[6] P.G. McLaren, M.A. Redfern, "Fourier-series techniques applied to distance protection," in Proc. Institution of Electrical Engineers Conf., 122(11): 1301-1305, 1975.

[7] J.W. Horton, "The use of walsh function for high-speed digital relaying," in IEEE PES Summer Meeting, Paper A 75 582 7: 1-9, 1975.

[8] R.G. Luckett, P.J. Munday, B.E. Murray, "A substation-based computer for control and protection," in IEE Conf. on Developments in Power System Protection, Pub. 125: 252-260, 1975.

[9] A.W. Brooks, "Distance relaying using least-squares estimates of voltage, current and impedance," in Proc. IEEE PICA Conf., 77CH 1131-2 PWR: 394-402, 1977.

[10] M.S. Sachdev, M.A. Baribeau, "A new algorithm for digital impedance relays," IEEE Trans. Power Appar., PAS-98(6): 2232-2240, 1979.

[11] A.A. Girgis, R.G. Brown, "Application of kalman filtering in computer relaying," IEEE Trans. Power Appar., PAS-100(7): 3387-3397, 1981.

[12] G. Benmouyal, "Removal of DC-offset in current waveforms using digital mimic filtering," IEEE Trans. Power Deliv., 10(2): 621-630, 1995.

[13] J.C. Gu, S.Li. Yu, "Removal of DC offset in current and voltage signals using a novel fourier filter algorithm," IEEE Trans. Power Deliv., 15(1): 73-79, 2000.

[14] S.L. Yu, J.C. Gu, "Removal of decaying DC in current and voltage signals using a modified fourier filter algorithm," IEEE Trans. Power Deliv., 16(3): 372-379, 2001.

[15] T. Sidhu, X. Zhang, F. Albas, M. Sachdev, "Discrete-Fourier-transform-based technique for removal of decaying DC offset from phasor estimates," IEE Proc. Gener. Transm. Distrib., 150(6): 745-752, 2003.

[16] Y. Guo, M. Kezunovic, D. Chen, "Simplified algorithms for removal of the effect of exponentially decaying dc-offset on fourier algorithm," IEEE Trans. Power Deliv., 18(3): 711-717, 2003.

[17] J.F. Miñambres Argüelles, M.A. Zorrozua Arrieta, J. Lázaro Domínguez, B. Larrea Jaurrieta, M. Sánchez Benito, "A new method for decaying DC offset removal for digital protective relays," Electr. Power Sys. Res., 76(4): 194-199, 2005.

[18] C.S. Chen, C.W. Liu, J.A. Jiang, "Application of combined adaptive fourier filtering technique and fault detector to fast distance protection," IEEE Trans. Power Deliv., 21(2): 619-626, 2006.

[19] S.H. Kang, D.G. Lee, S.R. Nam, P.A. Crossley, Y.C. Kang, "Fourier transform-based modified phasor estimation method immune to the effect of the DC offsets," IEEE Trans. Power Deliv., 24(3): 1104-1111, 2009.

[20] K.N.A. Al-Tallaq, H.D. Al-Sharai, M.E. El-Hawary, "Online algorithm for removal of decaying DC-Offset from fault currents," Electr. Power Sys. Res., 81(7): 1627-1629, 2011.

[21] K.M. Silva, B.F. Küsel, "DFT based phasor estimation algorithm for numerical digital relaying," Electron. Lett., 49(6): 412-414, 2013.

[22] M.R. Dadash Zadeh, Z. Zhang, "A new DFT-based current phasor estimation for numerical protective relaying," IEEE Trans. Power Deliv., 28(4): 2172-2179, 2013.

[23] S. Das, T. Sidhu, "A simple synchrophasor estimation algorithm considering IEEE standard C37.118.1-2011 and protection requirements," IEEE Trans. Instru., 62(10): 2704-2715, 2013.

[24] IEEE Standard for Synchrophasor Measurements for Power Systems, IEEE Standard C37.118.1, 2011 (Revision of IEEE Std C37.118, 2005).

[25] A. Rahmati, R. Adhami, "An accurate filtering technique to mitigate transient decaying DC offset," IEEE Trans. Power Deliv., 29(2): 966-968, 2013.

[26] A. Akbar Abdoos, S.A. Gholamian, M. Farzinfar, "Accurate and fast DC offset removal method for digital relaying schemes," IET Gener. Transm. Distrib., 10(8): 1769-1777, 2016.

[27] S.A. Gopalan, Y. Mishra, V. Sreeram, H.H.C. Iu, "An improved algorithm to remove DC offsets from fault current signals," IEEE Trans. Power Deliv., 32(2): 749-756, 2016.

[28] K.W. Min, S. Santoso, "DC offset removal algorithm for improving location estimates of momentary faults," IEEE Trans. Smart Grid, 9(6): 5503-5511, 2017.

[29] A.A. Yusuff, A.A. Jimoh, J.L. Munda, "Stationary wavelet transform and single differentiator based decaying DC-Offset filtering in post fault measurements," Measurement, 47: 919-928, 2013.

[30] C.D.L. da Silva, G. Cardoso Junior, L. Mariotto, G. Marchesan, "Phasor estimation in power systems using a neural network with online training for numerical relays purposes," IET Sci. Meas. Technol., 9(7): 836-841, 2015.

[31] R.K. Mai, L. Fu, Z.Y. Dong, K.P. Wong, Z.Q. Bo, H.B. Xu, "Dynamic phasor and frequency estimators considering decaying DC components," IEEE Trans. Power Sys., 27(2): 671-681, 2011.

[32] P. Banerjee, S.C. Srivastava, "An effective dynamic current phasor estimator for synchrophasor measurements," IEEE Trans. Instru., 64(3): 625-637, 2014.

[33] M. Tajdinian, M. Zareian Jahromi, K. Mohseni, S. Montaser Kouhsari, "An analytical approach for removal of decaying DC component considering frequency deviation," Electr. Power Sys. Res., 130: 208-219, 2015.

[34] A. Ahmadpour, S. Gholami Farkoush, "Gaussian models for probabilistic and deterministic Wind Power Prediction: Wind farm and regional," Int. J. Hydrogen Energy, 45(51): 27779-27791, 2020.

[35] J. Lee, W. Wang, F. Harrou, Y. Sun, "Wind Power prediction using ensemble learning-based models," IEEE Access, 8: 61517-61527, 2020.

[36] S. Sadeghi, H. Jahangir, B. Vatandoust, M. Aliakbar Golkar, A. Ahmadian, A. Elkamel, "Optimal bidding strategy of a virtual power plant in day-ahead energy and frequency regulation

markets: A deep learning-based approach," Int. J. Electr. Power Energy Syst., 127, 2020.

[37] J. Guan, J. Lin, J. Guan, E. Mokaramian, "A novel probabilistic short-term wind energy forecasting model based on an improved kernel density estimation," Int. J. Hydrogen Energy, 45(43): 23791-23808, 2020.

[38] L. Wang, "Frequency responses of phasor-based microprocessor relaying algorithms," IEEE Trans. Power Deliv., 14(1): 98-109, 1999.

[39] T.S. Sidhu, X. Zhang, V. Balamourougan, "A new half-cycle phasor estimation algorithm," IEEE Trans. Power Deliv., 20(2): 1299-1305, 2005.

[40] K.M. Silva, F.A.O. Nascimento, "Modified DFT-based phasor estimation algorithms for numerical relaying applications," IEEE Trans. Power Deliv., 33(3): 1165-1173, 2017.

[41] M. Tajdinian, A.R. Seifi, M. Allahbakhshi, "Half-cycle method for exponentially DC Components elimination applicable in phasor estimation," IET Sci. Meas. Technol., 11(8): 1032-1042, 2017.

[42] B. Ram, Power System Protection and Switchgear, Tata McGraw-Hill Education, 2011.

Biographies

Hamid Sardari received the B.Sc. and M.Sc. degrees in electrical engineering from Iran University of Science and Technology, Tehran, Iran, in 2003 and 2006, respectively, and Ph.D. from Islamic Azad University, Tehran, Iran, in 2020. He is currently pursuing research on fault location, digital protection, and phasor estimation in Islamic Azad University, Tehran, Iran.

- Email: sardari@iauet.ac.ir
- ORCID: 0000-0001-5032-5012
- Web of Science Researcher ID: NA
- Scopus Author ID: 36895082000
- Homepage: http://fani.iauet.ac.ir/fa/page/669/

Babak Mozafari received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Sharif University of Technology, Tehran, Iran, in 1998, 2001, and 2007, respectively. Currently, he is an associate professor in the Department of Electrical and Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran. His research interests include power system protection and power system dynamics.

- Email: mozafari@srbiau.ac.ir
- ORCID: 0000-0002-5699-2577
- Web of Science Researcher ID: AAT-5629-2021
- Scopus Author ID: 9743165700
- Homepage: https://faculty.srbiau.ac.ir/b-mozafari/fa

Heidar Ali Shayanfar received the B.Sc. and M.S.E. degrees in electrical engineering in 1973 and 1979, respectively. He received the Ph.D. degree in electrical engineering from Michigan State University, East Lansing, MI, USA, in 1981. Currently, he is a full professor in the Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran. His research interests include the application of artificial intelligence to power system control design, dynamic load modeling, power system observability studies, voltage collapse, and congestion management in a restructured power system, reliability improvement in distribution systems, and reactive pricing in deregulated power systems. He has published more than 490 technical papers in the international journals and conferences proceedings. Dr. Shayanfar is a member of the Iranian Association of Electrical and Electronic Engineers.

- Email: hashayanfar@iust.ac.ir
- ORCID: 0000-0002-2330-0546
- Web of Science Researcher ID: S-8857-2018
- Scopus Author ID: 55664571900
- Homepage: https://its.iust.ac.ir/profile/en/hashayanfar

How to cite this paper:

H. Sardari, B. Mozafari, H.A. Shayanfar, “Fast DC offset removal for accurate phasor estimation using half-cycle data window,” J. Electr. Comput. Eng. Innovations, 10(2): 341-350, 2022.

DOI: 10.22061/JECEI.2021.8205.492

URL: https://jecei.sru.ac.ir/article_1644.html

SRTTU

Journal of Electrical and Computer Engineering Innovations (JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Intelligent Transportation System based-on the Whale Algorithm in Internet of Things

Z. Boujarnezhad[1]***, M. Abdollahi***[2,*]

[1]*Department of Computer Engineering, Pooyesh Institute of Higher Education, Qom, Iran.*

[2]*School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran.*

Article Info

Article History:

Received 07 November 2021
Reviewed 17 December 2021
Revised 30 January 2022
Accepted 05 February 2022

Keywords:

Grey wolf optimizer
Whale optimization algorithm
Internet of things
Intelligent transportation system
Artificial intelligence
Harris Hawks optimizer

*Corresponding Author's Email Address:
meisam.abdolahi@ut.ac.ir

Abstract

Background and Objectives: As cities are developing and the population increases significantly, one of the most important challenges for city managers is the urban transportation system. An Intelligent Transportation System (ITS) uses information, communication, and control techniques to assist the transportation system. The ITS includes a large number of traffic sensors that collect high volumes of data to provide information to support and improve traffic management operations. Due to the high traffic volume, the classic methods of traffic control are unable to satisfy the requirements of the variable, and the dynamic nature of traffic. Accordingly, Artificial Intelligence (AI) and the Internet of Things (IoT) meet this demand as a decentralized solution.

Methods: This paper presents an optimal method to find the best route and compare it with the previous methods. The proposed method has three phases. First, the area should be clustered under servicing and, second, the requests will be predicted using the time series neural network. then, the Whale Optimization Algorithm (WOA) will be run to select the best route.

Results: To evaluate the parameters, different scenarios were designed and implemented. The simulation results show that the service time parameter of the proposed method is improved by about 18% and 40% in comparison with the Grey Wolf Optimizer (GWO) and Random Movement methods. Also, the difference between this parameter in the two methods of Harris Hawks Optimizer (HHO) and WOA is about 5% and the HHO has performed better.

Conclusion: The interaction of AI and IoT can lead to solutions to improve ITS and to increase client satisfaction. We use WOA to improve time servicing and throughput. The Simulation results show that this method can be increase satisfaction for clients.

Introduction

Nowadays, cities face complex and great challenges. For example, in smart cities, the conventional planning of transportation is not adequate [1]. In recent years, due to increasing demand for road safety, the ITS has been considered [2]. ITS is called as a system that helps transportation flow by using information, communication, and control techniques. ITS users are the network managers, transportation service providers, passengers, and owners of transportation fleets. Providing some of these services depends on decisions and policy-making about them. This system allows drivers and the Traffic Management Organizations (TMO) to exchange information in real-time. Therefore, road safety and efficiency are now becoming more important challenges.

Doi: 10.22061/JECEI.2022.8184.494

ITS provides solutions for cooperation, and a reliable transportation platform. ITS services can be considered as a data chain that consists of data acquisition, communications, processing, data distribution, utilization of information, and external factors. With the amount of information collected by RFID readers and their exchange between cars and digital interfaces, it is possible to extend the evaluation of transportation systems by actual processed data. This affects the decision-making process, planning, and implementation of public policies and makes them manageable [3]. In these systems, IoT can directly affect. The IoT connects various objects to each other according to a communication protocol through various sensor devices [4]. The main goal of the IoT is to do things faster, to automate all things, and control objects remotely. Also, AI along with the IoT can enhance the ITS quality and thus, improves the traffic situation, reduces congestion, trip delays and the delay of necessary services such as ambulance and police, decrease fuel consumption and ultimately increase the satisfaction of citizens.

Transportation systems are inherently large-scale and complex due to the considerable number of elements and interactions between the elements. Collaborative control is a method to manage multi-agent systems despite the complexity and a large number of elements [5]. On the other hand, recent advances in communication between machines and also communication between cars and transportation infrastructures, cause access to vehicles around and also general information about traffic flow. Access to the information of vehicles lead to emerged control methods such as collaborative control methods or Machine Learning (ML) methods that can be applied to transportation systems. Therefore, ML techniques can help to create an advanced and efficient ITS [6]. ML applies various techniques to facilitate learning in different devices of the network to make them automatic [7].

ITS includes a large number of traffic sensors that collect a high volume of data to provide information to support and improve traffic management operations [8]. Due to the high traffic volume, the classic methods of traffic control are unable to satisfy the requirements of the variable and dynamic nature of the traffic [9]. The main purpose of the traffic control system is to propose efficient management of transportation resources so that changes in traffic conditions are taken into account. AI fulfills the demand as a decentralized solution with the introduction of the concept of smart agent. Using smart agents to automatically sense traffic changes and perform appropriate actions by referring to the knowledge-based and meta-heuristic algorithms, the traffic control system can be effectively managed. This paper intends to minimize the distances taken by vehicles and service time to provide customer services.

In advance [10], we proposed an optimized path selection mechanism by road rescue servers based on a GWO algorithm. The GWO imitates the hunting behavior of grey wolves in nature [11]. In this paper, the WOA is used to optimize the problem of route selection in the transportation system. This research has been done to reduce customer service time. Since the WOA and the GWO are slightly different from each other, to determine the effect of these differences, we have compared the WOA with the GWO in this research. Also, according to the authors of this method, the GWO is one of the algorithms has been discovered in recent years, which has a high rate of convergence in obtaining the answer [11]. In this paper, the WOA is also has compared with the HHO. HHO is a novel population-based, nature-inspired optimization [12]. Using the HHO, the service time is reduced, but the throughput is almost equal to the WOA. According to the parameters of the problem, it seems that the WOA is suitable for this problem. In our future work, with considering the other design parameters such as energy and fuel consumption, applying the HHO algorithm can be a more efficient solution.

In this paper, we investigate the network of vehicles that have different sensors and are connected to the central server and each other via the internet. Firstly, we divide the service area by using the fuzzy clustering algorithm and take into account the cluster centers as the center of crowdedness. We also predict the number of requests using the time series neural network predictor. Finally, we find the optimal path using the whale algorithm.

The contributions of the paper are as follows:

- Using fuzzy clustering to divide the area under service.
- using time series to consider the behavioral patterns of people in the community to request vehicles.
- Applying an intelligent algorithm to find the best path.

Comparing the simulation results with previous methods indicates that the distance has been improved and thus the total time of servicing and increased the satisfaction of service providers.

The rest of the paper is organized as follows: first, the related work is investigated in second Section. In third Section, we will provide the proposed method and the simulation results are discussed in fourth Section. Finally, the conclusion of the paper is presented in fifth Section.

Related Work

Various researches have been conducted on ITS, which in most of them have been used only from one of the technologies of AI or IoT, and quantitative research has dealt with the combination of AI and IoT. In this section, some examples of related work will be described.

Kuppusamy et al. [13] proposed a framework for traffic control and data processing was provided using IoT. This framework includes a local server and a remote server that improve the processing time of the traffic signal. As a result, waiting times for vehicles, air pollution, and overtime are reduced at intersections. Dubey et al. [14] designed a system to control traffic signals so that signal lights are decided based on less wait time and less pollution. This system is designed for IoT applications. In [15] a smart traffic management system using the IoT is presented. A hybrid approach was used to optimize traffic flow on the road. Pyykonen et al. [16] offered a smart traffic control system for IoT applications. The roadside system measures and calculates a series of values that are stored in the database and sent to users through the 802.11 protocol. Bojan et al. [17] offered an IoT-based intelligent transportation system, that consists of three components: display, monitoring, and sensor. Geetha et al. [18] using IoT, provided a system for public transportation. Sutar et al. [19] have presented a framework for intelligent public transportation that deals with determining the position of buses and responding to passenger demand. Datta et al. [20] introduced a framework with data-driven architecture, which contributed to the design of a smart road rescue system in smart cities. Desai et al. [21] provided a vehicle regulatory system through software and hardware for routing and monitoring and supervision of the vehicles and finally, the cost savings were followed. Al-Dweik et al. [22] have presented a roadside unit for utilizing as the portion of a comprehensive ITS based on IoT. First, the information is collected by sensors and cameras. It is then sent to a central server for operations such as setting speed limits and issuing weather warnings. Jalaney et al. [23] reviewed the IoT-based architectures for intelligent public transport. Zhu et al. [24] discussed the role of IoT in creating parallel transportation systems and examining the impact of the system in several Chinese cities. Qureshi et al. [25] used various types of smart transport system applications and their technologies. Murad et al. [26] have proposed an integrated system IoT-based. This system simplifies the provision of information such as bus scheduling and online payment. Thakur et al. [27] investigated IoT-based solutions in the intelligent transportation system. Also in this research, the road safety techniques, communications between vehicles, and wireless communication techniques suitable for channels were studied. Sodhro et al. [28] proposed a QoS-aware algorithm to support multimedia transmission. Second, they proposed a novel QoS optimization scheme. Third, they proposed several QoS metrics to analyze the performance of V2V networks. Sodhro et al. [29] first, developed a system model for reliability and optimization of connection in the intelligent transportation system, and a SSLO algorithm. then, they proposed a reliability framework. Sodhro et al. [30] proposed 5G-based self-adaptive green and novel 5G-driven algorithms and, a reliable framework.

In recent investigations, the IoT has been used to improve the ITS. As follows, we review the researches in which AI has played the most important role. Odeh in [31] applied Genetic Algorithm (GA) to manage traffic signals. In this research, a video system was used to collect information, and a decision-making system based on the GA was applied. Using video images, the authors achieved the number of vehicles and ultimately optimized the time of green lights by applying GA. The comparison of the real and simulated results showed about a 40% decrease in the lights lag. Li et al. [32] have dealt with the optimization of traffic signals in smart cities by using the GA. In this research, a bi-level optimization framework was provided. The high-level problem reduces the travel time of the drivers, and the low-level problem, using the computational at the top level, helps to balance the network. If the lights are properly designed, it reduces the travel time of the drivers, traffic control, and congestion reduction which reduces environmental concerns. Zhou et al. [33] developed a signal timing system based on multi-objective optimization. The results of the implementation are: reducing the number of stops and latency and thus improving traffic. As previously discussed, AI can realize the interactive performance of information between objects and people, for example, help to expand intelligent transportation. In the following researches, the combination of these two technologies has been used.

Osuwa et al. [34] used AI including fuzzy logic and neural networks in IoT. Hamidouche et al. [35] applied the Grey Wolf and Whale algorithms, to exchange data on a heterogeneous wireless sensor network, taking into account the buffer overflow problem. Yadav et al. [36] proposed a traffic signal management system, through GA and the IoT. The purpose of designing this system is to reduce the waiting time of vehicles in traffic signals. It should be noted that a neural network has been trained to allocate green light time to each road. Liu et al. [37] designed a system for traffic emergency response by using IoT and data mining.

In the reviewed researches, a framework or architecture for traffic control using the IoT or AI was often presented. Less attention has been paid to emergency services and the use of AI in the IoT and ITS. Therefore, in this paper, using a combination of AI and IoT, we improve a routing algorithm and thus reduce the time to provide services to customers. The traditional ant colony algorithm has been often used for routing while we apply the Whale algorithm in this research. Due to the mechanisms that the whale algorithm has in prey encircling, the problem of transport optimization with this algorithm is investigated in this paper.

Proposed Method

The proposed method consists of three steps as follows:

1. Segmentation of the area under service.
2. Predicting the number of requests of each segment.
3. Finding the best path.

Fig. 1 shows the flowchart of the proposed method. The proposed method includes the number of vehicles in one or more deployment locations that must be referred to a set of customers and provide the service. These clients have a certain movement pattern. In fact, by predicting the number of transportation requests in each segment of the smart grid, we intend to move vehicles in such a way that the total distance traveled, the total travel time, and the number of required vehicles is minimized. At the same time, customer satisfaction is intended to be maximized. Network traffic is modeled through a set of links and nodes. For example, a simple traffic diagram is shown in Fig. 2.

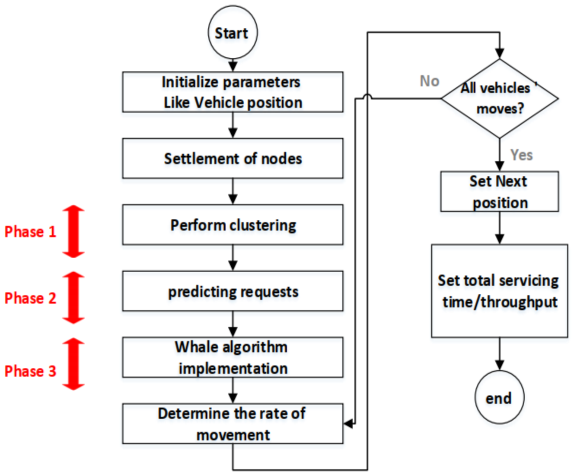

Start
Initialize parameters Like Vehicle position
Settlement of nodes
Perform clustering
predicting requests
Whale algorithm implementation
Determine the rate of movement
Phase 1
Phase 2
Phase 3
No
All vehicles moves?
Yes
Set Next position
Set total servicing time/throughput
end

Fig. 1: The flowchart of the proposed method.

In the proposed method, the service area is divided into blocks with variable sizes. The location of sending requests is calculated using the random function and will be inserted as a service location in a two-dimensional coordinate system and, we have tried to cluster these points. We consider the center of clusters as points with the request for high transport servicing, and it is called as a traffic point.

Fig. 2: A simple traffic diagram [38].

In the first step, Fuzzy C-Means (FCM) clustering is applied. In this method, each data belongs to a specific degree of each cluster and according to the degree of belonging, the presence of data to a cluster is determined [39]. This method is based on the C-Means function, which is a well-known unsupervised clustering algorithm and is successfully used to solve different clustering problems. This method is used in partitioning the network into some clusters [40]. This algorithm can divide the space of nodes into *K* clusters according to the distance between the cluster head and other nodes. This algorithm minimizes the objective function mentioned in the following equation which is a square error function.

$$J = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \,.\, d_{ik}^{2} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \,.\, \|x_k - v_i\|^2 \tag{1}$$

In (1), the exponent *m* is used to adjust the weighting effect of membership values. Large *m* will increase the fuzziness of the function. *m* is a real number *m>1* which is selected in most cases as *m=2*. X_k represents the K_{th} sample and V_i stands for the center of the i_{th} cluster. U_{ik} variable shows the amount of sample belongs to the *i* sample in cluster *k*. Mark ||*|| shows the similarity rate of the instance form the center of the cluster. Based on U_{ik} , a U-matrix can be defined that has *c* rows and *n* columns, and its components can take a value between 0 to 1.

In Fig. 3, the result of fuzzy clustering is shown in a two-dimensional environment. Colored stars, red squares and green squares indicate user requests, current requests, and server vehicles, respectively.

The second step of the proposed method is to consider the behavior patterns of community people to apply for vehicles in determining the optimal routing, which is used to measure the number of requests in each block of the intelligent transportation network. In this method, the traffic is computed and controlled by diffusion.

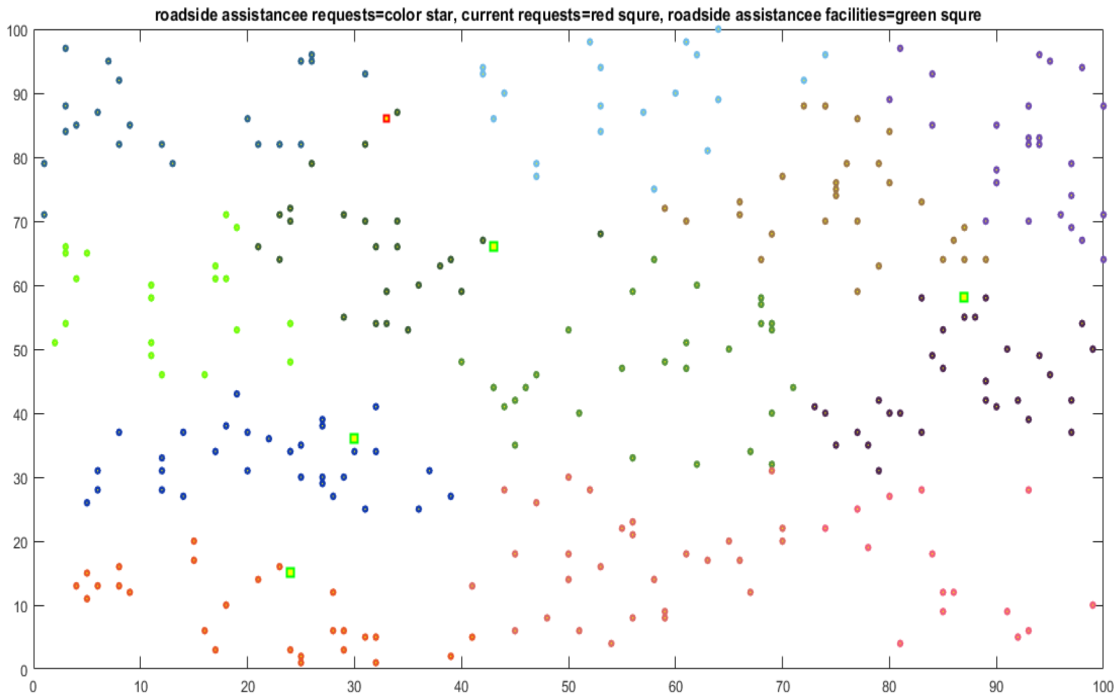


Fig. 3: Form clusters based on the location of the request in user points.

That is, we use the smart agents and network nodes to measure the number of future requests per block. We will use the time series neural network to predict the number of requests [41], [42]. In each science, the collected statistics are called time series, which are to be predicted and available in past periods. A time series of statistical data is collected at equal and regular intervals.

Neural networks can be used in various cases such as data storage and review, a general mapping of an input set into an output set, grouping, and classification of similar data, optimization, and prediction. Neural networks, such as regression, are tools for an approximation of functions and finding a relationship between independent and dependent variables. Neural networks are composed of units called Neuron, which is processing unit. The Neurons produce a value of output through the defined relationships between them and their weights. In the last step, we intend to determine the best movement direction of vehicles using the whale algorithm. So that it is closer to high traffic areas and can service more users in less time and with less distance. Therefore, we have provided the fitness function in which the parameters of moving vehicles and servicing rate, or the density of service demand in the scroll blocks have been included. The amount of service demand density was obtained based on the predicted amount of the time series neural network. Thus, it will be tried to move the vehicles to help blocks with greater demand density. According to the above-mentioned materials, the fitness function in the proposed method is defined as the following, which is tried to be maximized.

$$Fiteness(S_i) = \frac{\sum_{J\in S_i} SERV(J)^{\alpha}}{sumDis(S_i)^{\beta}} \tag{2}$$

Equation (3) calculates the distance between two points, and (4) shows the relationship between time and distance. The distance and time parameters are directly related. The Speed parameter is considered constant in this equation.

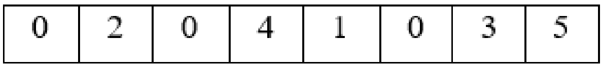
$$Dis = \sqrt{(X-x)^2 + (Y-y)^2} \tag{3}$$

$$t = \frac{Dis}{s} \tag{4}$$

In (3), *X, Y* represent the location parameter (X, Y) at the destination, and *x, y* represent the location parameter (X, Y) at the source. In (4), *t* represents the service time and *s* represents speed.

S_i represents a solution as a sequence of blocks met as a route moving vehicles, $SERV(J)$ is the rate of service in block *j* and the variable $sumDis(S_i)$ is the sum of the distances of the vehicles, α and β coefficients are the most important parameters for determining the level of fitting the solution. In the current method, each solution represents a moving block trail of the vehicles. In other words, the content of each home is equal to a block number that has passed or crosses an existing vehicle from that block. Therefore, the length of the solution is equal to the number of blocks in non-zero houses; the number of block meeting order is inserted by vehicles.

In Fig. 4, first, service to users in block 5 will then move in 2,7,4,8 blocks, respectively. In the following, how to determine the optimal solution in the whale algorithm will be described.



0	2	0	4	1	0	3	5

Fig. 4: A solution in Whale algorithm with 8 blocks.

Also, we can consider an array as a solution. For example, block 1 is serviced by vehicle (server) 1, block 2 is served by vehicle 2, and so on as depicted in Fig. 5.

Block Number	*1*	*2*	*3*	*4*
Service Number	*1*	*2*	*3*	*4*

Fig. 5: an array as a solution.

Fig. 6 shows how three users are served by two servers. The squares represent the servers and the circles represent the clients. Fig. 6(a) shows the initial state of placement of requests and servers. After clustering and executing the algorithm, the result is shown in the following figures. Fig. 6(b) shows the current request as a grey circle. Server 1 responds to the first request. Fig. 6(c) the second request is served by Server 2 (The answered Requests are displayed as black circles), and then in Fig. 6(d), the third request is served by Server 2.

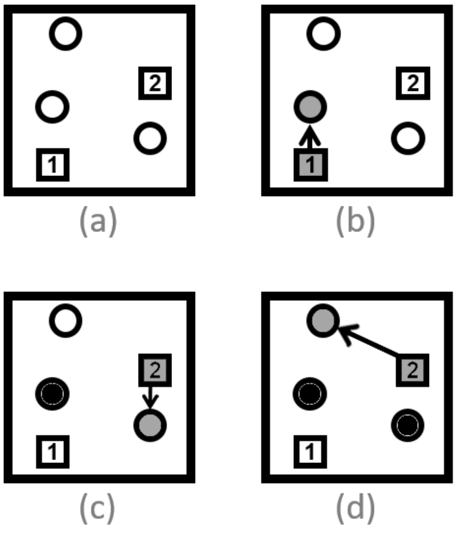


Fig. 6: Execution of whale algorithm for three clients and two servers.

The whale algorithm is a meta-heuristic optimization algorithm nature-inspired [43] by the hunting strategy in whales. The main difference between this algorithm and the GWO is simulated hunting behavior by using random techniques or the best search agent to chase prey and utilize a spiral to mimic the bubble-net attacking strategy of a whale. Randomization plays a very crucial role in exploration and exploitation. In each iteration, the number of search agents is generated and the optimal answer is selected. Mathematical modeling of the WOA divides into three phases of prey encircling, bubble network attacking method and, hunting search (exploration phase) that is provided in (5) to (10) [43].

In the prey encircling, search agents try to update their location towards the best search agent after each iteration. This behavior of whales is defined according to the following equations.

$$\vec{D} = \left|\vec{C}.\overrightarrow{X^*}(t) - \vec{X}(t)\right| \tag{5}$$

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A}.\vec{D} \tag{6}$$

In these equations, t denotes the current iteration while $\vec{A}$ and C are the coefficient vectors. $\overrightarrow{X^*}(t)$ is the position of the best and $\vec{X}(t)$ vector is the reference position. If there are better answers, $\overrightarrow{X^*}(t)$ it will be updated. Vectors $\vec{A}$ and $\vec{C}$ are calculated according to the following equations:

$$\vec{A} = 2\,\vec{a}.\vec{r} - \vec{a} \tag{7}$$

$$\vec{C} = 2.\vec{r} \tag{8}$$

where is $\vec{r}$ a random vector with the value in the range [0,1]. The values of $\vec{a}$ are reduced linearly from 2 to 0.

In the attack phase, two methods for modeling the behavior of the Whale Bubble Network are presented:

1) Shrinking encircling mechanism: This method is used to reduce the value of $\vec{a}$. According to (7) with a decrease in the amount of $\vec{a}$, the value of $\vec{A}$ is also reduced and, the value is placed in a range of [-a, a] that is reduced linearly by $\vec{a}$ from 2 to 0.
2) Spiral updating position: this method calculates the distance of the whale, which is in the position *(X, Y)* to make the hunt, which is the position *(X*, Y*).* Then the spiral path is formed between the position of the whale and the prey. The spiral update equation is given in the following equations:

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.cos(2\pi l) + \overrightarrow{X^*}(t) \tag{9}$$

$$\overrightarrow{D'} = \left|\overrightarrow{X^*}(t) - \vec{X}(t)\right| \tag{10}$$

It should be noted that whales swim around the prey in a spiral-shaped path simultaneously. To model this simultaneous behavior, we assume that there is a probability of 50% to choose between either the shrinking encircling or the spiral model. If *p<0.5*, the position is updated based on (6), and if *p>=0.5* the (9). *p* is randomly selected between 0 and 1. The value of p can determine the type of movement in the whale algorithm.

The third phase is the hunting search. This method uses the way to change $\vec{A}$ to search for hunting. Each of the whales is randomly searched for the position. Therefore, for a vector $\vec{A}$, we randomly assign values greater than 1 or less than -1 so that the search agent searches for positions farther from the reference whale. Unlike the exploitation phase, in this step, the search agent's position is randomly updated. Exploration follows two conditions. Mathematical equations related to this behavior of whales are given below:

$$\vec{D} = \left| \vec{C} \, . \, \overrightarrow{X_{rand}} - \vec{X} \right| \qquad (11)$$

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \, . \vec{D} \qquad (12)$$

where $\overrightarrow{X_{rand}}$ is a random position vector chosen from the current population. Finally, follows these conditions [44]:

- $|\vec{A}| \geq 1$ enforces exploration to WOA. This prevents local optimization to find the global optimal.
- $\overrightarrow{|A|} < 1$ for updating the position of current search agent/best solution is selected.

From the theoretical point, the whale algorithm can solve the optimization problems in different types, which is due to gradient-free mechanism, flexibility, and high local optima avoidance in this algorithm [45].

Results and Discussion

We have used MATLAB software to simulate the proposed method. To service customers' requests in the smart network of transportation, the number 500 transportation requests from 100 clients are collected in different parts of the area under monitoring and entered into a database.

The area is divided into blocks based on the frequency of past requests. We need to move four vehicles between these blocks (In scenario 4, the number of vehicles can be variable) so that the total traveled distance and the total travel time are minimized and the number of services provided per unit of time (throughput) is maximized.

This is shown as an example in Table 1. It should be noted that one of the sensors used in this research is the location sensor of vehicles.

Table 1: An example for throughput

Method Type	Round (min)	Distance (m)	Service Time(min)	Throughput
(1)	100	50	25	4
(2)	100	30	15	6

The System specifications to implement the proposed method, parameters of the simulation, and the meta-heuristic algorithm are shown in Table 2, Table 3, and Table 4, respectively.

Table 2: System specifications to implement the proposed method

Specifications	Value
CPU	Core i5
RAM	4 G
OS	Windows 10

Table 3: Simulation parameters

Simulation Parameters	Value
Number of simulation execution	5
Number of clients	100
Number of servers (Vehicle)	2,4,8,10,12,25
Environmental dimensions	100*100m
Number of transportation requests	100,200,300,400,500
Number of clusters	2,4,6,8,10,12,24,26
Request time period	50 minutes

Table 4: Meta-heuristic algorithms' parameters

Optimization Parameters	Value
Maximum number of repetitions	30
Primary population	30
Size of a solution	Number of blocks
The range of values allowed for a solution	Meeting Priority of Blocks

To evaluate the impact of different conditions, including the number of clusters, the number of vehicles, and the number of requests, we have defined different scenarios. In scenarios 1 and 2, the number of clusters is different and we have placed them in an array. In scenarios 3, 4 and, 5, we have considered the number of clusters to be constant and equal to 8.

The first scenario: Investigation of evaluation parameters during simulation time. To investigate the optimality of vehicle movement in the proposed (Whale) method, we have performed the simulation for 100 minutes and compared the performance of the three methods in terms of servicing time in fulfilling customers' transportation requests over time. The result is shown in Fig. 7. In the obtained results, we see that the total servicing time fluctuates so much that it is not possible to comment on the superiority of one method over another. Therefore, the cumulative value was calculated for this parameter. As it can be figured out in Fig. 8, the total servicing time in the proposed (Whale) method is less than the methods of Random Movement and GWO about 48% and 29%, respectively, and more than HHO about 8%. To evaluate the efficiency of the proposed method in the effective management of the transportation system, we have compared the throughput of the method in providing service to customers during the simulation run. In Fig. 9, we see that the number of services provided per unit time in the proposed (Whale) method is more than the methods of Random Movement and GWO about 67% and 1%, respectively, and but it is almost equal to HHO. Therefore, the proposed (Whale) method, using the spiral mechanism in simulating the bubble network attack, has been able to reduce the servicing time and increase throughput by selecting the optimal path of vehicles in the blocks resulting from the clustering of crowded points. The noteworthy point in all scenarios is that the difference in parameters in compared meta-heuristic methods is small. Population-based meta-heuristic optimization algorithms have one thing in common, regardless of their nature. The search process is divided into two phases: exploration and exploitation [43]. To prove their strength, these algorithms must focus on these two phases and strike a good balance between them.

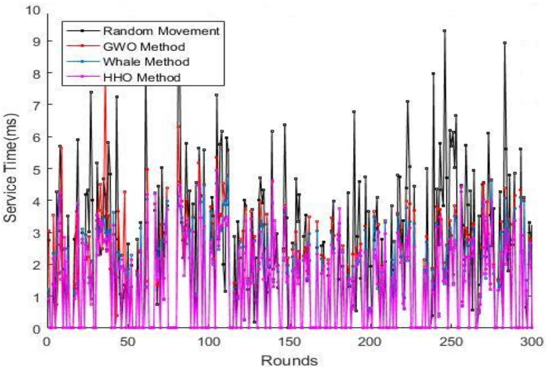


Fig. 7: The performance of Random Movement, GWO, Whale and HHO methods in terms of total service time in first scenario.

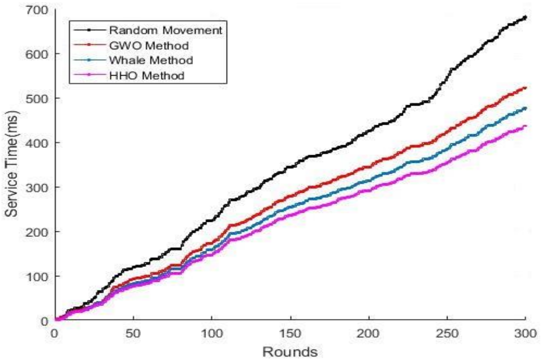


Fig. 8: The performance of Random Movement, GWO, Whale and HHO methods in terms of cumulative total service time in first scenario.

The main difference between the algorithms studied in this research is in the besiege of prey and its attack. In comparison to GWO and WOA, one of the most important factors of difference is the bubble attack in the whale method. Comparing HHO and WOA, different besiege mechanisms make the difference between the two methods. The HHO uses a series of search strategies based on prey energy and prey escape probability, and then selects the best move. Also in this method, the strength of random jumping helps to balance the phases of exploration and exploitation. Therefore, it can be said that these cases help to improve the parameters of the problem. As mentioned before, according to the conditions and parameters of the problem, the WOA is more suitable for this problem and the same throughput in these two methods confirms this.

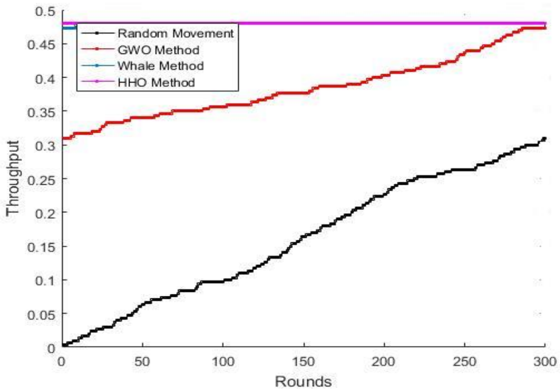


Fig. 9: Throughput comparison of Random Movement, GWO, Whale and HHO methods in first scenario.

The second scenario: Investigation of performance evaluation parameters with different clusters. In this scenario, we will investigate the performance of the proposed (Whale) method with some different cluster sizes. In Fig. 10, it can be concluded that the service time in the proposed (Whale) method is less than the methods of Random Movement and GWO about 35% and 18%, respectively, and more than HHO about 6%. Fig. 11 also specifies that the throughput of the model in providing customer service in the proposed (Whale) method is improved than Random Movement and GWO about 69% and 8%, respectively but it is almost equal to the HHO.

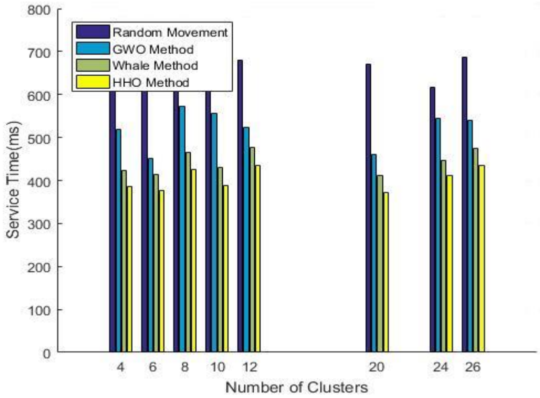


Fig. 10: The performance of Random Movement, GWO, Whale and HHO methods in terms of total service time in second scenario.

Since in this scenario four vehicles are considered to provide the service, increasing the clusters does not have much effect on the servicing time. The reason for the superiority of the Whale method towards the

Random Movement, the clever selection of path and, the GWO method, is the use of a spiral mechanism that reduces distance and ultimately reduces the service time. It should be noted that in all methods and scenarios, constant speed is considered.

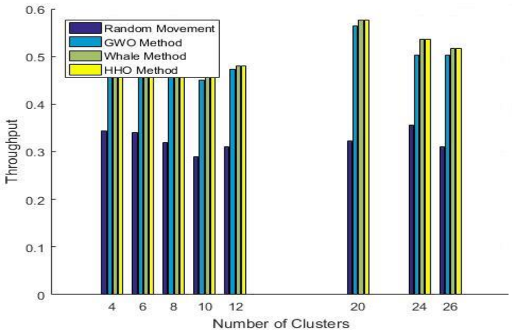


Fig. 11: Throughput comparison of Random Movement, GWO, Whale and HHO methods in second scenario.

The third scenario: Investigation of performance evaluation parameters with the number of different requests. In this scenario, we investigated the performance of the proposed (Whale) method with some different service requests (100 to 500 requests). As shown in Fig. 12 the total service time, despite the number of different requests in the proposed (Whale) method, is always less than Random Movement and GWO about 38% and 17%, respectively, and more than HHO about 5%. Also, as shown in Fig. 13, the method's throughput is increased compared to the Random Movement method of 84% and compared with the GWO method of 6%.

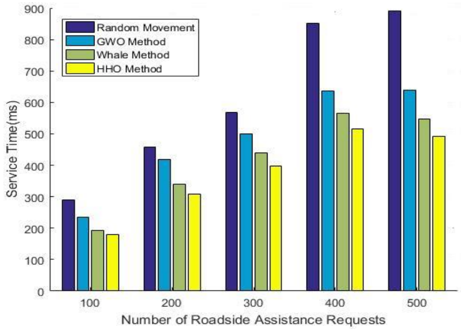


Fig. 12: The performance of Random Movement, GWO, Whale and HHO methods in terms of total service time in third scenario.

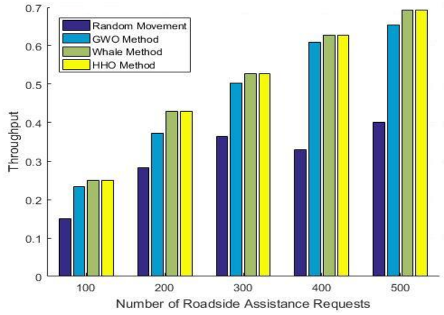


Fig. 13: Throughput comparison of Random Movement, GWO, Whale and HHO methods in third scenario.

This means that the proposed (Whale) method increases the efficiency of resources, due to the use of intelligent methods and the application of the spiral mechanism in the attack phase of the whale algorithm, and it can respond to the more number of requests in little time. The throughput in HHO and WHO is almost equal.

The fourth scenario: Investigation of performance evaluation parameters with different Vehicles (server). In this scenario, we investigated the performance of the proposed (Whale) method with different vehicles (2, 4, 8, 10, 12, 25). As shown in Fig. 14 and Fig. 15 the service time is lower than in previous methods despite the number of different vehicles in the proposed (Whale) method. However, when vehicles increase, the time of service of the Random Movement method decreases significantly, but the proposed (Whale) method still has better performance.

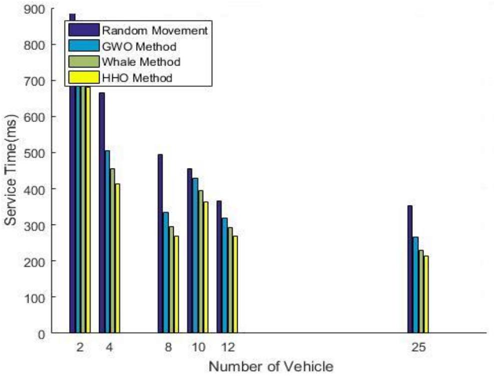


Fig. 14: The performance of Random Movement, GWO, Whale and HHO methods in terms of total service time in fourth scenario.

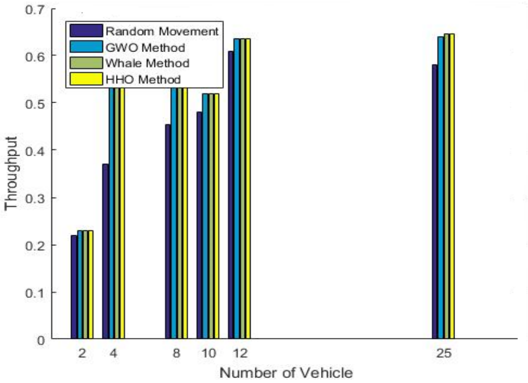


Fig. 15: Throughput comparison of Random Movement, GWO, Whale and HHO methods in fourth scenario.

Service time in the proposed method has been reduced compared to the Random Movement method by 23% and the GWO method by 11% and increased compared to the HHO by about 4%. The throughput parameter is not very different in the four methods when service vehicles increase but on average, this parameter increased by 20% compared to the Random Movement method and increased by 1% compared to the GWO method and it is almost equal to the HHO method. It can be said that the reason for improving parameters in this method, like in previous scenarios, is using a spiral mechanism in the attack phase of the Whale algorithm.

The fifth scenario: Investigation of performance evaluation parameters with different values the importance of fitness function variables. As stated before in the proposed scheme, using the whale optimization algorithm, the near-optimum route of vehicles is determined so that it can serve more users in less time and over a shorter distance. Hence, a fitting function is used which includes two indicators of the total traveled distances by vehicles and the density of demanded services in the tracked blocks. To determine the importance of each of these two parameters in determining the degree of the fitness of the solution, we used the values of α and β as the exponent values somehow the sum of them is equal to one (i.e., α+β=1).

In the last step, we evaluated the proposed intelligent transport system based on different exponent values of the coefficients. Fig. 16 shows that in our proposed method, despite the different values of the two effective indicators, the total service time is 35% less than the Random Movement method, 12% less than the GWO method, and 5% more than the HHO method. Since the whale algorithm uses a bubble method to attack, it reduces the distance and thus the service time. Fig. 17 also specifies that the throughput of the model in the proposed (Whale) method is more than Random Movement and GWO about 82% and 5%, respectively, and almost equal to the HHO method.

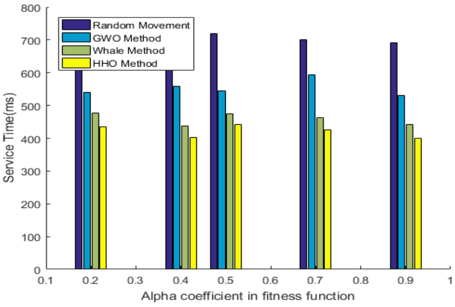


Fig. 16: The performance of Random Movement, GWO, Whale and HHO methods in terms of total service time in fifth scenario.

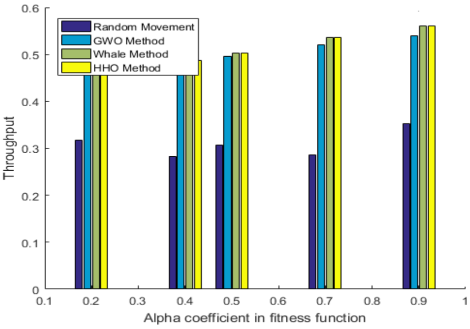


Fig. 17: Throughput comparison of Random Movement, GWO, Whale and HHO methods in fifth scenario.

Also, after performing several round of simulations, we found that the best answers are obtained when α=0.9 and β=0.1. Therefore, since these two exponent values show the effect of the parameters of the fitness function, the effect of the demanded service density is more than the sum of the travelled distances.

Conclusion

ITS is a system that uses information, communication, and control techniques to assist the flow of transportation. ITS tools have three basic and pivotal features: information, communication, combination, and cohesion, which these three characteristics help transportation and passengers to make better and more coordinated decisions. In this paper, we used the combination of AI and IoT to improve the ITS. The results of this research show that the interaction of AI and IoT can lead to solutions to improve ITS and increase client satisfaction. In this paper, we have used the FCM to block the area under servicing, the time series neural network to predict requests, and the whale algorithm to find the best path. The proposed method was evaluated under different conditions including the number of variable requests, the different times, the number of different vehicles, and the number of different clusters. The simulation results show that the throughput is increased by 5%, compared to the method that the grey wolf was used to optimize and 82%, compared to a randomly selected path. Throughput in HHO and Whale method is almost equal. In our future work, we intend to achieve an optimal vehicle movement plan in smart cities by considering other environmental characteristics and using new meta-heuristic methods, so that the convergence time will be reduced to the optimal response. Also, by applying a new fitness function, we will determine the number of optimal vehicles to service users' requests. In this method, we will be able to move only some vehicles at some point in time and others will remain in place to reduce the cost of service.

Author Contributions

Z. Boujarnezhad designed the experiments, collected data, and carried out the data analysis. Z. Boujarnezhad and M. Abdollahi interpreted the results and wrote the manuscript.

Acknowledgment

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

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

α	Level of significance
β	Observed value
AI	Artificial Intelligence
A	Coefficient vector
C	Coefficient vector
c	Number of rows
DLS	Damped Least-Squares
FCM	Fuzzy C-Means
GWO	Grey Wolf Optimizer
HHO	Harris Hawks Optimizer
IoT	Internet of Things
ITS	Intelligent Transportation System
K	Number of clusters
ML	Machine Learning
r	Random vector
S_i	a solution as a sequence of blocks met as a route moving vehicles
s	speed
t	Service time
TMO	Traffic Management Organizations
U_{ik}	shows the amount of sample belongs to the *i* sample in cluster *k*
V_i	stands for the center of the i_{th} cluster
WOA	Whale Optimization Algorithm
X_k	Represents the K_{th} sample
$\overrightarrow{X}(t)$	The position of the best
$\overrightarrow{X^*}(t)$	Reference position
$\overrightarrow{X_{rand}}$	Random position
(X,Y)	Whale position
(X,Y*)*	Hunt position

References

[1] K. Iqbal, M.A. Khan, S. Abbas, Z. Hasan, A. Fatima, “Intelligent transportation system (ITS) for smart-cities using Mamdani Fuzzy Inference System,” Int. J. Adv. Comput. Sci. Appl., 9(2): 94–105, 2018.

[2] A. Boukerche, Y. Tao, P. Sun, “Artificial intelligence-based vehicular traffic flow prediction methods for supporting intelligent transportation systems,” Comput. Networks, 182: 107484, 2020.

[3] R.M. Cardoso, N. Mastelari, M.F. Bassora, “Internet of things architecture in the context of intelligent transportation systems—a case study towards a web-based application deployment,” in Proc. 22nd International Congress of Mechanical Engineering (COBEM 2013): 7751–7760, 2013.

[4] C. Chen, H. Liu, Z. Wang, “Analysis and design of urban traffic congestion in urban intelligent transportation system based on big data and internet of Things,” in Proc. 2019 International Conference on Artificial Intelligence and Computer Science: 659–665, 2019.

[5] Z. Deng, T. Zhang, D. Liu, X. Jing, Z. Li, “A high-precision collaborative control algorithm for multi-agent system based on enhanced depth image fusion positioning,” IEEE Access, 8: 34842–34853, 2020.

[6] A.A. Brincat, F. Pacifici, S. Martinaglia, F. Mazzola, “The internet of things for intelligent transportation systems in real smart cities scenarios,” in Proc. 2019 IEEE 5th World Forum on Internet of Things (WF-IoT): 128–132, 2019.

[7] A. Ghosh, D. Chakraborty, A. Law, “Artificial intelligence in Internet of Things,” CAAI Trans. Intell. Technol., 3(4): 208–218, 2018.

[8] J. Guerrero-Ibáñez, S. Zeadally, J. Contreras-Castillo, “Sensor technologies for intelligent transportation systems,” Sensors, 18(4): 1212, 2018.

[9] K. Nellore, G.P. Hancke, “A survey on urban traffic management system using wireless sensor networks,” Sensors, 16(2): 157, 2016.

[10] Z. Boujarnezhad, M. Abdollahi, “Optimization of roadside assistance transportation based on the gray wolf algorithm in the internet of Things (in Persian),” in Proc. 28th Iran. Conf. Electr. Eng. (ICEE 2020): 683–689, 2020.

[11] S. Mirjalili, S.M. Mirjalili, A. Lewis, “Grey wolf optimizer,” Adv. Eng. Softw., 69: 46–61, 2014.

[12] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, “Harris hawks optimization: Algorithm and applications,” Futur. Gener. Comput. Syst., 97: 849–872, 2019.

[13] P. Kuppusamy, R. Kalpana, P.V. Venkateswara Rao, “Optimized traffic control and data processing using IoT,” Cluster Comput., 22: 2169–2178, 2019.

[14] A. Dubey, M. Lakhani, S. Dave, J.J. Patoliya, “Internet of Things based adaptive traffic management system as a part of Intelligent Transportation System (ITS),” in Oroc. 2017 International Conference on Soft Computing and its Engineering Applications (icSoftComp): 1–6, 2017.

[15] S. Javaid, A. Sufian, S. Pervaiz, M. Tanveer, “Smart traffic management system using Internet of Things,” in Proc. 2018 20th International Conference on Advanced Communication Technology (ICACT): 393–398, 2018.

[16] P. Pyykönen, J. Laitinen, J. Viitanen, P. Eloranta, T. Korhonen, “IoT for intelligent traffic system,” in Proc. 2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP): 175–179, 2013.

[17] T. Bojan, U. Kumar, V. Bojan, “An internet of things based intelligent transportation system,” in Proc. 2014 IEEE International Conference on Vehicular Electronics and Safety (ICVES 2014): 174–179, 2014.

[18] S. Geetha, D. Cicilia, “IoT enabled intelligent bus transportation system,” in 2017 2nd International Conference on Communication and Electronics Systems (ICCES): 7–11, 2017.

[19] S.H. Sutar, R. Koul, R. Suryavanshi, “Integration of smart phone and IOT for development of smart public transportation system,” in Proc. 2016 International Conference on Internet of Things and Applications (IOTA): 73–78, 2016.

[20] S.K. Datta, R.P.F. Da Costa, J. Harri, C. Bonnet, “Integrating connected vehicles in Internet of Things ecosystems: Challenges and solutions,” in Proc. 17th Int. Symp. a World Wireless, Mob. Multimed. Networks (WoWMoM 2016), 2016.

[21] M. Desai, A. Phadke, “Internet of Things based vehicle monitoring system,” IFIP Int. Conf. Wirel. Opt. Commun. Networks (WOCN): 1–3, 2017.

[22] A. Al-Dweik, R. Muresan, M. Mayhew, M. Lieberman, “IoT-based multifunctional scalable real-time enhanced road side unit for intelligent transportation systems,” in Proc. 2017 IEEE 30th Canadian conference on electrical and computer engineering (CCECE): pp. 1–6, 2017.

[23] J. Jalaney, R.S. Ganesh, “Review on IoT based architecture for smart public transport system,” Int. J. Appl. Eng. Res., 14(2): 466–471, 2019.

[24] F. Zhu, Y. Lv, Y. Chen, X. Wang, G. Xiong, F.-Y. Wang, “Parallel transportation systems: toward IoT-enabled smart urban traffic control and management,” IEEE Trans. Intell. Transp. Syst., 2019.

[25] K.N. Qureshi, A.H. Abdullah, “A survey on intelligent transportation systems,” Middle East J. Sci. Res., 15(5): 629–642, 2013.

[26] D.F. Murad, B.S. Abbas, A. Trisetyarso, W. Suparta, C.H. Kang, “Development of smart public transportation system in Jakarta city based on integrated IoT platform,” in Proc. 2018 Int. Conf. Inf. Commun. Technol. (ICOIACT 2018): 872–877, 2018.

[27] A. Thakur, R. Malekian, D.C. Bogatinoska, “Internet of Things based solutions for road safety and traffic management in intelligent transportation systems,” Commun. Comput. Inf. Sci., 778: 47–56, 2017.

[28] A.H. Sodhro et al., “Quality of service optimization in an iot-driven intelligent transportation system,” IEEE Wirel. Commun., 26(6): 10–17, 2019.

[29] A.H. Sodhro, J.J.P.C. Rodrigues, S. Pirbhulal, N. Zahid, A.R.L. de Macedo, V.H.C. de Albuquerque, “Link optimization in software defined IoV driven autonomous transportation system,” IEEE Trans. Intell. Transp. Syst., 22(6): 3511-3520, 2020.

[30] A.H. Sodhro et al., “Towards 5G-Enabled self adaptive green and reliable communication in intelligent transportation system,” IEEE Trans. Intell. Transp. Syst., 99: 1–9, 2020.

[31] S.M. Odeh, “Management of an intelligent traffic light system by using genetic algorithm,” J. Image Graph., 1(2): 90–93, 2013.

[32] Z. Li, M. Shahidehpour, S. Bahramirad, A. Khodaei, “Optimizing traffic signal settings in smart cities,” IEEE Trans. Smart Grid, 8(5): 2382–2393, 2016.

[33] P. Zhou, Z. Fang, H. Dong, J. Liu, S. Pan, “Data analysis with multi-objective optimization algorithm: A study in smart traffic signal system,” in 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), 307–310, 2017.

[34] A.A. Osuwa, E.B. Ekhoragbon, L.T. Fat, “Application of artificial intelligence in Internet of Things,” in Proc. 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN): 169–173, 2017.

[35] R. Hamidouche, Z. Aliouat, A.A.A. Ari, M. Gueroui, “An efficient clustering strategy avoiding buffer overflow in IoT sensors: a bio-inspired based approach,” IEEE Access, 7: 156733–156751, 2019.

[36] R.K. Yadav, R. Jain, S. Yadav, S. Bansal, “Dynamic traffic management system using neural network based iot system,” in Proc. the International Conference on Intelligent Computing and Control Systems (ICICCS 2020): 521–526, 2020.

[37] Z. Liu, C. Wang, “Design of traffic emergency response system based on Internet of Things and data mining in emergencies,” IEEE Access, 7: 113950–113962, 2019.

[38] M. Elhenawy, A.A. Elbery, A.A. Hassan, H.A. Rakha, “An intersection game-theory-based traffic control algorithm in a connected vehicle environment,” in Proc. IEEE Conf. Intell. Transp. Syst. (ITSC): 343–347, 2015.

[39] X. Wang, Y. Wang, L. Wang, “Improving fuzzy c-means clustering based on feature-weight learning,” Pattern Recognit. Lett., 25(10): 1123–1132, 2004.

[40] P. Arora, S. Varshney, et al., “Analysis of k-means and k-medoids algorithm for big data,” Procedia Comput. Sci., 78: 507–512, 2016.

[41] M. Rocha, P. Cortez, J. Neves, “Evolution of neural networks for classification and regression,” Neurocomputing, 70,(16–18): 2809–2816, 2007.

[42] F. Raue, W. Byeon, T.M. Breuel, M. Liwicki, “Parallel sequence classification using recurrent neural networks and alignment,” in Proc. Int. Conf. Doc. Anal. Recognition (ICDAR): 581–585, 2015.

[43] S. Mirjalili, A. Lewis, “The whale optimization algorithm,” Adv. Eng. Software, 95: 51–67, 2016.

[44] I.N. Trivedi, J. Pradeep, J. Narottam, K. Arvind, L. Dilip, “Novel adaptive whale optimization algorithm for global optimization,” Indian J. Sci. Technol., 9(38): 319–326, 2016.

[45] I. Aljarah, H. Faris, S. Mirjalili, “Optimizing connection weights in neural networks using the whale optimization algorithm,” Soft Comput., 22(1): 1–15, 2018.

Biographies

Zahra Boujarnezhad received the B.Sc. degree in computer science from Qom University, and the M.Sc. degree in computer engineering from Pooyesh Institute of Higher Education, in 2011 and 2020, respectively. Her research interests include the Internet of Things, meta-heuristic algorithms, and 5G Networks.

- Email: zahrboujar66@gmail.com
- ORCID: 0000-0002-7809-7872
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Meisam Abdollahi received his Ph.D. degree at the School of Electrical and Computer Engineering at the University of Tehran in 2020. He received his B.Sc. and M.Sc. degrees in computer hardware engineering (major computer architecture) from Iran University of Science and Technology and Sharif University of Technology, respectively. He is currently working on reliable challenges on the opto-electrical network-on-chip platforms as the assistant professor in the computer department of Iran University of science and Technology.

- Email: meisam.abdolahi@ut.ac.ir
- ORCID: 0000-0003-0187-6867
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://scholar.google.com/citations?user=IR2UM_sAAAAJ&hl=en&oi=ao

How to cite this paper:

Z. Boujarnezhad, M. Abdollahi, “Intelligent transportation system based-on the whale algorithm in internet of things,” J. Electr. Comput. Eng. Innovations, 10(2): 351-362, 2022.

DOI: 10.22061/JECEI.2022.8184.494

URL: https://jecei.sru.ac.ir/article_1675.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Enhancement of the Photoresponse in the Platinum Silicide Photodetector by a Graphene Layer

A.H. Mehrfar, A. Eslami Majd*

Faculty of Electrical and Computer Engineering, Malek Ashtar University of Technology, Tehran, Iran.

Article Info

Article History:

Received 15 November 2021
Reviewed 21 December 2021
Revised 02 February 2022
Accepted 05 February 2022

Keywords:

Two-dimensional
Near-infrared
Optoelectronic

*Corresponding Author's Email Address:

a_eslamimajd@mut-es.ac.ir

Abstract

Background and Objectives: The use of two-dimensional materials in the photodetector fabrication has received much attention in recent years. Graphene is a two-dimensional material that has been extensively researched to make photodetectors. The responsivity of graphene photodetectors was limited by the low optical absorption in graphene (~2.3% for single layer graphene). Therefore, graphene along with other materials has been used to fabricate a photodetector with the desired properties. The graphene is used for the improvement of the silicide platinum photodetector.
Methods: The platinum silicide photodetector with graphene has been experimentally fabricated and characterized, and all steps of the device fabrication and the characterization are completely provided in addition to required equations for device analysis is completely provided. A graphene layer is transferred on the platinum silicide layer, and the graphene layer creates the photoconductor gain in the platinum silicide photodetector. **Results:** In the proposed device, near-infrared light is detected in the platinum silicide, and by placing a layer of graphene on the platinum silicide, the optical current and responsivity increase compared to the platinum silicide photodetector without graphene. Experimental results show that the optical current, external quantum efficiency, and responsivity increase in the platinum silicide photodetector with graphene. The graphene not only functions as the charge transport channel, but also works as a photoconductor.
Conclusion: The optical current and responsivity are increased by the platinum silicide photodetector with graphene. In our photodetector, the highest responsivity is 120 $\frac{mA}{W}$ in the 1310 nm wavelength, and the optical current is 100 nA at the applied voltage of 8 V. Our photodetector has optical current, responsivity, and external quantum efficiency twice as much as platinum silicide photodetector. Experimental results show the good performance of graphene with platinum silicide photodetector.

Introduction

Graphene that is a two-dimensional material of carbon atoms is very attractive for optoelectronic applications [1]. Graphene has unique properties in mobility, conductivity, optical, and mechanical [1]-[4].

Graphene has a high photodetection potential due to its high speed and broad absorption spectrum [5]-[7]. Graphene can be an infrared photodetector operating at the room temperature. Graphene has potential

Doi: 10.22061/JECEI.2022.8487.517

applications in mid-IR spectroscopy [8]-[10] and biochemical sensing [11].

The responsivity of graphene photodetectors was limited by the low optical absorption in graphene (~2.3% for single layer graphene) and short carrier life time (few picoseconds) [6]-[12]. To compensate the low responsivity of graphene photodetectors, recent research has focused on the enhancement of the responsivity of graphene photodetecors. For example by using metallic plasmonics [13], [14], a waveguide [15], quantum dots [16]-[18], and microcavities [19], responsivity increases in the graphene photodetector. However, there are problems that the photodetector speed decreases with increasing the responsivity; we can refer to the structures that have high responsivity using adsorbent layer or quantum dots, but the speed of the photodetectors is reduced due to the trapping of the carriers in the absorbent material or quantum dots [16], [20]. The graphene photodetector was reported with small metal antenna structures that can be designed to improve both light absorption and photocarrier collection in graphene photodetectors [21]. The Graphene/silicon Schottky photodetector based on internal photoemission effect was reported [22]. There are several problems that limit the practical applications of graphene photodetectors such as low responsivity, the high dark current, and difficult manufacturing steps [22]. On the other hand, Schottky photodetectors based on an internal photoemission effect emit carriers from the metal layer to the semiconductor. These types of photodetectors can be used in different wavelengths. One of the Schottky photodetectors is platinum silicide (PtSi) photodetectors for the photodetection of the infrared wavelength. Platinum silicide photodetectors have many advantages including simple manufacturing process, high stability, high speed, low costs. However, because only a small fraction of the carriers are transferred to the semiconductor, low quantum efficiency was reported for these photodetectors [20]. Therefore, the design and the fabrication of a photodetector that can reduce the problems of graphene photodetectors and platinum silicide photodetectors is very attractive.

In this paper, we have experimentally investigated a silicide platinum photodetector with graphene, and the proposed device is fabricated and characterized. In addition, required equations are provided for the device analysis. A graphene layer is transferred on the platinum silicide layer. The Schottky barrier of silicon and platinum silicide is used as a photodetector by the internal photoemission effect, and the graphene layer is used as a charge transport of excited carriers. By radiating laser light on the proposed device, the electron-hole pair is created in the platinum silicide. By applying an external electrical voltage to the device, one of the carriers travels through silicon and another carrier travels through the graphene, and electrical current is created. The optical current, external quantum efficiency and responsivity are increased by proposed device. Our photodetector has optical current, responsivity, and external quantum efficiency twice as much as platinum silicon photodetector.

Fabrication Process

The fabrication process of the platinum silicide photodetector with graphene is shown in Fig. 1. N-type silicon with 8 to 10 Ω.cm resistivity is used. In addition, for n-type silicon and platinum silicide, the height of the Schottky barrier is approximately 0.84 eV. This Schottky barrier has detection up to a wavelength of 1470 nm. If p-type silicon is used, the height of the schottky barrier is approximately 0.3 eV, and this Schottky barrier has detection up to a wavelength of 4100 nm. P-type silicon has detection in the higher wavelength, but a platinum silicide photodetector with p-type silicon can't operate at the room temperature. Therefore, the choice of the lightly doped n-type silicon is a good choice for a detector operating at the room temperature.

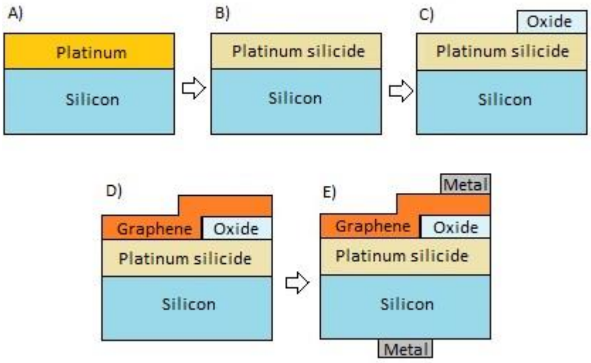


Fig. 1: The fabrication process of the platinum silicide photodetector with graphene.

The silicon surface must be clean. Therefore, the silicon surface is completely cleaned by RCA method. In the next step, a layer of platinum is deposited on the silicon by electron beam physical vapor deposition as shown in Fig. 1(A). After the deposition of the platinum layer, the sample is placed at 450 °C for one hour in the high vacuum (10^{-6} torr).

The formation of platinum silicide is illustrated in Fig. 2. The initial phase of a Pt_2Si layer begins to form at an interface layer, and silicon diffuses into platinum as shown in Fig. 2(B). Over time, the whole layer turns to Pt_2Si. The second phase of a platinum silicide layer is formed at the interface layer as shown in Fig. 2(D). At the end, the whole layer turns to platinum silicide as shown in Fig. 2(E).

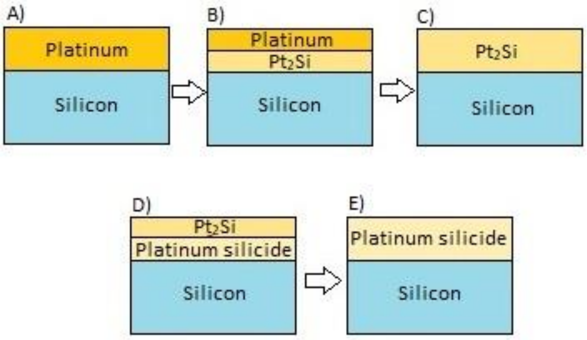


Fig. 2: The fabrication process of the platinum silicide.

Platinum silicide on the substrate is illustrated in Fig. 1(B). After the formation of the platinum silicide layer, the silicon oxide must be deposited on the platinum silicide. Therefore, the silicon oxide is deposited by sputtering with 150 nm thickness. The deposition is done by a shadow mask because the layer can be placed on the desired location as shown in Fig. 1 (C).

The real sample after the deposition of the silicon oxide is illustrated in Fig. 3. The location of silicon oxide and platinum silicide are also illustrated in Fig. 3.

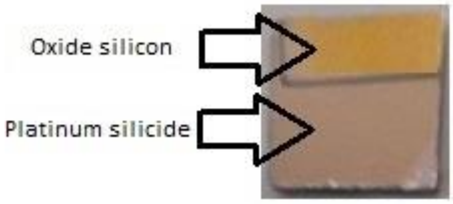


Fig. 3: A fabricated sample image after the deposition of the silicon oxide.

Graphene should be placed on the surface. In this structure, graphene grown by chemical vapor deposition method produced by Graphena Company is used. The graphene grown on copper as a single layer is illustrated in Fig. 4.

Fig. 4: Sample image of graphene on copper.

The graphene layer is a two-dimensional layer and is very thin, and this layer is not visible on a copper [23]. A PMMA layer that is used to protect and transfer is deposited on the graphene layer.

There are several ways to transfer graphene on different substrates. The most important of these methods include wet and dry. Among these methods, the wet method has more desirable properties, and it is more economical [24]. Therefore, a wet method is used to transfer graphene on the platinum silicide substrate. The transfer steps of the graphene by the wet method are illustrated in Fig. 5.

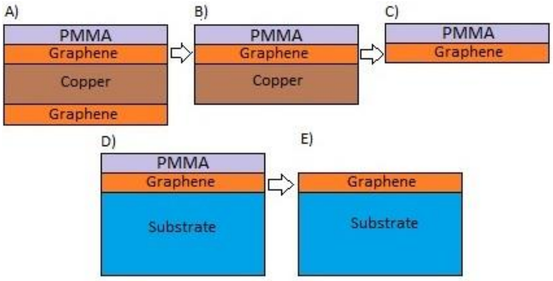


Fig. 5: Graphene transfer steps on the substrate.

The graphene on the copper from the side view is illustrated in Fig. 5(A).

There is a PMMA layer on the graphene, and there is graphene on the copper and under the copper. The graphene under the copper is not desirable and should be removed. Therefore, the sample is placed in dilute nitric acid solution to remove the graphene under the copper. Fig. 5(B) shows the sample after removing the graphene under the copper. In the next step, the sample is taken out from the nitric acid solution and the sample is rinsed by DI water. The sample is placed into an iron nitrate solution that has the suitable concentration to remove copper. Fig. 5(C) shows the sample after copper etching. After removing the copper, several rinsing steps are performed with DI water to eliminate completely the metal contamination. In the next step after cleaning the substrate surface with acetone, while the graphene sample is suspended on the water, the graphene is slowly placed on the substrate. The graphene on the substrate is shown in Fig. 5(D).

In the last step, after several annealing steps for graphene adhesion to the substrate, NMP solution is used to remove PMMA. Fig. 5(E) shows graphene on the substrate after PMMA etching.

Fig. 1(D) shows graphene on the platinum silicide and silicon oxide. Finally, metal contacts on the graphene and under silicon is deposited as shown in Fig. 1(E).

Theoretical Equations

Fig. 6 shows the energy band diagram of the graphene/PtSi/silicon structure. The incident photons are absorbed in the platinum silicide and generate electron-hole pairs. By applying a voltage between two metal contacts, the excited electrons randomly walk in the platinum silicide layer until they reach the interface between the platinum silicide and the silicon. The electrons surmount the barrier and are emitted into the silicon [20]. The generated holes move to a negative voltage that is connected to the metal contact on the graphene.

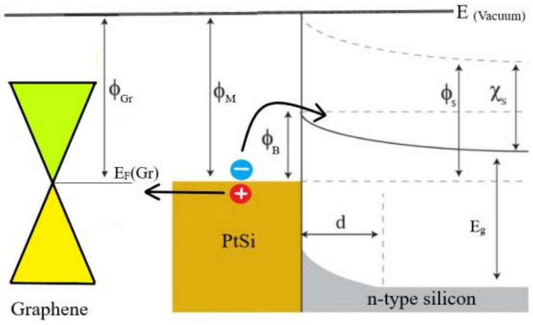


Fig. 6: The energy band diagram of the graphene/PtSi/silicon structure.

Fig.7 shows a side view of the structure to analyze the device. As shown in Fig. 7, an external voltage is applied to two metal contacts that are placed under the silicon and on the graphene that is placed on the silicon oxide. By applying an external voltage, the created electron-hole pair separates from each other and moves towards the metal contacts. In Fig. 7, the electron is shown in a black color and is moved towards the positive voltage, and its path to the metal contact under the silicon is shown with black arrows. The hole is shown in a white color and is moved towards the negative voltage, and its path from inside the graphene to the metal contact on the graphene is shown with white arrows. By applying an external voltage, the excited hole and electron are separated from each other and take two separate paths.

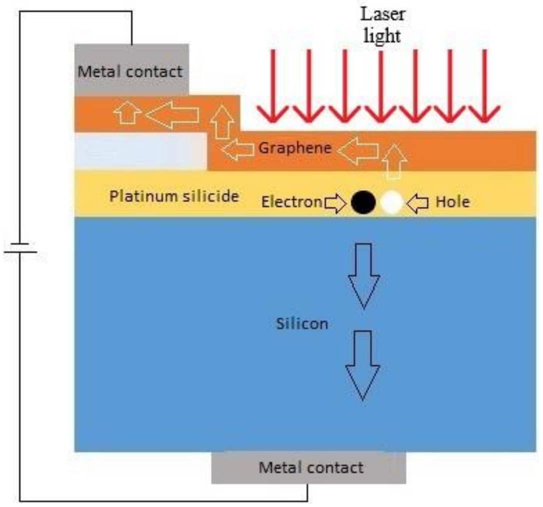


Fig. 7: The movement path of electrons and holes inside the structure after a laser radiation.

The electron travels vertically through the silicon, and the hole travels horizontally through the graphene. Because the mobility graphene is higher than silicon, holes reach the metal contact faster than electrons. The remained electrons cause extra holes to be injected into the device by external voltage, and more photocurrent is created. The graphene not only functions as the charge transport channel, but also works as a photoconductor [25].

The modified Fowler theory is an equation for the estimation of the internal quantum efficiency of Schottky photodetctors given by [26]-[30]:

$$\eta = \frac{1}{8\phi b}\frac{(hv-\phi b)^2}{hv} \qquad (1)$$

where h is plank constant, v is the frequency of the incident infrared light and ϕ_b is the Schottky barrier height of the platinum silicide and silicon. Equation (1) describes the internal quantum efficiency that is the number of photo-carriers collected by the photdetectors to the number of photons absorbed by the photodetector. Obviously, η depends on the incident photon energy and the Schottky barrier height.

The basic expression describing photocurrent in the photodetector is [25]:

$$\mathrm{I} = \alpha \times \varphi \times G \times e \times \eta \qquad (2)$$

where α is the adsorption rate of photodetector, φ is the incident photon flux, G is the photoconductive gain, and e is electron charge. The incident photon flux, φ, rate can be determined by the total optical power per photon energy [22]:

$$\varphi = \frac{\rho \times Wg \times Lg}{hv} \qquad (3)$$

where ρ is incident IR power density, W_g is the width of the channel, and L_g is the length of the channel. The photoconductive gain, G, is determined by the properties of the photodetector. The photoconductive gain can be defined as the number of carriers passing contacts per generated pair. The photoconductive gain is expressed by [25]:

$$G = \frac{\tau}{Tr} \qquad (4)$$

where τ is the recombination time and Tr is the transit time in the graphene channel. Tr is expressed by [25]:

$$\mathrm{Tr} = \frac{L^2}{\mu \times V} \qquad (5)$$

where L is the length of the graphene channel that generated carriers travel, μ is the mobility graphene, and V is applied voltage. The photoconductive gain for the high-quality graphene with the mobility of $1 \times 10^4 \frac{cm^2}{V.S}$, the 100 μm channel length, and 5 V bias voltage is less than one because the ultrafast photo-induced carrier recombination time in the graphene is in picosecond range. On the contrary, the photo-induced carriers are separated by the platinum silicide/silicon in our device, resulting in a much longer recombination time.

Results and Discussion

One of the best ways to measure graphene coverage on a surface is to take SEM images. Therefore, after graphene is transferred on the substrate, good informations can be obtained by taking the SEM image of graphene. Fig. 8 shows the SEM image of the graphene on the platinum silicide and silicon oxide substrate.

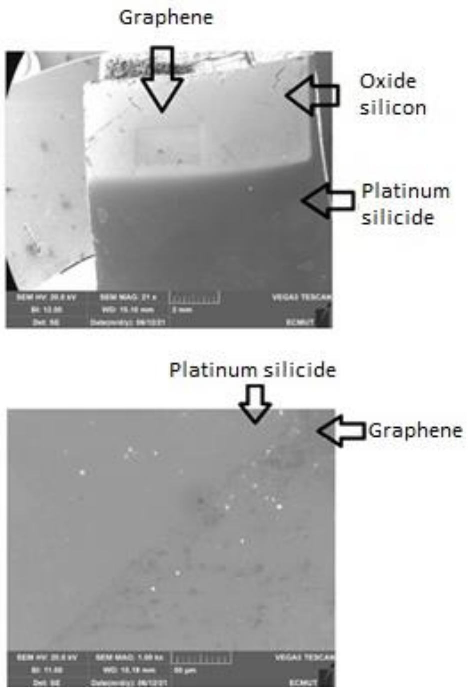


Fig. 8: SEM image of graphene on the surface of the platinum silicide.

SEM images were taken by the TESCAN VEGA3 tool. Fig. 8(A) shows the SEM image of the graphene on a surface with a magnification of 21. The part of the graphene is on the silicon oxide and another part on the platinum silicide. The silicon oxide in the image is shiny, and the platinum silicide is dark. The graphene has a very small thickness and does not have a specific color. The location of the graphene is shown by lines of the graphene edge. Fig. 8(B) shows the SEM image with magnification of 1000. This image is taken from the location of graphene on the surface of the platinum silicide. In the SEM images, the PMMA layer is removed from the graphene.

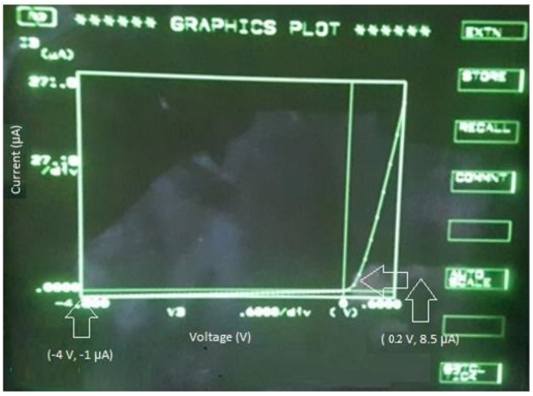


Fig. 9: I-V characteristic curve of Schottky diode between silicon and platinum silicide.

A Schottky diode is formed between silicon and platinum silicide. As shown in Fig. 9, I–V curve was obtained by an HP4450 semiconductor measurement device. To draw this curve, one terminal is connected to silicon and another terminal to platinum silicide. A sweep voltage from -4 to +0.6 V is applied to the platinum silicide, and the silicon is connected to a ground.

The Schottky diode is formed between silicon and platinum silicide. The diode anode is on the platinum silicide side, and the diode cathode is on the silicon side. The results are as expected; because n-type silicon is used, the silicon must act as a cathode for a Schottky diode. The characteristic curve of the diode I-V is theoretically obtained by [20]:

$$I = ABT^2 e^{-\frac{-q\phi b}{kT}}\left(e^{\frac{qv}{nkT}} - 1\right) \tag{6}$$

where A is the junction area, B is the Richardson constant, ϕ_b is the height of the Schottky barrier, k is Boltzmann constant, and T is the absolute temperature [20]. To characterize the photodetector, a wavelength of less than 1470 nm must be applied to the photodetector. The wavelength should not be in the visible range because silicon has a high absorption in in this range and produces a lot of optical current. For this reason, the detection of the Schottky photodetector of platinum silicide is not observed. Therefore, the wavelength should be greater than the visible range and less than 1470 nm. A good choice for the optical test of the photodetector is the use of a laser with a 1310 nm wavelength, which can show the detection of platinum silicide photodetector. By radiating laser light, the electron-hole pairs are created in the platinum silicide. By applying an external voltage, the carriers separate from each other and move towards the metal contacts. To calculate the optical current and responsivity of the photodetector, an optical setup must be prepared to minimize environment noise and to calculate the amount of optical current and responsivity. For this reason, an optical setup is prepared as shown in Fig. 10.

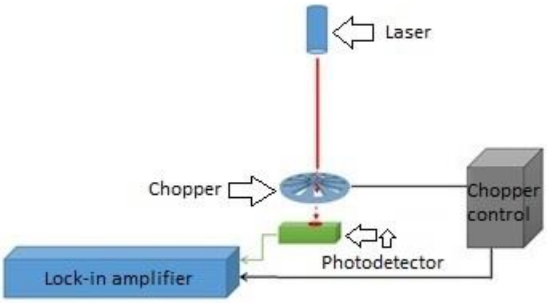


Fig. 10: Optical setup of the photodetector test.

In this optical setup, the 1310 nm wavelength is continuously radiated by the laser. After the light passes through a chopper with a 3 kHz frequency, the light is discrete and reaches to the photodetector. The laser spot diameter is 500 µm, and the laser power that reaches to the photodetector is 8.5 µW. Because the laser light reaches the photodetector discretely, the electrical current generated by the photodetector is discrete. The photodetector output and the chopper reference output are connected to a lock-in amplifier. In the lock-in amplifier, after electrical signals pass through various circuits, including frequency multiplier, Integrator circuit, amplifier, etc., the desired output is achieved with minimal noise.

There are two advantages at this optical setup; the first advantage is in reducing the environment noise, and the second advantage is in detection of the smallest electrical current. The current output from the photodetector includes the dark current and the optical current. The dark current is the current that passes through the photodetector in reverse bias without applying light. By applying laser light, the optical current is generated by the photodetector. If the dark current is subtracted from the photodetector output current, optical current is obtained [20]:

$$Iph = I(t) - I(dark) \tag{7}$$

where $I_{(t)}$ is the photodetector output current, $I_{(dark)}$ is the dark current, and I_{ph} is the optical current. The responsivity rate in terms of A/W for the photodetector is obtained [20]:

$$R = \frac{I_{ph}}{P} \tag{8}$$

where I_{ph} is the optical current, and P is the input power. External quantum efficiency of the photodetecor is given by [20]:

$$QE_{ex} = \frac{R}{\lambda} \times 1.245 \tag{9}$$

where R is responsivity, and λ is infrared wavelength. The fabricated photodetector is characterized in different ways. This photodetector is characterized by two cases of the platinum silicide photodetector with graphene and the platinum silicide photodetector without graphene. In the first case, the laser light is radiated on where the graphene is placed on the platinum silicide. The first case is shown in Fig. 11(A).

In the second case, the laser light is radiated on where platinum silicide is present, and there is no graphene. The second case is shown in Fig. 11(B). In this case, the platinum silicide photodetector works without graphene. In both cases, the contact must be taken from the structure to apply electrical voltage. In the first case, a contact is connected to silicon and another contact is connected to graphene on the silicon oxide. In the second case, the contact is connected to the platinum silicide, and another contact is connected to silicon.

In the Fig. 12, the optical current and the output voltage of the lock-in amplifier under different voltages are illustrated for both cases platinum silicide photodetector with graphene and platinum silicide photodetector without graphene. A resistor that help to calculation of the optical current is connected to photodetector in series.

After calculation of the optical current, responsivity is calculated by (8). In the Fig. 13, responisivity in different voltage are illustrated for both cases platinum silicide photodetector with graphene and platinum silicide photodetector without graphene.

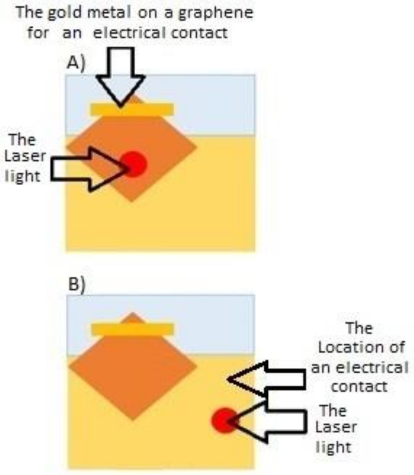

The gold metal on a graphene for an electrical contact
A)
The Laser light
B)
The Location of an electrical contact
The Laser light

Fig. 11: Location of laser light in different cases.

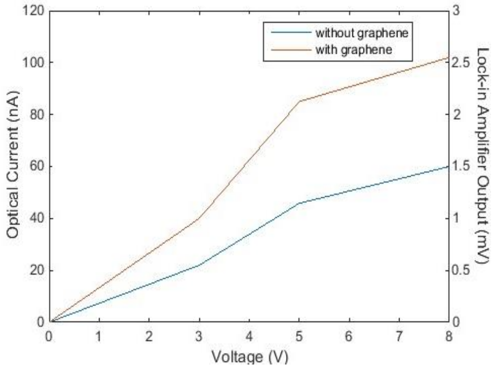


Fig. 12: The optical current and the output voltage of the lock-in amplifier under different voltages.

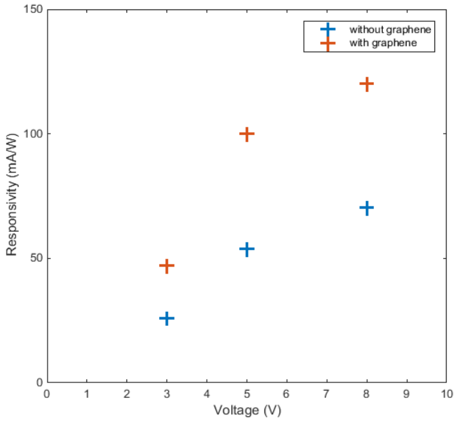


Fig. 13: photodetector responsivity under different voltages.

External quantum efficiency is calculated by (9). In the Fig. 14, external quantum efficiency in different voltage are illustrated for both cases platinum silicide photodetector with graphene and platinum silicide photodetector without graphene.

The platinum silicide photodetector with graphene has more optical current, responsivity, and external quantum efficiency than platinum silicon photodetector without graphene. In Table 1, our device is compared by several photodetectors previously reported. The use of graphene with platinum silicon has very interesting properties that can be considered in the future.

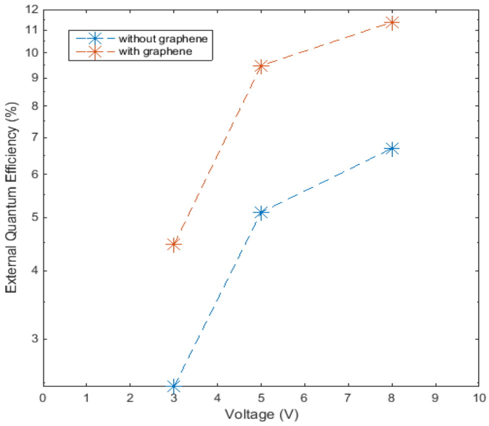


Fig. 14: External quantum efficiency of the photodetector under different voltages.

Table. 1: Summary of responsivity of several photodetectors previously reported

Structure	Responsivity	Wavelength	Ref
Platinum silicide with graphene	120 mA/W	1310 nm	This work
Platinum silicide	52 mA/W	1550 nm	[20]
Graphene	6.1 mA/W	1550 nm	[6]
Graphene/ silicon hetrostructure	2.29 mA/W	1550 nm	[22]
Graphene with cavity	20 mA/W	850 nm	[19]

Conclusion

In this paper, a photdetector structure of platinum silicide with graphene is presented. A graphene layer is placed on the platinum silicide. By radiating laser light, the electron-hole pairs are created in the platinum silicide. The photo-induced carriers separate from each other and move towards the metal contacts. The generated holes travel through the graphene and the generated electrons travels through the silicon. Because the mobility graphene is higher than silicon, holes reach the metal contact faster than electrons. The graphene not only functions as the charge transport channel, but also works as a photoconductor. Our device has more optical current, responsivity, and external quantum efficiency than platinum silicon photodetector. In our photodetector, the highest responsivity is 120 $\frac{mA}{W}$ in the 1310 nm wavelength, and the optical current is 100 nA at the applied voltage of 8 V. Our photodetector has optical current, responsivity, and external quantum efficiency twice as much as platinum silicide photodetector. This device can be a good candidate for optical communication applications.

Author Contributions

A.H. Mehrfar and A. Eslami Majd designed the experiments and collected the data. A. H. Mehrfar 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

PtSi	Platinum Silicide
PMMA	poly methyl methacrylate
NMP	N-Methyl-2-Pyrrolidone
SEM	Surface Imaging Method

References

[1] F. Bonaccorso, Z. Sun, T. Hasan, A.C. Ferrari, “Graphene photonics and optoelectronics,” Nat. photonics, 4(9): 611-622, 2010.

[2] R.R. Nair, P. Blake, A.N. Grigorenko, K.S. Novoselov, T. J. Booth, et al., “Fine structure constant defines visual transparency of graphene,” Science, 320(5881): 1308, 2008.

[3] K.I. Bolotin, K.J. Sikes, Z. Jiang, M. Klima, G. Fudenberg, et al., “Ultrahigh electron mobility in suspended graphene,” Solid state commun., 146(9): 351-355, 2008.

[4] V. Singh, D. Joung, L. Zhai, S. Das, S.I. Khondaker, et al., “Graphene based materials: past, present and future,” Prog. Mater. Scie., 56(8): 1178–1271, 2011.

[5] A.K. Geim, K.S. Novoselov, “The rise of graphene, in Nanoscience and technology: a collection of reviews from nature journals,” World Scientific, 11-19, 2010.

[6] T. Mueller, F. Xia, and P. Avouris, “Graphene photodetectors for high-speed optical communications,” Nat. photonics, 4(5): 297-301, 2010.

[7] F. Xia, T. Mueller, Y. Lin, A. Valdes-Garcia, and P. Avouris “Ultrafast Graphene Photodetector,” Nat. Nanotechnol., 4(12):839, 2009.

[8] B. Guo, Y. Wang, C. Peng, H.L. Zhang, G.P. Luo, et al., “Laser-based mid-infrared reflectance imaging of biological tissues,” Opt. express, 12(1): 208–219, 2004.

[9] C. Young, S.S. Kim, Y. Luzinova, M. Weida, D. Arnone, et al., “External cavity widely tunable quantum cascade laser based hollow waveguide gas sensors for multianalyte detection,” Sens. Actuators B, 140(1): 24–28, 2009.

[10] S. Cakmakyapan, P.K. Lu, A. Navabi, M. Jarrahi, “Gold-patched graphene nano-stripes for high-responsivity and ultrafast photodetection from the visible to infrared regime,” Light Sci. Appl., 7(1):20, 2018.

[11] A.B. Seddon, “Mid-infrared (IR)–A hot topic: The potential for using mid-IR light for non-invasive early detection of skin cancer in vivo,” phys. status solidi B, 250(5), 1020–1027 2013.

[12] J.M. Dawlaty, S. Shivaraman, M. Chandrashekhar, F. Rana, M.G. Spencer, “Measurement of ultrafast carrier dynamics in epitaxial graphene,” Appl. Phys. Lett., 92(4): 42116-42118, 2008.

[13] T. Low, P. Avouris, “Graphene plasmonics for terahertz to mid-infrared applications,” ACS nano, 8(2): 1086-1101, 2014.

[14] T.J. Echtermeyer, L. Britnell, P.K. Jasnos, A. Lombardo, R.V Gorbachev, et al., “Strong plasmonic enhancement of photovoltage in graphene,” Nat. Commun., 2(1): 1–5, 2011.

[15] R.J. Shiue, Y. Gao, Y. Wang, C. Peng, A.D. Robertson, et al., “High-responsivity graphene–boron nitride photodetector and autocorrelator in a silicon photonic integrated circuit,” Nano let., 15(11): 7288-7293, 2015.

[16] G. Konstantatos, M. Badioli, L. Gaudreau, J. Osmond, M. Bernechea, et al., “Hybrid graphene–quantum dot phototransistors with ultrahigh gain,” Nat. nanotechnol., 7(6): 363-368, 2012.

[17] X. Tang, K.W.C. Lai, “Graphene/HgTe quantum-dot photodetectors with gate-tunable infrared response,” ACS Appl. Nano Mater., 2(10): 6701–6706, 2019.

[18] Y. Chan, Z. Dahua, Y. Jun, T. Linlong, L. Chongqian, et al., “Fabrication of hybrid Graphene/CdS quantum dots film with the flexible photo-detecting performance,” Physica E, 124: 114216, 2020.

[19] M. Furchi, A. Urich, A. Pospischil, G. Lilley, K. Unterrainer, et al., “Microcavity-integrated graphene photodetector,” Nano let., 12(6): 2773–2777, 2012.

[20] A. Rogalski, Infrared and Terahertz Detectors, the third edition, CRC press, 2018.

[21] Y. Yao, R. Shankar, P. Rauter, Y. Song, J. Kong, et al., “High-responsivity mid-infrared graphene detectors with antenna-enhanced photocarrier generation and collection,” Nano let., 14(7): 3749–3754, 2014.

[22] X. Tang, H. Zhang, X. Tang, K.W.C. Lai, “Photoresponse enhancement in graphene/silicon infrared detector by controlling photocarrier collection,” Mater. Res. Express, 3(7): 76203-76214, 2016.

[23] A. Reina, X. Jia, J. Ho, D. Nezich, H. Son, et al., “Large area, few-layer graphene films on arbitrary substrates by chemical vapor deposition,” Nano Lett., 9 (1): 30–35, 2009.

[24] J.W. Suk, A. Kitt, C.W. Magnuson, Y. Hao, S. Ahmed, et al., “Transfer of CVD-grown monolayer graphene onto arbitrary substrates”, ACS nano, 5(9): 6916–6924, 2011.

[25] A. Rogalski, 2D materials for Infrared and Terahertz Detectors, the first edition, CRC press, 2020.

[26] J.M. Mooney, J. Silverman, “The theory of hot-electron photoemission in Schottky-barrier IR detectors,” IEEE Trans. Electron devices, 32(1): 33–39, 1985.

[27] V.E. Vickers, “Model of schottky barrier hot-electron-mode photodetection,” Appl. Opt., 10(9): 2190–92, 1971.

[28] V.L. Dalal, “Simple model for internal photoemission,” J. Appl. Phys., 42(6): 2274–2279, 1971.

[29] Z. Chen, Z. Cheng, J. Wang, X. Wan, C. Shu, et al., “High responsivity, broadband, and fast graphene/silicon photodetector in photoconductor mode,” Adv. Opt. Mater., 3(9): 1207–1214, 2015.

[30] M. Amirmazlaghani, F. Raissi, O. Habibpour, J. Vukusic, J. Stake, “Graphene-Si Schottky IR detector,” IEEE J. Quantum Electron., 49(7): 589–94, 2013.

Biographies

Amir Hossein Mehrfar was born in Tehran, Iran, 1991. He received the B.Sc. degree in Electrical Engineering from the University of Ghiaseddin Jamshid Kashani, in 2013. He received the M.Sc. degree in Electrical Engineering from Islamic Azad University South Tehran Branch. He is currently a Ph.D. student in the Malek-Ashtar University of Technology, Tehran, Iran. His current research interests include Graphene-Based Optoelectronics, Design and Modeling of Nano-Scale Semiconductor Devices, Design and Fabrication of IR Detectors.

- Email: Mehrfar.a.h@mut.ac.ir
- ORCID: 0000-0002-0501-0991
- 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 stary 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:

A.H. Mehrfar, A. Eslami Majd, “Enhancement of the photoresponse in the platinum silicide photodetector by a graphene layer,” J. Electr. Comput. Eng. Innovations, 10(2): 363-370, 2022.

DOI: 10.22061/JECEI.2022.8487.517

URL: https://jecei.sru.ac.ir/article_1676.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Robust Scheduling of Water and Energy Hub Considering CAES, Power-to-Gas Units, and Demand Response Programs

***S. Dorahaki*[1]*, S.S. Zadsar*[1]*, M. Rashidinejad*[1,*]*, M.R. Salehizadeh*[2]**

[1]*Department of Electrical Engineering, Shahid Bahonar University of Kerman,* Kerman, Iran.

[2]*Department of Electrical Engineering, Islamic Azad University, Marvdasht Branch,* Marvdasht, Iran.

Article Info

Article History:
Received 16 November 2021
Reviewed 27 December 2021
Revised 28 January 2022
Accepted 01 February 2022

Keywords:
Energy hub
Demand response programs
Power-to-gas (P2G)
Compressed air energy storage (CAES)
Robust optimization

*Corresponding Author's Email Address:
mrashidi@uk.ac.ir

Abstract

Background and Objectives: The smart energy hub framework encompasses physical assets such as thermal storage, boiler, wind turbine, PV panel, water storage and, water desalination unit to ensure continuity of electricity, water, thermal, and gas provision in the case of unexpected outages in the upstream networks. In this regard, the smart energy hub as an integrated structure provides a suitable platform for energy supply. Considering the drinking water resources in the smart hub structure can cause operational efficiency improvement.
Methods: This paper proposes an integrated scheduling model for energy and water supply. To address the issue of increasing operational flexibility, a set of new technologies such as Compressed Air Energy Storage (CAES) and Power-to-Gas (P2G) system are provided. Also, the energy price is modeled as an uncertain parameter using a robust optimization approach. The proposed model is established as a Mixed Integer Linear Function (MILP). The mentioned model is implemented using the CPLEX solver in GAMS software. The proposed model is simulated in different scenarios in the energy hub and the optimization results are compared with each other to validate the proposed method.
Results: The results show that using CAES technology and the P2G system can lead to reducing the operating costs to a desirable level. Moreover, the impact of the P2G unit on the operation cost is more than the CAES unit.
Conclusion: The energy hub operator should tradeoff between robustness and operation cost of the system. The obtained results ensured that the proposed methodology was robust, optimal, and economical for energy hub schedules.

Introduction

In the last decades, energy system has expanded from isolated energy carrier systems into integrated energy structure [1]. The integrated multi-carrier energy system play an important role in future smart grid power systems [2]. In any urban area, different energy carriers can be managed in an energy hub framework where the operation cost and emission mitigation issues are the main purposes [3]. In this regard, the energy hub system plays an important role in the field of energy conversion, generation, and storage in an efficient manner [4]. Due to the mentioned abilities of energy hub systems, input carriers of the energy hub system have a variety [5] where the electrical, thermal, gas and freshwater are the main energy input into the energy hub system [6].

Furthermore, Demand Response Programs (DRPs) are

Doi: 10.22061/JECEI.2022.8489.519

used in the energy hub system to increase operational efficiency [7]. In this regard, DRPs are categorized into electrical [8] and thermal [9] programs. The operating expenditure of energy systems can be decreased by the demand response programs [10]. The effects of electrical and thermal DRPs on the flexibility and reliability of the energy hub system are evaluated in [11]. Moreover, DRPs can decrease greenhouse gases emission in the energy hub system [12].

Furthermore, considering novel and efficient storage technologies in the energy hub system is caused to decrease in the operation cost [13]. One of the efficient Energy Storage Systems (ESS) technologies is Power-to-Gas (P2G), which is associated with challenges due to environmental problems as well as storage space [14]. The P2G storage system converts the extra electrical power into gas energy in the low electrical price hours and uses the stored gas energy in electrical peak hours [9], [15]–[18].

Researchers have recently been able to store excess energy by compressing air from electricity generated using renewable energy, a technology called Compressed Air Energy Storage (CAES) [19]. The cost of using this method is very low and its efficiency is much higher than power storage batteries [14]. Also, CAES is a low-cost method of energy storage that plays an important role in energy management, improving power quality, etc., and is the cheapest method of energy storage [20]. Some research papers climes that the CAES can cover the energy price uncertainty of the upstream network. In [21] a random optimization method is proposed to cover the uncertainty of energy prices in the electricity market. Profit maximization is the main objective of the mentioned paper.

In optimization problems, there are several ways for dealing with uncertainties, one of which is the robust optimization (RO) method [22]. In [23], the problem of unit commitment due to wind power uncertainty has been solved using the robust optimization method. This method has recently been introduced as an efficient method in mathematical programming in optimization problems for power system decision-makers. The future power grid, with the unprecedented infiltration of renewable energy sources, will face severe uncertainties that may cause problems in the operation of the grid. It is necessary to evaluate the uncertainty of system performance in this network. In [24], the uncertainty of renewable wind energy is investigated using a strong two-stage optimization method. The integrated electricity and heating system has been investigated in [25], [26] and the price uncertainty of electricity has been modeled using a robust optimization method.

In this paper, a novel robust energy and water optimization model is proposed. Also, the CAES unit, as well as P2G, are used to enhance the flexibility of the proposed energy hub system. Mixed-Integer Linear Programming (MILP) method is used to model the optimization of the proposed energy hub. Also, desired results are obtained using the CPLEX solver in the GAMS environment. Summary, the contributions of the paper are as follow:

- ✓ The role of novel energy storage technologies such as CAES and P2G units in the energy hub system is investigated.
- ✓ The water desalination units, as well as water storage, are considered in the energy hub system.
- ✓ The robust optimization method is used to model the upstream electrical price uncertainty.

The remaining of the paper is organized as follows. In section II, the proposed structure is stated. The formulation of the problem is specified in Section III. The case study is presented in Section IV. At the end of this study, the conclusion is given in Section V.

The Proposed Structure

In this paper, a novel Power and Water Robust Optimization (PWRO) framework has been proposed to decrease the effects of the parameter uncertainties in the energy hub structure. Furthermore, the uncertainty of price has been considered as an uncertainty parameter and has been modeled by the robust optimization method. Thermal, electricity, gas, and water carriers are inputs of the system. On the other hand, the demand for the proposed hub system should be satisfied. Furthermore, the boiler unit, thermal storage unit, and partial section of energy outputs of the Combined Heat and Power (CHP) unit are to receive the thermal energy of the energy hub system. Also, the wind turbine and the microturbine unit generate electrical power. The integrated structure of the energy hub test system is shown in Fig. 1.

Mathematical Formulation

The objective function of the proposed energy hub model is as follow:

$$Min\sum_{t=1}^{N_t}\left\{\begin{array}{l}\underbrace{\left(\pi_{net}^{E}(t)P_{net}^{E}(t)+\pi_{wind}^{E}P_{wind}^{E}(t)\right)}_{\text{Electrical Cost}}+\\ \underbrace{\left(\pi_{net}^{G}(t)P_{net}^{G}(t)\right)}_{\text{Gas Cost}}+\underbrace{\left(\pi_{net}^{T}(t)P_{net}^{T}(t)\right)}_{\text{Thermal Cost}}+\\ \underbrace{\left(\begin{array}{l}\pi_{Drink\_water}(t)W_{Drink\_water}(t)+\\ \pi_{Sea\ to\ drink}^{Des}W_{Sea\ to\ drink}^{Des}(t)\end{array}\right)}_{\text{Water Cost}}+\\ \underbrace{\left(\begin{array}{l}\pi_{DRP}^{E}\left(P_{down}^{E}(t)+P_{up}^{E}(t)\right)+\\ \pi_{DRP}^{T}\left(P_{down}^{T}(t)+P_{up}^{T}(t)\right)\end{array}\right)}_{DRP\ \text{Cos}t}\end{array}\right\} \tag{1}$$

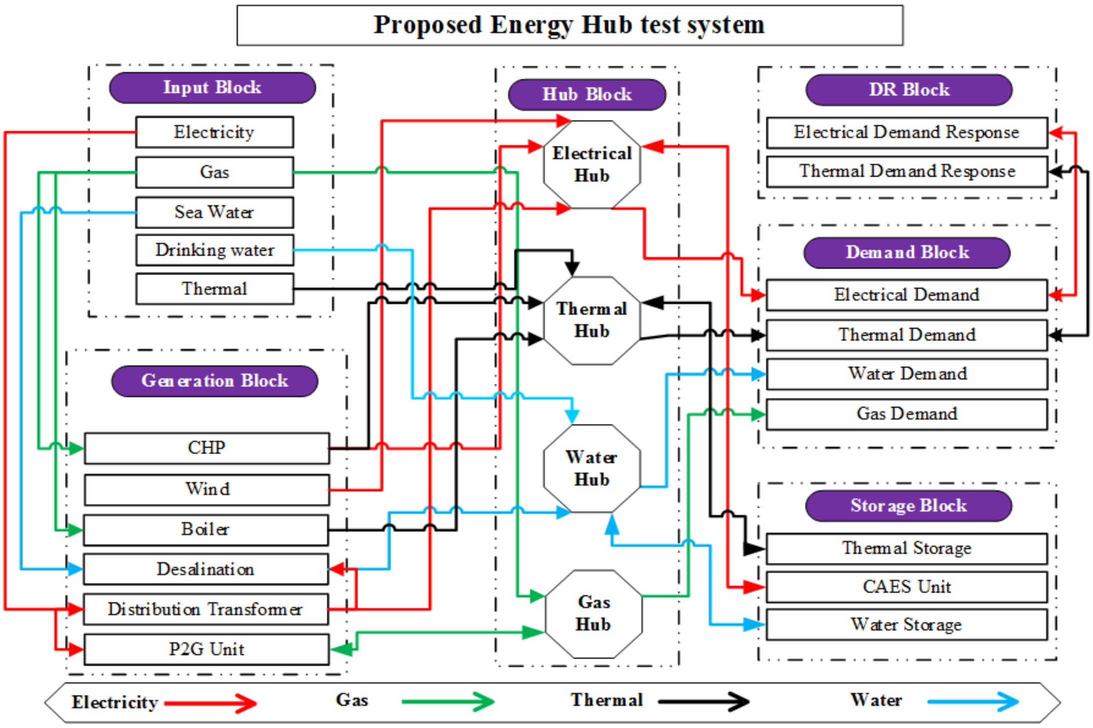


Fig. 1: The water and energy hub system.

The objective function is included different parts such as electrical cost, Gas Cost, Thermal Cost, Water Cost, and DRP Cost respectively.

$$\left[\eta^{EE}_{Trans} P^{E}_{net}(t)\right] + \left[\eta^{GE}_{CHP} P^{G}_{netCHP}(t)\right] + \left[\eta^{EE}_{Conv} P^{E}_{wind}(t)\right] + \left[P^{E}_{dis}(t) - P^{E}_{ch}(t)\right] + \left[P^{E}_{down}(t) - P^{E}_{up}(t)\right] + W^{Des}_{Sea\ to\ drink}(t)\sigma_{P2W} + P_{C,S}(t) - P_{CAES}(t) + P_{P2G}(t) - P_{G2P}(t) = P^{E}_{demand}(t) \tag{2}$$

$$P^{G}_{net}(t) - P^{G}_{netCHP}(t) - P^{G}_{netboil}(t) = P^{G}_{demand}(t) \tag{3}$$

$$\left[\eta^{GT}_{CHP} P^{G}_{netCHP}(t)\right] + \left[\eta^{GT}_{Boil} P^{G}_{netBoil}(t)\right] + P^{T}_{net}(t) + \left[P^{T}_{dis}(t) - P^{T}_{ch}(t)\right] + \left[P^{T}_{down}(t) - P^{T}_{up}(t)\right] = P^{T}_{demand}(t) \tag{4}$$

$$W^{Des}_{Sea\ to\ drink}(t) + W_{Drink_{water}}(t) + W_{dis}(t) - W_{Ch}(t) = p^{water}_{demand}(t) \tag{5}$$

The input electrical power and the wind turbine operation are considered as the electrical cost of the proposed system. The thermal, as well as gas cost, are the cost of input thermal and gas energy to the energy hub respectively. The water cost in the objective function consists of two parts namely: input drinking water from the upstream water network and the water desalination operation cost. Furthermore, the final part of the objective function is the cost of electrical and thermal demand response programs.

The balance limits of electric, thermal, gas, and water energy are as follows (2)-(5):

The input power, thermal, gas, and drinking water from the upstream network are limited by (6)-(9) respectively [27]:

$$0 \le P^{E}_{net}(t) \le P^{E}_{net-max} \tag{6}$$

$$0 \le P^{T}_{net}(t) \le P^{T}_{net-max} \tag{7}$$

$$0 \le P^{G}_{net}(t) \le P^{G}_{net-max} \tag{8}$$

$$0 \le W_{Drink_{water}}(t) \le W_{DW-max} \tag{9}$$

In addition, the input power of the distribution transformer is limited by (10) [28]:

$$0 \le P^{E}_{net}(t) \le \mathrm{P}^{\mathrm{input}}_{\mathrm{trans}} \tag{10}$$

Moreover, the input gas of CHP and boiler has been addressed as (11) and (12):

$$0 \le P^{G}_{netCHP}(t) \le \mathrm{P}^{\mathrm{input}}_{\mathrm{CHP}} \tag{11}$$

$$0 \le P^{G}_{netB}(t) \le \mathrm{P}^{\mathrm{input}}_{\mathrm{boiler}} \tag{12}$$

In the following, (13)-(19) models the operation of the CAES technology.

$$V^{inj}\left(t\right)=\alpha^{inj}P_{CAES}\left(t\right) \tag{13}$$

$$P_{C,S}\left(t\right)=\alpha^{P}V^{P}\left(t\right) \tag{14}$$

$$V_{min}^{inj}u^{inj}\left(t\right)\leq V^{inj}\left(t\right)\leq V_{max}^{inj}u^{inj}\left(t\right) \tag{15}$$

$$V_{min}^{P}u^{P}\left(t\right)\leq V^{P}\left(t\right)\leq V_{max}^{P}u^{P}\left(t\right) \tag{16}$$

$$u^{inj}\left(t\right)+u^{P}\left(t\right)\leq 1 \tag{17}$$

$$A\left(t+1\right)=A\left(t\right)+V^{inj}\left(t\right)-V^{P}\left(t\right) \tag{18}$$

$$A^{min}\leq A\left(t\right)\leq A^{max} \tag{19}$$

Equations (13) and (14) indicate the energy import and export in the CAES unit. The imported and exported energy in the CAES unit is limited by (15) and (16) respectively. The energy level of the CAES unit is obtained by (18). Furthermore, the capacity of the CAES unit is limited by (19). The thermal storage operation constraints have been provided in (20)-(25).

$$P_{s}^{T}\left(t\right)=P_{s}^{T}\left(t-1\right)+P_{ch}^{T}\left(t\right)-P_{dis}^{T}\left(t\right)-P_{loss}^{T}\left(t\right) \tag{20}$$

$$P_{loss}^{T}\left(t\right)=\vartheta_{loss}^{T}P_{s}^{T}\left(t\right) \tag{21}$$

$$\alpha_{min}^{T}P_{CAPA}^{T}\leq P_{s}^{T}\left(t\right)\leq\alpha_{max}^{T}P_{CAPA}^{T} \tag{22}$$

$$\beta_{min}^{T}P_{CAPA}^{T}I_{ch}^{T}\left(t\right)\leq P_{ch}^{T}\left(t\right)\leq\beta_{max}^{T}P_{CAPA}^{T}I_{ch}^{T}\left(t\right) \tag{23}$$

$$\beta_{min}^{T}P_{CAPA}^{T}I_{dis}^{T}\left(t\right)\leq P_{dis}^{T}\left(t\right)\leq\beta_{max}^{T}P_{CAPA}^{T}I_{dis}^{T}\left(t\right) \tag{24}$$

$$0\leq I_{dis}^{T}\left(t\right)+I_{ch}^{T}\left(t\right)\leq 1 \tag{25}$$

Equation (20) indicates the thermal storage status. Moreover, the loss of energy storage unit is modeled by (21). The capacity of the thermal storage is shown by (22). Charging and discharging of thermal storage are limited by (23) and (24). The status of the thermal energy storage unit in each hour is determined by (25).

The mathematical formulations of the water storage are as follow:

$$W_{storage}\left(t\right)=W_{storage}\left(t-1\right)+W_{ch}\left(t\right)-W_{dis}\left(t\right) \tag{26}$$

$$0\leq W_{storage}\left(t\right)\leq W_{storage-max} \tag{27}$$

$$0\leq W_{ch}\left(t\right)\leq W_{max-ch}I_{ch}^{W}\left(t\right) \tag{28}$$

$$0\leq W_{dis}\left(t\right)\leq W_{max-dis}I_{dis}^{W}\left(t\right) \tag{29}$$

$$0\leq I_{ch}^{W}\left(t\right)+I_{dis}^{W}\left(t\right))\leq 1 \tag{30}$$

The desalination unit has an efficiency coefficient that has been considered in (31):

$$W_{Sea\ to\ drink}^{Des}\left(t\right)=\eta_{sea\ to\ drink}W_{sea}\left(t\right) \tag{31}$$

In (32)-(35) and (36)-(40) the mathematical limitations of electrical and thermal energy storage are expressed [29].

$$\sum_{t=1}^{24}P_{down}^{E}\left(t\right)=\sum_{t=1}^{24}P_{up}^{E}\left(t\right) \tag{32}$$

$$0\leq P_{up}^{E}\left(t\right)\leq LPF_{up}^{E}P_{demand}^{E}\left(t\right)I_{up}^{E}\left(t\right) \tag{33}$$

$$0\leq P_{down}^{E}\left(t\right)\leq LPF_{down}^{E}P_{demand}^{E}\left(t\right)I_{down}^{E}\left(t\right) \tag{34}$$

$$0\leq I_{down}^{E}\left(t\right)+I_{up}^{E}\left(t\right)\leq 1 \tag{35}$$

$$\sum\nolimits_{t=1}^{24}P_{down}^{T}\left(t\right)=\sum\nolimits_{t=1}^{24}P_{up}^{T}\left(t\right) \tag{36}$$

$$0\leq P_{up}^{T}\left(t\right)\leq LPF_{up}^{T}P_{demand}^{T}\left(t\right)I_{up}^{T}\left(t\right) \tag{37}$$

$$0\leq P_{down}^{T}\left(t\right)\leq LPF_{down}^{T}P_{demand}^{T}\left(t\right)I_{down}^{T}\left(t\right) \tag{38}$$

$$0\leq I_{down}^{T}\left(t\right)+I_{up}^{T}\left(t\right)\leq 1 \tag{39}$$

Equations (32) and (36) indicate that the sum of downward and upward demand in a day should be equal (load shifting). Also, (33) and (34) as well as (37) and (38) show that the upward and downward DRP is limited to the partial loads. Equations (35) and (39) indicate that in each hour only one DRP strategy can be implemented (Upward or Downward). The P2G system is modeled as follow:

$$GS(t)=GS(t-1)+G_{P2G}^{ch}(t)-G_{P2G}^{dis}(t) \tag{40}$$

$$GS^{\min}\leq GS(t)\leq GS^{\max} \tag{41}$$

$$G_{P2G}^{ch,\min}\leq G_{P2G}^{ch}(t)\leq G_{P2G}^{ch,\max} \tag{42}$$

$$G_{P2G}^{dis,\min}\leq G_{P2G}^{dis}(t)\leq G_{P2G}^{dis,\max} \tag{43}$$

$$G_{P2G}^{ch}(t)=\eta_{P2G}P_{P2G}(t) \tag{44}$$

$$G_{P2G}^{dis}(t)=\eta_{G2P}P_{G2P}(t) \tag{45}$$

$$0\leq P_{G2P}(t)\leq P_{G2P}^{\max} \tag{46}$$

$$0 \le P_{P2G}(t) \le P_{P2G}^{\max} \tag{47}$$

Constraint (40) shows the charge level of the P2G system. The charge level of the P2G system is limited by (41). Charging and discharging the P2G system are limited by (42) and (43) respectively. Moreover, the energy conversion in the P2G system is modeled by (44) and (47).

A. Robust Optimization Modeling

The RO approach paves the way for system operators to act risk-aversely by changing the uncertainty budget. In this regard, the energy hub operator should tradeoff between operation cost and system robustness. It is clear that if the more uncertainty budget increases, the more risk-averse manner adopted.

The robust optimization approach compared with stochastic approaches has two main advantages:

- First, the implementation of robust optimization is simpler than the scenario-based approaches. This approach only requires the predicted values of the upper limit and the lower limit of the target variable.
- Second, unlike stochastic methods that use probabilistic guarantees to satisfy constraints, the proposed method is followed by optimal solutions that are safe against all changes in random variables.

In the following, the objective function of the energy hub problem is modeled based on the robust optimization approach that is proposed in [30]. The objective function of the deterministic problem can be rewritten as follow:

$$\sum_{t=1}^{24}[(\pi_{net}^{E}(t)P_{net}^{E}(t)] + \text{Other Costs} \tag{48}$$

In the above objective function, the electricity cost is separated from other operating costs to implement uncertainty. The other costs are gas, thermal, water, and demand response cost which were shown in (1).

Base on [30], the target of the operator is obtained to the worst solution and find a way to minimize the effects of the worst case. Therefore, the objective function can be rewritten as follow:

$$min \quad max\sum_{t=1}^{N_t}[(\pi_{net}^{RO,E}(t)P_{net}^{E}(t)] + \text{Other Costs} \tag{49}$$

where $\pi_{net}^{RO,E}(t)$ is the main grid price of electricity. The second term of objective function should be considered in solving the problem using the dual process. To model the price uncertainty, the uncertain price is modeled by forecasted value and deviation from forecasted value as follow:

$$\begin{aligned} & max\sum_{t=1}^{N_t}[(\pi_{net}^{E,\,forecasted}(t)(1+Z(t))P_{net}^{E}(t)] \\ & s.t \\ & Z(t) \le 1 \qquad :\beta_t \\ & \sum_{t=1}^{N_t} Z(t) \le \Gamma \qquad :\alpha \\ & Z(t) \ge 0 \end{aligned} \tag{50}$$

In the above formulation, α and β_t are the dual variables of constraints. Moreover, Γ is the uncertainty budget of the price of electricity. The objective function of the main problem is rewritten by considering the KKT condition as follow:

$$min \begin{Bmatrix} \sum_{t=1}^{N_t}[(\pi_{net}^{E,\,forecasted}(t)P_{net}^{E}(t)] + \\ \sum_{t=1}^{N_t}[\beta_t] + \alpha + \text{Other Costs} \end{Bmatrix} \tag{51}$$

$$\alpha + \beta_t \ge dev\,\pi_{net}^{E,\,forecasted}(t)P_{net}^{E}(t) \tag{52}$$

$$\beta_t \ge 0 \tag{53}$$

$$\alpha \ge 0 \tag{54}$$

$$constraints(2)-(47)$$

Fig. 2 shows the robust optimization algorithm in the energy hub framework.

In the first step, the uncertainty budget and the iteration index are considered equal to 0 and 1 respectively. In the second step, the proposed optimization problem will be solved and the energy hub variables are obtained. In the following, the uncertainty budget is updated and so, if the uncertainty budget is equal to 24 the obtained results are displayed.

Case Study

The energy management horizon time is considered 24 hours. Also, the electrical, thermal, gas, and water demands of the energy hub test system are shown in Fig. 3. The maximum and minimum electrical prices are shown in Fig. 4. Also, the thermal price of the energy hub test system is shown in Fig. 5. Furthermore, the input parameters of the energy hub test system are used from [18]. The effects of CAES and P2G units on the operation cost of the proposed energy hub system are shown in Table. 1.

In the base case scenario (scenario 1), the CAES and P2G units are neglected in the energy scheduling problem. In the second scenario, the CAES unit is considered and the P2G unit is neglected and vice versa in the third scenario. The simultaneity operation of the CAES unit and the P2G unit is considered in the fourth scenario.

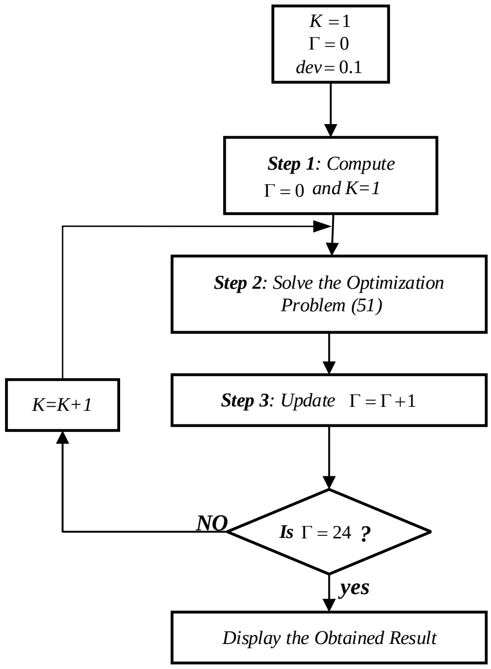

$K=1$
$\Gamma=0$
$dev=0.1$
Step 1: Compute $\Gamma=0$ and K=1
Step 2: Solve the Optimization Problem (51)
Step 3: Update $\Gamma=\Gamma+1$
K=K+1
NO
Is $\Gamma=24$?
yes
Display the Obtained Result

Fig. 2: The Proposed Robust optimization Algorithm.

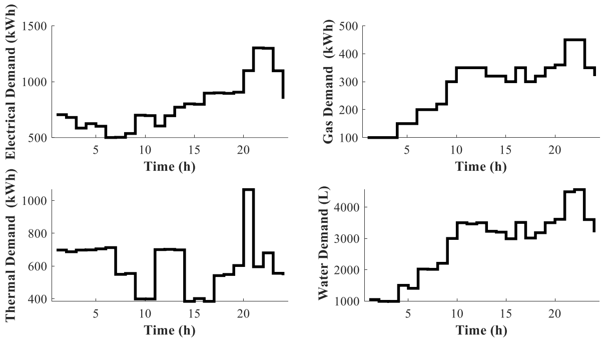


Fig. 3: Energy demands of the energy hub system.

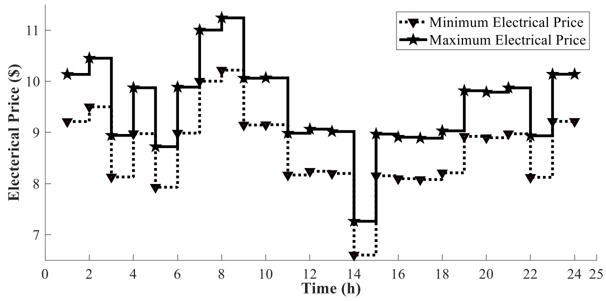


Fig. 4: The minimum and maximum upstream electrical market.

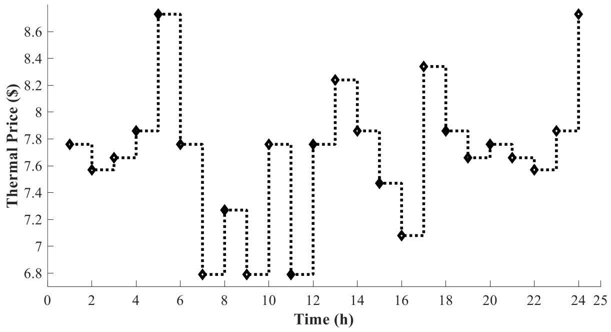


Fig. 5: Thermal price of energy hub test system.

Results show that the CAES unit can be used for operating cost reduction in the energy hub test system. however, the P2G unit is a more efficient device than the CAES unit. The final scenario is the best and the operation cost decreases 1.32% compared with the base scenario (scenario 1).

Table 1: The operation cost energy hub system

	CAES	P2G	Operation Cost ($)	Percentage (%)
Scenario 1	✗	✗	622039	-
Scenario 2	✓	✗	620861	-0.18
Scenario 3	✗	✓	614662	-1.18
Scenario 4	✓	✓	613801	-1.32

Fig. 6 and Fig. 7 show the operation of the CAES unit and P2G system. The results show that the CAES unit and P2G are appropriate for energy arbitrage between hours. In this regard, the system operator imports energy in the P2G and CAES units at the high energy price hours and exports the stored energy at the lower price hours.

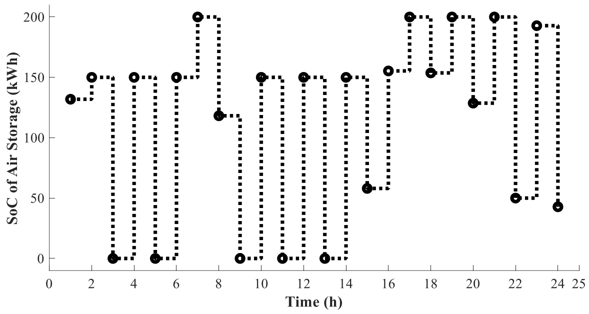


Fig. 6: The SOC of the CAES unit.

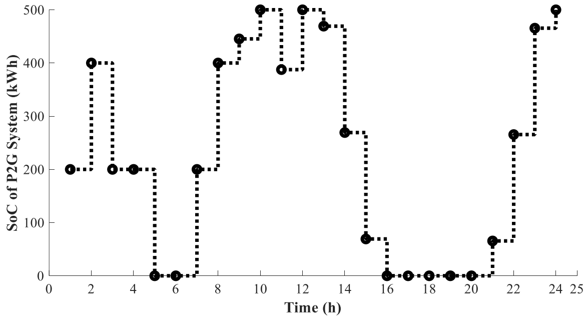


Fig. 7: The SOC of the P2G unit.

The results of electrical and thermal load shifting DRP are shown in Fig. 8 and Fig. 9 respectively. The positive values in the mentioned Figures are referred to the load decrement and vice versa. Results show that the electrical demand in the high price hours 1-2 and 7-10 is shifted down. Moreover, the electrical demand peak reduction in high peak hours 21-24 is more than other hours. Because the high electrical peak price and high electrical peak demand simultaneously occur. The thermal DRP works the same as the electrical one. For example, the thermal load is shifted down in high thermal price hours 12-14. Moreover, the thermal demand is shifted up in the low price and low thermal demand hour 11.

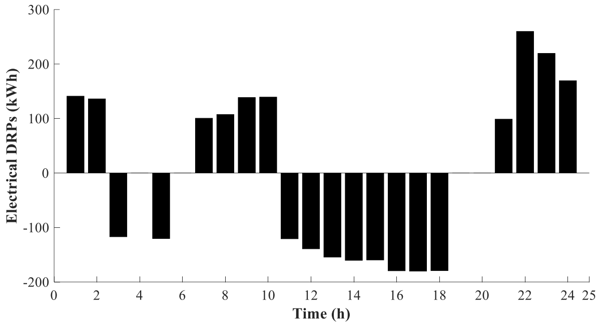


Fig. 8: The load shifting of the electrical DRP.

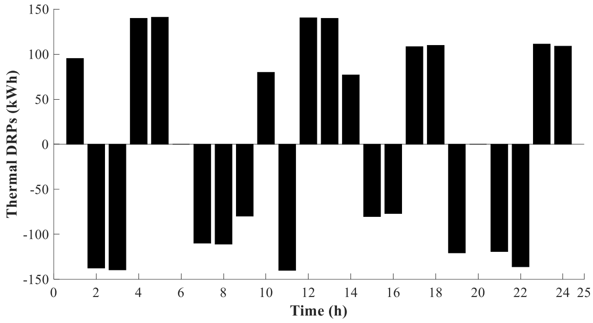


Fig. 9: The load shifting of the thermal DRP.

The effect of the uncertainty budget on the operational cost is presented in Fig. 10. By increasing the robust uncertainty budget, the total operation cost increases.

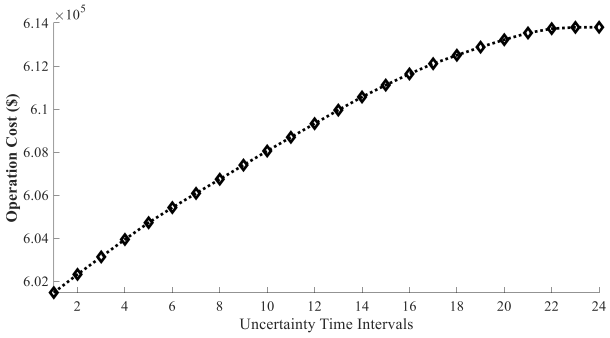


Fig. 10: The operation cost of the energy hub.

Conclusion

This paper proposes a novel robust energy nexus water optimization problem. The effects of uncertainty budget on the results of energy hub schedules were evaluated. The proposed approach was formulated as a Mixed Integer linear programming problem. The effects of the P2G unit and CAES units are evaluated on the operation cost. Results show that novel energy storage technologies such as P2G and CAES units can significantly decrease the daily operation cost (i.e., 1.32 %). However, the impact of the P2G unit (i.e., 0.18 %)is more than the CAES unit (i.e., 1.18 %). The robust optimization method was implemented to evaluate the uncertainty of upstream electricity prices. The results showed that the operation cost of the proposed system increased by increasing the robust uncertainty budget. However, the robustness of the proposed energy hub system was increased by considering a robust strategy (increasing the uncertainty budget). The energy hub operator should tradeoff between robustness and operation cost of the system. The obtained results ensured that the proposed methodology was robust, optimal, and economical for energy hub schedules. In future research, the electrical, thermal, water, and heating networks will be considered in the model.

Author Contributions

S. Dorahaki and S. S Zadsar proposed and designed the structure of the energy hub. Also, the optimization code has been implemented by S. Dorahaki. S. S Zadsar wrote the original draft of the manuscript. M. Rashidinejad and M. R. Salehizadeh reviewed the data analysis and results as well as reviewed and edited the manuscript.

Acknowledgment

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

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.

Abbreviations

CAES	Compressed Air Energy Storage
P2G	Power-to-Gas
MILP	Mixed Integer Linear Function
PWRO	power and water robust optimization
CHP	Combined Heat and Power
DRP	Demand Response Program

Sets and indices

t	Index of time.
$\boldsymbol{N_t}$	The number of the time periods.

Parameters

π_{net}^{E}	Electrical price.

π^{E}_{wind}	The operation cost of wind unit.
π^{G}_{net}	Gas price.
π^{T}_{net}	Thermal price.
π_{Drink_water}	Drinking water price.
$\pi^{Des}_{Sea\ to\ drink}$	The operation cost of the desalination unit.
η^{GT}_{Boil}	The efficiency of gas to thermal in Boiler.
$\eta_{sea\ to\ drink}$	The efficiency of the desalination unit.
η^{EE}_{Trans}	The efficiency of the transformer unit.
η^{GE}_{CHP}	The efficiency of the CHP unit.
η^{EE}_{Conv}	The efficiency of the converter unit.
η^{GT}_{CHP}	The efficiency of gas to thermal in CHP.
$\eta^{ch}_{HS},\eta^{dis}_{HS}$	The charge/discharge efficiency of thermal storage.
η_{G2P},η_{P2G}	The energy conversion efficiency of the P2G unit.
$\beta^{H}_{max},\beta^{H}_{min}$	Maximum/Minimum ratio of thermal charge.
ϑ^{H}_{loss}	The ratio of thermal storage loss.
$\alpha^{H}_{min},\alpha^{H}_{max}$	The min/max ratio of thermal storage.
α^{inj},α^{p}	Imported/exported efficiency to/from CAES.
$GS^{\min},GS^{\max}$	Minimum/Maximum SOC of the P2G unit.
$G^{ch,\min}_{P2G},G^{ch,\max}_{P2G}$	Minimum/Maximum charge of the P2G unit.
$G^{dis,\min}_{P2G},G^{dis,\max}_{P2G}$	Minimum/Maximum discharge of the P2G unit.
$V^{inj}_{min},V^{inj}_{max}$	Minimum/Maximum energy imported to CAES.
V^{P}_{min},V^{P}_{max}	Minimum/Maximum energy exported to CAES.
$P^{E}_{net-max}$	Maximum input electrical power.
$P^{T}_{net-max}$	Maximum input thermal energy.
$P^{G}_{net-max}$	Maximum input gas energy.
$W^{E}_{net-max}$	Maximum input drinking water.
W_{DW-max}	Maximum output water of desalination unit.
$\mathrm{P}^{\mathrm{input}}_{\mathrm{trans}}$	Input electrical power to the transformer.
$\mathrm{P}^{\mathrm{input}}_{\mathrm{CHP}}$	Input energy to CHP.
$\mathrm{P}^{\mathrm{input}}_{\mathrm{Boiler}}$	Input energy to the boiler.
$W_{storage-max}$	Maximum state of water storage.
$W_{max-ch},W_{max-dis}$	Maximum charge/discharge water.
$LPF^{E}_{up},LPF^{E}_{down}$	Shifted up/down electrical Demand.
P^{H}_{CAPA}	The thermal storage capacity.
P^{E}_{demand}	Electrical demand.
P^{G}_{demand}	Gas demand.
P^{T}_{demand}	Thermal demand.
P^{water}_{demand}	Water demand.

Variables

P^{E}_{net}	Input electrical power.
P^{E}_{wind}	Wind power.
P^{G}_{net}	Input gas power.
P^{T}_{net}	Input thermal power.
W_{Drink_water}	Input drinking water.
$W^{Des}_{Sea\ to\ drink}$	Output water of desalination unit.
P^{E}_{up},P^{E}_{down}	Electrical demand response up/down demand.
P^{T}_{up},P^{T}_{down}	Thermal up/down demand response.
$P_{CAES}(t),P_{C,S}(t)$	Imported/exported power to/from CAES.
$P_{P2G}(t),P_{G2P}(t)$	Energy conversion power of P2G unit.
P^{T}_{ch}	Thermal charge.
P^{T}_{dis}	Thermal discharge.
$W_{storage}$	State of water storage.
W_{sea}	Seawater.
$V^{inj}(t),V^{P}(t)$	Imported/exported energy to/from CAES.
$GS(t)$	The SOC of Gas energy in the P2G unit.
$G^{ch}_{P2G}(t),G^{dis}_{P2G}(t)$	Charging/discharging energy from the P2G unit.
dev	The upstream price deviation from the forecasted value.
W_{ch},W_{dis}	Charge/Discharge water from water storage.
P^{G}_{netCHP}	Input gas to CHP unit.
$P^{G}_{netboil}$	Input gas to Boiler unit.

Binary Variables

I^{H}_{ch},I^{H}_{dis}	The binary variable of thermal charge/discharge.
I^{E}_{up},I^{E}_{down}	The binary variable of shifted up/down DRPs.
I^{W}_{ch},I^{W}_{dis}	The binary variable of water charge/discharge.
$u^{inj}(t),u^{P}(t)$	Binary variables of imported and exported energy to the CAES.

References

[1] M. MollahassaniPour, I. Taheri, M. Hasani Marzooni, "Assessment of transmission outage Contingencies' effects on bidding strategies of electricity suppliers," Int. J. Electr. Power Energy Syst., 120: 106053, 2020.

[2] S. Dorahaki, M. Rashidinejad, M. Mollahassani-pour, A. Bakhshai, “An efficient hybrid structure to solve economic-environmental energy scheduling integrated with demand side management programs,” Electr. Eng., 101(4): 1249–1260, 2019.

[3] M. Mohammadi, Y. Noorollahi, B. Mohammadi-ivatloo, H. Yousefi, “Energy hub: From a model to a concept – A review,” Renew. Sustain. Energy Rev., 80: 1512–1527, 2017.

[4] W.S. Ho, S. Macchietto, J.S. Lim, H. Hashim, Z.A. Muis, W.H. Liu, “Optimal scheduling of energy storage for renewable energy distributed energy generation system,” Renew. Sustain. Energy Rev., 58: 1100–1107, 2016.

[5] S. Dorahaki, R. Dashti, H.R. Shaker, “The optimal energy management in the smart microgrid considering demand response program and energy storage,” in Proc. 2019 International Symposium on Advanced Electrical and Communication Technologies (ISAECT): 1–6, 2019.

[6] S. Dorahaki, M. Rashidinejad, S.F. Fatemi Ardestani, A. Abdollahi, M.R. Salehizadeh, “A home energy management model considering energy storage and smart flexible appliances: A modified time-driven prospect theory approach,” J. Energy Storage, 48: 104049, 2022.

[7] I.T. Emami, M. MollahassaniPour, M.R.N. Kalhori, M.S. Deilami, “Short-run economic–environmental impacts of carbon tax on bulk electric systems,” Sustain. Energy, Grids Networks, 26: 100480, 2021.

[8] S. Dorahaki, M. Rashidinejad, H. Farahmand, M. MollahassaniPour, M. Pourakbari-Kasmaei, J.P.S. Catalao, “An optimal flexible partitioning of smart distribution system considering electrical and gas infrastructure,” in Proc. 2021 IEEE Madrid PowerTech, 2021.

[9] M.A. Mirzaei et al., “An integrated energy hub system based on power-to-gas and compressed air energy storage technologies in presence of multiple shiftable loads,” IET Gener. Transm. Distrib., 2020.

[10] M. Mollahassani-pour, M. Rashidinejad, A. Abdollahi, “Spinning reserve contribution using unit responsibility criterion incorporating preventive maintenance scheduling,” Int. J. Electr. Power Energy Syst., 73: 508–515, 2015.

[11] A. Dini, A. Hassankashi, S. Pirouzi, M. Lehtonen, B. Arandian, A.A. Baziar, “A flexible-reliable operation optimization model of the networked energy hubs with distributed generations, energy storage systems and demand response,” Energy, 239: 121923, 2022.

[12] Y. Cao, Q. Wang, J. Du, S. Nojavan, K. Jermsittiparsert, N. Ghadimi, “Optimal operation of CCHP and renewable generation-based energy hub considering environmental perspective: An epsilon constraint and fuzzy methods,” Sustain. Energy, Grids Networks, 20: 100274, 2019.

[13] Z. Zeng, T. Ding, Y. Xu, Y. Yang, Z. Dong, “Reliability evaluation for integrated power-gas systems with power-to-gas and gas storages,” IEEE Trans. Power Syst., 35(1): 571–583, 2020.

[14] L. Chen, T. Zheng, S. Mei, X. Xue, B. Liu, Q. Lu, “Review and prospect of compressed air energy storage system,” J. Mod. Power Syst. Clean Energy, 4(4): 529–541, 2016.

[15] M. Ban, J. Yu, M. Shahidehpour, Y. Yao, “Integration of power-to-hydrogen in day-ahead security-constrained unit commitment with high wind penetration,” J. Mod. Power Syst. Clean Energy, 5(3): 337–349, 2017.

[16] A. Mazza, F. Salomone, F. Arrigo, S. Bensaid, E. Bompard, G. Chicco, “Impact of Power-to-Gas on distribution systems with large renewable energy penetration,” Energy Convers. Manag. X, 7: 100053, 2020.

[17] C. Wulf, J. Linssen, P. Zapp, “Power-to-gas—concepts, demonstration, and prospects,” Hydrog. Supply Chain Des. Deploy. Oper., 2018: 309–345, 2018.

[18] S. Clegg and P. Mancarella, “Storing renewables in the gas network: modelling of power-to-gas seasonal storage flexibility in low-carbon power systems,” IET Gener. Transm. Distrib., 10(3): 566–575, 2016.

[19] Z. Soltani, M. Ghaljehei, G.B. Gharehpetian, H.A. Aalami, “Integration of smart grid technologies in stochastic multi-objective unit commitment: An economic emission analysis,” Int. J. Electr. Power Energy Syst., 100: 565–590, 2018.

[20] X. Luo, J. Wang, M. Dooner, J. Clarke, C. Krupke, “Overview of Current development in compressed air energy storage technology,” Energy Procedia, 62: 603–611, 2014.

[21] W. Cai, R. Mohammaditab, G. Fathi, K. Wakil, A.G. Ebadi, N. Ghadimi, “Optimal bidding and offering strategies of compressed air energy storage: A hybrid robust-stochastic approach,” Renew. Energy, 143: 1–8, 2019.

[22] A.L. Soyster, “Technical note—convex programming with set-inclusive constraints and applications to inexact linear programming,” Oper. Res., 21(5): 1154–1157, 1973.

[23] Y. Zhang, N. Gatsis, G.B. Giannakis, “Robust energy management for microgrids with high-penetration renewables,” IEEE Trans. Sustain. Energy, 4(4): 944–953, 2013.

[24] M. Yan, N. Zhang, X. Ai, M. Shahidehpour, C. Kang, J. Wen, “Robust two-stage regional-district scheduling of multi-carrier energy systems with a large penetration of wind power,” IEEE Trans. Sustain. Energy, 10(3): 1227–1239, 2019.

[25] M. Alipour, K. Zare, H. Zareipour, H. Seyedi, “Hedging Strategies for heat and electricity consumers in the presence of real-time demand response programs,” IEEE Trans. Sustain. Energy, 10(3): 1262–1270, 2019.

[26] M. Nazari-Heris, B. Mohammadi-Ivatloo, G.B. Gharehpetian, M. Shahidehpour, “Robust short-term scheduling of integrated heat and power microgrids,” IEEE Syst. J., 13(3): 3295–3303, 2019.

[27] S. Dorahaki, R. Dashti, H.R. Shaker, “Optimal outage management model considering emergency demand response programs for a smart distribution system,” Appl. Sci., 10(21): 7406, 2020.

[28] S. Dorahaki, M. Rashidinejad, S.F.F. Ardestani, A. Abdollahi, M.R. Salehizadeh, “A Peer-to-Peer energy trading market model based on time-driven prospect theory in a smart and sustainable energy community,” Sustain. Energy, Grids Networks: 100542, 2021.

[29] S. Dorahaki, A. Abdollahi, M. Rashidinejad, M. Moghbeli, “The role of energy storage and demand response as energy democracy policies in the energy productivity of hybrid hub system considering social inconvenience cost,” J. Energy Storage, 33: 102022, 2021.

[30] M. Ghahramani, M. Nazari-Heris, K. Zare, B. Mohammadi-ivatloo, “Robust Short-term scheduling of smart distribution systems considering renewable sources and demand response programs,” Robust Optimal Planning and Operation of Electrical Energy Systems, Cham: Springer International Publishing, 2019, pp. 253–270.

Biographies

Sobhan Dorahaki received M.Sc. degrees in Electrical Engineering from Shahid Bahonar University of Kerman, Kerman, Iran in 2017. He is currently pursuing a Ph.D. degree in the Faculty of Electrical and Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. His research interests include optimization, energy hub modeling, and Peer-to-Peer energy trading.

- Email: Sobhandorahaki@gmail.com
- ORCID: 0000-0003-1899-4210
- Web of Science Researcher ID: AAQ-8942-2021

- Scopus Author ID: 57193502686
- Homepage: NA

Seyedeh Soudabeh Zadsar received the B.S. degree in Electrical Engineering from Shahid Bahonar University of Kerman, Iran, in 2011, and M.S. degree from University of Kerman Graduate University of Technology, Iran in 2016. She is currently a Ph.D. student in the Department of Electrical Engineering, Shahid Bahonar University of Kerman, Iran. Her current research interests include smart grids, energy hub systems, renewable energy, and optimization algorithms.

- Email: zadsars@yahoo.com
- ORCID: 0000-0001-7522-1551
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Masoud Rashidinejad received his B.Sc. degree in Electrical Engineering and M.Sc. degree in Systems Engineering from the Isfahan University of Technology, Isfahan, Iran. He received his Ph.D. degree in Electrical Engineering from Brunel University, London, UK, in 2000. He is currently a professor in the Department of Electrical Engineering, Shahid Bahonar University of Kerman, Kerman, Iran. His areas of interest are power system optimization, power system planning, electricity restructuring, and energy management

- Email: mrashidi@uk.ac.ir
- ORCID: 0000-0002-3046-7088
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://elec.uk.ac.ir/en/~mrashidi

Mohammad Reza Salehizadeh received the B.Sc. degree from the Power and Water University of Technology, Tehran, Iran, in 2003 and the M.S. degree from Shahrood University of Technology, Shahrood, Iran, in 2004 and a Ph.D. degree from Islamic Azad University, Science & Research Branch, Tehran, Iran, 2014, all in electrical engineering. He is currently an Assistant Professor of electrical engineering and the head of the Electrical Engineering Department in Islamic Azad University, Marvdasht Branch, Iran. He was also the head of industry and society relations office for three years at Islamic Azad University, Marvdasht Branch. His research interests include power system operation and planning, Electricity market modeling and bidding strategies in dynamic energy markets, game theory and machine learning, Multi-criteria decision-making approaches, optimal control, and smart grid. He serves as an invited reviewer for several journals in the area of power and energy systems. He was also a guest editor in Energies-MDPI, IET Smart cities, and Sustainable Energy Grids and Networks in Elsevier.

- Email: salehizadeh@miau.ac.ir
- ORCID: 0000-0002-1708-6862
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

How to cite this paper:

S. Dorahaki, S.S. Zadsar, M. Rashidinejad, M.R. Salehizadeh, “Robust scheduling of water and energy hub considering CAES, power-to-gas units, and demand response programs” J. Electr. Comput. Eng. Innovations, 10(2): 371-380, 2022.

DOI: 10.22061/JECEI.2022.8489.519

URL: https://jecei.sru.ac.ir/article_1677.html

SRTTU

Journal of Electrical and Computer Engineering Innovations

(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

A Survey of Deep Learning Techniques for Maize Leaf Disease Detection: Trends from 2016 to 2021 and Future Perspectives

H. Nunoo-Mensah[1,], S. Wewoliamo Kuseh[1], J. Yankey[1], F. A. Acheampong[2]*

[1]*Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.*

[2]*Department of Computer Science and Technology, University of Electronic Science and Technology of China, Chengdu, China.*

Article Info

Article History:
Received 14 November 2021
Reviewed 16 December 2022
Revised 04 February 2022
Accepted 08 February 2022

Keywords:
Convolutional neural network
Deep learning
Maize disease
Transfer learning

*Corresponding Author's Email Address:

hnunoo-mensah@knust.edu.gh

Abstract

Background and Objectives: To a large extent, low production of maize can be attributed to diseases and pests. Accurate, fast, and early detection of maize plant disease is critical for efficient maize production. Early detection of a disease enables growers, breeders and researchers to effectively apply the appropriate controlled measures to mitigate the disease's effects. Unfortunately, the lack of expertise in this area and the cost involved often result in an incorrect diagnosis of maize plant diseases which can cause significant economic loss. Over the years, there have been many techniques that have been developed for the detection of plant diseases. In recent years, computer-aided methods, especially Machine learning (ML) techniques combined with crop images (image-based phenotyping), have become dominant for plant disease detection. Deep learning techniques (DL) have demonstrated high accuracies of performing complex cognitive tasks like humans among machine learning approaches. This paper aims at presenting a comprehensive review of state-of-the-art DL techniques used for detecting disease in the leaves of maize.

Methods: In achieving the aims of this paper, we divided the methodology into two main sections; Article Selection and Detailed review of selected articles. An algorithm was used in selecting the state-of-the-art DL techniques for maize disease detection spanning from 2016 to 2021. Each selected article is then reviewed in detail taking into considerations the DL technique, dataset used, strengths and limitations of each technique.

Results: DL techniques have demonstrated high accuracies in maize disease detection. It was revealed that transfer learning reduces training time and improves the accuracies of models. Models trained with images taking from a controlled environment (single leaves) perform poorly when deployed in the field where there are several leaves. Two-stage object detection models show superior performance when deployed in the field.

Conclusion: From the results, lack of experts to annotate accurately, Model architecture, hyperparameter tuning, and training resources are some of the challenges facing maize leaf disease detection. DL techniques based on two-stage object detection algorithms are best suited for several plant leaves and complex backgrounds images.

Introduction

The importance of maize production cannot be overemphasized. Maize is ranked among the most heavily grown and consumed cereals in the world [1],

Doi: 10.22061/JECEI.2022.8602.531

[2]. However, the contributions of maize to people's economic well-being are hindered by disease-ridden crops that affect the yield, thus reducing income and affecting food security. It is estimated that about 14 million tonnes of maize were lost to Northern leaf Blight (NLB) in the United States between 2012 and 2015, which translates to about $1.9 billion [3]. The gravity of the situation warrants those efficient measures to control or mitigate any threat that may hamper the growth and production of maize are found. These control or mitigation strategies must be proposed to ensure food security globally.

In the hope to mitigate disease infestation on maize fields, preventive techniques need to be employed. Nevertheless, plant diseases can still strike even when all preventive protocols are in force. Therefore, it is imperative to know that an accurate diagnosis of a disease is an essential first step to timely control plant diseases. Plant diseases have long been studied. There are well-established control mechanisms for controlling plant disease, that is if they are detected early enough. The timely diagnosis and classification of plant diseases is a critical aspect in preventing yield loss and improving product quality [4], [5]. A relevant characteristic of a sound disease detection system will be its ability to identify early signs and symptoms of the disease. Notably, the system must include containment strategies that prevent or limit the disease spread once it is detected [6]. Plant disease phenotyping is one crucial process that allows early detection of a particular kind of plant diseases, enabling growers, breeders, and researchers to effectively apply the appropriate control measures to mitigate the disease's effects.

Plant phenotyping in the past involved human experts visually inspecting diseased plants to observe defects in various parts of the plant: leaves, stems, roots, in other to predict the presence of a particular kind of disease [4]. This detection technique by human experts is often time-consuming, subject to erroneous decisions, and impractical for largescale fields [7], [8]. Microscopic evaluation of morphology features like spores, mycelium to identify pathogens is another plant disease detection technique in literature [9]. Computer vision and machine learning can solve these issues by enabling high accuracy and scalable plant phenotyping. Recent techniques have focused on using automated systems to detect plant diseases in agriculture accurately. Computer-aided methods combined with crop images (image-based phenotyping) have become very dominant for plant disease detection [10]. Numerous image-based plant disease detection techniques have been developed, which shows better accuracy and precision than visual inspection [8]. Machine learning (ML) has been applied to many computer vision problems, including face recognition, speech processing, and disease tissue classification in medicine. The success of ML techniques is as a result of their ability to identify a hierarchy of features and generalized trends from available data [11].

In narrowing down on machine learning approaches, deep learning techniques have demonstrated high accuracies of performing complex cognitive tasks like humans [7]. Deep Learning (DL) is the state-of-the-art ML approach widely used to address problems in health care, agriculture, audio and speech processing [12]. Convolutional Neural Networks (CNNs) are state-of-the-art deep learning algorithms used to address computer vision problems recently, especially image classification tasks. Traditional ML approaches require a manual selection of features that are thought to be helpful in a classification task. However, CNNs can learn which features are most important and which are not. The usage of DL in agriculture and plant disease detection have proven to give very high accuracies enabling better agriculture and crop management quality [13].

This survey aims to present a comprehensive review of state-of-the-art deep learning techniques used for detecting disease in the leaves of maize. The survey documents all relevant proposals in the domain to enable readers to understand maize disease detection using deep learning methods proposed from 2016 to 2021. Most recent survey papers mainly focus on plant disease detection. Which encompasses many plants and not necessarily maize thus do not provide an in-depth discourse of the subject matter. This paper will act as a primary source for discussing maize leaf disease detection using deep learning methods to the best of our knowledge. The paper details concepts, approaches, available datasets, and the strengths or shortfalls of DL techniques for maize leaf disease detection.

The remainder of the paper is organized as follows. Next Section discusses the concept of deep learning, transfer learning and highlights some plant leaf disease datasets. In third Section, the methodology used in acquiring the candidate papers for review have been highlighted and detailed review of deep learning-based proposals for maize leaf disease detection is outlined. Open issues and future research directions are provided in fourth Section. Conclusions are made fifth Section.

Deep Learning and Plant Leaf Disease Datasets

This section discusses the concept of deep learning. It also elucidates some models that have been adopted for transfer learning. Finally, the section provides dataset used in the design of plant leaf disease detection to serve as a primer for new researchers in the field.

A. Deep Learning

Deep learning is recently gaining popularity and momentum because of its success in various applications. Deep learning is a sub-field of machine learning. It extends classical Machine learning by adding more depth (layers) to a model. Successive layered learning or a hierarchical way of representing data abstractly emphasizes deep learning. Deep learning does not mean any more profound understanding for using this approach. Instead, it refers to the successive layers of representation. The depth of the model is characterized by the number of layers in the model [14]. Modern deep learning approaches consist of tens or even hundreds of successive layers for data representation. All these layers learn automatically from data. These layers learn through neural networks models; the layers are stacked on top of each other. Figure 1 illustrates the basic layered structure of a deep learning model [15].

B. Convolutional Neural Networks

Convolutional neural networks have become very dominant in the field of deep learning and are the approach used for visual object recognition and other computer vision problems. CNN was first introduced over twenty years ago. However, they have become widely used today due to improvements in hardware and the development of very deep CNNs. CNNs are not only applied to images but show better results in speech recognition, and natural language processing problems [16]. Convolution is the essential operation of a convolutional neural network (CNN). This convolution operation is achieved by applying filters (also known as kernels) to input data, mostly an image. Convolutional filters are composed of two-dimensional matrices of real values: the dimensions of a filter are smaller than the dimensions of the input data used in training. The aim of convolution operations is to extract features from an input image and thereby preserving the spatial relationship between pixels [17].

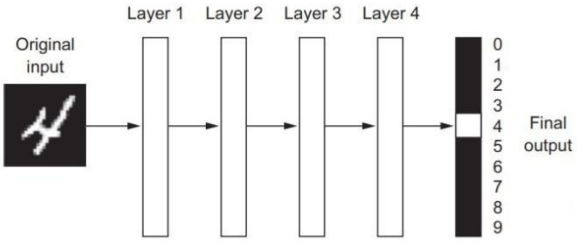


Fig. 1: Example of digit classification using deep learning [15].

A standard convolutional neural network structure comprises several essential building blocks that represent the layers of the network. The number of layers and combinations of building blocks varies depending on the architecture. Fig. 2 represent a standard convolutional neural network which consists of convolutional layers (Conv layers) classification layer (Softmax layer), fully-connected layers (FC layers), and compression layers (Pool layers) [18].

Feature maps (output) from previous layers are convolved with distinct filters and a scalar product is calculated over the entire length and the width of the given filter. The output of the filters is then pass through either a linear or most of the time a non-linear function [19]. It is very important to select good kernels to be able to capture salient and important information from the data. This allows for strong inferences about the content of the input data [20]. The result of the convolution operation is the output feature maps which the filters find.

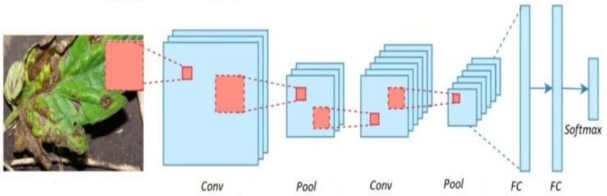


Fig. 2: CNN for plant disease detection [18].

The compression layer is a filter that non-linearly reduces the number of pixels, or compresses image dimension (down-sampling). This filter does not contain learned weights. The pooling layer's is responsible for secondary feature extraction by reducing the dimensions of the feature maps. It also increases the robustness of feature extraction.

For convolutional neural networks there is normally one or two fully-connected layers used as classifiers. All neurons in this layer are connected to all neurons in the previous layer. There is a last fully-connected layer before the output layer. In classification problems, the output of the convolutional neural network is reduced to activation function, the most commonly used function is Softmax. This is because it generates a well-performed distribution of the outputs. Support vector machines (SVM) can also be combined with CNNs for classification problems.

Activation functions play a very important role in the success of training deep neural networks. The role of activation function is to “mimic” the behaviour of a biological neuron by deciding if a neuron should turn off or on. Most commonly used and successful activation function is the Rectifier linear unit (ReLU). ReLu is very simple and effective and has become the default activation function used in deep learning. Other activations functions have been proposed to replace ReLu but the performance improvements tend to be inconsistence with different models and datasets. Other derivatives of ReLu are: Parametric ReLu (PReLu), Exponential linear unit (ELU) and Leaky ReLu (LReLu).

C. Modern Architectures and Transfer Learning

Deep learning algorithms, unlike typical machine learning algorithms, can automatically extract features either through semi-supervised or unsupervised learning and attempt to learn high-level features from huge amounts of data. One challenge about deep learning is the massive dependency on data because it needs large data to better understand patterns in data [21].

Transfer learning uses knowledge from a source domain to improve the learning ability of a target domain by transferring information between the two domains. One important requirement that will enable successful knowledge transfer is that both the source and target domains should be related closely [22]. Transfer learning is needed where there is limited amount of targeted training data: this may as result of expensive data collection and labelling, data being rear or data being inaccessible. Transfer learning has been applied successfully in many applications including image classification, software defect classification, text sentiment classification [21].

As early as 2012, deep neural networks were achieving significant results in tasks classification and detection of objects over large image datasets. For example, in the likes of the ImageNet (ImageNet Large Scale Visual Recognition Challenge) competitions. ImageNet image dataset has more than 20; 000 categories (classes) with over 80 million images. From 2012 to 2017, when the last competition was held, the winning architects were convolutional neural networks. It was the first time a deep learning technique, i.e., convolutional neural networks, showed a significant improvement over previous results obtained by standard machine learn ing techniques and manual processing of features. Over time, these architects have become more successful than man himself in tasks classifications and detection over the ImageNet image dataset. Thus, these architectures have become standard architectures that have proven successful not only over the ImageNet dataset but on significantly wider range of problems. This is ensured through the tech- niques of transfer learning or as a basis or idea for new architectures. Some of these modern architectures include: GoogleNet [23], AlexNet [24], ResNet [25], and VGGNet [26], DenseNet [27], EfficientNet [28] etc.

D. Datasets

Automated diagnosis and identification of plant diseases may allow for more rapid advances in plant breeding as well as easier monitoring of farmers' fields. However, given the multiple differences in lighting and direction, it is challenging for a simple algorithm to differentiate between the specific disease and other causes of dead plant tissue in a normal field. A vast amount of high-quality human-generated training data is required to train a machine learning system to accurately detect a certain disease from photographs obtained in the field. Therefore, datasets become an integral part of any machine leaning algorithm, because the amount and quality of the dataset goes a long way to affect the performance of an algorithm. This section of the review takes a look at some available datasets for plants disease detection.

The largest public database of leaf images is PlantVillage, [29]. collected and maintained by a non - profit project run by Penn State University in United States and EPFL in Switzerland. The database consists of 54309 pictures, of 14 types of plants, divided into 38 classes (healthy and diseased leaves). These, however, were captured with detached leaves on a simple background, and CNNs trained on them can't perform well on field photos. Fig. 3 shows examples of images from each class of the PlantVillage dataset.

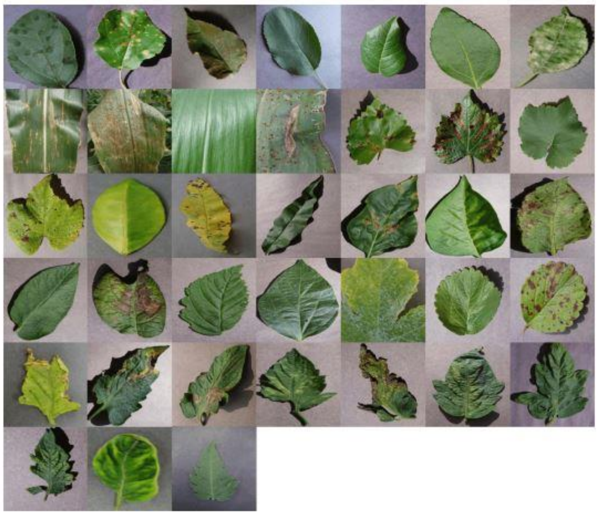


Fig. 3: Sample images of from PlantVillage dataset [29].

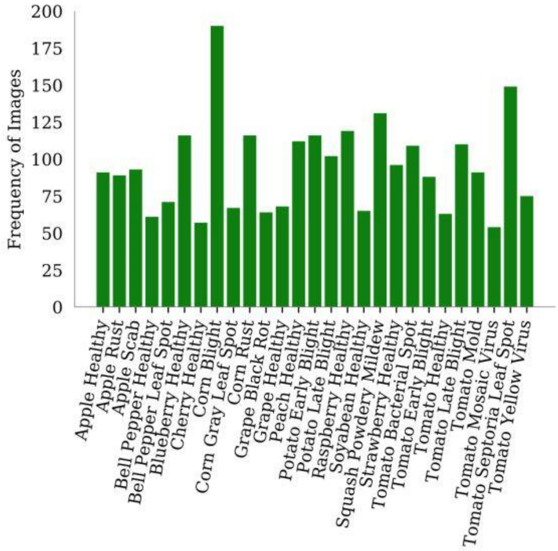


Fig. 4: Statistics from the PlantDoc dataset [30].

PlantDoc by Singh et al. [30] contains 2; 598 data points in total over 27 classes; 17 diseases and 10 healthy classes the dataset authors purport that

PlantDoc is a first of its kind that contains non-controlled image settings. This is envisaged to enhance performance of trained models used in practical real-life applications. Fig. 4 shows statistics of the PlantDoc dataset with a sizeable maize content

The dataset by Wiesner-Hanks et al. [31] is made up of full images of maize leaves shot in three different ways: using a handheld camera which produced 1787 images with 7669 annotations, camera mounted on a boom consisting of 8766 images with 55; 919 annotations and those pictures taken by a drone which is made up of 7669 images with 42; 117 annotations. There is no way to indicate the confidence of annotations. Some lesions are easily visible, which others are partially or entirely occluded from the focal plane. Other factors affecting the confidence of annotation are a heavy shade or being washed out by bright sunlight. It was reported that even experts found it difficult to distinguish between NLB and similar-looking diseases. The generalizability of the data is affected since image samples were taken in a single field in New York State. However, symptoms of the same disease can present or develop differently. Thus, the performance is hindered by the limitations mentioned earlier.

Since popular datasets cut across multiple plant leaves, the CMLD dataset [32] on the other hand, combines PlantVillage and PlantDoc maize or corn related data points to form a new dataset. Unrelated data samples were ignored in the creation of this new hybrid dataset. CMLD contains 4188 data points in total. The distributions are 1306 images for common rust, 574 images for grey leaf spot, 1146 images for blight, and 1162 healthy images.

Existing Deep Learning-based Proposals for Detecting Maize Leaf Disease

A. Methodology for Selecting State-of-the-Art Models

Following research works done by [33]-[35] a search was done in the following databases: IEEE Xplore, Scopus, ResearchGate, and Google Scholar. The keywords used in searching for articles were: "plant leaf disease detection", "Maize leaf disease detection", and "deep learning-based maize leaf detection". The year range was limited to 2016 - 2021. The procedure for selecting the existing candidate works for this study is presented in Algorithm 1. The search results in IEEE Xplore using the keys words shows that in 2018, 77 research papers were published. Out of these three were for maize leaf diseases detection and none for Deep learning techniques for maize leaf detection There was an increase in the number of conducted research in the years 2019 and 2020. The publications for 2019 and 2020 were 148 and 208, respectively. This shows the growing interest in plant leaf disease detection and the desire to maximize gains in agriculture. However, the number of publications involving maize leaf disease detection remains at three. Deep learning techniques for maize leaf detection in 2019 and 2020 were one and two, respectively, in the IEEE Xplore. As of July 2021, there were 91 publications, with only one involving maize leaf detection using deep learning techniques.

Algorithm 1: Article Selection for the Survey

```
Result: researchArticles
databases = ["IEEE Xplore", "Scopus",
 "ResearchGate", "Google Scholar"];
keywords = ["plant leaf disease detection", "maize
 leaf disease detection", "deep learning-based maize
 leaf detection"];
for database in databases do
    for keyword in keywords do
        if publication_year >= 2016 && <= 2021 then
            return researchArticles;
        else
            return None;
        end
    end
end
```

B. Current State-of-the-Art Models

There have been many approaches and techniques to detect maize disease in plants accurately. This section of the review highlights DL approaches used, datasets used in the study, contributions and limitations of the existing DL techniques. This review seeks to review state-of-the-art works from 2016 to 2021.

Richey et al. [36], used supervised transfer learning: based the ResNet50 model for the identification of Northern Corn Leaf Blight disease in maize plants. Two publicly available datasets were used for training and validation. Tensorflow with Keras high-level API public deep learning libraries are used. The model performance included an F1 score of 0.99, Accuracy of 0.99 and precision of 0.98 and a Recall of 1.00. The model was then served to a mobile application for practical field purposes. Esgario et al.

Esgario et al. [37] also used transfer learning for classification and severity estimation of four biotic stresses in coffee leaves. Two different datasets were generated, leaf dataset and symptom datasets using standard and mixup image augmentation techniques. AlexNet, GoogleNet, VGG19, Resnet50, MobileNetV2 were trained using single-task and multi-task CNN architectures. The GoogleNet, ResNet50, and AlexNet performed better with multi-task learning. ResNet50 achieved the best results. Multi-task learning made learning much faster because only a single model was trained.

A limitation identified by the authors involved the low representativity of the dataset that covered only the principal biotic stresses that affect coffee trees.

Nevertheless, this could be improved by increasing the number of images, thus adding new kinds of stress to the dataset.

Sambuddha et al. [7] proposed an explainable machine learning framework for the identification of stress in soybean with remarkable accuracy. Their proposed xPlNet framework comprised two main phases: the deep convolutional neural network (DCNN) and explanation phasesThe classification accuracy achieved by their model was 94.13%.

Xihai et al. [38] used improved deep learning models, GoogleNet and CIFAR10 (transfer learning), to identify disease in leaves of maize plants. The GoogleNet model achieved the highest accuracy of 98.9% as compared to the CIFAR10 of 98.8% accuracy. However, the authors claimed that with their improved CIFAR10, the testing accuracy could be improved by 0:7% and the loss reduced by 10:2%.

Wu et al. [39] proposed a three-stage pipeline CNN to detect the presence of NLB in field images of maize plants using images acquired from an unmanned aerial vehicle (UAV). Their model achieved an accuracy of 95.1%. Liang et al. [10] proposed a Deep Convolutional Neural for the detection of rice blast disease. Three feature extraction methods were used for feature extraction: Convolutional neural network (CNN), Harr-wavelet (Haar-WT), and local binary pattern histograms (LBPH). Using t-Distributed Stochastic Neighbor Embedding (t-SNE) as a criterion for evaluating the performance of the three feature extractors, CNN showed a better performance than the other two handcrafted features. Support Vector Machines (SVM) when combined with all the feature extractors for classification, the CNN-based feature extractor shows far superior performance than LBPH and Haar-WT. The results showed that the quantitative analysis accuracy, receiver operating characteristic curve (ROC), and area under the ROC Curve (AUC) agreed with qualitative analysis using t-SNE. CNN and

CNN+SVM showed superior performance than LBPH+SVMand Haar-WT+SVM. It can, however, be alluded to that CNN and CNN+SVM are better for rice blast identification. CNN+SVM was a solid competitor for rice blast detection, but the latter is preferred. Their technique was limited to detection and failed to address the issue of the severity of the disease.

Panigrahi et al. [40] proposed a CNN-based model for the detection of three major corn diseases: northern leaf blight, common rust, and Cercospora leaf spot. The authors used images from the PlantVillage dataset for the study. A CNN model was proposed, which consisted of 3 convolutional layers and two fully connected dense layers. The dropout layer is used to prevent overfitting. The proposed model achieved an accuracy of 98.78% with less convergence time. Sibiya and Sumbwanyambe [41] designed a CNN model for detecting leaf disease of corn plants using a Java-based neural network framework, Neuroph. The training set included personal images captured from the field and the PlantVillage dataset. Their model had 50 hidden layers of CNN built for the classification of three maize diseases. The overall accuracy of the CNN was 92.85% but achieved individual accuracies ranging from 87% to 99.9%. The model could easily overfit because small amounts of data points were used in training.

Garg et al. [42] proposed a deep framework (cascaded CNN) to detect and automatically quantify the presence of a disease in plants. The model was trained using field images captured by using unmanned aerial vehicles (UAVs). Their framework extracted phenotypic traits to detect and estimate the severity of a leaf disease at the leaf level. The results of their experiment gave a severity correlation of 73%. A modified LeNet architecture proposed by Priyadharshini et al. [43] for the classification of three-leaf diseases of maize is discussed. The study was carried out by using images of maize leaves from the PlantVillage dataset. Principal component analysis (PCA) was used for preprocessing. To improve the classification accuracy of their proposed model, the authors adjusted the framework by varying the depth and kernel size. The model accuracy was 97.89%. The simulation results for maize leaf disease classification demonstrated the proposed method's potential in maize disease classification.

Richey and Shirvaikar [31] used an object detection algorithm, YoloV4, to detect the presence of NLB in the leaf of maize plants. Using a subset of the publicly available dataset by Wiesner-Hanks et al. [44] with augmentation techniques, 5699 images and a test set of 1251 images were used. Evaluating the model based on Intersection-over-Union (IoU) and Mean Average Precision (mAP), the model reported a 93.55%mAP with an average of 77.13% IoU.

Bhatt et al. [45], proposed a novel CNN technique for classifying corn leaves into Healthy, Common Rust, Late Blight, and Leaf Spot by using adaptive boosting and other classifiers to train on features from four CNN architectures (i.e., VGG-16, ResNet-50, Inception-v2, and MobileNetv1). Adaptive Boosting assisted the classifiers in developing a solid rule for class labels. An accuracy of 98% was achieved together with classification scores of 0:97, 0:98, 0:97 for precision, recall, and f1-score, respectively.

To improve the performance of CNNs in the detection of maize leaf disease classification, da Rocha et al. [46] used Bayesian optimization to help find optimal hyperparameter for training using the PlantVillage dataset. The significant contribution of their study was finding the best hyperparameters using Bayesian optimization. The authors employed K-fold cross-validation for training three CNN architectures: AlexNet, SqueezeNet and ResNet-50. Interestingly the three models obtained 97% accuracy, indicating that optimization produced improved generalization throughout all the models.

Waheed et al. [47], using an optimized DenseNet architecture proposed a novel technique for the recognition and classification of three maize leaf diseases. In determining optimal hyperparameter values, the authors used a grid search to find these optimal values. However, but may present a curse of dimensionality. DenseNet uses significantly fewer parameters as compared to other CNN architectures used in the experiment. Experimental results showed that DenseNet achieved an accuracy of 98:06% with fewer parameters and training time.

Lin et al. [48] proposed a novel multi-channel convolutional neural network (MCNN) to improve the identification of five maize leaf diseases. Their proposal employed techniques used in video saliency detection that imitates human visual behavior. Their model achieved an average accuracy of 92:31%,

Liu et al. [49] used transfer learning based on EfficientNet for the automatic recognition of maize leaf diseases. The model parameters trained on the ImageNet dataset were maintained during training, and the fully connected layers and Softmax were optimized. Images collected from the internet were used for training. The training speed was significantly improved with a recognition accuracy of 98:52%.

A transfer learning approach based on the Inceptionv3 and Inception-v4 approach was designed by Sun et al. [50] to classify maize diseases. The pre-trained model was fine-tuned, providing a new approach for maize disease identification. The dataset used was from AI challenger and consisted of eight categories. An experimental result indicated that transfer learning could help reduce the training time of the network.

Syarief and Setiawan [51] analyzed four classes of diseased maize leaf images using seven CNN architectures and three classification methods. The data was obtained from the PlantVillage dataset. The best classification method identified by the authors were AlexNet and SVM, with an accuracy of 93:5%.

Sumita et al. [52] proposed a real-time deep learning-based model that is deployed onto a raspberry pi for identifying and classifying major corn diseases. The bulk of the dataset used is from PlantVillage dataset, but few images were captured from corn plantations. Live images of an infected or healthy corn plant are captured by a Smartphone camera and sent to the raspberry pi for processing through a Wi-Fi network. The average accuracy of the model is 98:40%, but the accuracy reduces to 88:66% when deployed.

Tian et al. [53] also proposed a multi-layer deep neural network for the recognition of six different diseases of corn plants. Dataset used in the study is from experimental fields. VGG-16 is used for feature extraction. Smut and rust disease achieved 100% accuracy but with an overall accuracy of 96:8%. Several methods for classifying plant diseases that can learn from small amounts of data are proposed in [54].

PlantVillage dataset and coffee leaf datasets were used in the study. Transfer learning, triplet networks, and Deep Adversarial Metric Learning (DAML) [55] are the main building blocks of these methods. Very high accuracy of 99% was achieved, thus demonstrating the efficiency of transfer learning.

A summary of the discussed state-of-the-art models and proposals have been outlined in Table 1.

Open Issues and Future Research Directions

This section highlights some challenges in plant disease detection. It outlines some directions for future research in using deep learning techniques for plant disease identification and detection in intelligent agriculture, especially diseases in cereal crops.

From the discussions mentioned earlier, it can be found that one of the challenges facing plant disease detection is the lack of experts to annotate accurately. The problem arises when experts cannot rightly differentiate between dead tissues and diseases when compiling a dataset. This task requires experts and experienced professionals to identify plant diseases that are difficult and costly, especially for new or rare diseases. Furthermore, crop diseases vary in severity. The data collection is unquestionably important when using deep learning technologies to identify crop pests and illnesses.

Model architecture, hyperparameter tuning, and training resources also throw another challenge in plant disease detection. Shallow architectures are best suited for small datasets. Most recent models for object detection offer another angle to consider in selecting or building a model for disease detection and classification. The adaptive boosting (AdaBoost) technique is a choice to be considered to enhance the performance of detection models. Most DL techniques are focused mainly on the detection and classification of maize leaf disease. The paper recommends that future research on maize leaf disease detection, classification, and quantification of disease severity will help improve smart agriculture. Quantification is an area that is least explored by researchers in the field but has the possibility of providing more insightful data for rapid decision-making during farming.

Table 1: A summary of the discussed state-of-the-art models and proposals have been outlined.

	Authors	DL Algorithm	Dataset	Contribution	Performance	Limitation
1	Afifi et al. [54]	▪ ResNet18, ResNet34, ResNet50, ▪ Triplet networks ▪ Deep Adversarial Metric Learning	PlantVillage	Demonstrates the efficiency of transfer learning for corn diseases detection	An accuracy of 99% was reported by the authors	The proposed models have low accuracy under varied conditions
2	Richey and Shirvaikar [31]	▪ YoloV4	Subset of [44]	The authors used an object detection algorithm for the detection of NLB in the leaf of maize plants Evaluation	Their model reported a 93:55% mAP with an average of 77:13% IoU.	Their work did not consider multiple regions of interest.
3	Kanish et al. [42]	▪ A self-trained cascaded CNN model	They captured field images using UAVs	The authors proposed a framework for the detection and estimation of leaf disease severity	Experiments gave a severity correlation of 73%.	The dataset used by the authors were not extensive thus the reduced accuracy levels.
4	Liu et al. [49]	▪ EfficientNet	The authors sourced images from the internet	The authors fine-tuned EfficientNet for the automatic recognition of maize leaf diseases	They achieved a recognition accuracy of 98:52%	The authors used non-standardized images for their dataset
5	Sun et al. [50]	▪ Inception-v3 ▪ Inception-v4	AI challenger	The authors leveraged transfer learning capabilities to aid in classifying maize leaf diseases.	The transfer learning procedure reduced the training time of the network significantly	The proposed framework might not perform well on images that contain several leaves, due to the kind of images used for training
6	Syarief and Setiawan [51]	▪ AlexNet ▪ VGG16 ▪ VGG19 ▪ GoogleNet ▪ Inception-V3 ▪ ResNet50 ▪ ResNet101	PlantVillage	The authors classified maize leaf diseases using pre-trained models	AlexNet achieved the best average classification accuracy of 93:5%.	The best results of AlexNet with an SVM classifier recorded lower accuracies than the state-of-the-art
7	Sumita et al. [52]	▪ Self-trained CNN model	PlantVillage	A real-time corn disease identification and classification using a Raspberry Pi was designed and implemented by authors.	An average accuracy of 98:40% was recorded during model training, however, the accuracy reduced to 88:66% when deployed	Model overfitting on training data likely cause for reduction in the implementation accuracy.
8	Panigrah et al. [40]	▪ Self trained CNN model	PlantVillage	The authors proposed a CNN-based model for the detection of three major corn diseases	Their proposed model achieved an accuracy of 98:78% with little convergence time.	The model can under-fit due to small data samples used for training.
9	Blake et al. [36]	▪ ResNet50	PlantVillage	The authors proposed a real-time maize disease detection model using transfer learning.	They achieved an accuracy of 99%.	Their model however, performed poorly on field images.
10	Esgario et al. [37]	▪ AlexNet ▪ GoogleNet ▪ VGG19 ▪ Resnet50 ▪ MobileNet-v2	The authors used images captured using smartphones	A multi-task learning technique for classification and severity estimation of four biotic stresses in coffee leaves were proposed by the authors.	ResNet50 achieved the highest results among the candidate models.	In-field images were not used which could have positively impacted the model.

11	da Rocha et al. [46]	▪ AlexNet ▪ SqueezeNet ▪ ResNet-50	PlantVillage	The authors determined the optimum hyper-parameters using Bayesian optimization for disease classification.	All three CNNs obtained a 97% accuracy.	Model was not tested with field data.
12	Waheed et al. [47]	▪ DenseNet	The authors used images manually gathered from different sources	A novel technique for recognition and classification of three maize leaf diseases was proposed by authors.	The DenseNet model used achieved an accuracy of 98:06% with less parameters and training time.	Their framework may present a curse of dimensionality.
13	Wu et al. [39]	▪ Self-trained CNN model	Images acquired from [31]	The authors proposed a three-stage pipeline CNN to detect the presence of NLB in field images of maize plants.	Their model achieved an accuracy of 95:1%.	Their proposed model cannot detect severity of NLB in maize plants.
14	Liang el al. [10]	▪ Self-trained CNN model	Images were acquired from the Institute of Plant Protection, Jiangsu Academy of Agricultural Sciences.	A deep convolutional neural network for the detection of rice blast disease was proposed by authors	CNN and CNN+SVM showed superior performance.	Reliability and robustness of the model needs improvements.
15	Sibiya and Sumbwanyambe [41]	▪ Self-trained CNN model	Field images + PlantVillage database	The authors proposed a CNN for the classification of three maize disease.	An overall accuracy of 92:85%.	The model can easily overfit due to the small amount of data used in training the model.
16	Priyadharshini et al. [43]	▪ Modified LeNet architecture	PlantVillage	The authors proposed a LeNet method's potential in maize leaf disease classification.	The reported model accuracy was 97:89%	It is expected that the model might perform poorly on field images; this is due to the controlled nature of the images used.
17	Bhatt et al. [45]	▪ VGG-16 ▪ ResNet-50 ▪ Inception-v2 ▪ MobileNet-v1	PlantVillage	The authors used a CNN for classifying corn diseases using adaptive boosting techniques.	An accuracy of 98% was achieved in their work	Some of models used in the ensemble had larger parameters and took longer periods to train. The accuracy of the ensemble was not verified with field image
18	Tian et al. [53]	▪ VGG16	Field images	A multi-layer deep neural network for the recognition of six different disease of corn was proposed by the authors.	They recorded an overall accuracy of 96:8%.	Their proposed method did not take into account the different characteristics of the plant at different stages of the diseased journey.

19	Lin et al. [48]	▪ Multichannel CNN	Images collected from maize planting bases	A novel multi-channel convolutional neural network (MCNN) for improving the identification of maize leaf disease was proposed by the authors.	An average model accuracy of 92:31% was recorded by the authors.	Their procedure involved complex image prepossessing steps.
20	Sambuddha et al. [7]	▪ Self-trainedCNN model	The authors used images from the field under controlled conditions.	An explainable ML framework for the identification of stress in soybean was proposed by the authors.	They achieved a classification accuracy of 94:13%	There was a high level of confusion among bacterial blight, bacterial pustule, and Septoria brown spots during the labeling process.
21	Xihai et al. [38]	▪ GoogleNet ▪ Cifar10 model	PlantVillage	The authors proposed an improved deep learning model based on transfer learning.	The GoogleNet model achieved the highest accuracy of 98:9%	The model required much training time and might not perform well on uncontrolled field conditions

Conclusion

The early detection of a plant disease enables stakeholders to apply the appropriate controlled measures to mitigate against the disease effectively. Recent techniques have focused on automated techniques using deep learning to detect diseases in maize plants accurately. This review details DL techniques that are used for automated maize leaf diseases detection and classification. The paper introduces plant disease detection and some of the shortfalls of traditional techniques used. Recent automated techniques, some essential datasets, and some DL architectures were also highlighted. The paper further gave a detailed account of recent DL techniques used to detect diseases in the leaves of maize plants and a discussion of their significant contributions and limitations. Challenges and future research directions in maize leaf disease detection are also presented in the paper.

Author Contributions

S. W. Kuseh and H. Nunoo-Mensah conceptualized, developed the algorithm for selecting the state-of-the-art DL techniques and wrote the initial draft of the paper. J. Yankey and F. A. Acheampong proofread and improved the structure of paper.

Acknowledgment

This study was supported by Kwame Nkrumah University of Science and Technology (KNUST) Engineering Education Project (KEEP) of the College of Engineering.

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

ML	Machine Learning
DL	Deep Learning
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
NLB	Northern Leaf Blight
LBPH	Local Binary Pattern Histogram
t-SNE	t-Distributed Stochastic Neighbor Embedding
SVM	Support Vector Machine
ROC	Receiver Operating Characteristics
UAV	Unmanned Aerial Vehicle

References

[1] C.A. Wongnaa, D. Awunyo-Vitor, A. Mensah, F. Adams, “Profit efficiency among maize farmers and implications for poverty alleviation and food security in ghana,” Sci. Afr., 6: e00206, 2019.

[2] B. Shiferaw, B. M. Prasanna, J. Hellin, M. Banziger, “Crops that feed the world 6. past successes and future challenges to the role played by maize in global food security,” Food Secur., 3(3): 307, 2011.

[3] S.P. Mohanty, D.P. Hughes, M. Salathe, “Using deep learning for image-based plant disease detection,” Front. plant scie., 7: 1419, 2016.

[4] C. Manjunath, H. Karthik, N. Reddy, S. Sahana, et al., “A review on detection of plant diseases through various methodologies,” Int. J. Res. Appl. Scie. Eng. Tech., 6(5): 74–80, 2018.

[5] A. Singh, B. Ganapathysubramanian, A.K. Singh, S. Sarkar, “Machine learning for high-throughput stress phenotyping in plants,” Trends plant scie., 21(2): 110–124, 2016.

[6] F. Martinelli, R. Scalenghe, S. Davino, S. Panno, G. Scuderi, P. Ruisi, P. Villa, D. Stroppiana, M. Boschetti, L.R. Goulart, et al., “Advanced methods of plant disease detection. a review,” Agron. Sustainable Dev., 35(1): 1–25, 2015.

[7] S. Ghosal, D. Blystone, A.K. Singh, B. Ganapathysubramanian, A. Singh, S. Sarkar, “An explainable deep machine vision framework for plant stress phenotyping,” PNAS, 115(18): 4613–4618, 2018.

[8] C. DeChant, T. Wiesner-Hanks, S. Chen, E.L. Stewart, J. Yosinski, M.A. Gore, R.J. Nelson, H. Lipson, “Automated identification of northern leaf blight-infected maize plants from field imagery using deep learning,” Phytopathology, 107(11): 1426 1432, 2017.

[9] A.K. Mahlein, “Plant disease detection by imaging sensors–parallels and specific demands for precision agriculture and plant phenotyping,” Plant dis., 100(2): 241–251,2016.

[10] W.J. Liang, H. Zhang, G.f. Zhang, H.x. Cao, “Rice blast disease recognition using a deep convolutional neural network,” Sci. rep., 9(1): 1–10, 2019.

[11] A. Singh, B. Ganapathysubramanian, A.K. Singh, S. Sarkar, “Machine learning for high-throughput stress phenotyping in plants,” Trends plant sci., vol. 21, no. 2, pp. 110–124, 2016.

[12] R.I. Hasan, S.M. Yusuf, L. Alzubaidi, “Review of the state of the art of deep learning for plant diseases: A broad analysis and discussion,” Plants, 9(10): 1302, 2020.

[13] A.K. Singh, B. Ganapathysubramanian, S. Sarkar, A. Singh “Deep learning for plant stress phenotyping: trends and future perspectives,” Trends plant sci., 23(10): 883–898, 2018.

[14] A. Kamilaris, F.X. Prenafeta-Boldu, “Deep learning in agricul- ture: A survey,” Comput. Electron. Agric., 147: 70–90, 2018.

[15] F. Chollet et al., Deep learning with Python, vol. 361. Manning New York, 2018

[16] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, “Deformable convolutional networks,” in Proc. IEEE International Conference On Computer Vision: 764–773, 2017.

[17] N.K. Manaswi, N.K. Manaswi, S. John, Deep learning with applications using python. Springer, 2018.

[18] 2020 [Online]. Available: https://missinglink.ai/guides/convolutional-neuralnetworks/convolutional-neural-network-architectureforging-pathways-future/..

[19] M.Z. Alom, T.M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M.S. Nasrin, M. Hasan, B.C. Van Essen, A.A. Awwal, V. K. Asari, “A state-of-the-art survey on deep learning theory and architectures,” Electronics, 8(3): 292, 2019.

[20] I. Hadji, R.P. Wildes, “What do we understand about convo lutional networks?,” arXiv preprint arXiv:1803.08834, 2018.

[21] K. Weiss, T.M. Khoshgoftaar, D. Wang, “A survey of transfer learning,” J. Big data, 3(1): 1–40, 2016.

[22] B. Tan, Y. Zhang, S. Pan, Q. Yang, “Distant domain transfer learning,” in Proc. the AAAI Conference on Artificial Intelli gence, 2017.

[23] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE conference on computer vision and pattern recognition: 1–9, 2015.

[24] A. Krizhevsky, I. Sutskever, G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Adv. neural inf. Process. Syst., 25(2): 1097–1105, 2012.

[25] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” in Proc. the IEEE conference on computer vision and pattern recognition: 770–778, 2016.

[26] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556, 2014.

[27] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, “Densely connected convolutional networks,” in Proc. the IEEE conference on computer vision and pattern recognition: 4700–4708, 2017.

[28] M. Tan, Q. Le, “Efficientnet: Rethinking model scaling for con volutional neural networks,” in Proc. International Conference on Machine Learning (PMLR): 6105–6114, 2019.

[29] D. Hughes, M. Salathe, et al., “An open access repository of images on plant health to enable the development of mobile disease diagnostics,” arXiv preprint arXiv:1511.08060, 2015.

[30] D. Singh, N. Jain, P. Jain, P. Kayal, S. Kumawat, N. Batra, “Plantdoc: a dataset for visual plant disease detection,” in Proc. the 7th ACM IKDD CoDS and 25th COMAD: 249–253, 2020.

[31] B. Richey, M.V. Shirvaikar, “Deep learning based real-time detection of northern corn leaf blight crop disease using yolov4,” in Proc. Real-Time Image Processing and Deep Learning: 1173606, 2021.

[32] S. Ghose, “Corn or maize leaf disease dataset,” Nov 2020.

[33] B. Gokulnath, G. Usha Devi, “A survey on plant disease prediction using machine learning and deep learning techniques,” Inteligencia Artificial, 23(65): 136-154, 2020.

[34] A.K. Singh, B. Ganapathysubramanian, S. Sarkar, and A. Singh, “Deep learning for plant stress phenotyping: trends and future perspectives,” Trends plant sci., 23(10): 883–898, 2018.

[35] A. Singh, B. Ganapathysubramanian, A.K. Singh, S. Sarkar, “Machine learning for high-throughput stress phenotyping in plants,” Trends plant sci., 21(2): 110–124, 2016.

[36] B. Richey, S. Majumder, M. Shirvaikar, N. Kehtarnavaz, “Real time detection of maize crop disease via a deep learning-based smartphone app,” in Proc. Real-Time Image Processing and Deep Learning: 114010A, 2020.

[37] J.G. Esgario, R.A. Krohling, J.A. Ventura, “Deep learning for classification and severity estimation of coffee leaf biotic stress,” Comput. Electron. Agric., 169: 105162, 2020.

[38] X. Zhang, Y. Qiao, F. Meng, C. Fan, M. Zhang, “Identification of maize leaf diseases using improved deep convolutional neural networks,” IEEE Access, 6: 30370–30377, 2018.

[39] H. Wu, T. Wiesner-Hanks, E.L. Stewart, C. DeChant, N. Kaczmar, M.A. Gore, R.J. Nelson, H. Lipson, “Autonomous detection of plant disease symptoms directly from aerial imagery,” Plant Phenome J., 2(1): 1–9, 2019.

[40] K.P. Panigrahi, A.K. Sahoo, H. Das, “A CNN approach for corn leaves disease detection to support digital agricultural system,” in Proc. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184): 678–683, 2020.

[41] M. Sibiya, M. Sumbwanyambe, “A computational procedure for the recognition and classification of maize leaf diseases out of healthy leaves using convolutional neural networks,” AgriEngi neering, 1(1): 119–131, 2019.

[42] K. Garg, S. Bhugra, B. Lall, “Automatic quantification of plant disease from field image data using deep learning,” in Proc. the IEEE/CVF Winter Conference on Applications of Computer Vision: 1965–1972, 2021.

[43] R.A. Priyadharshini, S. Arivazhagan, M. Arun, A. Mirnalini, “Maize leaf disease classification using deep convolutional neural networks,” Neural Comput. Appl., 31(12): 8887–8895, 2019.

[44] T. Wiesner-Hanks, E.L. Stewart, N. Kaczmar, C. DeChant, H. Wu, R. J. Nelson, H. Lipson, M. A. Gore, “Image set for deep learn ing: field images of maize annotated with disease symptoms,” BMC research notes, 11(1): 1–3, 2018.

[45] P. Bhatt, S. Sarangi, A. Shivhare, D. Singh, S. Pappula, “Iden tification of diseases in corn leaves using convolutional neural networks and boosting,” in ICPRAM: 894–899, 2019.

[46] E.L. da Rocha, L. Rodrigues, J.F. Mari, “Maize leaf disease classification using convolutional neural networks and

hyperparameter optimization," in Proc. Anais do XVI Workshop de Vis˜ao Computacional: 104–110, SBC, 2020

[47] A. Waheed, M. Goyal, D. Gupta, A. Khanna, A. E. Hassanien, H.M. Pandey, "An optimized dense convolutional neural network model for disease recognition and classification in corn leaf," Comput. Electron. Agric., 175: 105456, 2020.

[48] Z. Lin, S. Mu, A. Shi, C. Pang, X. Sun, et al., "A novel method of maize leaf disease image identification based on a multichannel convolutional neural network," Trans. ASABE, 61(5): 1461–1474, 2018.

[49] J. Liu, M. Wang, L. Bao, X. Li, "Efficientnet based recognition of maize diseases by leaf image classification," J. Phys Conf. Ser., 1693: 012148, IOP Publishing, 2020.

[50] X. Sun, J. Wei, "Identification of maize disease based on transfer learning," J. Phys. Conf. Ser., 1437: 012080, IOP Publishing, 2020.

[51] M. Syarief, W. Setiawan, "Convolutional neural network for maize leaf disease image classification," Telkomnika, 18(3): 1376–1381, 2020.

[52] S. Mishra, R. Sachan, D. Rajpal, "Deep convolutional neural network based detection system for real-time corn plant disease recognition," Procedia Comput. Sci., .167: 2003–2010, 2020.

[53] J. Tian, Y. Zhang, Y. Wang, C. Wang, S. Zhang, T. Ren, "A method of corn disease identification based on convolutional neu ral network," in Proc. 2019 12th International Symposium on Computational Intelligence and Design (ISCID): 245–248, 2019.

[54] A. Afifi, A. Alhumam, A. Abdelwahab, "Convolutional neural network for automatic identification of plant diseases with limited data," Plants, 10(1): 28, 2021.

[55] Y. Duan, W. Zheng, X. Lin, J. Lu, J. Zhou, "Deep adversarial metric learning," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2780–2789, 2018.

Biographies

Henry Nunoo-Mensah is a Lecturer at the Department of Computer Engineering, KNUST-Kumasi. and a researcher at the Connected Devices (CoDe) Lab. He had his BSc, MPhil, and Ph.D. in Computer Engineering, all from the Kwame Nkrumah University of Science and Technology (KNUST). He is interested in research covering wireless sensor networks, network security, intelligent agents, and algorithm design and optimization.

- Email: hnunoo-mensah@knust.edu.gh
- ORCID: 0000-0002-8965-4371
- Web of Science Researcher ID: H-9154-2019
- Scopus Author ID: NA
- Homepage: NA

Simon Wewoliamo Kuseh is an MPhil student at the Department of Computer Engineering, Kwame Nkrumah University of Science and Technology (KNUST). He had his BSc in Computer Engineering also from KNUST, Kumasi. His research interests include smart agriculture, privacy, security and trust management in IoT, and energy efficient routing algorithms in wireless sensor networks.

- Email: kusehsw@gmail.com
- ORCID: NA
- Web of Science Researcher ID: ADC-4264-2022
- Scopus Author ID: NA
- Homepage: NA

Jepthah Yankey is a Lecturer at the Department of Computer Engineering, KNUST-Kumasi. He had his BSc and MPhil. in Computer Engineering from the Kwame Nkrumah University of Science and Technology (KNUST). His research interests include artificial intelligence for development (AI4D) software verification and testing, and IoTs.

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

Francisca Adoma Acheampong had her DEng and MEng degrees from the University of Electronic Science and Technology of China (UESTC), Chengdu where she was part of the Computational Intelligence Lab, at the School of Computer Science and Engineering. She has a Bachelors in Computer Engineering from KNUST. Her research interest is intelligent agents, emotion recognition and natural language processing.

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

How to cite this paper:

H. Nunoo-Mensah, S.W. Kuseh, J. Yankey, F.A. Acheampong, "A survey of deep learning techniques for maize leaf disease detection: Trends from 2016 to 2021 and future perspectives" J. Electr. Comput. Eng. Innovations, 10(2): 381-392, 2022.

DOI: 10.22061/JECEI.2022.8602.531

URL: https://jecei.sru.ac.ir/article_1678.html

SRTTU

Journal of Electrical and Computer Engineering Innovations

(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

A General Approach for Operational Bandwidth Extension in Spherical Microphone Array

M. Kalantari*[*]*, M. Mohammadpour Tuyserkani, S.H. Amiri

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

Article Info

Article History:

Received 06 October 2021
Reviewed 23 November 2021
Revised 06 December 2021
Accepted 03 February 2022

Keywords:

Spherical microphone arrays
Aliasing
Beamforming
Operational bandwidth

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

Abstract

Background and Objectives: Operating frequency range of a microphone array is limited by the array configuration. Spatial aliasing occurs at frequencies considered to be out of the microphone array operating range that leads to side-lobes in the array beam pattern and consequently degrades the performance of the microphone array. In this paper, a general approach for increasing the operational bandwidth of the spherical microphone array without physical changes to the microphone array is proposed.

Methods: Recently, Alon and Rafaely proposed a beamforming method with aliasing cancellation and formulated it for some well-known beamformers such as maximum directivity (MD), maximum white noise gain (WNG), and minimum variance distortionless response (MVDR) which have been called MDAC, MGAC, MVDR-AC beamformer respectively. In this paper, we derive MDAC method from different point of view. Then, based on our perspective, we propose a new method that is easily applicable for any beamforming algorithms.

Results: Comparing with MDAC and MGAC beamformers, performance measures for our approach show improvement in directivity index (DI) and white noise gain (WNG) by nearly 19% and 15% respectively.

Conclusion: Aliasing and, in consequence, unwanted side lobe formation is the main factor in spherical microphone arrays operational bandwidth determination. Most of the methods previously presented to reduce aliasing demanded physical changes in the array structure which comes at a cost. In this paper we propose a new method based on Alon and Rafaely's approach via designing a constrained optimization problem using orthogonality property of spherical harmonics, to achieve better performance.

Introduction

Spherical microphone arrays are a type of microphone arrays that have a spherical array structure in which microphones are placed on the surface of a sphere. This kind of microphone arrays has been an interesting field of study for the past decade. Because of their symmetry they can steer the beam pattern over any desired direction in the space [1].

One of the main concepts central to spherical microphone arrays is its operational bandwidth [2]-[9]. The operational bandwidth of the spherical microphone arrays is determined by their lower and upper frequency limits [10], [11]. The lower frequency limit is bounded by some factors such as sensor noise and the upper frequency limit is bounded by spatial aliasing [8], [12]. In fact, analyzing the array bandwidth limitations by

Doi: 10.22061/JECEI.2021.8304.501

decomposing the sound field into spherical harmonics shows that with the increase of the frequency, the sound pressure function of the sound field around the sphere will be of a higher order [10], [13]. In many cases this order is higher than the array's maximum order, which is determined by the number of microphones that leads to spatial aliasing [14], [15].

Spatial aliasing and consequently unwanted side-lobe formation is the main cause of performance degradation of spherical microphone arrays at high frequencies [16], [17]. Some solutions have been presented for reducing aliasing effects [18]-[20], but they can only perform well in a sound field with certain characteristics [21], [22]. Other methods for increasing arrays performance in high frequencies tend to minimize side-lobe levels [23]. These methods include increasing the number of microphones in the array, using other types of directional microphones, or using microphones with wider surface [7], [24]-[26], [10], [19], [27], [28]. All the methods mentioned above need to make physical changes in the array structure that comes with a cost in many cases.

Recently Alon and Rafaely proposed a new spherical microphone array beamforming with an aliasing model for describing high sound field orders, aliased into the lower array orders. For that, they include the effect of spatial aliasing in the definition of desired objective, such as directivity factor, and develop a new version of beamformers with aliasing cancellation capabilities. Their method was found to be valuable for injecting aliasing cancellation capability to some of the well-known beamformers such as maximum directivity (MD) beamformer, after which called maximum directivity beamformer with aliasing cancellation (MDAC). This new beamformer achieves higher directivity index (DI) with a narrower main lobe and lower side lobes, compared with standard MD beamformer. This method was also used to develop maximum white noise gain with aliasing cancellation (MGAC), and minimum variance distortionless response with aliasing cancellation (MVDR-AC) beamformers [1].

In this paper we aim to look at the method presented by Alon and Rafaely from a different point of view. The main contribution of this work is to design a constrained optimization problem with some appropriate constraints to find the closest signal to the desired unaliased signal. These constraints are attained by using the orthogonality property of spherical harmonics. Then this estimation of unaliased signal can be used in beamforming process using ordinary beamforming coefficient of a high order beamfomer.

This paper is organized as follows. The second section reviews the spherical array processing fundamentals. The third section presents the proposed method for aliasing cancellation beamforming. Simulation results and comparisons with the rival method are presented in the fourth section, and the end section concludes the paper.

Spherical Array Processing

This section shortly explains the theory of spherical microphone array processing [16], [29]. The formulation provided in this section will be utilized in the third section to develop the proposed beamformer.

A. Spherical Array Processing

Consider a sound field composed of multiple "single frequency plane wave" each with amplitude density denoted by $a(k,\theta_k,\phi_k)$ arriving from direction (θ_k,ϕ_k) with a wave vector $\tilde{\mathbf{k}} = -\mathbf{k} = (k,\theta_k,\phi_k)$ and wave number k. The sound pressure at $\mathbf{r} = (r,\theta,\phi)$ due to this sound field can be written as follows [16]

$$p(k,r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} p_{nm}(k,r)Y_n^m(\theta,\phi) \quad (1)$$

where $p_{nm}(k,r)$ are the spherical harmonic coefficients of the sound pressure, and $Y_n^m(\theta,\phi)$ are the spherical harmonics. The relation between the pressure on the sphere and the amplitude of the plane waves composing the sound field in the spherical harmonic domain is

$$p_{nm}(k,r) = b_n(kr)a_{nm}(k) \quad (2)$$

where $a_{nm}(k)$ is the spherical Fourier transform of $a(k,\theta_k,\phi_k)$, i.e.,

$$a_{nm}(k) = \int_0^{2\pi}\int_0^{\pi} a(k,\theta_k,\phi_k)[Y_n^m(\theta_k,\phi_k)]^* \sin\theta_k d\theta_k d\phi_k \quad (3)$$

and $b_n(kr)$ defines the projection of the sound field onto the sphere surface. The expression for $b_n(kr)$ depends on the array configuration. For example, in the case of a single open sphere, we have

$$b_n(kr) = 4\pi i^n j_n(kr) \quad (4)$$

where $j_n(x)$, is the spherical Bessel function of the first kind.

If the pressure function is order-limited, meaning that $p_{nm}(k,r) = 0 \ \forall n > N$, then we can represent the function by a finite number of spherical harmonics, so we have

$$p(k,r,\theta,\phi) = \sum_{n=0}^{N}\sum_{m=-n}^{n} p_{nm}(k,r)Y_n^m(\theta,\phi) \quad (5)$$

Equation (1) is, in fact, the inverse spherical Fourier transform of the pressure function [14]. So, we have

$$p_{nm} = \int_0^{2\pi}\int_0^{\pi} p(\theta,\phi)[Y_n^m(\theta,\phi)]^* \sin\theta d\theta d\phi \quad (6)$$

which is the spherical Fourier transform of $p(\theta,\phi)$. For the sake of simplicity, parameters k,r have been omitted.

According to the Cubature method, it can be possible to compute the multiple integrations of a given function using a summation over sample of the function [29]. So,

$$p_{nm} \approx \sum_{q=1}^{Q} \alpha_q\, p(\theta,\phi)[Y_n^m(\theta,\phi)]^* = \hat{p}_{nm} \tag{7}$$

where Q, is the total number of samples and α_q is the sampling weight whose value depends on the sampling scheme. For order-limited function, the approximation becomes equality, given a sufficiently large Q. In this case, $p(\theta,\phi)$ can be reconstructed perfectly on the sphere using the inverse spherical Fourier transform. But, in the case of p_{nm} of infinite order, perfect reconstruction is not possible due to aliasing.

Several sampling methods, such as equal-angle, Gaussian, and uniform sampling, have been previously presented [16], for which the sampling weight α_q and sampling points (θ_q,ϕ_q) have been derived such that (6) is maintained with equality for order-limited functions.

Due to some constraints, we may want to use any arbitrary given sampling set. So, assume that the samples of the function, $p(\theta_q,\phi_q)$, are given, together with the positions of the samples, (θ_q,ϕ_q), for $q = 1,\dots,Q$. Using (5) we have

$$p(\theta_q,\phi_q) = \sum_{n=0}^{N}\sum_{m=-n}^{n} p_{nm} Y_n^m(\theta_q,\phi_q), \quad 1 \le q \le Q, \tag{8}$$

This equation can be written in matrix forms as

$$\mathbf{p} = \mathbf{Y}\mathbf{p_{nm}} \tag{9}$$

where column vectors $\mathbf{p}$ of length Q and $\mathbf{p_{nm}}$ of length $(N+1)^2$ are defined as

$$\mathbf{p} = \left[p(\theta_1,\phi_1), p(\theta_2,\phi_2),\dots,p(\theta_Q,\phi_Q)\right]^T \tag{10}$$

and

$$\mathbf{p_{nm}} = \left[p_{00}, p_{1(-1)}, p_{10}, p_{11},\dots,p_{NN}\right]^T \tag{11}$$

and the matrix $\mathbf{Y}$ of dimensions $Q \times (N+1)^2$ is shown in (12),

$$\mathbf{Y} = \begin{bmatrix} Y_0^0(\theta_1,\phi_1) & Y_1^{-1}(\theta_1,\phi_1) & Y_1^0(\theta_1,\phi_1) & Y_1^1(\theta_1,\phi_1) & \dots & Y_N^N(\theta_1,\phi_1) \\ Y_0^0(\theta_2,\phi_2) & Y_1^{-1}(\theta_2,\phi_2) & Y_1^0(\theta_2,\phi_2) & Y_1^1(\theta_2,\phi_2) & \dots & Y_N^N(\theta_2,\phi_2) \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ Y_0^0(\theta_Q,\phi_Q) & Y_1^{-1}(\theta_Q,\phi_Q) & Y_1^0(\theta_Q,\phi_Q) & Y_1^1(\theta_Q,\phi_Q) & \dots & Y_N^N(\theta_Q,\phi_Q) \end{bmatrix} \tag{12}$$

Equation (9) is called inverse discrete spherical Fourier transform. Also,

$$\mathbf{p_{nm}} = \mathbf{Y}^{\dagger}\mathbf{p} \tag{13}$$

is called discrete spherical Fourier transform, where $\mathbf{Y}^{\dagger} = (\mathbf{Y}^H\mathbf{Y})^{-1}\mathbf{Y}^H$ is the pseudo-inverse of $\mathbf{Y}$. This equation can be written in the following form

$$p_{nm} = \sum_{q=1}^{Q} \alpha_q^{nm}\, p(\theta_q,\phi_q) \tag{14}$$

where the sampling weights, α_q^{nm}, are the elements of matrices $\mathbf{Y}^{\dagger}$, having a row index given by (n^2+n+m) and a column index is given by q.

B. Spatial Aliasing

As we have already mentioned, sampling of order-limited functions on the sphere with an appropriate sampling scheme should lead to an exact and aliasing-free computation of the spherical harmonic coefficient. But for non-order-limited functions, errors may occur due to spatial aliasing. For analyzing and describing these errors, we can rewrite (7) as follows [24]:

$$\begin{aligned} \hat{p}_{nm} &= \sum_{q=1}^{Q} \alpha_q^{nm} \sum_{n'=0}^{\infty}\sum_{m'=-n'}^{n'} p_{n'm'} Y_{n'}^{m'}(\theta_q,\phi_q) \\ &= \sum_{n'=0}^{\infty}\sum_{m'=-n'}^{n'} \left[\sum_{q=1}^{Q} \alpha_q^{nm} Y_{n'}^{m'}(\theta_q,\phi_q)\right] p_{n'm'} \\ &= \sum_{n'=0}^{\infty}\sum_{m'=-n'}^{n'} \epsilon_{nm}^{n'm'} p_{n'm'}, \end{aligned} \tag{15}$$

where

$$\epsilon_{nm}^{n'm'} = \sum_{q=1}^{Q} \alpha_q^{nm} Y_{n'}^{m'}(\theta_q,\phi_q) \tag{16}$$

In an ideal, aliasing-free sampling, $\epsilon_{nm}^{n'm'}$ equals one for $(n,m) = (n',m')$ and zero elsewhere. If we assume that the spherical harmonic coefficients of the original function before sampling, $p_{n'm'}$, is order-limited but to very high order denoted by $\tilde{N}$, we can represent (14) in a matrix form as follows

$$\mathbf{\hat{p}_{nm}} = \mathbf{E}\mathbf{\tilde{p}_{nm}} \tag{17}$$

where $\mathbf{\hat{p}_{nm}}$ of length $(N+1)^2$ holds the approximated spherical harmonic coefficients $\hat{p}_{nm}$, $\mathbf{\tilde{p}_{nm}}$ of length $\left(\tilde{N}+1\right)^2$ holds the spherical harmonic coefficients p_{nm} of the original function, with $\tilde{N} \ge N$, and matrix $\mathbf{E}$ of dimensions $(N+1)^2 \times \left(\tilde{N}+1\right)^2$, having elements $\epsilon_{nm}^{n'm'}$ with row index (n^2+n+m) and column index $\left(n'^2+n'+m'\right)$, called aliasing matrix. Matrix $\mathbf{E}$ can be written as

$$\mathbf{E} = \mathbf{Y}^{\dagger}\mathbf{\tilde{Y}} \tag{18}$$

where matrix $\mathbf{Y}$ of dimensions $Q \times (N+1)^2$ has been defined in (12) and matrix $\mathbf{\tilde{Y}}$ of dimensions $Q \times \left(\tilde{N}+1\right)^2$, holding the values of $Y_{n'}^{m'}(\theta_q,\phi_q)$ as in (16).

C. Spherical Array Beamforming

Array equations or beamforming equations is as follows [29]

$$y = \int_0^{2\pi}\int_0^{\pi} w^*(k,\theta,\phi)p(k,r,\theta,\phi)\sin\theta d\theta d\phi = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} w_{nm}^*(k)p_{nm}(k,r) \tag{19}$$

where $w^*(k,\theta,\phi)$, is the beamforming coefficients. The standard discrete form of beamforming in the space domain is

$$y = \mathbf{w}^{\mathbf{H}}\mathbf{p} \tag{20}$$

where $\mathbf{p}$ is as (10) with a little modification in notation

$$\mathbf{p} = [p_1(k), p_2(k), \dots, p_Q(k)]^T \tag{21}$$

with $p_q(k) = p(k,r,\theta_q,\phi_q)$, $q = 1,\dots,Q$, and $\mathbf{w}$ is the $Q \times 1$ weight vector as follows

$$\mathbf{w} = [w_1(k), w_2(k), \dots, w_Q(k)]^T \tag{22}$$

Assuming $w_{nm} = 0\ \forall n > N$, the discrete form of beamforming in spherical harmonic domain is as

$$y = \mathbf{w}_{\mathbf{nm}}^H\mathbf{p}_{\mathbf{nm}} \tag{23}$$

where the $(N+1)^2 \times 1$ vector $\mathbf{w}_{\mathbf{nm}}$ is given by

$$\mathbf{w}_{\mathbf{nm}} = [w_{00}(k), w_{1(-1)}(k), w_{10}(k), w_{11}(k)\ \dots, w_{NN}(k)]^T, \tag{24}$$

and the $(N+1)^2 \times 1$ vector $\mathbf{p}_{\mathbf{nm}}$ is given by

$$\mathbf{p}_{\mathbf{nm}} = [p_{00}(k,r), p_{1(-1)}(k,r), p_{10}(k,r), \dots, p_{NN}(k,r)]^T \tag{25}$$

In these equations $p_{nm}(k)$ and $w_{nm}(k)$ are the spherical Fourier transform of $p(k,r,\theta,\phi)$ and $w(k,\theta,\phi)$ respectively and N is called the effective order of the array.

Array output due to a unit-amplitude plane-wave sound field or array beam pattern is defined as

$$y = \mathbf{w}_{\mathbf{nm}}^{\mathbf{H}}\mathbf{v}_{\mathbf{nm}} \tag{26}$$

where $\mathbf{v}_{\mathbf{nm}}$, is a $(N+1)^2 \times 1$ column vector as

$$\mathbf{v}_{\mathbf{nm}} = [v_{00}(k,r), v_{1(-1)}(k,r), v_{10}(k,r), \dots, v_{NN}(k,r)]^T \tag{27}$$

with v_{nm} represents the array input due to the plane wave sound field. Since for unit amplitude plane wave we have [29]

$$a_{nm}(k) = [Y_n^m(\theta_k,\phi_k)]^* \tag{28}$$

according to (2) we have

$$v_{nm}(k,r) = b_n(kr)[Y_n^m(\theta_k,\phi_k)]^* \tag{29}$$

where (θ_k,ϕ_k), is the arrival direction of the plane wave.

Using a different set of beamforming coefficients, different beam patterns can be designed. For instance, axis-symmetric beamformers with $w_{nm}^*(k) = \frac{d_n(k)}{b_n(kr)}Y_n^m(\theta_l,\phi_l)$ [29], of which two famous beamformers, namely the maximum directivity (MD) beamformer and the maximum white noise gain (WNG) beamformer described in the sequel. Note that the beamformer coefficients are function of look direction which denoted by (θ_l,ϕ_l) in the above relation.

- **Maximum Directivity Beamformer**

The directivity factor (DF) is the ratio between the array response in the look direction and the average response across all directions and is defined mathematically as follows

$$\mathrm{DF} = \frac{|y(\theta_l,\phi_l)|^2}{\frac{1}{4\pi}\int_0^{2\pi}\int_0^{\pi}|y(\theta,\phi)|^2\sin\theta d\theta d\phi} = \frac{\mathbf{w}_{\mathbf{nm}}^{\mathrm{H}}\mathbf{A}\mathbf{w}_{\mathbf{nm}}}{\mathbf{w}_{\mathbf{nm}}^{\mathrm{H}}\mathbf{B}_{\mathrm{D}}\mathbf{w}_{\mathbf{nm}}}$$

$$\mathbf{A} = \mathbf{v}_{\mathbf{nm}}\mathbf{v}_{\mathbf{nm}}{}^{\mathrm{H}} \tag{30}$$

$$\mathbf{B}_{\mathrm{D}} = \frac{1}{4\pi}diag(|b_0|^2, |b_1|^2, |b_1|^2, \dots, |b_N|^2)$$

$$\mathbf{v}_{\mathbf{nm}} = [v_{00}, v_{1(-1)}, v_{10}, v_{11}, \dots, v_{NN}]^T$$

The explicit dependency of $b_n(kr)$ on kr has been omitted for notation simplicity. Note that in (30) $\mathbf{v}_{\mathbf{nm}}$ is the unit-amplitude plane wave arriving from look direction.

The maximum directivity (MD) beamformer is designed to satisfy

$$\underset{\mathbf{w}_{\mathbf{nm}}}{\text{minimize}}\ \mathbf{w}_{\mathbf{nm}}^H\mathbf{B}_D\,\mathbf{w}_{\mathbf{nm}} \quad \text{subject to } \mathbf{w}_{\mathbf{nm}}^H\mathbf{v}_{\mathbf{nm}} = 1 \tag{31}$$

Solving the above optimization problem leads to the following beamforming coefficients

$$\mathbf{w}_{\mathbf{nm}}^{MD^H} = \frac{\mathbf{v}_{\mathbf{nm}}^H\mathbf{B}_D^{-1}}{\mathbf{v}_{\mathbf{nm}}^H\mathbf{B}_D^{-1}\mathbf{v}_{\mathbf{nm}}} \tag{32}$$

or equivalently

$$w_{nm}^*(k) = \frac{4\pi}{(N+1)^2}\frac{1}{b_n(kr)}Y_n^m(\theta_l,\phi_l) \tag{33}$$

which is an axis-symmetric beamformer with $d_n(k) = \frac{4\pi}{(N+1)^2}$.

- **Maximum WNG Beamformer**

WNG is a general measure for array robustness which is defined as the improvement in SNR at the array output compared to the array input. Mathematically,

$$\mathrm{WNG} = \frac{\mathbf{w}_{\mathbf{nm}}^H\mathbf{A}\mathbf{w}_{\mathbf{nm}}}{\mathbf{w}_{\mathbf{nm}}^H\boldsymbol{B}_G\mathbf{w}_{\mathbf{nm}}}$$

$$\mathbf{A} = \mathbf{v}_{\mathbf{nm}}\mathbf{v}_{\mathbf{nm}}^H \tag{34}$$

$$\mathbf{B}_G = \mathbf{Y}^{\dagger}\mathbf{Y}^{\dagger H}$$

The maximum WNG beamformer (MG) is designed to satisfy the problem below

$$\underset{\mathbf{w_{nm}}}{\text{minimize}}\ \mathbf{w_{nm}^H B_G\, w_{nm}} \qquad \text{subject to } \mathbf{w_{nm}^H v_{nm}} = 1 \tag{35}$$

Solving the above optimization problem leads to the following beamforming coefficients

$$\mathbf{w_{nm}^{MG^H}} = \frac{\mathbf{v_{nm}^H B_G^{-1}}}{\mathbf{v_{nm}^H B_G^{-1} v_{nm}}} \tag{36}$$

For uniform or nearly uniform sampling scheme, this leads to

$$w_{nm}^*(k) = \frac{b_n^*(kr)Y_n^m(\theta_l,\phi_l)}{\sum_{n=0}^{N}\frac{2n+1}{4\pi}|b_n(kr)|^2} \tag{37}$$

which is an axis-symmetric beamformer with $d_n(k) = \frac{|b_n(kr)|^2}{\sum_0^N \frac{2n+1}{4\pi}|b_n(kr)|^2}$.

The Proposed Method

The operational bandwidth of a spherical microphone array is defined by its upper and lower frequency limits. The lower frequency limit is bounded by sensor noise and other errors, such as a mismatch in microphone gain and phase response, inaccurate positioning of microphones and limited computational accuracy [29], that is not our concern in this paper. The upper frequency limit is bounded by spatial aliasing [24]. To avoid significant error due to spatial aliasing we must have $N \geq kr$ [24]. In fact, the upper frequency is determined by $k \leq N/r$. However, at a higher frequency, the performance of the microphone array degrades due to spatial aliasing. For example, in beamforming problem, because of spatial aliasing we are not able to compute $p_{nm}(k,r)$ precisely, and according to (19) we will have inaccurate beamforming which in turn, for instance, degrades DF in MD beamformer.

A. The Proposed Method for Maximum Directivity Beamformer

Recently, Alon and Rafaely proposed a beamforming method with aliasing cancellation and formulated it for some well-known beamformers such as maximum-directivity, maximum WNG, and minimum variance distortion less response (MVDR), which is called MDAC (maximum directivity with aliasing cancellation), MGAC (maximum WNG with aliasing cancellation), and MVDR-AC (MVDR with aliasing cancellation) respectively. In this section, first, we derive their MDAC beamformer from a different point of view (See Theorem 1). Then, based on our perspective we propose a new beamforming method with aliasing reduction to acquire better performance.

Theorem 1. The MDAC beam pattern is equivalent to the MD beam pattern corresponding to the minimum norm solution of (17).

Proof: In [1], Alon and Rafaely proved that the MDAC beam pattern is as

$$A_{MDAC}(k,\theta_k,\phi_k) = \mathbf{\tilde{w}_{nm}^{MD^{\boldsymbol{H}}}}\,\mathbf{E}^H\left(\mathbf{EE}^H\right)^{-1}\mathbf{\hat{v}_{nm}} \tag{38}$$

where

$$\mathbf{\tilde{w}_{nm}^{\mathit{MD}}} = \left[w_{00}(k), w_{1(-1)}(k), w_{10}(k), w_{11}(k)\ \ldots, w_{\tilde{N}\tilde{N}}(k)\right]^T \tag{39}$$

and $\mathbf{\hat{v}_{nm}} = \mathbf{E\tilde{v}_{nm}}$, is the aliased version of

$$\mathbf{\tilde{v}_{nm}} = \left[v_{00}(k,r), v_{1(-1)}(k,r), v_{10}(k,r), \ldots, v_{\tilde{N}\tilde{N}}(k,r)\ \right]^T \tag{40}$$

$w_{nm}(k)$ is as (33) and $v_{nm}(k,r)$ is as (29).

Now, we derive this result from a different point of view as follows. If we rewrite (17) for unit-amplitude plane wave we have

$$\mathbf{\hat{v}_{nm}} = \mathbf{E\tilde{v}_{nm}} \tag{41}$$

It is an underdetermined linear equations, i.e., there are fewer equations than unknowns. Therefore, this equation has many solutions. The minimum norm solution of this equation is

$$\mathbf{\tilde{v}_{nm}^{\boldsymbol{min\_norm}}} = \mathbf{E}^H\left(\mathbf{EE}^H\right)^{-1}\mathbf{\hat{v}_{nm}} \tag{42}$$

where $\mathbf{E}^H(\mathbf{EE}^H)^{-1}$ is the pseudo-inverse of matrix $\mathbf{E}$.

Now, we have a plane wave, $\mathbf{\tilde{v}_{nm}^{\boldsymbol{min\_norm}}}$, arriving from the direction (θ_k,ϕ_k), for which we can construct a beam pattern using beamforming coefficients $\mathbf{\tilde{w}_{nm}^{\mathit{MD}}}$. Note that $\mathbf{\tilde{v}_{nm}^{\boldsymbol{min\_norm}}}$ is an estimation of a unit-amplitude plane wave of order $\tilde{N}$. So, the beam pattern is as

$$\begin{aligned} A_{min\_norm}(k,\theta_k,\phi_k) &= \mathbf{\tilde{w}_{nm}^{MD^{\mathit{H}}}}\,\mathbf{\tilde{v}_{nm}^{\boldsymbol{min\_norm}}} \\ &= \mathbf{\tilde{w}_{nm}^{MD^{\boldsymbol{H}}}}\,\mathbf{E}^H\left(\mathbf{EE}^H\right)^{-1}\mathbf{\hat{v}_{nm}} \end{aligned} \tag{43}$$

which is the same as $A_{MDAC}(k,\theta_k,\phi_k)$. □As we have mentioned before, (41) is an underdetermined linear equation and has many solutions. If we obtained the desired solution, $\mathbf{\tilde{v}_{nm}}$, then by multiplying it with $\mathbf{\tilde{w}_{nm}^{MD^{\boldsymbol{H}}}}$ we could obtain the best beamforming pattern, i.e. beam pattern with maximum directivity. So, in our proposed method, the aim is to find a solution as close as to $\mathbf{\tilde{v}_{nm}}$. For that, we use the following theorem.

Theorem 2: The best beamforming pattern for maximum directivity beamformer can be obtained using $\mathbf{\tilde{v}_{nm}}$ via solving the following optimization problem

$$\begin{gathered} \underset{\mathbf{\tilde{v}_{nm}}}{\text{minimize}}\ \|\mathbf{E\tilde{v}_{nm}} - \mathbf{\hat{v}_{nm}}\| \\ \text{subject to} \\ \left|\sum_{\theta,\phi}\kappa_i^*(\theta,\phi)\kappa_j(\theta,\phi)\right| = 0, \forall i \neq j \\ \left|\sum_{\theta,\phi}\kappa_i^*(\theta,\phi)\kappa_i(\theta,\phi)\right| = 1, \\ (\theta,\phi) \in \{(\theta_1,\phi_1),(\theta_2,\phi_2),\ldots.,(\theta_Q,\phi_Q)\} \end{gathered} \tag{44}$$

where $\boldsymbol{\kappa}^* = \widetilde{\mathbf{B}}_D^{-1}\tilde{\mathbf{v}}_{\mathbf{nm}}$, and $\boldsymbol{\kappa} = [\kappa_1, \kappa_1, \dots, \kappa_{\widetilde{N}}]$. $\widetilde{\mathbf{B}}_D$ is as $\mathbf{B}_D$ in (30) with N replaced by $\widetilde{N}$ and (θ_i, ϕ_i), are the positions of the i^{th} sample.

Proof: We know that the spherical harmonics have orthogonality property [29], i.e,

$$\int_0^{2\pi}\int_0^{\pi}[Y_n^m(\theta,\phi)]^* Y_{n'}^{m'}(\theta,\phi)\sin\theta d\theta d\phi = \delta_{nn'}\delta_{mm'} \tag{45}$$

where $\delta_{nn'}$ is equal to unity for $n = n'$ and zero otherwise. Using this property as constraints, (41) can be solved using the following optimization problem

$$\begin{gathered}\underset{\tilde{\mathbf{v}}_{\mathbf{nm}}}{\text{minimize}} \quad \|\mathbf{E}\tilde{\mathbf{v}}_{\mathbf{nm}} - \hat{\mathbf{v}}_{\mathbf{nm}}\| \\ \text{subject to} \\ \left|\int_0^{2\pi}\int_0^{\pi}\kappa_i^*(\theta,\phi)\,\kappa_j(\theta,\phi)\sin\theta\, d\theta\, d\phi\right| = 0 \\ \forall i \neq j \\ \left|\int_0^{2\pi}\int_0^{\pi}\kappa_i^*(\theta,\phi)\,\kappa_i(\theta,\phi)\sin\theta\, d\theta\, d\phi\right| = 1\end{gathered} \tag{46}$$

where $\boldsymbol{\kappa}^*$ and $\widetilde{\mathbf{B}}_D$ is as defined above.

Considering the discrete form of these constraints results (44). Now, the MD beamforming with aliasing reduction is as follows

$$A_{MDAC_{Proposed}}(k, \theta_k, \phi_k) = \widetilde{\mathbf{w}}_{\mathbf{nm}}^{MD^H}\tilde{\mathbf{v}}_{\mathbf{nm}} \tag{47}$$

where $\tilde{\mathbf{v}}_{\mathbf{nm}}$ is the solution of (46).

B. *The Proposed Method for Arbitrary Beamformer*

We can apply the proposed method to an arbitrary beamformer as follows. First, we can solve the following optimization problem

$$\begin{gathered}\underset{\tilde{\mathbf{p}}_{\mathbf{nm}}}{\text{minimize}} \quad \|\mathbf{E}\tilde{\mathbf{p}}_{\mathbf{nm}} - \hat{\mathbf{p}}_{\mathbf{nm}}\| \\ \text{subject to} \\ \left|\int_0^{2\pi}\int_0^{\pi}\kappa_i^*(\theta,\phi)\,\kappa_j(\theta,\phi)\sin\theta\, d\theta\, d\phi\right| = 0 \\ \forall i \neq j \\ \left|\int_0^{2\pi}\int_0^{\pi}\kappa_i^*(\theta,\phi)\,\kappa_i(\theta,\phi)\sin\theta\, d\theta\, d\phi\right| = 1\end{gathered} \tag{48}$$

where $\boldsymbol{\kappa}^* = \widetilde{\mathbf{B}}_D^{-1}\tilde{\mathbf{p}}_{\mathbf{nm}}$. We assume that the solution to the above problem is $\tilde{\boldsymbol{\rho}}_{\boldsymbol{nm}}$. Then the desired beamforming with reduced aliasing is as follows

$$A_{proposed}(k, \theta_k, \phi_k) = \widetilde{\mathbf{w}}_{\mathbf{nm}}^H\tilde{\boldsymbol{\rho}}_{\boldsymbol{nm}} \tag{49}$$

where $\widetilde{\mathbf{w}}_{\mathbf{nm}}^H$ is the beamforming coefficients of the ordinary beamformer. For example, for maximum WNG beamformer $\widetilde{\mathbf{w}}_{\mathbf{nm}}^H$ is as introduced in (36) with N replaced by $\widetilde{N}$.

Simulations Results

In this section we compare the proposed methods, MDAC_{Proposed} and MGAC_{Proposed}, with counterparts in Alon and Rafaely's method, namely MDAC and MGAC. For this purpose, a microphone array with 50 microphones has been considered. These microphones are placed on the surface of a rigid sphere with radius of $r = 10$ cm, based on gaussian sampling scheme. The maximum order of an order-limited function that can be constructed from this configuration is $N = 4$ [29].

The operational bandwidth of this microphone array can be determined by the condition $kr \leq N$ which results $f_{max} = 2.1$ kHz. So, the array can not handle higher frequency without aliasing. The simulation is designed for three different frequencies, $k_1 r = 3.6$ ($\widetilde{N} = N = 4 > k_1 r$), $k_2 r = 8.1$ ($\widetilde{N} = 9 > k_2 r$), and $k_3 r = 14.2$ ($\widetilde{N} = 15 > k_3 r$) where each kr represents different frequency corresponds to $f_1 = 1.9$ kHz, $f_2 = 4.4$ kHz, and $f_3 = 7.6$ kHz.

A. *MDAC_{Proposed} Simulation*

The look direction for every beamformer in this simulation is equal. The sound field around the sphere is composed of a single unit-amplitude plane wave of orders $\widetilde{N} = 4, 9, 15$ respectivly, and arrives from the same angle as look direction, $(\theta_0, \phi_0) = (\theta_l, \phi_l) = (90^\circ, 80^\circ)$. Beam patterns are compared over three different frequencies, $k_1 r = 3.6$, $k_2 r = 8.1$, and $k_3 r = 14.2$.

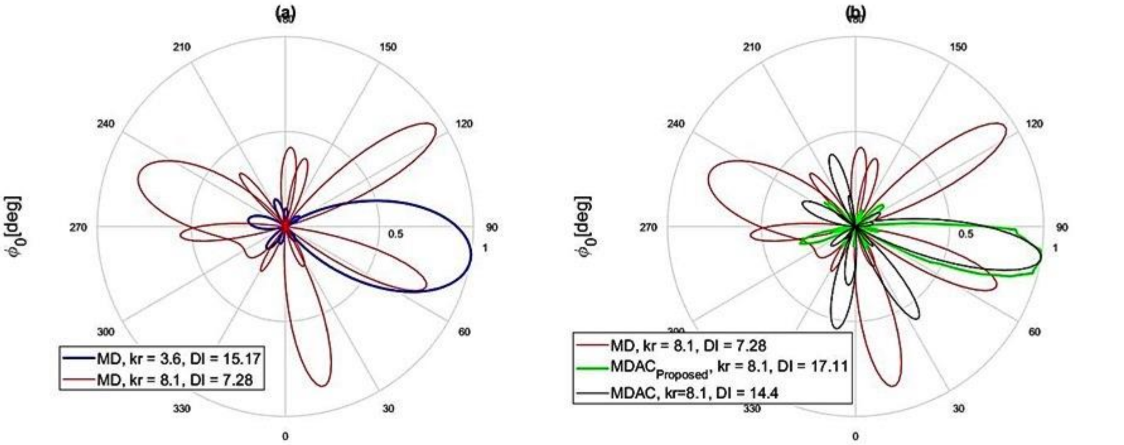


Fig. 1: The beam pattern of a fourth-order array with look direction $(\theta_l, \phi_l) = (90^\circ, 80^\circ)$ (a) comparison between MD beam patterns at frequencies $k_1 r = 3.6$ (f = 1.9 kHz) and $k_2 r = 8.1$ (f = 4.4 kHz) and (b) comparison between MD, MDAC_{Proposed}, and MDAC beamformers at $k_2 r = 8.1$ (f = 4.4 kHz).

Fig. 1(a) shows two beam patterns of MD beamformer at two different frequencies. At lower frequency, $k_1r = 3.6$, a beam pattern is obtained for sound field of order $\tilde{N} = 4$, so beamforming is done without any aliasing, however at a higher frequency, $k_2r = 8.1$, side-lobe levels are increased and the beam pattern is not directional anymore. This performance degradation between k_1r and k_2r can be explained by the array's operational bandwidth. Here, the maximum frequency that can be handled by microphone array must be lower than $f_{max} = 2.1$ kHz as mentioned in the fourth section. Therefore, because of the $k_1r \leq N$ ($f_1 \leq f_{max}$), aliasing free condition is satisfied and the array output is without aliasing. For the second part, because of the $k_2r > N$, condition is not satisfied, and array response suffers from aliasing. Fig. 1(b) shows array beam patterns for three different beamformers (MD, MDAC, and MDAC$_{\text{Proposed}}$) in frequency of $k_2r = 8.1$. The arrays performance for the MD beamformer is the same as part (a) of Fig. 1, but a comparison between MDAC and MDAC$_{\text{Proposed}}$ beam patterns shows that side-lobes are at much lower levels in MDAC$_{\text{Proposed}}$ than MDAC, in addition, main-lobe of MDAC$_{\text{Proposed}}$ is narrower than MDAC.

Another comparison between MD, MDAC, and MDAC$_{\text{Proposed}}$ beamformers is shown in Fig. 2, in which the plot balloon of these beamformers shows that MDAC$_{\text{Proposed}}$ has lower side-lobe levels compared to MDAC.

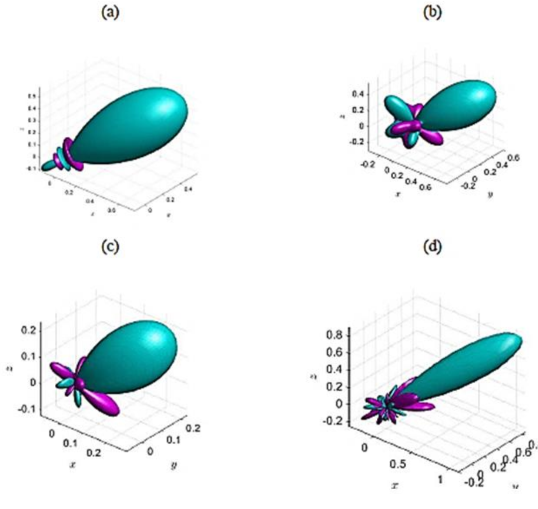


Fig. 2: Plot balloon comparison between MD, MDAC, and MDAC$_{\text{Proposed}}$ beamformers with look direction $(\theta_l, \phi_l) = (90^{\circ}, 80^{\circ})$ (a) MD beamformer plot balloon at $k_1r = 3.6$ ($f = 1.9$ kHz) (b) MD beamformer at $k_2r = 8.1$ ($f = 4.4$ kHz) (c) MDAC beamformer at $k_2r = 8.1$ ($f = 4.4$ kHz) (d) MDAC$_{\text{Proposed}}$ at $k_2r = 8.1$ ($f = 4.4$ kHz).

A more computational manner for comparison between MD, MDAC, and MDAC$_{\text{Proposed}}$ beamformers is by comparing their directivity indexes calculated with (30). Table 1 shows the directivity index of each beamformer at k_1r and k_2r. At $k_1r = 3.6$, all beamformers used in this simulation have the same DI. The reason is that for each one the condition $kr \leq N$ is satisfied. However, at $k_2r = 8.1$ the MDAC beamformer has $\text{DI} = 14.4$ but the DI for MDAC$_{\text{Proposed}}$ is 17.11. This shows that MDAC$_{\text{Proposed}}$ has improved directivity by nearly 19%.

Table 1: A comparison between DI of MD, MDAC, and MDAC$_{\text{Proposed}}$ beamformers in two different frequencies $k_1r = 3.6$ and $k_2\,r = 8.1$

Beamformer	DI (dB) with $kr = 3.6$	DI (dB) with $kr = 8.1$
MD	13.9	7.28
MDAC	13.9	14.4
MDAC$_{\text{Proposed}}$	13.9	17.11

Fig. 3 shows a better comparison between MDAC and MDAC$_{\text{Proposed}}$ beamformers, where both are operating in the sound field with order of $\tilde{N} = 15 > N$ and for $k_3r = 14.2$ ($f_3 = 7.6$ kHz). In this figure, it is clear that MDAC beamformer has side-lobes of high levels, almost as high as the main-lobe, but the MDAC$_{\text{Proposed}}$ beamformer managed to keep the side-lobe levels as low as in k_2r. This shows that the MDAC beamformer's performance decays as the sound field order increases but MDAC$_{\text{Proposed}}$ is robust enough to keep the side-lobe levels low. So, despite the low computational complexity in MDAC beamformer, the accuracy of this method decays when the frequency is extremely higher than f_{max}. In other words, when $kr \gg N$ it cannot perform well, but MDAC$_{\text{Proposed}}$ beamformer performed exactly as in $k_2r = 8.1$. Because MDAC$_{\text{Proposed}}$ always uses a nearly optimum solution, the frequency does not affect its performance. In this case, MDAC$_{\text{Proposed}}$ improved DI by 42%, even higher improvement than it has at lower frequencies.

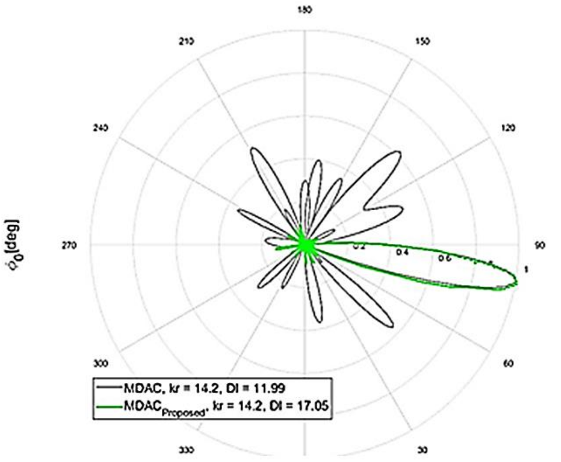


Fig. 3: A comparison between MDAC and MDAC$_{\text{Proposed}}$ beam patterns at the frequency $k_3r = 14.2$ ($f_3 = 7.6$ kHz). The gray plot represents MDAC beam pattern with $\text{DI} = 11.99$ and green plot represents MDAC$_{\text{Proposed}}$ beam pattern with $\text{DI} = 17.05$.

B. MGAC_{Proposed} Simulation

For this simulation, the same configuration as in the previous section for microphone array is used. Beamformers are compared at $k_1 r = 3.6$ and $k_2 r = 8.1$ which represent frequencies of $f_1 = 1.9\,\text{kHz}$ and $f_2 = 4.4\,\text{kHz}$ respectively. Table 2 shows the WNG of maximum WNG beamformer (MG), MGAC, and $\text{MGAC}_{\text{Proposed}}$ beamformers. As it shows, all of the beamformers attained the WNG of 15.8 at $k_1 r$. This is because of the fact that the condition $kr < N$ is satisfied at $k_1 r = 3.1$. But in $k_2 r = 8.1$ the WNG drops for MG beamformer due to the beamformer's bandwidth limitation. It can be seen that WNG is improved by nearly 15% in $\text{MGAC}_{\text{Proposed}}$ compared to MGAC beamformer.

For more comparison between MGAC and $\text{MGAC}_{\text{Propsed}}$ beamformers, Table 3 shows DI of these two beamformers at frequencies $k_1 r$ and $k_2 r$ in which it can be seen that $\text{MGAC}_{\text{Propseed}}$ improved DI by 3%. Although this is not a major improvement, it should be noted that improving WNG does not necessarily improve DI [30], but it is worthwhile to note that while improving the WNG, $\text{MGAC}_{\text{Proposed}}$ also improved DI.

Table 2: A comparison between WNG of MG, MGAC, and $\text{MGAC}_{\text{Proposed}}$ beamformers in two different frequencies $k_1 r =$ 3.6 and $k_2\, r = 8.1$

Beamformer	WNG with $kr = 3.6$	WNG with $kr = 8.1$
MG	15.8	10.93
MGAC	15.8	15.91
$\text{MGAC}_{\text{Proposed}}$	15.8	16.54

Table 3: A comparison between DI of MG, MGAC, and $\text{MGAC}_{\text{Proposed}}$ beamformers in two different frequencies $k_1 r =$ 3.6 and $k_2\, r = 8.1$

Beamformer	DI (dB) with $kr = 3.6$	DI (dB) with $kr = 8.1$
MG	12.3	8.71
MGAC	12.3	13.7
$\text{MGAC}_{\text{Proposed}}$	12.3	14.21

Results and Discussion

The simulation results show that the proposed method has the ability to extend the operational bandwidth of a spherical microphone array and consequently achieves higher DI and lower side lobe level compared to MDAC and MD beamformers in high frequencies. In addition to the MDAC beamformer, this method can be extended to other beamforming methods such as maximum WNG beamformer. Simulation results show improvement over those beamformers as well.

Conclusion

Aliasing and, in consequence, unwanted side lobe formation is the main factor in spherical microphone arrays operational bandwidth determination. Most of the methods previously presented to reduce aliasing demanded physical changes in the array structure which comes at a cost.

Recently Alon and Rafely presented a method to reduce aliasing in high frequencies without demanding a physical change in array. This method is based on including the effect of spatial aliasing in the definition of desired objective. Their approach found to be useful in developing aliasing cancellation capability for beamformers. In this paper we proposed a new method based on Alon and Rafaely's approach via designing a constrained optimization problem using orthogonality property of spherical harmonics, to achieve better performance.

The proposed method achieves higher DI and lower side lobe level compared to MDAC and MD beamformers in high frequencies and results show that it has improved the MDAC beamformer performance by nearly 19%. In addition to the MDAC beamformer, this method can be extended to other beamforming methods such as maximum WNG beamformer. Simulation results show improvement of WNG by nearly 15% over MGAC beamformer.

Because the optimization problem in this method can be of very high dimension, obtaining the unaliased signal can be time consuming. Thus, the presented method cannot be applied to real time purposes. For future work it can be considered to develop this approach for real time applications. Also, the derivation of formula in this work is based on the plane wave assumption for arriving waves.

So, we assume implicitly far field waves in our proposed method. Aliasing cancellation beamformer for near field sound waves is also suggested for future work.

Author Contributions

M. Kalantari proposed the method. M. Kalantari and M. Mohammadpour Tuyserkani implemented the proposed method, interpreted the results and wrote the manuscript. Solving the optimization problems carried out by S. H. Amiri. He also contributed in writing the manuscript.

Acknowledgments

This work was supported by Shahid Rajaee Teacher Training University under contract number 15947.

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

MVDR	Minimum variance distortionless response
MD	Maximum directivity
MG	Maximum white noise gain
DI	Directivity index
MDAC	Maximum directivity beamformer with aliasing cancellation
MGAC	Maximum white noise gain beamformer with aliasing cancellation
MVDR-AC	Minimum variance distortionless response with aliasing cancellation
SNR	Signal-to-noise ratio
WNG	White noise gain
α_q	Sampling weights
$\boldsymbol{\alpha}$	Vector of sampling weights
θ	Elevation angle
ϕ	Azimuth angle
$a(\cdot)$	Plane-wave decomposition in the space domain
a_{nm}	Plane-wave decomposition in the spherical-harmonics domain
$b_n(\cdot)$	Function relating pressure to plane-wave decomposition
DF	Directivity factor
d_n	Axis-symmetric beamforming weighting function
$j_n(\cdot)$	Spherical Bessel function of the first kind
k	Wave number
$\mathbf{k}$	Wave vector denoting propagation direction
$\tilde{\mathbf{k}}$	Wave vector denoting arrival direction
N	Order of spherical harmonics
p	Sound pressure in the space domain
p_{nm}	Sound pressure in the spherical harmonics domain
$\mathbf{p}$	Sound pressure vector in the space domain
$\mathbf{p_{nm}}$	Sound pressure vector in the spherical harmonics domain
Q	Number of samples or microphones
$\mathbf{r}$	Vector of spherical coordinates
$\mathbf{v}$	Steering vector in the space domain
$\mathbf{v_{nm}}$	Steering vector in the spherical harmonics domain
$w(\cdot)$	Beamforming weighting function in the space domain
w_{nm}	Beamforming weighting function in the spherical harmonics domain
$\mathbf{w}$	Beamforming weighting vector in the space domain
$\mathbf{w_{nm}}$	Beamforming weighting vector in the spherical harmonics domain
$Y_n^m(\cdot)$	Spherical harmonics
$\mathbf{Y}$	Matrix of Spherical harmonics

References

[1] D. Alon, B. Rafaely, "Beamforming with optimal aliasing cancellation in spherical microphone arrays," IEEE/ACM Trans. Audio Speech Lang. Process., 24(1): 196-210 2016.

[2] B. Bernschütz, "Bandwidth extension for microphone arrays," in Proc. 133th AES Convention, San Francisco, USA: 1–10, 2012.

[3] B. Bernschütz, "Microphone arrays and sound field decomposition for dynamic binaural recording", Ph.D. thesis, Technische Universität Berlin, 2016.

[4] W.H. Liao, Y. Mitsufuji, K. Osako, K. Ohkuri, "Microphone array geometry for two dimensional broadband sound field recording," in Proc. Audio Engineering Society Convention, 2018.

[5] T.D. Abhayapala, D.B. Ward, "Theory and design of high order sound field microphones using spherical microphone array," in Proc. ICASSP: 1949–1952, 2002.

[6] B. Rafaely, "Spatial sampling and beamforming for spherical microphone arrays," in Proc. Hands-Free Speech Communication and Microphone Arrays (HSCMA): 5-8, 2008.

[7] J. Meyer, G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the sound field", in Proc. IEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP): II–1781–II–1784, 2002.

[8] Z. Li, R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," IEEE Trans. Audio Speech Lang. Process., 15(2): 702–714, 2007.

[9] H. Beit-On, B. Rafaely, "Focusing and frequency smoothing for arbitrary arrays with application to speaker localization," IEEE/ACM Trans. Audio Speech Lang. Process., 28: 2184-2193, 2020.

[10] M.R.P. Thomas, "Practical concentric open sphere cardioid microphone array design for higher order sound field capture," in Proc. ICASSP-IEEE International Conf. Acoustics, Speech and Signal Processing (ICASSP): 666-670, 2019.

[11] D.P. Jarret, E.A.P. Habets, P.A. Naylor, Theory and Applications of Spherical Microphone Array Processing, Springer, 2017.

[12] B. Rafaely, "Plane-Wave decomposition of the sound field on a sphere by spherical convolution," J. Acous. Soc. Am., 116(4): 2149–2157, 2004.

[13] V. Pulkki, S. Delikaris-Manias, A. Politis, "Spatial decomposition by spherical array processing," in Proc. Parametric Time-Frequency Domain Spatial Audio, IEEE: 25-47, 2018.

[14] J. R. Driscoll, D. M. Healy, "Computing Fourier transforms and convolutions on the 2-sphere," Adv. Appl. Math., 15(2): 202–250, 1994.

[15] A.H. Moore, M. Brookes, P.A. Naylor, "Robust spherical harmonic domain interpolation of spatially sampled array manifolds," in Proc. IEEE International Conf. Acoustics, Speech and Signal Processing (ICASSP), 2017.

[16] B. Rafaely, "Analysis and design of spherical microphone arrays," IEEE Trans. Speech Audio Process., 13(1): 135–143, 2005.

[17] U. Elahi, Z. Khalid, R.A. Kennedy, "Design of a spatially constrained anti-aliasing filter using slepian functions on the sphere," in Proc.

13th International Conf. Signal Processing and Communication Systems (ICSPCS): 1-6, 2019.

[18] B. Bernschutz, “Bandwidth extension for microphone arrays”, in Proc. Audio Engineering Society Convention 133, Audio Engineering Society, 2012.

[19] U. Elahi, Z. Khalid, R.A. Kennedy, “Spatially constrained anti-aliasing filter using slepian eigenfunction window on the sphere,” in Proc. 12th International Conf. Signal Processing and Communication Systems (ICSPCS): 1-6, 2018.

[20] J. Lin, X. Wu, T. Qu, “Anti spatial aliasing HOA encoding method based on aliasing projection matrix,” in Proc. IEEE 3rd International Conf. Information Communication and Signal Processing (ICICSP): 321-325, 2020.

[21] J. Dmochowski, J. Benesty, S. Affes, “On spatial aliasing in microphone arrays,” IEEE Trans. Signal Process., 57(4): 1383–1395, 2009.

[22] X. Zhao, G. Huang, J. Chen, J. Benesty, “An improved solution to the frequency-invariant beamforming with concentric circular microphone arrays,” in Proc. IEEE International Conf. Acoustics, Speech and Signal Processing (ICASSP): 556-560, 2020.

[23] A. Macovski, “Ultrasonic imaging using arrays,” in Proc. IEEE, 67(4): 484–495, 1979.

[24] B. Rafaely, B. Weiss, E. Bachmat, “Spatial aliasing in spherical microphone arrays,” IEEE Trans. Signal Process., 55(3): 1003–1010, 2007.

[25] M. Agmon, B. Rafaely, J. Tabrikian, “Maximum directivity beamformer for spherical-aperture microphones,” in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA): 153–156, 2009.

[26] V. Tourbabin, B. Rafaely, “Sub-Nyquist spatial sampling using arrays of directional microphones,” in Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA): 76–80, 2011.

[27] S. Brown, V. Sethu, D. Taubman, “Spatial wiener filter to reduce spatial aliasing with spherical microphone arrays,” J. Acous. Soc. Am., 145(4): 2254-2264, 2019.

[28] A.D. Firoozabadi, P. Irarrazaval, P. Adasme, H. Durney, M.S. Olave, “A novel quasi-spherical nested microphone array and multiresolution modified SRP by gammatone filterbank for multiple speakers localization,” in Proc. Signal Process.: Algorithms, Architectures, Arrangements, and Applications (SPA): 208-213 2019.

[29] B. Rafaely, Fundamentals of Spherical Arrays Processing, Springer, 2nd edn., 2018.

[30] B. Rafaely, “Phase-Mode versus delay-and-sum spherical microphone array processing,” IEEE Signal Process. Lett., 12(10): 713–716, 2005.

Biography

Mohammad Kalantari received B.Sc. degree in Computer Engineering from Iran University of Science and Technology (IUST), Tehran, Iran and M.Sc. and Ph.D. in Computer Engineering from Amirkabir University of Technology (AUT), Tehran, Iran in 2001 and 2009 respectively. He is currently working as Assistant Professor at Signal Processing Laboratory in Computer Engineering Department at Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran. His area of interest includes, statistical signal processing, spherical array processing, sampling theory and compressed sensing.

- Email: mkalantari@sru.ac.ir
- ORCID: 0000-0002-6852-9344
- Web of Science Researcher ID: NA
- Scopus Author ID: 55893680300
- Homepage: https://www.sru.ac.ir/kalantari/

Mohammadmehdi Mohammadpour Tuyserkani received B.Sc. degree in Computer Engineering from Shahed University, Tehran, Iran in 2016. He is currently working as a Research Assistant in Signal Processing Laboratory in the faculty of Computer Engineering at Shahid Rajaee Teacher Training University (SRTTU), Tehran, Iran. His area of interest includes spherical array processing, statistical signal processing and machine learning.

- Email: mm.mohammadpour@sru.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Seyed Hamid Amiri received B.Sc. degree in Software Engineering from Shahid Bahonar University, Kerman, Iran and M.Sc. and Ph.D. in Artificial Intelligence from Sharif University of Technology (SUT), Tehran, Iran in 2009 and 2015 respectively. He is currently working as Assistant Professor in Computer Engineering Department at Shahid Rajaee Teacher Training University. His area of interest includes computer vision and video processing, statistical signal processing, statistical pattern recognition and numerical optimization in artificial intelligence.

- Email: s.hamidamiri@sru.ac.ir
- ORCID: 0000-0002-4723-896X
- Web of Science Researcher ID: NA
- Scopus Author ID: 57197243231
- Homepage: https://www.sru.ac.ir/hamidamiri/

How to cite this paper:

M. Kalantari, M. Mohammadpour Tuyserkani, S.H. Amiri, “A general approach for operational bandwidth extension in spherical microphone array,” J. Electr. Comput. Eng. Innovations, 10(2): 393-402, 2022.

DOI: 10.22061/JECEI.2021.8304.501

URL: https://jecei.sru.ac.ir/article_1680.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Design and Optimization of a Dual Polarized Hat Feed Reflector Antenna

M. Bod*, F. Geran Gharakhili

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

Article Info

Article History:

Received 04 November 2021
Reviewed 15 December 2021
Revised 12 January 2022
Accepted 26 February 2022

Keywords:

Self-supported Feed
Ring-focus reflector
Genetic algorithm (GA) optimization
Orthogonal mode transducer

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

Abstract

Background and Objectives: Self-supported rear-radiating feeds have been widely used as reflector antenna feeds for mini terrestrial and satellite links. While in most terrestrial and satellite links a dual-polarized antenna for send and receive applications are required, all of the reported works regarding this topic are presenting a single polarized self-supported reflector antenna. In this paper, a dual-polarized hat feed reflector antenna with a low sidelobe and low cross-polarization level is presented.
Methods: The proposed antenna consists of an orthogonal mode transducer (OMT), a 60 cm ring focus reflector, and a rear radiating waveguide feed known as the hat feed. 21 parameters of hat feed structure are selected and optimized with a genetic algorithm (GA). A predefined ring focus curve is used as a reflector in the optimization procedure. Dual polarization for send and receive applications is also obtained by an OMT at the rear side of the reflector antenna.
Results: A prototype of the proposed hat feed reflector antenna is fabricated and the measurement results are compared with simulation ones. The proposed antenna has return loss better than 15 dB at both polarizations in the 17.7~19.7 GHz frequency range. The 60cm reflector antenna has 40dBi gain which means that the proposed antenna has about 70% radiation efficiency. About 20dB sidelobe level and more than 40 dB cross-polarization have also been realized in the measurement patterns of the proposed antenna.
Conclusion: A dual-polarized hat feed reflector antenna with excellent radiation efficiency, high sidelobe, and low cross-polarization level is proposed. The proposed antenna can be a good candidate for high-frequency terrestrial and satellite communications.

Introduction

Self-supported rear-radiating feeds have been widely used over the past years as reflector antenna feeds for mini terrestrial and satellite links [1]-[3]. In these feeds, mechanical support is provided by the feeding waveguides that extend from the reflector vertex to the feed, and additional struts are omitted from the feed structures.

This geometry makes it possible to simply locate the transmitter and receiver at the rear side of the reflector. Until now Different types of Self-supported feeds such as splash plate feed [4]-[6], cup feed [7]-[9], hat feed [10]-[20], and even some types of microstrip feed [21]-[24], have been designed and used for the reflector antenna.

Hat feeds are one of the compact well-known self-supporting feeds for reflector antennas. The hat feeds usually consist of a waveguide neck, a dielectric head,

Doi: 10.22061/JECEI.2022.8484.525

and a corrugated hat. Due to the uniform radiation pattern, hat feeds have low cross-polarization levels, low far-out sidelobe, and high efficiency. However, these excellent performances are usually provided in a very narrow bandwidth (less than 10%) which is limited the hat feeds applications. In recent years some approaches, to improve the impedance bandwidth of the hat feed reflector antenna are proposed [15]-[18].

In [16], a hat feed reflector antenna was designed by genetic algorithm (GA) optimization. The designed antenna has 33%, impedance bandwidth, however, the radiation performance of the antenna was not very well in the designed frequency band. Another hat feed reflector antenna without the dielectric head and 26% bandwidth is reported in [17].

However, due to the use of a non-symmetrical structure, the body of revelation (BOR) efficiency of the designed antenna was very low.

A recent development in the hat feed reflector antenna is reported in [18]. In that paper, a single polarized Ku-band hat feed reflector antenna is designed via a comprehensive GA optimization. A ring focus reflector has also been used for obtaining nearly 100% phase efficiency. The final antenna has a low sidelobe level and low cross-polarization in the whole operational impedance bandwidth and fulfills the requirements of the ETSI EN 302 standard for terrestrial fixed radio systems [25].

It should be said that all of the papers mentioned above are presenting a single polarized hat feed reflector antenna. While in most terrestrial and satellite links a dual-polarized high gain antenna for send and receive applications is required 0-[29]. The effects of dual-polarization on the performance of an optimized hat feed reflector antenna are not studied well until now.

In this paper, a novel dual-polarized hat feed reflector antenna for 17.7~19.7 GHz frequency band is presented. The proposed hat feed structure is optimized via GA with a novel strategy that considers reflector efficiency at two different polarization and feeds impedance bandwidths simultaneously. The final antenna has a 60 cm ring focus reflector, a corrugated hat feed, a self-supported transition waveguide, and an orthogonal mode transducer (OMT) at the rear side of the reflector. A prototype of the proposed antenna is fabricated and the measurement results are compared with the simulations.

The effects of dual-polarization on the measured reflection coefficient and the measured radiation pattern are carefully studied. The measurement results show the proposed dual-polarized hat feed reflector antenna is a good candidate for high-frequency point-to-point communication.

This paper is organized as follows. The antenna structure is introduced in the second section. The optimization producer of the proposed antenna is explained in the third section. In the fourth section, results and discussion have been presented. The conclusion is provided in the fifth section.

Antenna Design

The configuration of the proposed dual-polarized hat feed reflector antenna is shown in Fig. 1. As it can be seen in this figure the proposed antenna consists of a compact hat feed structure, a 60 cm ring focus reflector, and an OMT structure at the rear side of the reflector to create the dual-polarization.

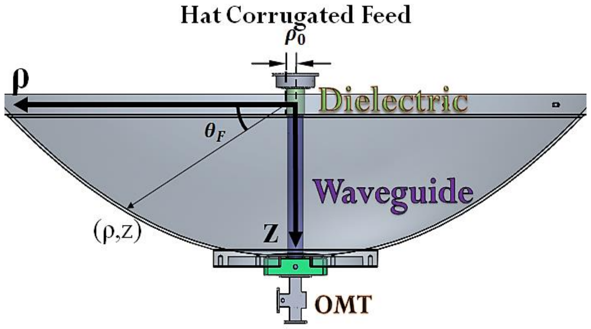


Fig. 1: Configuration of the proposed dual-polarized hat feed reflector antenna which consists of an orthogonal mode transducer (OMT), a 60 cm ring focus reflector, and a rear radiating hat feed.

A cross-section of the hat feed structure is shown in Fig. 2 (a). As it can be seen the hat feed antenna is consisting of a metallic head, a circular waveguide, and a Teflon dielectric which is attached to these parts.

The metallic head of the feed antenna has several symmetrical metallic corrugations in its body. These corrugations are used to improve the reflector illuminations and reduce the sidelobe level of the reflector antenna in the operational bandwidth. The width and the height of these corrugations are determined from GA optimization.

A dielectric head, so-called the antenna neck, is made of a Teflon dielectric with $\varepsilon_r = 2.2$ and $tan\delta = 0.001$. This neck surrounds the circular waveguide to provide a rigid structure.

A cylindrical air transition is also created in this dielectric neck to improve the impedance matching of the hat feed antenna. The parameters of this transition are also determined by GA optimization.

The metallic waveguide in this structure has two main roles. First, it is used as the mechanical supporter of the feed structure, and second, it is used as an electromagnetic transition between the OMT output and the dielectric head. The inner diameter of this cylindrical waveguide at the hat end and OMT output is 14mm and

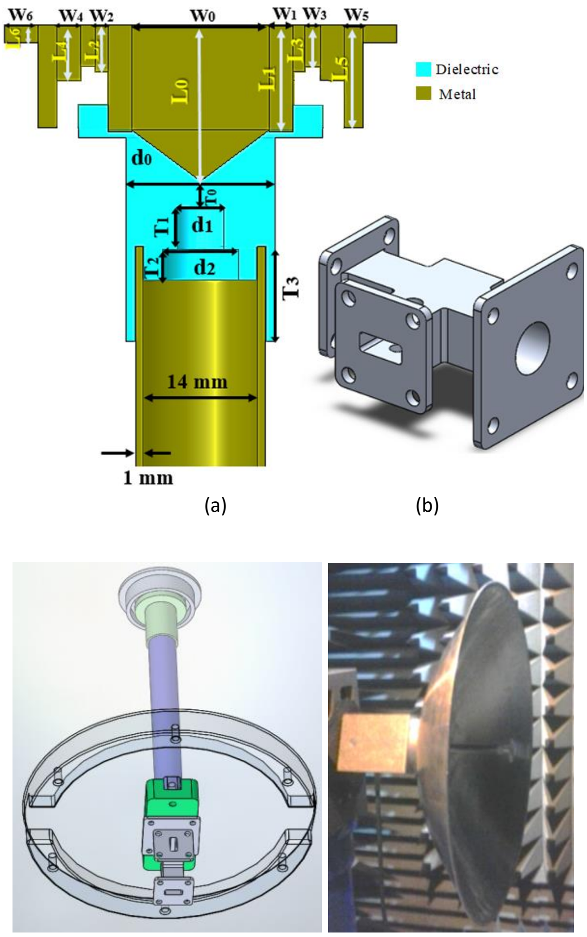
11mm respectively, while the outer diameter is fixed at 15mm.

To improve the radiation efficiency of the proposed antenna, a ring focus parabolic antenna is used as the antenna reflector. The shape of the ring focus reflector antenna is defined as follows:

$$\rho = 2F\tan(\theta_f\ /2)+\rho_0 \qquad (1)$$

$$z\ = F - F\tan^2\left(\theta_f\ /2\right)$$

In which ϑ_f is the polar angle of the feed as shown in Fig. 1, *F* is the focal length of the ring focus reflector and is chosen to be 15 cm, and ρ_0 is the radius of the reflectors focus ring and is selected as 14 mm. The diameter of this reflector is about 60cm and the value of the *F/D* is chosen to be 0.25. This value for *F/D* is chosen according to [16], and [18] to have optimum radiation efficiency in the designed antenna.



(a) (b)

(c) (d)

Fig.2: Configuration of the proposed hat feed reflector structure. (a) The cross-section of that feed antenna, (b) the configuration of the OMT for creating dual-polarization, (c) the alignment of the hat feed and OMT, (d) the fabricated prototype of the proposed hat feed reflector antenna.

To create a dual-polarization radiation with the proposed hat feed reflector antenna, an orthogonal mode transducer (OMT) [29] is designed and placed at the rear side of the proposed reflector antenna. As shown in Fig. 2 (b) and (c) this Ku band OMT consists of two WR51 rectangular flanges and an 11mm cylindrical output. About 35 dB isolation between OMT ports is considered and the cylindrical output of the OMT excites the cylindrical waveguide of the hat feed antenna with two different polarizations. To have tuned in the fabrication process of the dual-polarized hat feed some metallic screws are inserted at the cylindrical waveguide of the OMT.

Table 1: The final optimized value of the proposed dual-polarized hat feed reflector antenna. [mm]

Para.	Value	Para.	Value	Para.	Value
L_0	19.7	W_0	16.85	d_0	18.4
L_1	13.1	W_1	2.95	d_1	5.65
L_2	5.72	W_2	1.49	d_2	9.31
L_3	5.13	W_3	1.86	T_0	3.18
L_4	6.85	W_4	2.95	T_1	5.17
L_5	12.7	W_5	2.33	T_2	3.78
L_6	2.26	W_6	4.08	T_3	11.65

Optimization Procedure

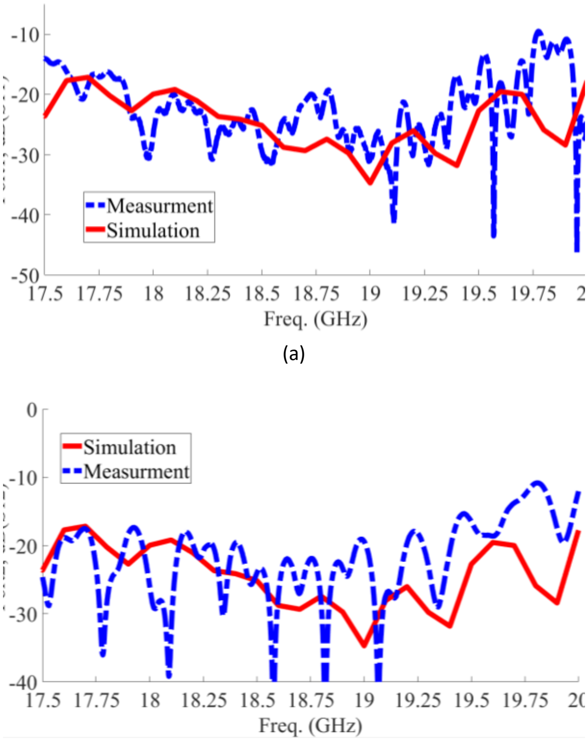
The proposed dual-polarized antenna should cover a 17.7 ~ 19.7 GHz frequency band with a reflection coefficient less than -15dB. The boresight gain of this reflector antenna should be greater than 38.7 dBi at both polarizations of 18.7 GHz frequency. Furthermore, more than 20 dB sidelobe level in the whole frequency band of operation is needed.

To fulfill these requirements, as shown in Fig. 2 (a), 21 parameters of the hat feed reflector antenna are optimized with a genetic algorithm. These parameters change the corrugations and the dielectric transitions of the hat feed antenna.

The GA is implemented in MATLAB software. A population of 50 randomly generated chromosomes is created in the GA routine in MATLAB. These chromosomes are interpreted as the value of the 21 parameters of the hat feed antenna. The hat feed antenna and the OMT are simulated with an FEM solver in the well-known HFSS software. The link between HFSS and MATLAB is created by Visual Basic scripting (VB) [30]. The reflection coefficient and the radiation pattern of the hat feed antenna and OMT are extracted from the

HFSS. The phase center of the simulated feed is also calculated in this step.

At the next step, the feed radiation pattern is used to illuminate the predefined 60 cm ring focus reflector antenna in HFSS-IE. The feed phase center is placed at the focal center of the reflector antenna and the radiation pattern of the reflector antenna is calculated with the method of moments in HFSS. The handoff process between FEM and IE solver of HFSS is also done with VB scripting.



(a)

(b)

Fig. 3: The measured and simulated return loss of the proposed dual-polarized hat feed reflector antenna from OMT output ports (a) first OMT output port, (b) second OMT output port.

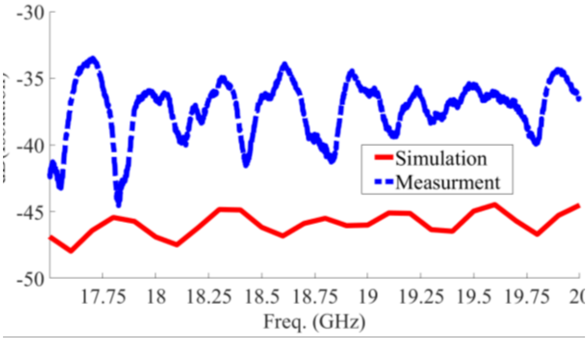


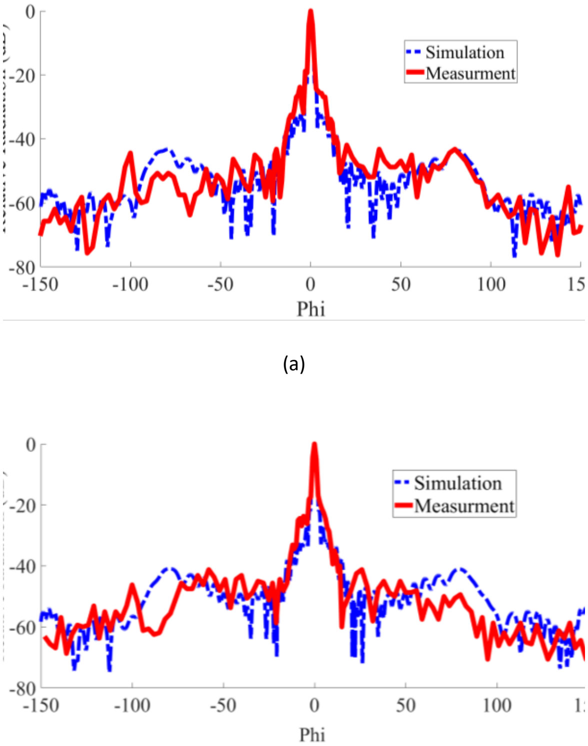
Fig. 4: The measured and simulated isolation of the proposed dual-polarized hat feed reflector antenna.

After obtaining the radiation pattern of the whole reflector antenna, the generated chromosome should be evaluated with a cost function. Therefore, a multi-criteria function is defined as follows:

$$F = \sum_{i=1}^{N}\left[w_{1i}\left(E_1(f_i)\right)^2 + w_{2i}\left(E_2(f_i)\right)^2\right] \tag{2}$$

in which N is the number of the sampling frequency, w_i represents the weighting value at the ith sample and f_i is the ith sampling frequency. E_1 and E_2 are the error functions for the reflection coefficient and peak gain of the evaluated antenna respectively and can be defined as follows:

$$E_1(f_i) = \begin{cases} |S_{11}(f_i)| - 0.09 & |S11| > 0.09 \\ 0 & |S11| \le 0.09 \end{cases} \tag{3}$$



(a)

(b)

Fig. 5: The simulated and measured normalized radiation pattern of the proposed hat feed reflector antenna at each polarization at 18 GHz (a) first OMT output port, (b) second OMT output port.

$$E_2(f_i) = \left(15 + 20\log(f_i)\right) - PeakGain(f_i) \tag{4}$$

In (3) the value of 0.09 corresponds to the return loss of -20 dB and the value of 15 corresponds to the peak gain of 60 cm reflector antenna with 80% aperture efficiency. The aperture efficiency of the reflector antenna is related to the antenna peak Gain with the following formula [31]:

$$\eta(\%) = \frac{G\lambda_0^2}{4\pi A_{phy}} \qquad (5)$$

in which η is the aperture efficiency of the proposed antenna, *G* is the value of antenna peak gain at the corresponding wavelength λ_0, and A_{phy} is the physical aperture area of the reflector antenna.

To accelerate the GA optimization, the cost function (2) is evaluated at six equally spaced frequencies sampled in the 17.7~ 19.7 GHz frequencies band. After evaluating each chromosome of a generation, the next generation is obtained from the previous ones by 80% crossover function, 14% tournament selections, and 6% mutation function.

The GA is converged to the desired results after passing 70 generations. The final optimized value for each parameter is shown in Table .1.

Results and Discussion

A prototype of the proposed dual-polarized hat feed reflector antenna is fabricated as shown in Fig. 2 (d). The metallic head is inserted in the dialectic neck and sticks to the dielectric neck and the metallic waveguide with a high-temperature glue.

This metallic waveguide is then screwed to the rear side of the ring focus reflector and the OMT as shown in Fig. 2 (c).

The fabrication accuracy of the hat feed antenna and the surface accuracy of the reflector antenna are less than 0.1 mm and 0.4mm respectively.

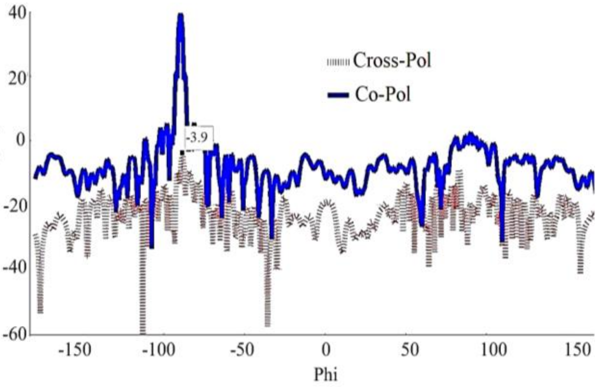


Fig. 6: The measured co and cross-polarization Gain of the proposed dual-polarized hat feed reflector antenna at 19 GHz frequency.

The return loss of the proposed antenna from OMT outputs is measured and compared with the simulation results in Fig. 3. The measured results in this figure are reported after some tuning with the tuning of metallic screws at the OMT. These screws are fixed in their optimum positions. From Fig. 3 it is clear that the proposed antenna has a return loss better than 15 dB in the whole frequency band of 17.7~19.7 GHz and for both OMT ports. The measured and simulated isolation between each port of the proposed antenna is shown in Fig. 4. As it can be seen more than 30 dB isolation between antenna ports is obtained in the measurement result.

The far-field radiation pattern of the proposed antenna for each polarization is also measured in the anechoic chamber. Fig. 5 compares the simulated and measured normalized radiation pattern of the proposed antenna at each polarization at 18 GHz frequency. As it can be seen in this figure the proposed antenna has a sidelobe level of about 20 dB at both polarizations. The front-to-back ratio better than 50 dB can also be realized from Fig. 5.

The measured peak gain and the cross-polarization of the proposed antenna are shown in Fig. 6. As it can be seen the proposed antenna has about 40 dBi Gain at 19 GHz frequency. According to (5) the proposed has about 70% aperture efficiency at 19 GHz frequency. This radiation efficiency can be considered as high efficiency in the Ku bands. From this figure, it is obvious that the proposed antenna has also more than 40 dB cross isolation.

Table 2: Comparison of the reported hat feed reflector antenna with the proposed antenna.

Ref.	Sim./ Meas.	Polarization	Freq. band	η(%) @Midpoint	SLL. @Midpoint	Bandwidth (%)
[13]	Sim.	Single	ka	78	40dB	N/A
[14]	Meas.	Single	K/Ka	70	10dB	N/A
[15]	Meas.	Single	Ku	74	20dB	47
[17]	Meas.	Single	Ku	63	15dB	26
[18]	Meas.	Single	Ku	65	25dB	30
This work	**Meas.**	**Dual Pol.**	**Ku**	**70**	**20dB**	**13**

Sim. /Meas.: Declaration of the nature of work as simulation or experiment.

N/A: Not available.

In order to make a fair assessment of the performance between the proposed hat feed reflector antenna and other similar antenna, Table 2 is also given. As can be seen in this table the proposed antenna is the first reported dual-polarized hat feed reflector antenna.

This antenna has an acceptable aperture efficiency and low side lobe level in compared to the previous works. It should be mentioned that the bandwidth of the proposed antenna is limited to the performance of the OMT structure and is different from single polarized antennas. From all of these results, it can be concluded that the proposed dual-polarized antenna is a good candidate for high-frequency terrestrial and satellite communications.

Conclusion

This paper describes the concept and design of a novel dual-polarized hat feed reflector antenna. The proposed hat feed antenna has a 60 cm ring focus reflector and is designed for 17.7~19.7 GHz frequency with a reflection coefficient less than -15 dB and radiation efficiency greater than 70%. To fulfill these requirements 21 parameters are defined at the metallic head and dielectric transition of the hat feed antenna. A two-step GA optimization process is also created with the HFSS MATLAB link.

A prototype of the optimized antenna with an orthogonal mode transducer is fabricated and the measured results are compared with the simulation ones.

The measured results show that for both input ports of the hat feed antenna a low reflection coefficient and high radiation efficiency can be realized. The proposed dual-polarized antenna can be used for high-frequency terrestrial and satellite communications.

Author Contributions

Both M. Bod and F. Geran developed the proposed antenna idea and performed the analytic simulations and measurements. M. Bod has written the manuscript and F. Geran 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

A_{phy}	the physical aperture area
BOR	body of revelation
E_1	the error functions of the reflection coefficient
E_2	the error functions of the peak gain
F	focal length
f_i	the ith sampling frequency
GA	genetic algorithm
G	antenna peak gain
OMT	orthogonal mode transducer
w_i	the weighting value at the ith sample
ϑ_f	polar angle of the feed
ρ_0	radius of the reflectors focus ring
η	aperture efficiency
λ_0	corresponding wavelength

References

[1] S. Rao, S. K. Sharma, L. Shafai., Handbook of reflector antennas and feed systems. Artech House, 2013.

[2] A. Rebollo, Á.F. Vaquero, M. Arrebola, M.R. Pino, "3D-Printed Dual-Reflector antenna with self-supported dielectric subreflector," IEEE Access, 8: 209091-209100, 2020.

[3] H.D. Lang, S. Flepp, H. Mathis, "Self-Supporting circularly polarized backfire helix feed antenna with reflector and director for deep dish reflector antennas in the L-Band," in Proc. 2021 XXXIVth General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS): 1-4, 2021.

[4] C. Liu, S. Yang, Z. Nie, "Design of a parabolic reflector antenna with a compact splash-plate feed," in Proc. Cross Strait Quad-Regional Radio Science and Wireless Technology Conference: 24-244, 2013.

[5] Y. Asci, E. Curuk, K. Yegin, C. Ozdemir, "Improved splash-plate feed parabolic reflector antenna for Ka-Band VSAT applications," in Proc. 46th European Microwave Conference: 1283-1286, 2016.

[6] P. Tang, "A design of high performance splash-plate feed for parabolic reflector antenna," in Proc. Sixth Asia-Pacific Conference on Antennas and Propagation: 1-3, 2017.

[7] G.T. Poulton, T.S. Bird, "Improved rear-radiating waveguide cup feeds," in Proc. International Symposium on Antennas and Propagation, 24 :79-82, 1986.

[8] R. Schwerdtfeger, "A coaxial dual mode feed system," in Proc. 9th European Microwave Conference: 196-200, 1979.

[9] A. Moallemizadeh, R. Sarraf-Shirazi, M. Bod, "Design of a novel compact cup feed for parabolic reflector antennas," Prog. Electromag. Res. Lett., 64: 81-86, 2016.

[10] M.K. Dwivedi, A.K. Sharma, "Compact reflector antenna system for ku-band satcom on the move (SOTM)," in Proc. 6th International Conference on Signal Processing and Integrated Networks: 254-257,2019.

[11] P.S. Kildal, "The hat feed: A dual-mode rear-radiating waveguide antenna having low cross polarization," IEEE Trans. Antennas Propag., 35(9):1010-1016, 1987.

[12] J. Yang, P.S. Kildal, "Calculation of ring-shaped phase centers of feeds for ring-focus paraboloids," IEEE Trans. Antennas Propag., 48(4): 524-528, 2000.

[13] Y. Zhao, B. Zhu, Y. Meng, Z. Yan, "Design of a hat feed for ring focus reflector antenna," in Proc. International Conference on Microwave and Millimeter Wave Technology, 2019.

[14] F. Greco, L. Boccia, E. Arnieri, G. Amendola, "K/Ka-Band cylindrical reflector antenna for compact satellite earth terminals," in Proc. IEEE Trans. Antennas Propag., 67(8): 5662-5667, 2019.

[15] E. Geterud, J. Yang, T. Ostling, “Wideband hat-fed reflector antenna for satellite communications,” in Proc. 5th Eur. Conf. on Antennas Propag., Rome, 2011.

[16] M. Denstedt, T. Ostling, J. Yang, P.S. Kildal, “Tripling bandwidth of hat feed by genetic algorithm optimization,” in Proc. IEEE AP-S Int. Symp., Hawaii, 2007.

[17] W. Wei, J. Yang, T. Ostling, T. Schafer, “A new hat feed for reflector antennas realized without dielectrics for reducing manufacturing cost and improving reflection coefficient,” IET Microwaves Antennas Propag., 5(7): 837–843, 2011.

[18] E.G. Geterud, J. Yang, T. Ostling, P. Bergmark, "Design and optimization of a compact wideband hat-fed reflector antenna for satellite communications," IEEE Trans. Antennas Propag., 61(1): 125-133, 2013.

[19] F. Greco, G. Amendola, L. Boccia, E. Arnieri, "A dual band hat feed for reflector antennas in Q-V band," in Proc. 10th Eur. Conf. on Antennas Propag., 2016.

[20] Y.J. Zhao, B.C. Zhu, Y.B. Meng, Z.H. Yan, "Design of a hat feed for ring focus reflector antenna," in Proc. International Conference on Microwave and Millimeter Wave Technology, 2019.

[21] A.A. Kishk, L. Shafai, "Optimization of microstrip feed geometry for prime focus reflector antennas," IEEE Trans. Antennas Propag., 37(4): 445-451, 1989.

[22] M.Q.E. Maula, L. Shafai, "Low-Cost, microstrip-fed printed dipole for prime focus reflector feed," IEEE Trans. Antennas Propag., 60: 5428-5433, 2012.

[23] T. Chen, H. Wu, "Dual-Polarized planar reflector feed for direct broadcast satellite systems," IEEE Antennas Wireless Propag. Lett., 9: 693-696, 2010.

[24] M.Q.E. Maula, L. Shafai, Z. A. Pour, "A Corrugated printed dipole antenna with equal beamwidths," IEEE Trans. Antennas Propag., 62: 1469-1474, 2014.

[25] Fixed Radio Systems Characteristics and Requirements; Final Draft, ETSI EN 302 217-4-2 V1.5.1 (2009-09), ETSI [Online].

[26] Z. Zhang, Y. Zhao, N. Liu, L. Ji, S. Zuo, G. Fu, "Design of a dual-beam dual-polarized offset parabolic reflector antenna," IEEE Trans. Antennas Propag., 67(2): 712-718, 2019.

[27] A. Mehrabani, L. Shafai, "Compact dual circularly polarized primary feeds for symmetric parabolic reflector antennas," IEEE Antennas Wireless Propag. Lett., 15: 922-925, 2016,

[28] R. Olsson, P. Kildal, S. Weinreb, "The eleven antenna: a compact low-profile decade bandwidth dual polarized feed for reflector antennas," IEEE Trans. Antennas Propag., 54(2): 368-375, Feb. 2006.

[29] S. Manshari, S. Koziel, L. Leifsson, "Compact dual-polarized corrugated horn antenna for satellite communications," IEEE Trans. Antennas Propag., 68(7): 5122-5129, July 2020.

[30] A. R. Mallahzadeh, M. Bod" Method for designing low-pass filters with a sharp cut-off," IET Microwaves Antennas Propag. 8: 10–15, 2014.

[31] C.A. Balanis, Antenna Theory: Analysis and Design, John Wiley & Sons, 2005.

Biographies

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 with the Amirkabir University with a fellowship awarded by the Iran National Science Foundation (INSF). He is currently an Assistant Professor with the Electrical Engineering Department, Shahid Rajaee Teacher Training University, Tehran. He has authored or co-authored 16 journal papers and one Persian book. 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
- Web of Science Researcher ID: AAH-5551-2019
- Scopus Author ID: 54918504900
- Homepage: https://www.sru.ac.ir/mohammadbod/

Fatemeh Geran was born in Ghaemshar, Iran in 1977. She received her BSc degree in Electrical Engineering (Telecommunication) from Tehran University, Tehran, Iran in 1999. Also, she received her M.Sc. and Ph.D. degrees in Electrical Engineering (Telecommunication) from Tarbiat Modares University, Tehran, Iran, in 2003 and 2009, respectively. She is currently an associate professor in the Faculty of Electrical Engineering at Shahid Rajaee Teacher Training University, Tehran, Iran. Her research interest fields are antenna, RF subsystems in the microwave and mm-wave bands, and RF energy harvesting.

- Email: f.geran@sru.ac.ir
- ORCID: 0000-0002-7845-8391
- Web of Science Researcher ID: AAN-8757-2020
- Scopus Author ID: 16038985700
- Homepage: https://www.sru.ac.ir/geran-2/

How to cite this paper:

M. Bod, F. Geran Gharakhili, "Design and optimization of a dual polarized hat feed reflector antenna," J. Electr. Comput. Eng. Innovations, 10(2): 403-410, 2022.

DOI: 10.22061/JECEI.2022.8484.525

URL: https://jecei.sru.ac.ir/article_1683.html

SRTTU

Journal of Electrical and Computer Engineering Innovations (JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Improving the Diagnosis of COVID-19 using a Combination of Deep Learning Models

***I. Zabbah*[1]*, K. Layeghi*[1,*]*, R. Ebrahimpour*[2]**

[1]*Department of Computer Engineering, Islamic Azad University, North Tehran Branch, Tehran, Iran.*

[2]*Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.*

Article Info

Article History:

Received 03 November 2021
Reviewed 17 December 2021
Revised 13 February 2022
Accepted 14 February 2022

Keywords:

COVID-19
Deep models
Data mining
Machine learning
Mixture of experts

*Corresponding Author's Email Address: *K_layeghi@iau-tnb.ac.ir*

Abstract

Background and Objectives: COVID-19 disease still has a devastating effect on society health. The use of X-ray images is one of the most important methods of diagnosing the disease. One of the challenges specialists are faced is no diagnosing in time. Using Deep learning can reduce the diagnostic error of COVID-19 and help specialists in this field.

Methods: The aim of this model is to provide a method based on a combination of deep learning(s) in parallel so that it can lead to more accurate results in COVID-19 disease by gathering opinions. In this research, 4 pre-trained (fine-tuned) deep model have been used. The dataset of this study is X-ray images from Github containing 1125 samples in 3 classes include normal, COVID-19 and pneumonia contaminated.

Results: In all networks, 70% of the samples were used for training and 30% for testing. To ensure accuracy, the K-fold method was used in the training process. After modeling and comparing the generated models and recording the results, the accuracy of diagnosis of COVID-19 disease showed 84.3% and 87.2% when learners were not combined and experts were combined respectively.

Conclusion: The use of machine learning techniques can lead to the early diagnosis of COVID-19 and help physicians to accelerate the healing process. This study shows that a combination of deep experts leads to improve diagnosis accuracy.

Introduction

COVID-19 is a family of SARS viruses that was first observed in January 2020 in Wuhan, China [1]. Most coronaviruses are originated from animals and can be transmitted to humans due to their zoonotic nature. Among these viruses, SASR-COV and MERS-COV can cause death in humans [2]. COVID-19's mutated structure makes it difficult to find an effective solution to the disease, and this has led to the death of many people in various countries, including the United States, China, Italy, Iran, etc. [2], [3]. Unlike SARS, COVID-19 affects internal organs such as the liver and kidneys in addition to the respiratory system.

The virus can eventually lead to death by weakening the immune system [5]. Recently, the use of artificial intelligence methods with a deep learning approach in various areas of machine learning such as image recognition, image classification and segmentation has been considered by researchers [6], [7].

Attention to deep learning in the diagnosis of various diseases has led to valuable results [8], including diagnosis of coronary heart disease using deep learning [9], diagnosis of tumor and its volume in the lungs and Breast [12] diagnosis brain tumor [11] and classification of diabetic retinopathy [10].

Doi: 10.22061/JECEI.2022.8200.491

The use of deep learning in the diagnosis of diseases from x-ray images has also been considered, for example, a study conducted to diagnose pneumonia using this type of image [13]. Although accurate CT images are reliable and can detect lung infections, pneumonia and tumors and produce clearer images of tissues and organs, but using x-rays images is faster, easier, more accessible, cheaper and less harmful. Therefore, the diagnosis of COVID-19 from these images can help speed up the timely treatment of this disease [14], [15]. There are three general methods for applying deep learning from x-ray images: 1- using fine-tuned models, 2- using unfine-tuned models and 3- pre-trained models [16]. For example, in the one study by using Resnet Deep Learner, labels such as age and gender were used to classify the dataset, and finally, a classification was performed using an MLP classifier [17]. In another study, pneumonia was diagnosed using VGG-16 and inceptionV3 models. In this study, the dataset was divided into three classes: Partial Pneumonia, Viral Pneumonia and Normal Pneumonia, and the classifier SVM was used [18]. The use of deep convolutional learner with innovative architecture and using CT images has been done in some studies to diagnose COVID-19 [19]. Also, the use of two-dimensional and three-dimensional images and its detection by deep learners has been considered. [20]. In an another study a network called COVID-net [21], which expands and compresses layers in deep learning, it was claimed that it could increase the accuracy of diagnosing COVID-19 [22]. In another research, pre-trained networks such as inceptionv3, ResNet50, Resnet72 using the transfer method (Transfer Learning Method) were used to diagnose COVID disease from X-ray images [23]. Transfer learning is the reuse of a pre-trained model on a problem. It's currently very popular in deep learning because it can train deep neural networks with comparatively little data. This is very useful in the data science field since most real-world problems typically do not have millions of labeled data points to train such complex models. One of the most important challenges we face in diagnosing COVID 19 disease is how to increase the accuracy of learners [24].

In this study, an attempt has been made to increase the accuracy of detection of COVID-19 from X-ray images of The model presented here uses a combination of AlexNet, GoogleNet, SqueezeNet, and MobileNetv2 networks to provide a framework called deep experts combination. By summarizing the opinions of the above networks, the model can increase the accuracy of diagnosing COVID-19 disease. It may summarize the opinions of the above networks, and enhance the accuracy in the diagnosis of COVID-19.

Databases, Models, Proposed Method

A. Dataset

The dataset used in this study was obtained from the GitHub open-source repository shared by Dr. Joseph Cohen et al [25]. In this dataset, we only considered the x-ray images, and in total, there were 127 X-Ray images diagnosed with COVID-19, 500 X-ray images of healthy humans, and 500 cases identified as pneumonia. We studied 43 and 84 images of x-ray due to COVID-19 infected males and females respectively. They had an average age of 55 years. Fig. 1 shows examples of people with COVID-19 identified by professionals. In all deep models, 70% of the data is used for training and 30% for testing. K-fold cross-validation technique was also used to increase accuracy.

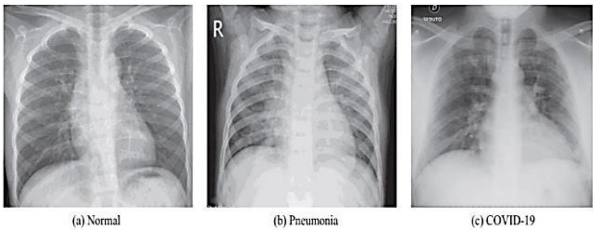


Fig. 1: An example of a normal image, pneumonia and COVID19.

B. MobileNet Deep Network (MobileNetv2)

MobileNet-v2 Network is used to image detection and segmentation. It can provide higher recognition accuracy than MobileNet while no more parameters and computation costs are needed. MobileNetv2 is very similar to the original MobileNet (MobileNetv1). In the other hand, MobileNetV1 has 13 depth-wise separable convolution whereas MobileNetV2 has got 17 of them. MobileNetV2 uses depth-wise separable convolution as efficient building blocks.

Table 1: Structure and parameters of MobileNet network

Type	Stride	Filter Size	Output Size
Convolution	2 ×2	3×3×3×32	224×224×۳
Convolution DW	1×1	3×3×32	112×112×32
Convolution	1×1	1×1×32×64	112×112×32
Convolution DW	2 ×2	3×3×64	112×112×64
Convolution	1×1	1×1×64×128	56×56×64
Convolution DW	1×1	3×3×128	56×56×128
Convolution	1×1	1×1×64×128	56×56×128
Convolution DW	2 ×2	3×3×128	56×56×128
Convolution	1×1	1×1×128×258	28×28×128
Convolution DW	1×1	3×3×256	28×28×256
Convolution	1×1	1×1×256×256	28×28×256
Convolution DW	2 ×2	3×3×256	28×28×256
Convolution	1×1	1×1×256×512	14×14×256
Convolution DW	1×1	3×3×512	14×14×512
Convolution	1×1	1×1×512×512	14×14×512

However, linear bottlenecks between the layers and shortcut connections between the bottlenecks are two new features of MobileNetV2. The basic structure is shown in Table 1.

The network has an image input size of 224×224×3. It is divided into two separate layers, which are depth wise convolution and pointwise convolution. Depth wise layer is responsible for filtering while pointwise layer for combining feature maps coming from different channels. By this splitting operation, the computation cost is greatly decreased [26]. After the depth-wise separable convolution, the ReLU operation on the low-dimensional features is easy to cause information loss [27].

C. Squeezenet Model

Now let us to describe the SqueezeNet architecture. SqueezeNet is a smaller CNN architecture that uses fewer parameters while obtaining accurate [28]. Table 2 shows that SqueezeNet begins with a convolution layer, followed by 8 Fire modules, ending with a final conv layer. The number of filters per fire module gradually is increased from the beginning to the end of the network. SqueezeNet performs max-pooling with a stride of 2 after layers first conv, fire4, fire8, and final conv. Fire modules architecture consist of two layers, squeeze layer and expand layer, both of them are the main key to SqueezeNet architecture. Squeeze layer is a layer composed of three convolution layers with each size 1 × 1. Expand layer is a layer composed of a combination of four lxl convolution layers and four 3x3 convolution layers. The full architecture is presented in Table 1.

Table 2: SqueezeNet architecture

Type	Stride	Filter Size	Output Size
Input	-	-	224×224×3
Convolution	2	96×96×7	109×109×96
Pooling	2	3×3	54×54×96
Fire 2	-	16×16×1, 64×1×1, 64×3×3	54×54×128
Fire 3	-	16×16×1, 64×1×1, 64×3×3	54×54×128
Fire 4	-	3×3×1, 128×1×1, 128×1×32	54×54×256
Pooling	2	3×3	27×27×256
Fire 5	-	32×1×1, 128×1×1, 128×3×3	27×27×256
Fire 6	-	48×1×1, 192×1×1, 192×3×3	27×27×384
Fire 7	-	48×1×1, 192×1×1, 192×3×3	27×27×384
Fire 8	-	64×1×1, 256×1×1, 256×3×3	27×27×512
Pooling	2	3×3	13×13×128
Fire 9	-	64×1×1, 256×1×1, 256×3×3	13×13×512
Convolution	1	6×13×13	13×13×6
Pooling	-	13×13	1×1×6

D. Alexnet Network Model [29]

AlexNet has eight layers; the first five were convolutional layers. It can be seen from Table 3 that the back of the first, second and fifth convolutional layers is the pooling layer. The calculation process of the convolutional layer is as follows and the last three were fully connected layers. It has been proven that the network has learned rich feature representations for a wide range of images. The network has an image input size of 227-by-227(is 224 × 224 × 3). Table 3 shows the parameters used in this study in the Alex network.

Table 3: Structure and parameters of AlexNet network

layer	Stride	Filter Size	Kernel
Convolution	4	55×55×96	11×11
Max Pooling	2	27×27×96	3×3
Convolution	1	27×27×256	5×5
Max Pooling	2	13×13×256	3×3
Convolution	1	13×13×384	3×3
Convolution	1	13×13×384	3×3
Convolution	1	13×13×256	3×3
Max Pooling	2	6×6×256	3×3
FC	-	9216	-
FC	-	4096	-
FC	-	4096	-
FC	-	1000	-

E. Google Net network model (GoogleNet) [30]

GoogleNet is a type of convolutional neural network based on the Inception architecture. Table 4 shows that, Convolution and max-pooling operations are performed on the input, respectively. Then sent into the next inception module. Fig. 2 shows a part of the Google Net model called Inception-v3, introduced in 2015 by Szegedy et al [31]. In this model, Kernels with dimensions of 1 × 1, 3 × 3 and 5 × 5 are applied to the input. The outputs of all these layers are merged together to be considered as the input of the next layer.

Table 4: The general structure and parameters of the Google Net network

layer	Stride	Filter Size
Convolution1	4	3×224×224
Max Pooling1	2	64×112×112
Convolution2	1	64×56×56
Max Pooling2	2	192×56×56
Inception3a	1	192×28×28
Inception3b	1	256×28×28
Max Pooling3	1	480×28×28
Inception4a	2	480×14×14
Inception4b	-	512×14×14
Inception4c	-	512×14×14
Inception4d	-	512×14×14
Inception4e	-	528×14×14
Max Pooling4		832×14×14
Inception5a		832×7×7
Inception5b		832×7×7
Max Pooling5		1024×7×7
FC		1024×1×1

F. The Classification Method

In a learning system, there must be a balance between the accuracy and the generalization ability. Feature selection and data validation in this learning system is done for the classification [32].

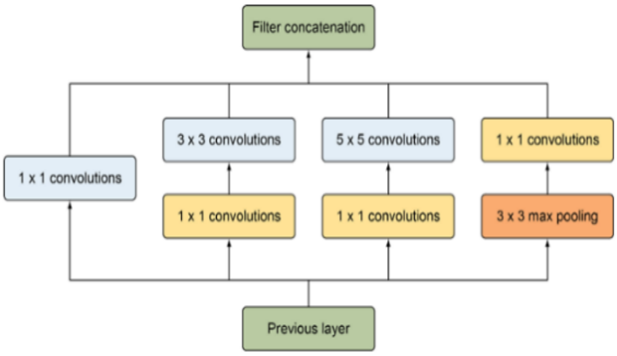


Fig. 2: Inception model.

Proposed Method

A crucial issue in CNN is to determine how much training data is needed to achieve a good level of the accuracy, especially in the field of medicine which is subjected to the lack of standardized data. For example, to train AlexNet, we need about 60 million images. In addition, training a Convolutional Neural Network (CNN) requires powerful hardware. To solve this problem, researchers use two methods: 1- fine-tuned networks 2- apply CNN to extract features and then send them to classifiers [32].

In the first method, the CNN network may not be able to train and predict the parameter because the numbers of the image are not enough. In this case, the network parameters can be reset. In the second method, (when we use a large amount of training data for the network) the number of images are enough and CNN can extract features from images and classifying them by a classifier such as SVM or MLP.

In this study, in order to diagnose COVID-19, the first method has been used which called transfer is learning. In deep learning, transfer learning is a technique whereby a neural network model is first trained on a problem similar to the problem that is being solved.

There are many models which have been trained with large amount of training data. For example, Transform Learning (TL) is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. Therefore, it allows rapid progress and improved performance.

The process of TL consists of two stages: 1) Choose a pre-trained Deep Learning model .2) rearranging the designed model based on the size of the dataset and its features. Fig. 3 shows how to use the TL technique in this study.

Nowadays, CNN models are widely used. This is because of its strong ability of high-speed parallel processing. In this study, we present combing the ideas of deep models to increase the accuracy of COVID-19 detection because the total experimental results of classifiers indicate increase in overall classification accuracy.

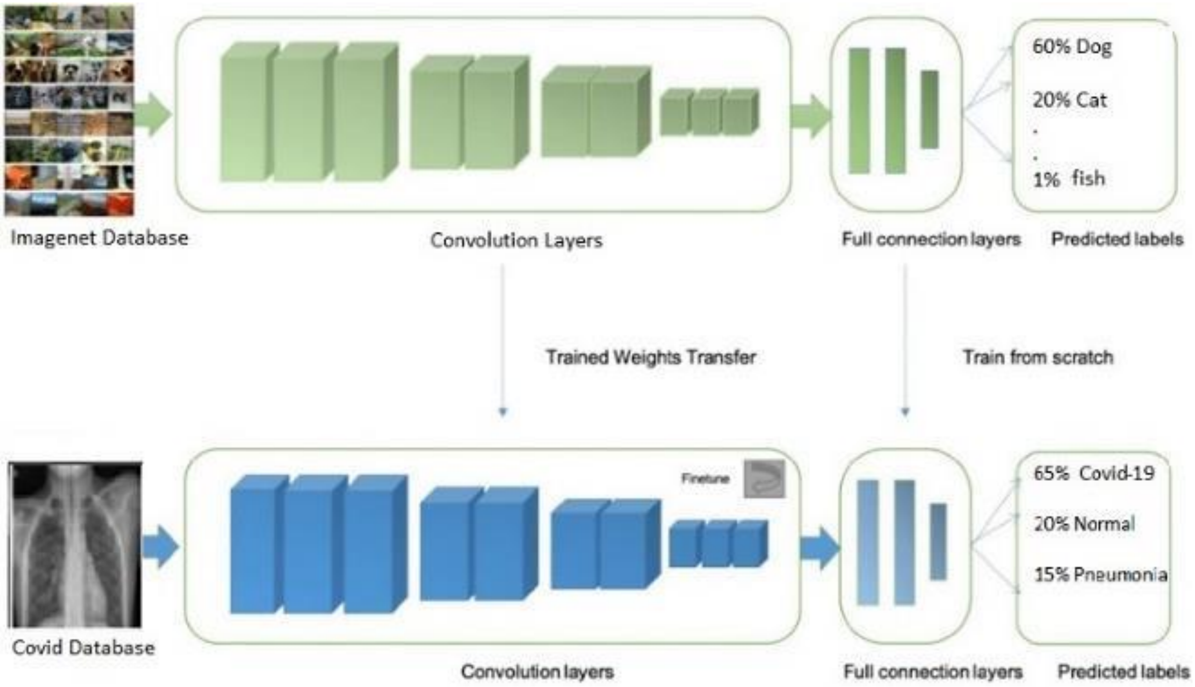


Fig. 3: Using the transfer learning technique to transfer weights to deep nets to diagnose COVID-19.

The proposed method used a combination of deep models for classification. This mechanism, called the combination of experts, divides the input space into subsets and then assigns each subset to a classifier [34]. These subsets are divided based on the target. The feature vector for the input was represented by the fully connected (FC) layer where each input is connected to all neurons. If present, FC layers are usually found towards the end of CNN architectures and can be used to optimize objectives [35]. The output is a vector with N components. N is the number of classes that classifiers predict them. In this study there are 3 classes include normal, Pneumonia and COVID-19. The purpose of CNN is to generate an output vector with N components. Each number shows the probability of belonging to the desired class. Probability can also be written as a percentage ,and the highest percentage indicates the final result.

The ultimate goal of the deep model is to reduce the error between predicted values and expected values which is referred to as the Cost Function. Generally, there are various methods for estimation of a cost function such as Gradient Descent, batch Gradient Descent, Stochastic Gradient Descent etc. In this study, a Stochastic descending gradient algorithm has been used to optimize the cost function in the diagnosis of COVID-19. Training a CNN classifier on small datasets does not work well. In contrast to this problem, various techniques are used such as the production of Synthetic Data or data augmentation. In this study, rotating, zooming and moving X-ray images have been used to augment data. In the proposed method, 4 deep models were used and after using the weight transfer technique in all deep models, the probability percentage was collected from the fully connected layer of each network. Finally, probability percentage was sent to a module called Majority Voting. A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN) used in this model for classification method. The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons (with threshold activation. Fig. 4 shows the proposed method. It should be mentioned that the Support Vector Machine (SVM), Decision Tree, or any other classification method can be used as an alternative. For example, if the probability of sample x belonging to class j in GoogleNet, Alex Net, MobileNetv2 and in Squeeze Net is p1, p2, p3 and p4 respectively, the majority vote of the experts is calculated based on (1):

$$p_{xe}^{j} = \frac{\sum_{i=1}^{n}(px_i^j)}{n} \qquad (1)$$

where px_i^j is the probability of sample x belongs to class j from point of view expert i, n number of experts (in this study n = 4). p_{xe}^{j} is the total expert opinion on the probability of sample x belongs to class j. It should be mentioned that p_{xe}^{j} maybe greater than 100. In order to solve this problem, (2) is used:

$$p_{xne}^{j} = \frac{100 \times p_{xe}^{j}}{\sum_{j=1}^{c}(p_{xe}^{j})} \qquad (2)$$

where p_{xe}^{j} is a total expert opinion about p_{xne}^{j} is the normalized probability. For example, in sample x if GoogleNet, AlexNet, MobileNetv2 and Squeeze Net with probabilities of 40%, 40%, 50% and 40% respectively, belonging to the normal class and with probabilities of 30%, 20%, 30% and 40% belonging to the pneumonia class and with a probability of 30%, 40%, 20% and 20% belonging to the COVID-19 class, total expert opinion to the normal class, the pneumonia class and the COVID-19 class is 42.5%, 30% and 27.5% respectively. Therefore, this sample belonging to the normal class. The training parameters used in the four networks are shown in Table 5.

Table 5: Parameters used in the training phase in deep networks

Learning Parameters	Google Net	Alex Net	Squeeze Net	Mobile Netv2
Learning rate	3e-4	3e-4	3e-4	3E-4
Batch size	10	10	10	10
Optimizer	NG*	NG*	SGD**	SGD**
Loss Function	C	C	C	C
Epochs per each Training Phase	100	100	100	100
Horizontal/Vertical flipping	yes	Yes	Yes	YES
Zoom Range	5٪	5٪	5٪	5٪
Rotation Range	٠	٠	٠	٠
Width/Height shifting	5٪	5٪	5٪	5٪
Shift Range	5٪	5٪	5٪	5٪
Re-scaling	1/25	1/25	1/25	1/25

SGD: Stochastic Gradian Descend NG: Nadam Categorical C: Crossentopy

Results and Discussion

The software MATLAB 2019 was used to train a deep model and a computer with Microsoft Windows 10 64-bit version was used for the experiment. Its specifications were as follows: Intel 3687 @ i7 processor, 8G of RAM, and 1G of graphics memory.

To assess the reliability of the proposed method we considered the following standard metrics: Accuracy, Sensitivity, Specificity and F1_score. These metrics are calculated on the concept of the true-positive (TP), true-negative (TN), false-positive (FP), and false-negative (FN) scores:

-TP is the amount of positive COVID-19 that were correctly labelled as positive.

- FP is the amount of negative (healthy) COVID-19 that was mislabeled as positive.

- TN is the amount of negative (healthy) COVID-19 that was correctly labelled as healthy.

- FN is the amount of positive COVID-19 that were mislabeled as negative (healthy).

Accuracy: Accuracy is one metric for evaluating classification models .it is computed as (3):

$$Accuracy = \frac{\sum True\ positive + True\ negative}{\sum Total\ population} \tag{3}$$

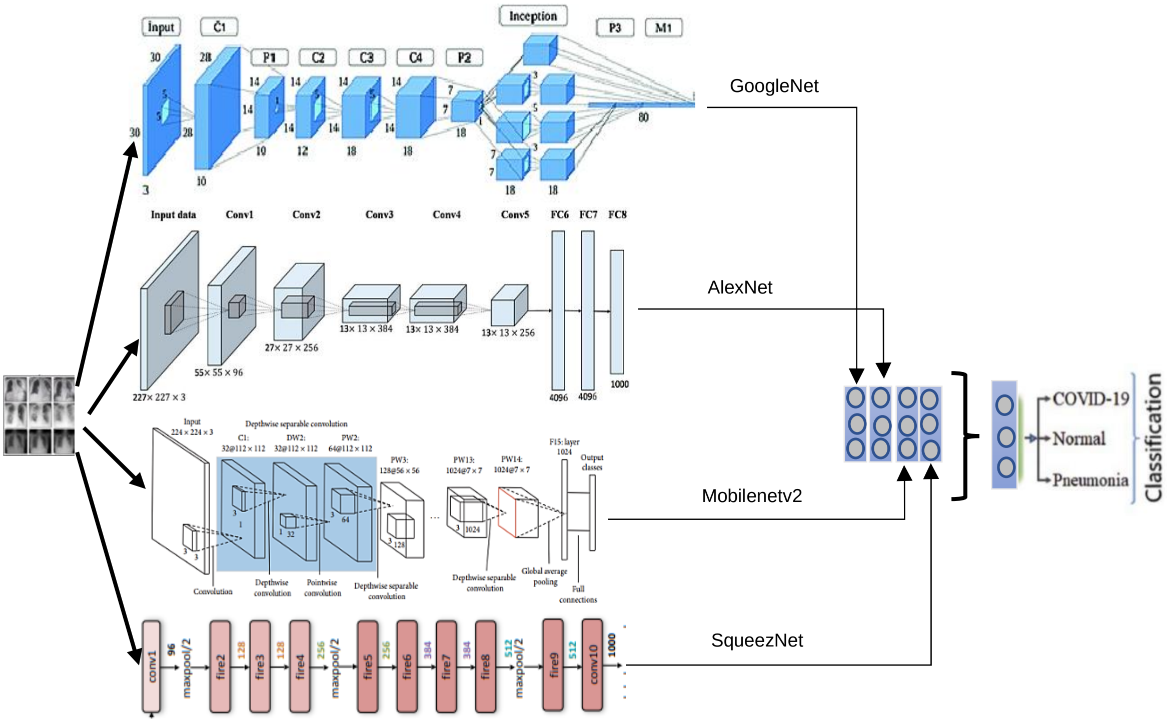


Fig. 4: Proposed method.

Sensitivity and specificity are metrics of the performance of classification test that is widely used in (4) and (5):

$$(Sensivity) = \frac{TP}{TP + FN} \tag{4}$$

$$(\text{Specificity}) = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{5}$$

The F1-score is a measure of a test's accuracy. It is calculated as (6):

$$(F1_{score}) = \frac{Sensivity \times Precision}{Sensivity + Precision} \times 2 \tag{6}$$

The obtained results according to the metrics mentioned are shows in Table 7 and Table 8. As mentioned, in order to evaluate the model k-fold cross-validation technique with k=5 has been used. At the end of 5-fold, the average accuracy was calculated. Fig. 5 shows how to use the k-fold method for five-class. The results obtained in some folds in the deep models of Google Net, AlexNet, MobileNet and SqueezeNet are shown in Fig. 6.

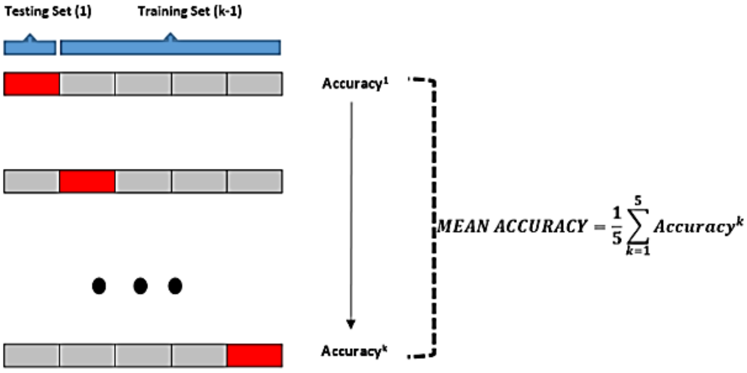
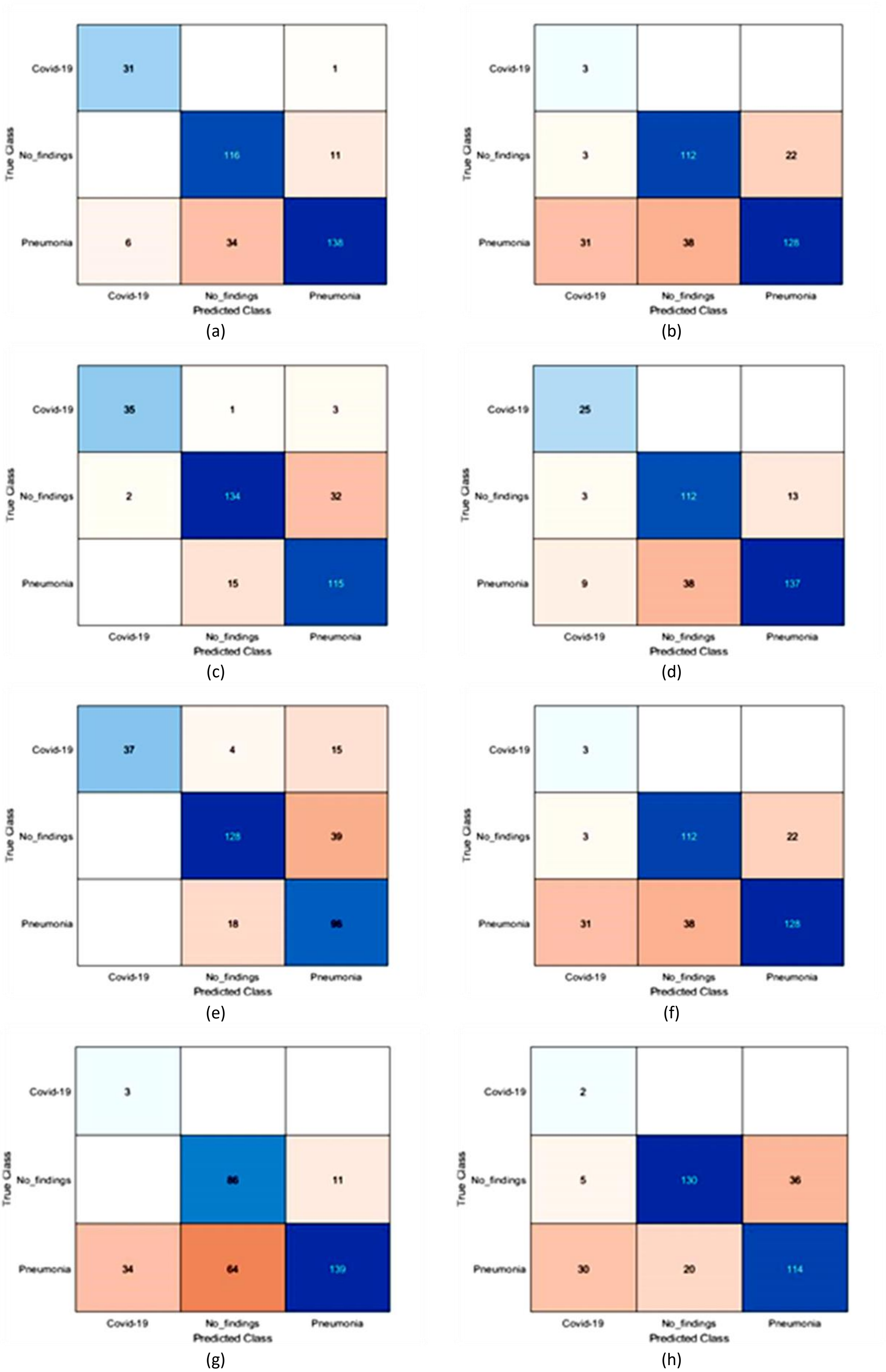


Fig. 5: k-fold cross-validation (k = 5).

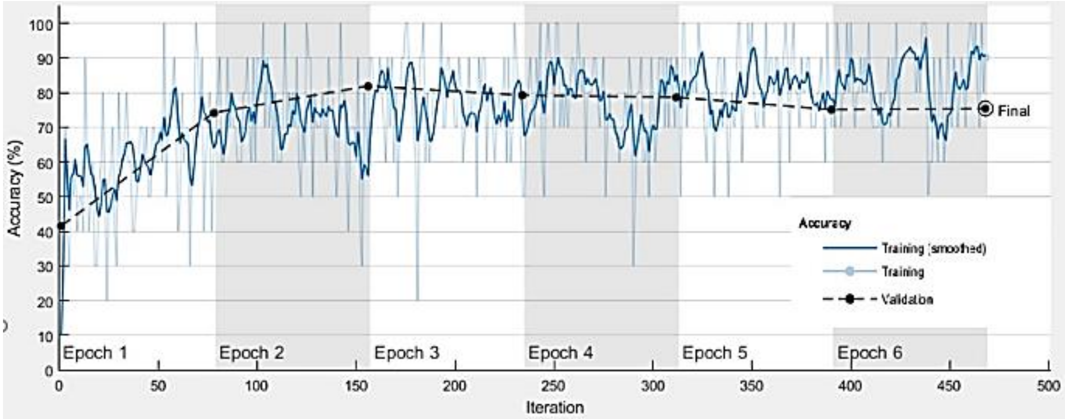


(a) (b) (c) (d) (e) (f) (g) (h)

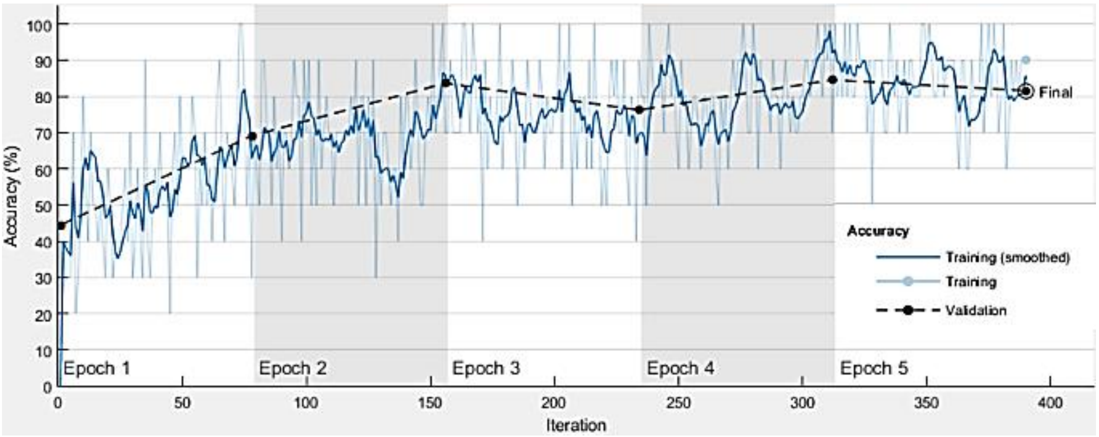
Fig. 6: The confusion matrix of the network for fold 1 and 3. (a): GoogleNet fold 3, (b): GoogleNet fold 1, (c) AlexNet fold 3, (d): AlexNet fold 1, (e): MobileNet fold 3, (f): MobileNet fold 1, (g): SqueezeNet fold 3, (h): SqueezeNet fold 1.

Table 6: Results of different folds in each deep models of Google Net, AlexNet, MobileNet and Skiesnet

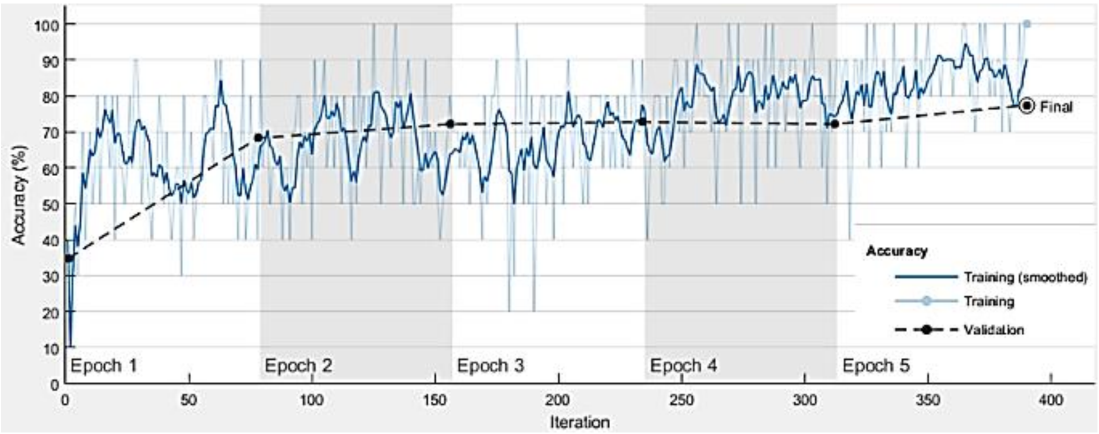
	Fold	COVID19 Correct detected	COVID19 Not detected	COVID19 Wrong detected	Pneumonia Correct detected	Pneumonia Not detected	Pneumonia Wrong detected	Normal Correct detected	Normal Not Detected	Normal Wrong detected	Acracy
MobileNet	1	33	3	4	112	25	28	129	28	21	٪84/3
MobileNet	2	36	7	1	111	11	39	129	33	11	٪84/9
MobileNet	3	32	1	5	127	27	23	127	23	23	٪84/9
MobileNet	4	35	3	2	113	14	37	137	35	13	٪84/6
MobileNet	5	36	6	1	106	14	44	136	39	14	٪82/5
GoogleNet	1	34	0	3	136	27	14	136	14	24	٪76
GoogleNet	2	31	1	6	138	40	12	138	34	11	٪75/1
GoogleNet	3	35	4	2	115	15	35	134	34	16	٪74/3
GoogleNet	4	34	3	3	130	16	20	129	31	21	٪76/9
GoogleNet	5	35	2	2	133	14	27	125	23	25	٪76/9
Alex Net	1	35	6	2	117	14	33	135	30	25	٪85/2
Alex Net	2	35	4	2	115	15	35	134	34	16	٪84/3
Alex Net	3	25	0	12	137	47	13	112	38	16	٪84/2
Alex Net	4	35	2	4	100	16	50	132	34	18	٪81/1
Alex Net	5	34	3	3	115	10	35	133	30	17	٪83/9
Squeeze Net	1	2	0	35	114	50	36	130	41	20	٪73
Squeeze Net	2	3	0	34	139	11	98	86	64	11	٪67/7
Squeeze Net	3	3	0	34	128	69	22	112	25	38	٪72/1
Squeeze Net	4	4	1	33	130	45	30	117	20	33	٪74/4
Squeeze Net	5	3	1	34	132	41	28	115	23	35	٪75



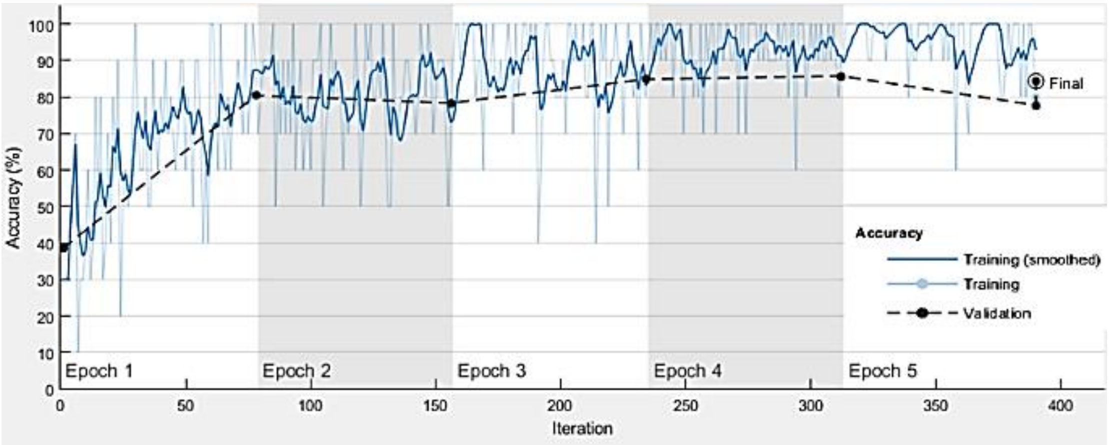
(a)



(b)



(c)



(d)

Fig. 7: Training and testing processes in the 4 network. (a): AlexNet, (b): GoogleNet, (c): SqueezsNet, (d): MobileNet.

Table 7: The number of true and false positives and false negatives for each network

	COVID19 Correct Detected	COVID19 Not Detected	COVID19 Wrong Detected	Pneumonia Correct Detected	Pneumonia Not Detected	Pneumonia Wrong Detected	Normal Correct Detected	Normal Not Detected	Normal Wrong Detected
GoogleNet	22	2	15	142	8	75	90	60	7
AlexNet	30	3	7	18	41	136	136	41	14
MobileNet	35	2	8	22	30	121	121	29	15
SqueezeNet	37	۰	19	54	18	128	128	22	39
Mixture of Experts	36	1	3	18	25	126	126	24	15

Table 8: Evaluation metrics

		Sensitivity	**Specificity**	**F1-Score**	**Final Accuracy**
GoogleNet	Covid19	0.59	0.99	0.72	**75.4%**
	Normal	0.60	0.96	0.72	
	Pneumonia	0.94	0.60	0.77	
AlexNet	Covid19	0.54	1	0.70	**81.6%**
	Normal	0.91	0.77	0.83	
	Pneumonia	0.84	0.85	0.77	
MobileNetv2	Covid19	0.94	0.97	0.87	**84.3%**
	Normal	0.80	0.91	0.85	
	Pneumonia	0.85	0.83	0.83	
SqueezeNet	Covid19	1	0.93	0.79	**77.4%**
	Normal	0.85	0.79	0.80	
	Pneumonia	0.64	0.90	0.72	
Mixture of Experst	Covid19	0.89	0.99	0.92	**87.2%**
	Normal	0.80	0.93	0.86	
	Pneumonia	0.91	0.81	0.85	

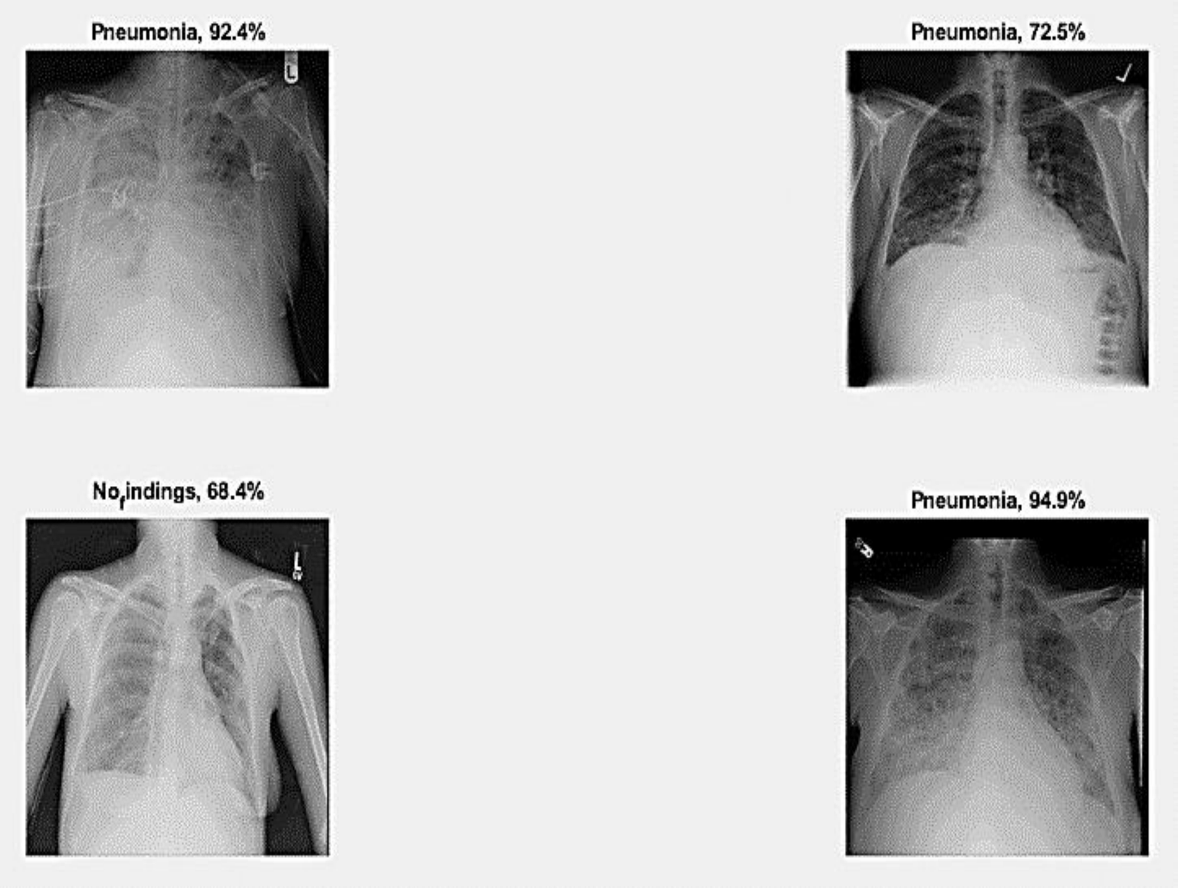


Fig. 8: The sample of Mixture of experts.

Conclusion

Classification with Deep Convolutional Neural Networks proposed by Alex Krizhevsky et al. built a new landscape in Computer Vision by destroying old ideas in one masterful stroke.

The paper used a CNN to get a Top-5 error rate (rate of not finding the true label of a given image among its top 5 predictions) of 15.3%. The next best result trailed far behind (26.2%).

This architecture is popularly called AlexNet. Number of parameters for alexnet is 60M parameters. One focus of GoogLeNet was to address the question of which sized convolution kernels are best. Previous popular networks employed choices as small as 1×1 and as large as 11×11. One insight in GoogLeNet was that sometimes it can be advantageous to employ a combination of variously-sized kernels.

The parameter for google net is 4 M parameters. MobileNetV2 is very similar to the original MobileNet, except that it uses inverted residual blocks with bottlenecking features. It has a drastically lower parameter count than the original MobileNet.

MobileNets support any input size greater than 32 x 32, with larger image sizes offering better performance.

Since the classification algorithms are not suited for all problems alone, combining expert's opinions can be a solution [36].

An idea in this research is the combination of deep learning networks in parallel. This method is also called majority vote or collective wisdom.

The results show that the combination of deep learning networks and the use of the majority voting method is more effective and efficient than one deep model.

The cause of improvement can be described as follows: Although each of the deep models can predict the disease COVID-19 alone, but the use of multiple convolution layers and various kernels that exist in pre-trained deep models, extract different features in different networks.

Therefore, in order to increase the accuracy of detection COVID-19, deep models are first performed alone, then their opinions are combined. One of the limitations of the proposed method is the use of MLP networks.

MLP networks used to recognize the real sample from the synthetic ones. Because of the lack of data in the minority class, MLPs might not be the best choice.

Regarding the advantages of the proposed method, it can be mention that: Although using the proposed architecture will give you better results, but it requires more time, which is one of the disadvantages of the method.

This issue can be considered in future studies. In the proposed model, synthetic samples are being generated in the feature space, in future work can used input space instead feature space. And in propose method we generated synthetic samples with MLP network.

We suggest using stronger, more diverse or a greater number of classifiers to solve this problem.

Author Contributions

I. Zabbah, K. Layeghi, and R. Ebrahimpour presented improving the Diagnosis of COVID-19 using a combination of Deep Learning Models. I. Zabbah examined each policy and wrote the manuscript. K. Layeghi and R. Ebrahimpour interpreted the results and improved the structure of paper.

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

MLP	Multi-Layer Perceptron
SVM	Support Vector Machine
CNN	Convolution Neural Network
TL	Transfer Learning
CTS	Computed Tomography Scan

References

[1] Z. Wu, J.M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention," Jama, 323(13): 1239-1242, 2020.

[2] W. Kong, P.P. Agarwal, "Chest imaging appearance of COVID-19 infection, radiology," Radiol. Cardiothorac. Imaging, 2(1): 1-4, 2020.

[3] T. Lancet, Editorial COVID-19: too little, too late? The Lancet 395(10226): 755, 2020.

[4] M.S. Razai, K. Doerholt, S. Ladhani, P. Oakeshott, “Coronavirus disease 2019 (COVID-19): a Guide for UK GPs” BMJ, 368(800): 1-5, 2020.

[5] X. Peng, X. Xu, Y. Li, L. Cheng, X. Zhou, B. Ren, "Transmission routes of 2019-nCoV and controls in dental practice," Int. J. oral sci., 12(1): 1-6, 2020.

[6] M. Togaçar, B. Ergen, Z. Comert, "Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders," Med. Hypotheses, 135, 2020.

[7] X. Liu, Z. Deng, Y. Yang, Liu, Xiaolong, Zhidong Deng, Yuhan Yang, "Recent progress in semantic image segmentation," Artif. Intell. Rev., 52(2): 1089-106, 2019.

[8] M. Zhang, X. H. Wang, Y. L. Chen, K. L. Zhao, Y. Q. Cai, C. L. An, M. G. Lin, X. D. Mu, "Clinical features of 2019 novel coronavirus pneumonia in the early stage from a fever clinic in Beijing," Zhonghua Jie He He Hu Xi Za Zhi, 43(0): 215-218, 2020.

[9] M. Zreik, N. Lessmann, R.W. Van Hamersvelt, J.M. Wolterink, M. Voskuil, M.A. Viergever, T. Leiner, I. sgum, "Deep learning analysis of the myocardium in coronary ct angiography for identifcation of patients with functionally significant coronary artery stenosis," Med. Image Anal., 44 (1): 72-85, 2018.

[10] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sanchez, "A survey on deep learning in medical image analysis," Med. Image Anal. 42(1): 60-88, 2017.

[11] S. Lakshmanaprabu, S.N. Mohanty, K. Shankar, N. Arunkumar, G. Ramirez, "Optimal deep learning model for classifcation of lung cancer on ct images," Future Generat. Comput. Syst., 9(2): 374–82, 2019.

[12] J.Z. Cheng, Y.H. Chou, J. Qin, C.M. Tiu, Y.C Chang, C.S Huang, D. Shen, C.M. Chen, "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans," Sci. rep., 6(1): 1–13, 2016.

[13] A.K. Jaiswal, P. Tiwari, S. Kumar, D. Gupta, A. Khanna, J.J.P.C. Rodrigues, "Identifying pneumonia in chest X-rays: a deep learning approach", Measurement, 145: 511–518,2019.

[14] P. An, H. Chen, X. Jiang, J. Su, Y. Xiao, Y. Ding, H. Ren, M. Ji, Y. Chen, W. Chen, et al., “Clinical characteristics of coronavirus disease (COVID-19) patients with gastrointestinal symptoms: A report of 164 cases,” Dig. Liv. Dis., 52(10): 1076-1079, 2020.

[15] M. Sherief, "Hepatic and gastrointestinal involvement in coronavirus disease (COVID-19): What do we know till now," Arab Gastroenterol., 21(1): 3-8. (2020).

[16] I.M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," Sci. rep., 9(1): 1-10, 2019.

[17] E.E.D. Hemdan, M.A. Shouman, M.E. Karar, "A framework of deep learning classifiers to diagnose COVID-19 in X-Ray images," arXiv preprint arXiv:2003.11055, 2020.

[18] S.S. Yadav, S.M. Jadhav, “Deep convolutional neural network based medical image classification for disease diagnosis,” J. Big Data, 6(1): 1-18, 2019.

[19] S. Wang, B. Kang, J. Ma, X Zeng, M. Xiao, J. Guo, M. Cai, J. Yang, Y. Li, X. Meng, et al., “A deep learning algorithm using ct images to screen for corona virus disease (COVID-19),” Eur. Radiol., 31(8): 6096-6104, 2021.

[20] O. Gozes, M. Frid-Adar, H. Greenspan, P.D. Browning, H. Zhang, W. Ji, A. Bernheim, E. Siegel, " Rapid ai development cycle for the coronavirus (COVID-19) pandemic: initial results for automated detection & patient monitoring using deep learning ct image analysis," arXiv:2003.05037; 1-22, 2020.

[21] L. Wang, ZQ. Lin, A. Wong, "COVID-net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images," arXiv preprint arXiv: 2003.09871.

[22] L. Wang, A. Wong, "A tailored deep convolutional neural network design for detection of covid-19 cases from chest radiography images," J. Network Comput. Appl. 20(1): 1-12, 2020.

[23] A. Laghi, "Cautions about radiologic diagnosis of COVID-19 infection driven by artificial intelligence," Lancet Digital Health 2(5): e225, 2020.

[24] K.S. Lee, J.Y. Kim, E.T. Jeon, W. Choi, N. Kim, K. Lee, "Evaluation of scalability and degree of fine-tuning of deep convolutional neural networks for COVID-19 screening on chest X-ray images using explainable deep-learning algorithm," J. Pers. Med., 10(4), 213, 2019.

[25] J.P. Cohen, P. Morrison, L. Dao, “COVID-19 Image Data Collection,” arXiv:2003.11597, 2020.

[26] J. Zhang, Y. Xie, Y. Li, C. Shen, Y. Xia, "COVID-19 screening on chest x-ray images using deep learning based anomaly detection," arXiv preprint arXiv:2003.12338

[27] A. Akbari, H. Farsi, S. Mohamadzadeh, "Deep neural network with extracted features for social group detection," J. Electr. Comput. Eng. Innovations (JECEI), 9(1): 47-56, 2021.

[28] F. Ucar, U. Korkmaz, M. Ferhat, K. Deniz. "COVIDiagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease (COVID-19) from X-ray images," Med. Hypotheses 140: 109761, 2020.

[29] A. Krizhevsky, l. Sutskever, G. Hinton, “Imagenet classification with deep convolutional neural networks” part of Advances in Neural Information Processing Systems 25, 1097-1105, 2012.

[30] C. Szegedy, W. Liu,Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, “Going deeper with convolutions,” in Proc. the IEEE conference on computer vision and pattern recognition: 1-9, 2015.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, “Rethinking the inception architecture for computer vision,” in Proc. the IEEE conference on computer vision and pattern recognition: 2818-2826, 2016.

[32] J.P. Cohen, COVID-19 Image Data Collection, 2020.

[33] J. Khosravi, M. Shams Esfandabadi, R. Ebrahimpour, “Image registration based on sum of square difference cost function,” J. Electr. Comput. Eng. Innovations (JECEI), 6(2): 273-281, 2018.

[34] T. Zhou, H. Lu, Z. Yang, S. Qiu, B. Huo, Y. Dong, "The ensemble deep learning model for novel COVID-19 on CT images," Appl. Soft Comput. 98: 106885, 2021.

[35] E. Pazouki, M. Rahmati, "A variational level set approach to multiphase multi-object tracking in camera network base on deep features," J. Electr. Comput. Eng. Innovations (JECEI), 9(2): 203-214, 2021.

[36] S. Masoudnia, R. Ebrahimpour, “Mixture of experts: a literature survey,” Artif. Intell. Rev., 42(2): 275-293, 2014.

Biographies

Iman Zabbah is a member at the Faculty of computer engineering, Islamic Azad University of torbat-e-Heydariye, Iran. He obtained M.S. degree in the field of Artificial intelligence from the Islamic Azad University of Mashhad, Iran in July 2008. Currently, he is a Ph.D. student at Islamic Azad University of Tehran North Branch University, from 2015 to 2020. His area of interest includes artificial intelligence, data processing and Deep Learning.

- Email: imanzabbah@gmail.com
- ORCID: 0000-0003-1630-444X
- Web of Science Researcher ID: Na
- Scopus Author ID: Na
- Homepage: Na

Kamran Layeghi is professor at the Faculty of Computer Engineering, Azad University, Tehran North Branch, Tehran-Iran. He is Ph.D. graduate in the field of Robotics and Artificial Intelligence from Keele University, England in 1991. Dr. Layeghi is the Author or co-author of many national and international Journal and conference publications in his research areas such as Kinematics and dynamics of Robotics, Cognitive robotics, AI fields which include Machine Learning, Deep Learning, Neural Network, Genetic Algorithm, Fuzzy Logic, Pattern Recognition and Computer Vision.

- Email: K_layeghi@iau-tnb.ac.ir
- ORCID: 0000-0003-3238-8436
- Web of Science Researcher ID: Na
- Scopus Author ID: Na
- Homepage: Na

Reza Ebrahimpour is a professor at the Faculty of computer engineering, Shahid Rajaee Teacher Training University, Tehran, Iran. He obtained Ph.D. degree in the field of Cognitive Neuroscience from the SCS, IPM, Tehran, Iran in July 2007. Dr. Ebrahimpour is the author or co-author of more than 100 international journal and conference publications in his research areas, which include Cognitive and Systems Neuroscience, Human and Machine Vision, Decision Making and Object Recognition.

- Email: ebrahimpour@ipm.ir
- ORCID: 0000-0002-7013-8078
- Web of Science Researcher ID: AAH-8531-2019
- Scopus Author ID: 14021180400
- Homepage: https://www.sru.ac.ir/en/school-of-computer/reza-ebrahimpour/

How to cite this paper:

I. Zabbah, K. Layeghi, R. Ebrahimpour, "Improving the diagnosis of COVID-19 using a combination of deep learning models," J. Electr. Comput. Eng. Innovations, 10(2): 411-424, 2022.

DOI: 10.22061/JECEI.2022.8200.491

URL: https://jecei.sru.ac.ir/article_1684.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

An Adaptive Cubature Kalman filter for Target Tracking

R. Havangi*

Faculty of Electrical Engineering and Computer, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 20 December 2021
Reviewed 24 January 2022
Revised 13 February 2022
Accepted 12 March 2022

Keywords:
Target tracking
Cubature Kalman filter
Square root cubature Kalman filter

*Corresponding Author's Email Address: *Havangi@Birjand.ac.ir*

Abstract

Background and Objectives: The target tracking problem is an essential component of many engineering applications.The extended Kalman filter (EKF) is one of the most well-known suboptimal filter to solve target tracking. However, since EKF uses the first-order terms of the Taylor series nonlinear extension functions, it often makes large errors in the estimates of state. As a result, target tracking based on EKF may diverge.

Methods: In this manuscript, an adaptive square root cubature Kalman filter (ASRCKF) is poposed to solve the maneuvering target tracking problem. In the proposed method, the covariance of process and measurement noises is estimated adaptively. Thus, the performance of proposed method does not depend on the noise statistics and its performance is robust with unknown prior knowledge of the noise statistics. Morover, it has a consistently improved numerical stability why the matrices of covariance are guaranteed to remain semi- positive. The performance of the proposed method is compared with EKF, and the unscented Kalman filter (UKF) for target tracking problem.

Results: To evaluate the proposed method, many experiments is performed. The proposed method is evaluated on the non-maneuvering and maneuvering target tracking.

Conclusion: The results show that the proposed method has lower estimation errors with faster convergence rate than other methods. The proposed method can track the tates of moving target effectively and improve the accuracy of the system.

Introduction

The problem of target tracking is a basic problem in the fields of the civil and military. The purpose of target tracking problem is to estimate the velocity and position of a moving target from noisy measurements [1]-[2]. In the target tracking problem, the estimation of state is confronted with two problems: one is that the measurement and process noise cannot be accurately described and are usually not accurate. The second is that the measurement and process noise cannot be accurately described and are usually not accurate the nonlinearity of measurement and motion model [3]-[4]. Various nonlinear Bayesian approach are developed for the problem of target tracking in the literature, which aims to estimate the velocity and position of the target using measurements.

The EKF is a widely used nonlinear filter to target tracking [5]-[6]. An online adaptive Kalman filter is proposed for target tracking with unknown noise statistics in [7].

In this paper, the expectation maximization algorithm is employed to construct the noise can effectively estimate the one-step prediction mean vector, the one-step prediction error covariance matrix. In [8], a robust filter is presented to shape estimation of a maneuvering star-convex extended target based on adaptive extended

Doi: 10.22061/JECEI.2022.8168.486

Kalman filter. Basically, since EKF uses the Taylor first order approximation for nonlinear functions, it makes large errors [8]-[9]. Therefore, if it is very nonlinear, such as a target tracking problem, the error of estimation can be large or even divergent, and the filter is unstable [9]-[10].

To increase accuracy, UKF based on target tracking is introduced in literatures [11]-[14]. In this method, there is no need to calculate the Jacobin matrix of the nonlinear state and measurement equation [14]-[15]. In [16], target tracking based on square-root unscented Kalman filters is presented. This method, propagate not the covariance matrix itself but its singular value decomposition (SVD) factors instead.

Compared to EKF, UKF has better accuracy. However, UKF does not use for non-Gaussian distributions [17]-[18]. In addition, it's the computation load is heavy for high-dimensional systems such as target tracking consequently thus, the filter can be converged slowly.

In 2009, the cubature Kalman filter (CKF) is proposed [19]-[20].

The CKF creates cubature integral points, and these points are used to calculate the posterior probability of the system. In [21], a Gaussian-sum cubature Kalman filter (GSCKF) is proposed for the problem of tracking and it has excellent performance from the point of view of filter accuracy and consistency. In [22], a strong tracking cubature Kalman filter is proposed for target tracking problem.

A limitation of target tracking based on traditional CKF is that statistical characteristics of noises are assumed to known [23]-[25].

As a result, the development of CKF method is limited. To solve these problems, in this paper, the problem of target tracking based on ASRCKF is proposed. The target tracking based on ASRCKF is updated repeatedly by propagating square root factors of the mean and covariance of the state variable, which ensures the positive semi-definiteness and symmetry of the covariance matrix and thus improves numerical accuracy and stability.

The main contribution of the paper is that proposed adaptive algorithm has good filtering accuracy and strong robustness. This method improves the tracking ability of the SRCKF method for a maneuvering target. The proposed method can prevent potential filter divergence and enhances the numerical stability. Moreover, in the proposed method, the covariance of process and measurement noises is estimated adaptively.

Thus, the performance of proposed approach does not depend on the noise statistics and its performance is robust with unknown prior knowledge of the noise statistics. Simulation results show that the proposed method has a superior tracking performance.

The rest of manuscript is as follows. The target tracking formulation is presented in the second Section. In the next Section, the target tracking based on ASRCKF is proposed.

The results are given in the fourth Section. In the fifth Section, the conclusion is presented.

Target Tracking Formulation

The discrete-time dynamic equation of the target motion is as [26]-[27]:

$$X_k = F^r(X_{k-1})X_{k-1} + G\omega_{k-1}$$

$$G = \begin{bmatrix} \frac{T^2}{2} & T & 0 & 0 \\ 0 & 0 & \frac{T^2}{2} & T \end{bmatrix}^T \tag{1}$$

where G is the input matrix, the (x_k, y_k) is position components, $(\dot{x}_k, \dot{y}_k)$ is velocity components, T is a sampling interval and X_k is as $X_k = [x_k, \dot{x}_k, y_k, \dot{y}_k]^T$, ω_k is the process noise with covariance matrices Q_t. Moreover, F^r is transition matrix corresponding to mode r. The transition matrix for non-maneuvring target is as follows:

$$F^1(X_{t-1}) = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The coordinated turn model is the most common model for maneuvering targets. In this case, F^r is as [28]-[29]:

$$F^r(X_{t-1}) = \begin{bmatrix} 1 & \frac{\sin\Omega_t^{(r)}T}{\Omega_t^{(r)}} & 0 & \frac{1-\cos\Omega_t^{(r)}T}{\Omega_t^{(r)}} \\ 0 & \cos\Omega_t^{(r)}T & 0 & -\sin\Omega_t^{(r)}T \\ 0 & \frac{1-\cos\Omega_t^{(r)}T}{\Omega_t^{(r)}} & 1 & \frac{\sin\Omega_t^{(r)}T}{\Omega_t^{(r)}} \\ 0 & \sin\Omega_t^{(r)}T & 0 & \cos\Omega_t^{(r)}T \end{bmatrix}, \text{ r=2,3} \tag{3}$$

where $\Omega_t^{(3)} > 0$, $\Omega_t^{(2)} < 0$ are is anticlockwise and clockwise turn maneuver, respectively. Target observation model is as:

$$Z_k = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ \arctan(\frac{y_k}{x_k}) \end{bmatrix} + v_k$$

where_ Z_k denotes the measurement at k^{th} instant and v_k is measurement noise with covariance matrices R_t.

Cubature Rule

Calculating the Gaussian nonlinear transfer density is the most important step for Bayesian filtering in Gaussian domain [18]. It is as follows:

$$I = \int g(x) \mathrm{N}(x, P_x) dx \tag{4}$$

where n_x is dimension of state x, $\mathrm{N}(x, P)$ is the Gaussian prior density of x and g is the nonlinear function.

The third-degree spherical cubature rule to numerically calculate the integral I with $2n_x$ equal weighted cubature points is used in CKF [30]:

$$I = \frac{1}{2n_x} \sum_{j=1}^{2n_x} g(\sqrt{P_x} \xi_j + x) \tag{5}$$

where the cubature point ξ_j is as:

$$\xi_j = \begin{cases} \sqrt{n_x} e_i^T & \mathrm{i=1,2,\dots,n}_x \\ -\sqrt{n_x} e_i^T & \mathrm{i=n}_x\mathrm{+1,n}_x\mathrm{+2,\dots,2n}_x \end{cases}$$

where $e_i^T \in R^{n_x}$ is the ith column vector of $I_{n_x \times n_x}$.

Target Tracking Based ASRCKF

Assume at time k-1, $s_{K-1|k-1}$ is the square-root of the covariance matrix $P_{K-1|k-1}$, i.e. $P_{K-1|k-1} = s_{K-1|k-1} s_{K-1|k-1}^T$, the current state cubature points are computed as follows:

$$\chi_{k-1|k-1}^i = s_{k-1|k-1} I(i) + \hat{x}_{k-1|k-1} \qquad i = 1, \dots 2n_x \tag{6}$$

where $I(i)$ is as

$$I(i) = \begin{cases} \sqrt{n_x}\,[1]_i\,, & \mathrm{i=1,\dots,n}_x \\ -\sqrt{n_x}\,[1]_{i-n_x}\,, & i = n_x + 1, \dots, 2n_x \end{cases}$$

With $[1]_i$ is the i-th column vector of the $n \times n$ matrix I. The cubature points are transmitted in state equation as:

$$\chi_{k|k-1}^{i*} = f(\chi_{k-1|k-1}^i) \tag{7}$$

The predicted mean $\hat{x}_{k|k-1}$ is calculated using the transformed cubature points $\chi_{k|k-1}^{*i}$ as follows:

$$\hat{x}_{k|k-1} = \frac{1}{2n_x} \sum_{i=1}^{2n_x} \chi_{k|k-1}^i \tag{8}$$

$$S_{k|k-1} = \mathrm{Tria}\left\{ X_{k|k-1}, \sqrt{Q_k} \right\} \tag{9}$$

where $\mathrm{Tria}(.)$ is a general triangularization algorithm and $X_{k|k-1}$ is as:

$$X_{k|k-1} =$$
$$\frac{1}{\sqrt{2n_x}} \left[\chi_{k|k-1}^{1*} - \hat{x}_{k|k-1} \quad \chi_{k|k-1}^{2*} - \hat{x}_{k|k-1} \quad \dots \quad \chi_{k|k-1,}^{2n*} - \hat{x}_{k|k-1} \right] \tag{10}$$

When measurement is revisited, the cubature point set is calculated as follows:

$$\chi_{k|k-1}^i = s_{k|k-1} I(i) + \hat{x}_{k|k-1} \qquad i = 1, \dots 2n_x \tag{11}$$

The transformed cubature points are transmitted in measurement equation:

$$\chi_{k|k-1}^{i**} = h(\chi_{k|k-1}^i) \tag{12}$$

The mean values and square-root of the covariance matrix of predicted measurement points are estimated:

$$\hat{z}_{k|k-1} = \frac{1}{2n} \sum_{i=0}^{2n} \chi_{k|k-1}^{**i} \tag{13}$$

$$S_{zz,k|k-1} = \mathrm{Tria}\left\{ Z_{K|k-1}, S_R \right\} \tag{14}$$

where $Z_{K|k-1}$ is as:

$$Z_{k|k-1} =$$
$$\frac{1}{\sqrt{2n}} \left[\chi_{k|k-1}^{1**} - \hat{z}_{k|k-1} \quad \chi_{k|k-1}^{2**} - \hat{z}_{k|k-1} \quad \dots \quad \chi_{k|k-1}^{2n**} - \hat{z}_{k|k-1} \right]$$

The cross covariance S_{xz} between the states and measurements are:

$$S_{xz,k|k-1} = X_{k|k-1} Z_{k|k-1}^T \tag{15}$$

$$X_{k|k-1} =$$
$$\frac{1}{\sqrt{2n_x}} \left[\chi_{k|k-1}^{1} - \hat{x}_{k|k-1} \quad \chi_{k|k-1}^{2} - \hat{x}_{k|k-1} \quad \dots \quad \chi_{k|k-1,}^{2n} - \hat{x}_{k|k-1} \right] \tag{16}$$

The Kalman gain K_k is calculated by

$$K_k = (S_{xz,k|k-1} / S_{zz,k|k-1}^T) / S_{zz,k|k-1} \tag{17}$$

The updated state $\hat{x}_{k|k}$ and the square-root of covariance $S_{k|k}$ are obtained as follows:

$$\hat{x}_{k|k} = \hat{x}_{k|k-1} + K_k (z_k - \hat{z}_{k|k-1}) \tag{18}$$

$$S_{k|k} = Tria([X_{k|k-1} - K_k Y_{k|k-1}, K_k \sqrt{R_k}\,]) \tag{19}$$

According to the equations $S_{k|k-1}$ and $S_{k|k}$ in (9) and (19), respectively, it can be seen that Q_k and R_k have a great impact on their values. However, the noise statistics can change over time in the target tracking application.

The set of unknown statistical of noise needs to estimate with the error covariance and the state of system.

Any mismatch between real noises that affect the system and those that are assumed in SRCKF reduce the performance of SRCKF that can also have divergence. As a result, it is necessary to know that the Q_k and R_k matrices exactly.

In this paper, an adaptive SRCKF is proposed method that solves the SRCKF problems by estimating the Q_k and R_k matrices. Suppose the process and measurement noise are defined as $w_w \sim \mathrm{N}(0,Q_k)$ and $v_k \sim \mathrm{N}(0,R_k)$. To estimate the covariance of process and measurement noises, the function of posteriori density is assumed as follows:

$$J^* = p(X_k, Q_k, R \mid Z_k) \tag{20}$$

where $Z_k = [z_1 \ \ z_2 \ \ \ldots \ \ z_k]$ is the measurement vector and $X_k = [x_1 \ \ x_2 \ \ \ldots \ \ x_k]$ is the state vector. According to the properties of conditional probability, the J^* function can be written as:

$$J^* = \frac{P(Z_k \mid X_k, Q_k, R_k) P(X_k \mid Q_k, R_k) p(Q_k, R_k)}{P(Z_k)} \tag{21}$$

where $p(Q_k, R_k)$ is depend on with the priori information, which can be considered as constant value. As $p(Z_k)$ is not involved in the problem of optimization. As a result, the function of J^* can be written as follows:

$$J = P(Z_k \mid X_k, Q_k, R_k) P(X_k \mid Q_k, R_k) p(Q_k, R_k) \tag{22}$$

the term $p(Q_k, R_k)$ is a constant value why it calculates based on a priori information. The term $p(X_k \mid Q_k, R_k)$ in (22) can be calculated using the multiplicative theorem of conditional probability as follows.

$$\begin{aligned} p(X_k \mid Q_k, R_k) &= p(x_0) \prod_{j=1}^{k} p(x_j \mid x_{j-1}, Q_k) = \\ &\frac{1}{(2\pi)^{n/2} \left|P_{0|0}\right|^{1/2}} \exp(-\frac{1}{2} \|x_0 - \hat{x}_0\|^2_{P_{0|0}^{-1}}) \\ &\prod_{j=1}^{k} \frac{1}{(2\pi)^{n/2} |Q_k|^{1/2}} \exp(-\frac{1}{2} \|x_j - f(x_{j-1})\|^2_{Q_k^{-1}}) = \\ &M_1 |Q_k|^{-k/2} \exp\left\{ -\frac{1}{2} \left[\|x_0 - \hat{x}_0\|^2_{P_{0|0}^{-1}} + \sum_{j=1}^{k} \|x_j - f(x_{j-1})\|^2_{Q_k^{-1}} \right] \right\} \end{aligned} \tag{23}$$

where $M_1 = \dfrac{1}{(2\pi)^{n/2} \left|P_{0|0}\right|^{1/2}}$ is a constant, n is the process dimension. Also, the term $p(Z_k \mid X_k, Q_k, R_k)$ can be calculated as follows.

$$\begin{aligned} p(Z_k \mid X_k, Q_k, R_k) &= \prod_{j=1}^{k} p(z_j \mid x_j, R_k) \\ &= \prod_{j=1}^{k} \frac{1}{(2\pi)^{m/2} |R_k|^{1/2}} \exp(-\frac{1}{2} \|z_j - h(x_j)\|^2_{R_k^{-1}}) \\ &= M_2 |R_k|^{-k/2} \exp(-\frac{1}{2} \sum_{j=1}^{k} \|z_j - h(x_j)\|^2_{R_k^{-1}}) \end{aligned} \tag{24}$$

where m represents the measurement dimension, and $M_2 = \dfrac{1}{(2\pi)^{mk/2}}$ is a constant. By considering (23) and (24), the problem of estimation can be reformulated as an optimization problem with the cost function J:

$$\begin{aligned} J &= M_1 M_2 |P_0|^{-1/2} |Q_k|^{-k/2} p(Q_k, R_k) \\ &\exp\left\{ -\frac{1}{2} \left[\|x_0 - \hat{x}_0\|^2_{P_{0|0}^{-1}} + \sum_{j=1}^{k} \|x_j - f(x_{j-1})\|^2_{Q_k^{-1}} + \sum_{i=1}^{k} \|z_j - h(x_j)\|^2_{R_k^{-1}} \right] \right\} \\ &= C |Q_k|^{-k/2} |R_k|^{-k/2} \exp\left\{ -\frac{1}{2} \left[\sum_{i=1}^{k} \|x_j - f(x_{j-1})\|^2_{Q_k^{-1}} + \sum_{j=1}^{k} \|z_j - h(x_j)\|^2_{R_k^{-1}} \right] \right\} \end{aligned} \tag{25}$$

where

$$C = M_1 M_2 \left|P_{0|0}\right|^{-1/2} p(Q_k, R_k) \exp\left\{ -\frac{1}{2} \|x_0 - \hat{x}_0\|^2_{P_{0|0}^{-1}} \right\}$$

As the logarithm operation for both sides of (25) cannot change the extreme points of the cast function J, to find the maximized parameter of coast function J, firstly, take a logarithm from both sides of (25):

$$\begin{aligned} \ln J = &-\frac{k}{2} \ln|Q_k| - \frac{k}{2} \ln|R_k| - \\ &\frac{1}{2} \sum_{i=1}^{k} \|x_i - f(x_{i-1})\|^2_{Q_k^{-1}} - \frac{1}{2} \sum_{i=1}^{k} \|z_i - h(x_i)\|^2_{R_k^{-1}} + \ln C \end{aligned} \tag{26}$$

Using the derivative of J relative to Q_k and R_k, the noise covariance values are calculated as:

$$\left.\frac{\partial \ln J}{\partial Q_k}\right|_{Q_k = \hat{Q}_k} = 0, \quad \left.\frac{\partial \ln J}{\partial R_k}\right|_{R_k = \hat{R}_k} = 0 \tag{27}$$

Consequently, covariance values Q_k and R_k can be calculated as:

$$\hat{Q}_k = \frac{1}{k} \sum_{j=1}^{k} \left\{ \left(\hat{x}_j - f(\hat{x}_{j-1})\right) \left(\hat{x}_j - f(\hat{x}_{j-1})\right)^T \right\} \tag{28}$$

$$\hat{R}_k = \frac{1}{k} \sum_{j=1}^{k} \left\{ \left(z_j - h(x_j)\right) \left(z_j - h(x_j)\right)^T \right\} \tag{29}$$

The terms $f(\hat{x}_{j-1})$ and $h(\hat{x}_j)$ can be calculated from the SRCKF as follows:

$$f_{j-1}(\hat{x}_{j-1}) = \frac{1}{2n}\sum_{i=1}^{2n} f(\chi^i_{j-1|j-1}) \qquad (30)$$

$$h_j(\hat{x}_j) = \frac{1}{2n}\sum_{i=1}^{2n} h(\chi^i_{j|j-1}) \qquad (31)$$

By substituting (30) and (31) into (28)-(29), the $\hat{Q}_k$ and $\hat{R}_k$ is obtained as follows:

$$\hat{R}_k = \frac{1}{k}\sum_{j=1}^{k}\left\{\left(z_j - \frac{1}{2n}\sum_{i=1}^{2n} h(\chi^i_{j|j-1})\right)\left(z_j - \frac{1}{2n}\sum_{i=1}^{2n} h(\chi^i_{j|j-1})\right)^T\right\} \qquad (32)$$

$$\hat{Q}_k = \frac{1}{k}\sum_{j=1}^{k}\left\{\left(\hat{x}_j - \frac{1}{2n}\sum_{i=1}^{2n} f(\chi^i_{j-1|j-1})\right)\left(\hat{x}_j - \frac{1}{2n}\sum_{i=1}^{2n} f(\chi^i_{j-1|j-1})\right)^T\right\} \qquad (33)$$

Results and Discussion

The proposed method is evaluated on the non-maneuvering and maneuvering target tracking. In simulations, the total number of Monte Carlo runs is 100, the initial state of the target is $x_0 = \begin{bmatrix}0 & 1 & 0 & -0.5\end{bmatrix}^T$, and the corresponding covariance is $P_0 = diag(\begin{bmatrix}0.1 & 0.1 & 0.1 & 0.1\end{bmatrix})$ the process noise covariance is $Q_k = diag(\begin{bmatrix}0.001 & 0.001 & 0.001 & 0.001\end{bmatrix})$, and the covariance of measurement noise is $R_k = diag(\begin{bmatrix}1 & 1\end{bmatrix})$.

Non-Maneuvering Target

For a non-maneuvering target, its course and velocity, both of which are assumed to remain constant throughout the observation duration. In Non-maneuvering target, the target moves with velocity (1m/s, -0.5m/s) starting from the (0m, 0m). T sampling period is T=0.26, The proposed method is compared with that of other methods under different conditional.

Scenario 1: Performance with known statistics noise

First, the proposed method is evaluated under effect of a noise with known statistics and the performance of it is compared with EKF and UKF.

Fig. 1 shows results by EKF, UKF and the proposed method, and Fig. 2 shows the tracking performances of the methods on X and Y. Obviously, the tracking performance of the proposed method is better than that of EKF and UKF. Fig. 1 depicts that EKF and UKF lose the target overtime, and the error of estimation increased mainly.

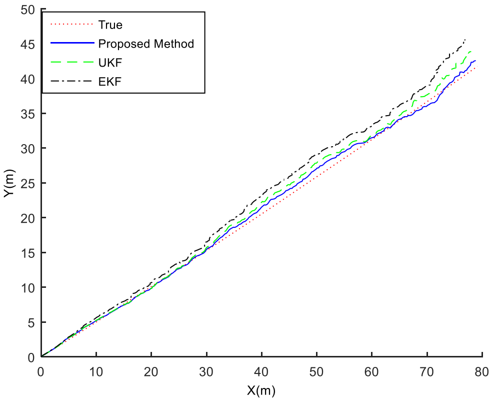


Fig. 1: Results of trajectory.

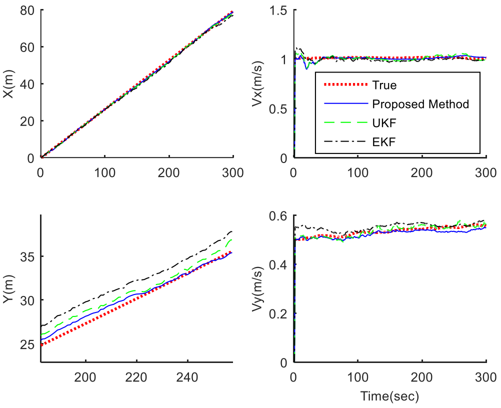


Fig. 2: True states and estimated states values.

Moreover, although EKF and UKF keep the tracking, the error of estimation is large and non-convergent. However, the proposed method follows the target with small estimation error.

Then, in order to more evaluate, the root-mean square error (RMSE) is calculated.

The RMSE of position and velocity for N times simulation is as:

$$RMSE_{pos}(k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((x_k^i - \hat{x}_k^i)^2 + (y_k^i - \hat{y}_k^i)^2)} \qquad (34)$$

where *N* is the total number of Monte Carlo simulation, *k* is the k-th discrete time point of the total simulation time, (x_t, y_t) and $(\hat{x}_t, \hat{y}_t)$ are the true and estimated positions. Similarly, to the RMSE of position, the formulation of RMSE of velocity is as follows:

$$RMSE_{vel}(k) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((\dot{x}_k^i - \hat{\dot{x}}_k^i)^2 + (\dot{y}_k^i - \hat{\dot{y}}_k^i)^2)} \qquad (35)$$

where $(\dot{x}_t, \dot{y}_t)$ and $(\hat{\dot{x}}_t, \hat{\dot{y}}_t)$ are the true and estimated velocities.

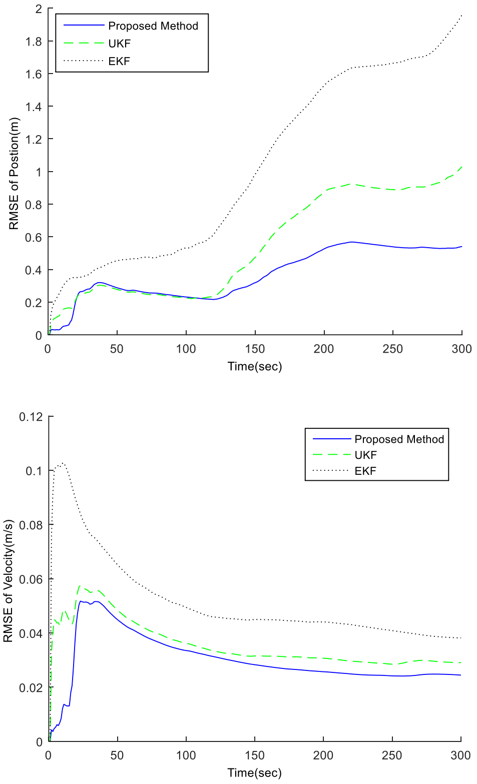


Fig. 3: RMSE over time with known noise statistics.

The RMSE of estimations over time is shown in Fig. 3. Fig. 3 clearly shows that the RMSE of the proposed algorithm are smaller than that of other methods. To further quantify the performance of the methods, the mean and variance of the RMSE algorithms are investigated in Fig. 4. It can be seen that the RMSE in target tracking based on the proposed method is lower than that of EKF and UKF.

Scenario 2: Performance with unknown statistics noise

In this subsection, to illustrate the further benefits of the proposed method, it is assumed that statistics noises are unknown. Assume the initial values of the noise statistics are $Q_k = diag([0.001 \quad 0.001 \quad 0.001 \quad 0.001])$ and $R_k = diag([0.1 \quad 0.1])$, which are different from the true noise covariances. The comparison of methods is shown in Figs. 5-7. The tracking results by EKF, UKF and the proposed method are shown if Fig. 5 and the tracking performances of the methods on X and Y is depicted in Fig. 6. The RMSE of estimation is shown in Fig. 7. It can be observed that the performance of proposed method is almost close to the previous case, while the performance of EKF and UKF is worse than the performance of EKF and UKF in the previous case. The better performance of the proposed method is because that it can estimate the covariance of noises, whereas the other methods depend on the fixed prior knowledge about the process noises.

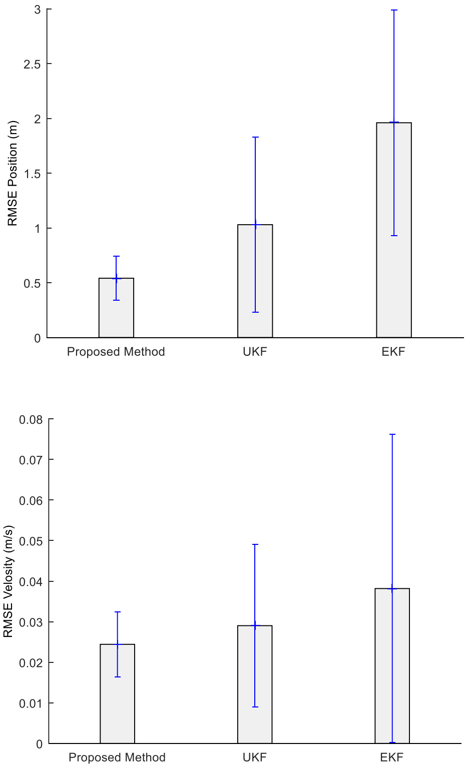


Fig. 4: RMSE of algorithms with known noise statistics.

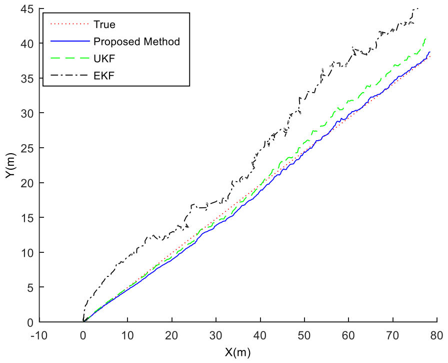


Fig. 5: Results of trajectory.

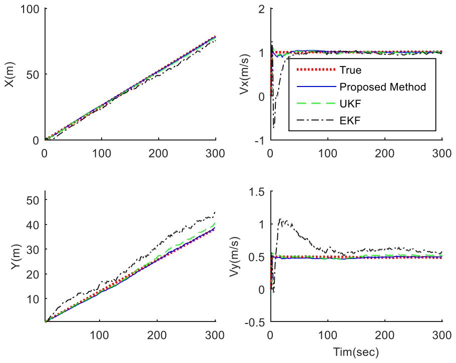


Fig. 6: True states and estimated states values.

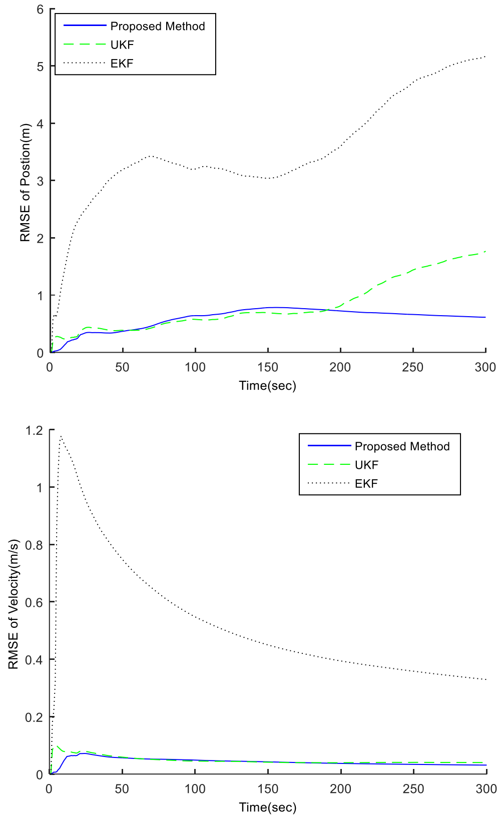


Fig.7: RMSE over time with unknown noise statistics.

Maneuvring Target

The performance of the proposed method is evaluated for tracking of the maneuver target. In maneuvring target, the target moves at an even acceleration and the target motion state will varies, which has to account for the variation of acceleration. In simulations, it is assumed that the target is (0, 0) and the sampling period is T=0.26.

For 100 s, the target starts to make a turn rate of 5°/s. Then it turns for 200s with -8 °/s.

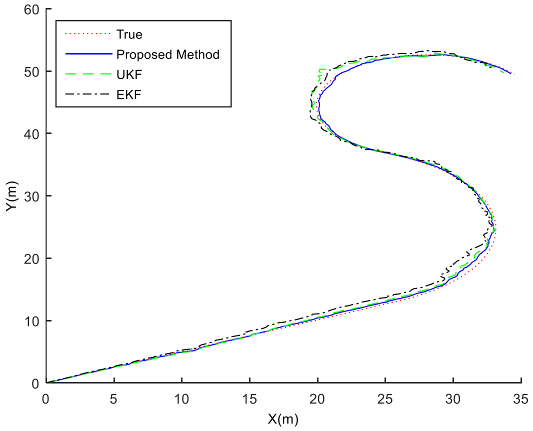


Fig. 8: Results of trajectory.

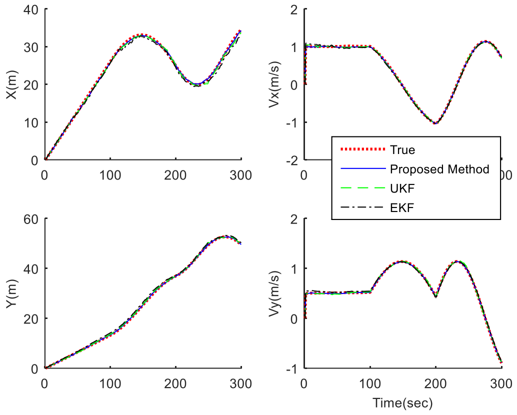


Fig. 9: True states and estimated states values.

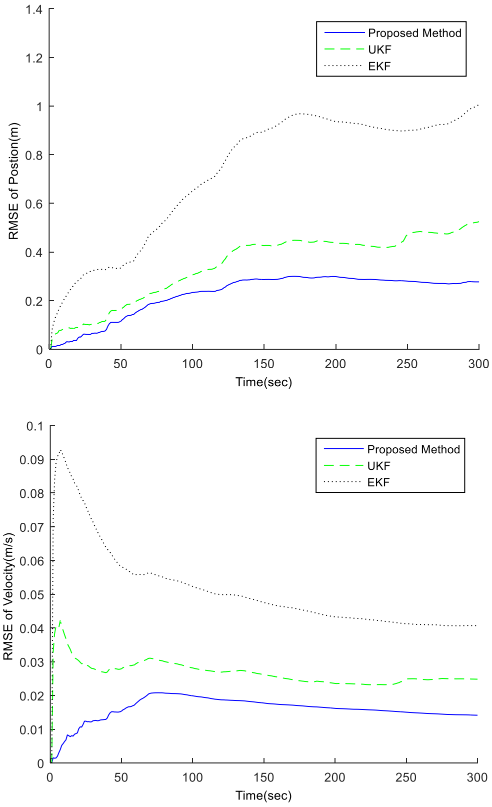


Fig. 10: RMSE of over time with known noise statistics.

Scenario 1: Performance with known statistics noise

In this scenario, first, the performance of the proposed method under the influence of a noise with known statistics is evaluated and compared with performance of UKF and EKF. The estimated trajectories of methods are shown in Fig. 8 and the tracking results by various methods on Y and X are shown in Fig. 9. It observed that tracking accuracy of the proposed method is better than that of EKF and UKF and converges faster. In fact, the estimate value of proposed method is the closest with the true value. The EKF and UKF lose the target and the error of tracking mainly increases.

However, the proposed method traces the target in the whole scenario with small error of estimation. The RMSE of methods are respectively shown in Figs. 10-11. Obviously, the result indicated that the proposed algorithm has better performance in accuracy of estimation compared to EKF and UKF.

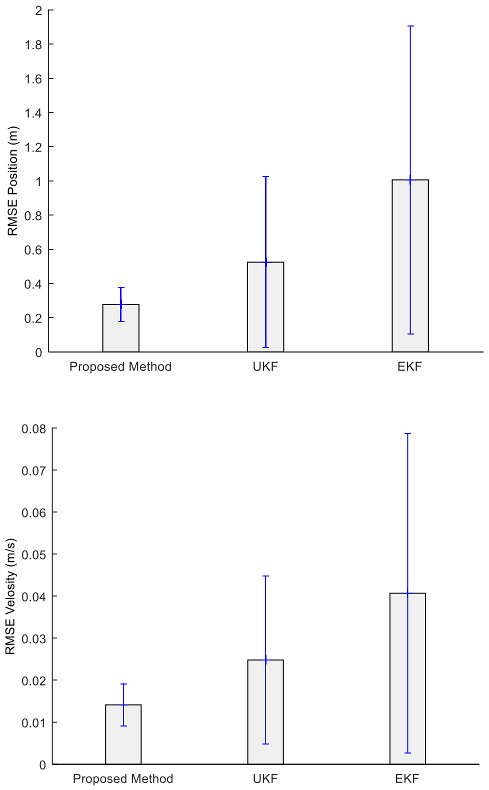


Fig. 11: RMSE of with known noise statistics.

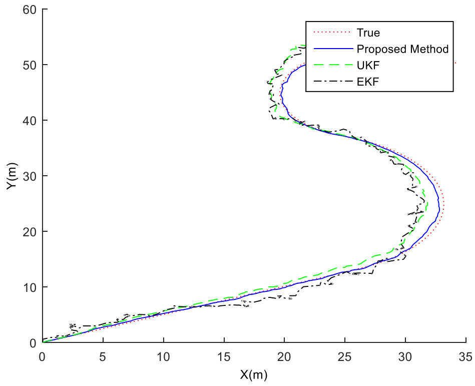


Fig. 12: Results of trajectory.

Scenario 2: Performance with unknown statistics noise

In this sub scenario, the robustness and adaptively of the proposed method is tested when statistics noises are considered wrongly. Figs. 12-14 show the results. Similar to Non-maneuvering case, it observes that the performance of other methods is worse than the performance of them in the previous case, while the performance of proposed algorithm is almost similar to that of the previous case. The RMSE of the position and velocity of the proposed method, UKF and the EKF are shown in Fig. 14.

The proposed method shows the best performance and the range of error is even lower than the UKF and EKF for the RMSE in velocity and position.

The proposed method with noise statistic estimator has very well performance in the complicated conditions, which shows a good adaptability and robustness.

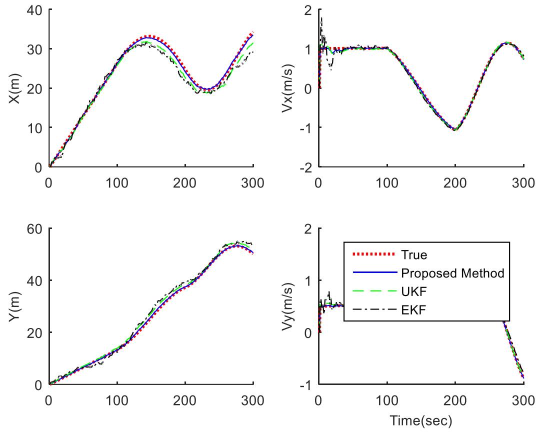


Fig. 13: True states and estimated states values.

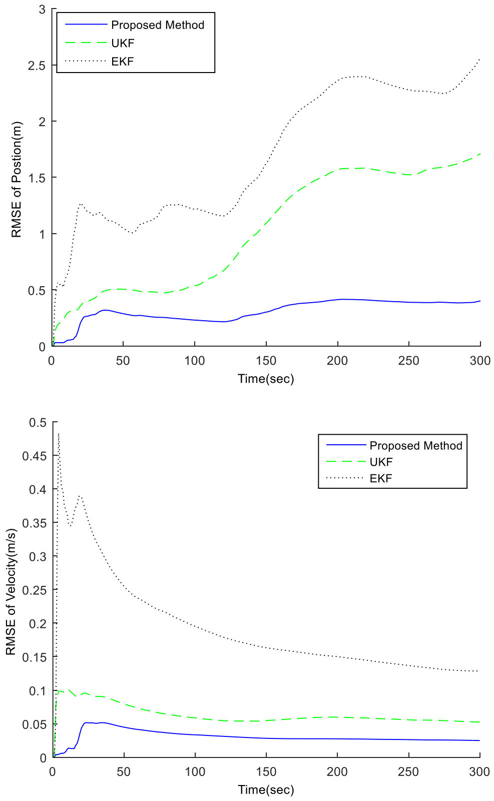


Fig. 14: RMSE of over time with unknown noise statistics.

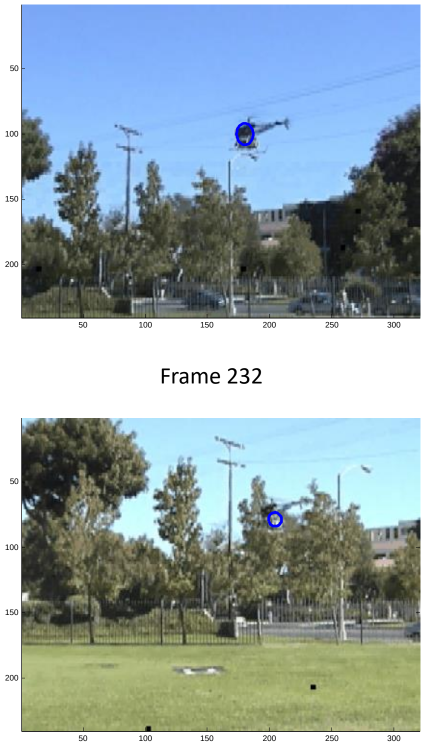
Table 1 shows the RMSE algorithms. From Table 1, It can be seen that the performance of the proposed method is superior to other methods.

Table 1: Performance of algorithms

Statistics noise	Method	RMSE of Position	RMSE of velocity
Unknown	**Proposed Method**	**0.38**	**0.02**
	CKF	1.6	0.05
	UKF	1.71	0.06
	EKF	2.56	0.13
know	**Proposed Method**	**0.27**	**0.015**
	CKF	0.45	0.21
	UKF	0.52	0.025
	EKF	1	0.045

Helicopter Tracking

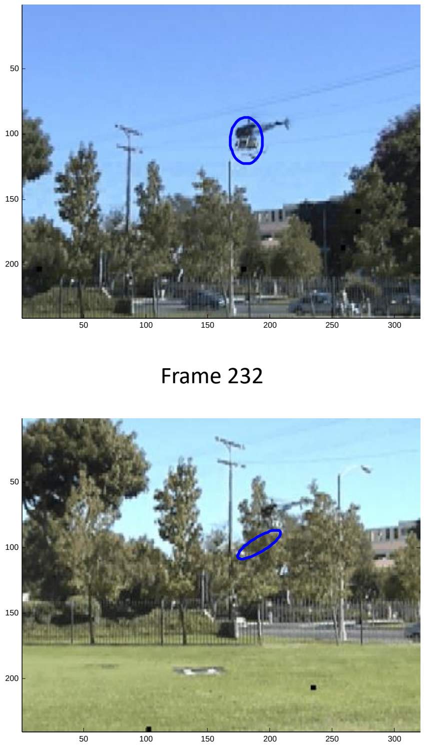
For further investigation, the proposed method in tracking the purpose of video sequences is examined.



Frame 232

Frame 499

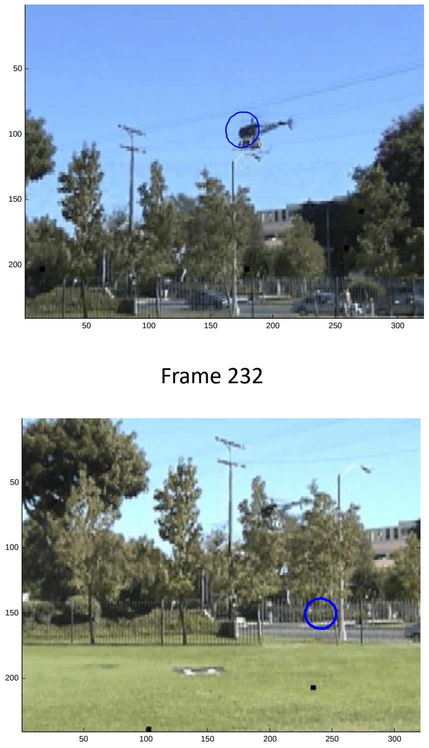
Fig. 15: Tracking by the proposed method.



Frame 232

Frame 499

Fig. 16: Tracking by UKF.



Frame 232

Frame 499

Fig. 17: Tracking by EKF.

The results are shown can be seen in Figs. 15-17. The results show the superior performance of proposed method. From the video in Figs. 15-16, it was observed that EKF and UKF can initially successfully track the target. However, when the helicopter moves in front of the leaves, the trackers miss the target. While in the proposed method, the target is successfully tracked during the video.

Conclusion

In this paper, the target tracking based on adaptive square root cubature Kalman filter is proposed. The proposed method does not need to know the statistics of noises, while other traditional cubature Kalman filters need the noise statistics. Moreover, the proposed method has better numerical characteristics and guaranteed positive semi-definiteness of error covariance matrix.

Instead of decomposing Cholesky at each step, the proposed method updates and propagates square-root of the covariance of error. It has a consistently improved numerical stability. The effectiveness and feasibility of the proposed method is evaluated by the Monte Carlo simulations in different scenarios.

The RMSE of the position and velocity have been evaluated using the UKF, EKF, and proposed method. It has been observed that the RMSE of position and velocity are less in the proposed method compared to the EKF and the UKF.

The results imply the superiority of the proposed method compared to the EKF and the UKF. The proposed method provides better performance in tracking accuracy than other methods.

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 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

ω_k	Process noise
Q_t	Process noise covariance
F^r	Transition matrix corresponding to mode r
Z_k	Measurement at kth instant
v_k	Measurement noise
R_t	Measurement noise covariance
$\sqrt{P_x}$	Square-root factor of P_x
$\chi^{*i}_{k\|k-1}$	Transformed cubature points
$S_{k\|k-1}$	Square root of the covariance matrix
X_k	State vector
S_{xz}	Cross covariance between the states and measurements
J^*	Function of posteriori density
n	Process dimension
$\|A\|$	Determinant of a square matrix A
x_0	Initial state of the target
P_0	Initial covariance
(x_k, y_k)	Position components
$(\dot{x}_k, \dot{y}_k)$	Velocity components
T	Sampling interval
$\Omega_t^{(3)}$	Anticlockwise turn maneuver
$\Omega_t^{(2)}$	Clockwise turn maneuver
$\mathrm{N}(x,P)$	Guassian prior density of x
g	Nonlinear function
n_x	State dimension
ξ_j	Cubature point
e_i^T	The ith column vector of $I_{n_x \times n_x}$
$\mathrm{Tria}(.)$	General triangularization algorithm
$\hat{x}_{k\|k}$	The updated state
N	Total number of Monte Carlo simulation

G	Input matrix
$\hat{x}_{k\|k-1}$	Predicted mean
K_k	Kalman gain
$(\hat{x}_t, \hat{y}_t)$	Estimated positions
$RMSE_{pos}$	RMSE of position
$RMSE_{vel}$	RMSE of velocity

References

[1] M. Baradaran Khalkhali, A. Vahedian, H. Sadoghi Yazdi, "Multi-target state estimation using interactive Kalman filter for multi-vehicle tracking," IEEE Trans. Intell. Transp. Syst., 21(3), 2020.

[2] M. Eltoukhy, M. Omair Ahmad, M.N.S. Swamy, "An adaptive turn rate estimation for tracking a maneuvering target," IEEE Access, 8: 94176-94189, 2020.

[3] Z. Gong, G. Gao, M. Wang, "An adaptive particle filter for target tracking based on double space-resampling," IEEE Access, 9: 91053-91061, 2021.

[4] X.R. Li, V.P. Jilkov, "Survey of maneuvering target tracking. Part II: Motion models of ballistic and space targets," IEEE Trans. Aerosp. Electron. Syst. , 46(1): 96-110, 2010.

[5] S. Koteswara Rao, "Modified gain extended Kalman filter with application to bearings-only passive manoeuvring target tracking," IEE Proc.-Radar Sonar Navig., 152(4): 239-244, August 2005 .

[6] L. Yang, L. Can ,L. Man, H. Xueyao ,W. Yanhua, "Cascaded Kalman filter for target tracking in automotive radar," J. Eng., 2019(19), 2019.

[7] Y. Chen, W. Li , Y. Wang, "Online adaptive Kalman filter for target tracking with unknown noise statistics," IEEE. Sens. Lett., 5(3), 2021.

[8] T. Ma, Q. Zhang, C. Chen, S. Gao, "Tracking of maneuvering star-convex extended target using modified adaptive extended Kalman filter," IEEE Access, 4: 21430-21438, 2016.

[9] K. Doğançay, W.Y. Wang , N.H. Nguyen, "Bias-compensated diffusion pseudolinear Kalman filter algorithm for censored Bearings-only target tracking," IEEE Signal Process. Lett., 26(11): 1703-1707, 2019.

[10] Y. Yang , X. Fan , Z. Zhuo , S. Wang , J. Nan , Y. Xu, "Amended Kalman filter for maneuvering target tracking," Chinese J. Electron., 25(6): 1166-1171, 2016.

[11] J. Li, M. Ye, S. Jiao, W. Meng, X. Xu, "A novel state estimation approach based on adaptive unscented Kalman filter for electric vehicles," IEEE Access, 8: 185629-185637, 2020.

[12] S.K. Raoa, K.R. Rajeswari, K. S.Lingamurty, "Unscented Kalman filter with application to bearings-only target tracking," IETE J. Res., 55(2): 63-67, 2009.

[13] H. Zhang, G. Dai, J. Sun, Y. Zhao, "Unscented Kalman filter and its nonlinear application for tracking a moving target," Optik, 124(20): 4468-4471, 2013.

[14] W. Zhoui, J. Hou, "A new adaptive robust unscented Kalman filter for improving the accuracy of target tracking," IEEE Access, 7: 77476-77489, 2019.

[15] W. Zhou, J. Hou, "A new adaptive high-order unscented Kalman filter for improving the accuracy and robustness of target tracking," IEEE Access, 7: 118484-118497, 2019.

[16] G. Yu. Kulikov, M.V. Kulikova, "Hyperbolic-SVD-Based square-root unscented Kalman filters in continuous-discrete target tracking scenarios," IEEE Trans. Autom. Control, 67(1): 366-373, 2021.

[17] B. Ge, H. Zhang, L. Jiang, Z. Li, M. M.Butt, "Adaptive unscented Kalman filter for target tracking with unknown time-varying noise covariance," sensors, 19(6): 1371, 2019.

[18] C. Liu, P. Shui, G. Wei, S. Li "Modified unscented Kalman filter using modified filter gain and variance scale factor for highly maneuvering target tracking," J. Syst. Eng. Electron., 25(3): 380–385, 2014.

[19] I. Arasaratnam , S. Haykin, "Cubature Kalman filters," IEEE Trans. Automat. Contr., 54(6):1254–1269, 2009.

[20] Q. Chen, C. Yin, J. Zhou, Y. Wang, X. Wang, C. Chen, ''Hybrid consensus-based cubature Kalman filtering for distributed state estimation in sensor networks,'' IEEE Sensors J., 18(11): 4561–4569, 2018.

[21] P.H. Leong, S. Arulampalam, T.A. Lamahewa, T.D. Abhayapala, "A Gaussian-sum based cubature Kalman filter for bearings-only tracking," IEEE Trans. Aerosp. Electron. Syst., 49(2):1161–1176, 2013.

[22] H.W. Zhang, J.W. Xie, J.A. Ge, W.L. Lu, B.Z. Liu, "Strong tracking SCKF based on adaptive CS model for manoeuvring aircraft tracking," IET Radar Sonar Navig., 12: 742–749, 2018.

[23] B. Gao, G. Hu, Y. Zhong, X. Zhu, "Cubature Kalman filter with both adaptability and robustness for tightly-coupled GNSS/INS integration," IEEE Sensors J. , 21(13): 14997-15011, 2021.

[24] M. Yan, F. Fang, Y. Cai, "Maneuvering target tracking based on adaptive cooperative cubature Kalman filter," in Proc. Chinese Automation Congress (CAC), 2019.

[25] A. Roy, D. Mitra, "Multi-target trackers using cubature Kalmanfilter for Doppler radar tracking in clutter", IET Signal Process., 10(8): 888-901, 2016.

[26] B. Ristic, S. Arulampalam, N. Gordon, "Beyond the Kalman filter–particle filters for tracking applications," Norwood, MA: Artech House, 2004.

[27] Y. Bar-Shalom, X.R. Li, T. Kirubarajan, Estimation with applications to tracking and navigation: theory algorithms and software. New York, John Wiley & Sons, 2004.

[28] M. Eltoukhy, M. Omir Ahmad, M.N.S. Swamy, "An adaptive turn rate estimation for tracking a maneuvering target," IEEE Access, 8: 94176 – 94189, 2020.

[29] X.R. Li, V.P. Jilkov, ''Survey of maneuvering target tracking. Part I. Dynamic models,'' IEEE Trans. Aerosp. Electron. Syst., 39(4): 1333–1364, 2003.

[30] A. Zhang, S. Bao, F. Gao, W. Bi, "A novel strong tracking cubature Kalman filter and its application in maneuvering target tracking," Chin. J. Aeronaut., 32(11): 2489–2502, 2019.

Biographies

Ramazan Havangi received his M.S. and Ph.D. degrees from the K.N. Toosi University of Technology, Tehran, Iran, in 2003 and 2012, respectively. He is currently an Associate Professor of control systems with the Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran. His main research interests are inertial navigation, integrated navigation, estimation and filtering, evolutionary filtering, simultaneous localization and mapping, fuzzy, neural network, and soft computing.

- Email: Havangi@Birjand.ac.ir
- ORCID: NA
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: https://cv.birjand.ac.ir/havangi/fa

How to cite this paper:

R. Havangi, "An adaptive cubature kalman filter for target tracking," J. Electr. Comput. Eng. Innovations, 10(2): 425-436, 2022.

DOI: 10.22061/JECEI.2022.8168.486

URL: https://jecei.sru.ac.ir/article_1691.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

A SEPIC-Cuk-CSCCC Based SIMO Converter Design Using PSO-MPPT For Renewable Energy Application

S. Mukherjee [*]

Department of Electrical Engineering, RCC institute of information technology, Kolkata, India.

Article Info

Article History:
Received 06 December 2021
Reviewed 25 January 2022
Revised 27 February 2022
Accepted 13 March 2022

Keywords:
Particle swarm optimization - MPPT
SPV based application
Single input multiple outputs
DC-DC power conversion

[*]Corresponding Author's Email Address:
sarbo.1234@gmail.com

Abstract

Background and Objectives: The increasing requirement of different voltage and power levels in various power electronics applications, especially based on renewable energy, is escalating the growth of the different DC-DC converter topologies. Besides single-input single-output (SISO), multi-input multi-output (MIMO) type topologies become famous. So, in this paper, a Single-Ended Primary Inductance Converter (SEPIC), Cuk and Canonical Switch Cell (CSC) based single-input multi-output (SIMO) boost converter is proposed with a maximum power point tracking (MPPT) controller.
Methods: The Design of the three different DC-DC converter-based SIMO topology has been developed and thereafter the operation of the proposed converter is verified with Solar Photovoltaic (SPV), connected as an input to the converter. To extract maximum power from the SPV and MPPT controller is also developed. Finally, the converter's transfer function is developed using small-signal analysis and the system's stability is analyzed with and without compensation.
Results: A MATLAB simulation has been done to verify the theoretical analysis. Successful extraction of the maximum power from the SPV panel (65W, V_{mpp} 18.2V, I_{mpp} = 3.55A) with Particle Swarm Optimization (PSO) is verified. SEPIC and Cuk-based DC-DC converter can successfully operate in boost mode with a gain of 2.66. A significant reduction in the Cuk converter capacitor voltage ripple is also established.
Conclusion: So, this paper represents an SPV-fed SIMO boost converter based on SEPIC Cuk CSC topology. In addition to that, a PSO-based MPPT controller is also introduced for maximum power extraction. Verification of the theoretical analysis with simulation results is also described.

Introduction

To meet the growing power demand without any environmental issues, renewable energy sources are the most eligible option in the recent era. Continuous decrement of the Solar Photo Voltaic (SPV) cost makes itself more valuable for future research and implementation [1]. Besides that, the SPV system has the advantages of the absence of rotating parts, minimum maintenance, and almost zero environmental loss [2].

The power characteristics of the renewable energy source are nonlinear because of the dependency on solar irradiance, ambient temperature. The typical P-V and V-I characteristics of any solar photo-voltaic cell are shown in Fig. 1. There exists one operating point where they generate maximum power. To take full advantage of available energy resource and achieve maximum utilization efficiency, maximum power point tracking (MPPT) control techniques that extract maximum power from the renewable source is essential.

Doi: 10.22061/JECEI.2022.8379.509

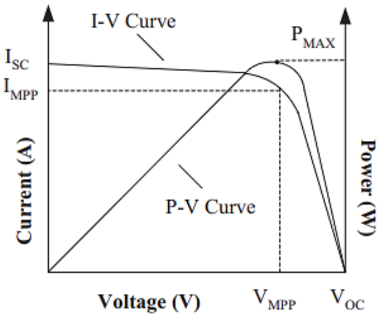


Fig. 1: Typical P-V and V-I characteristics of SPV cell [3].

Several researchers prescribed different MPPT methods for extracting maximum power [3]-[6]. Based on the involvement of several control variables, types of control strategies, nature of the available circuitry, and cost of applications are the main factors to select the MPPT algorithm. The PSO-based MPPT control method is one of the most effective techniques which can accurately track the maximum power point under variable ambient conditions. A swarm intelligence-based algorithm PSO is used to operate by finding the global optimal solution [7], [8]. The primary reason behind the popularity of this algorithm is the presence of very few numbers adjustable parameters.

The evolution of developing DC-DC converter has been constant over the past year. The requirement of different voltage and power levels across the power electronics devices in different applications enhances the growth of the development of the DC-DC converter. So not only the single quadrant converter, multi-level [9]-[11] and multi quadrant converter [12] has become famous day by day for their new control strategy and topologies, which cause the efficiency improvement and the reduction of size. In this paper [9], the advantage and disadvantages of several types of multi-output converter were described. It has been observed that the efficiency of the system is reduced if multiple active switches are used with high-frequency switching. Single input multiple output (SIMO) converter is one of the well-known multi-port converters used for such applications. i.e. Hybrid/ Electric Vehicles (EHV), microelectronics, lighting, telecommunication, channel multiplexing, and digital communication [13], [14], etc. This topology consists of a single inductor, responsible for controlling the current and voltage of different output ports. So, sometimes this topology is named “Single Inductor Multiple Output” (SIMO). Increased efficiency with reduced cost can be achieved in the SIMO converter as less number component is used.

In this paper, a SEPIC Cuk CSC combination converter-based SIMO converter, connected with an SPV, is proposed. A single switch is shared by all of these converters which provides simplification in control. A Particle Swarm Optimization (PSO) based MPPT control is implemented to extract maximum power from the SPV. In the proposed converter SEPIC converter provide positive voltage output. Besides this Cuk and CSC converter produced a negative output voltage. A simulation model is developed in MATLAB/Simulink software and the verification of PSO –MPPT based SIMO converter output is verified. So, implementation of the MPPT, simultaneous regulated bipolar voltage generation and minimum switching loss are the major advantages of the proposed converter.

The rest of the paper is organized as follows: the next section is named as a proposed converter which explains the construction of the proposed topology. The next section is named as an operational principle where the operation of the PSO-based MPPT controller and the different DC-DC converter is explained. In the next section, named as a design consideration, the selection of the inductor and capacitor is discussed. After that, the stability analysis and the effect of the controller in the proposed system are discussed. In the next section details of the simulation results are discussed along with the comparison of the proposed topology with other SIMO converters. Finally, the last section contains the conclusion of the proposed study.

Proposed Topology

In [15], [16], basic topologies of the non-isolated converter are introduced for solar PV application. Besides the conventional topologies, a combination of these converters can sometimes be found advantageous in many applications as given in [17]-[20].

A comparative study of different MPPT techniques for a basic converter with their performance analysis is also described in [4]. PSO-based MPPT algorithm is preferred over conventional techniques like perturb and observe (P and O) and incremental conductance (IC). In this paper, a SEPIC-Cuk-CSC combinational SIMO converter is proposed in Fig. 2, where a PSO-based MPPT technique is applied to extract maximum solar PV power.

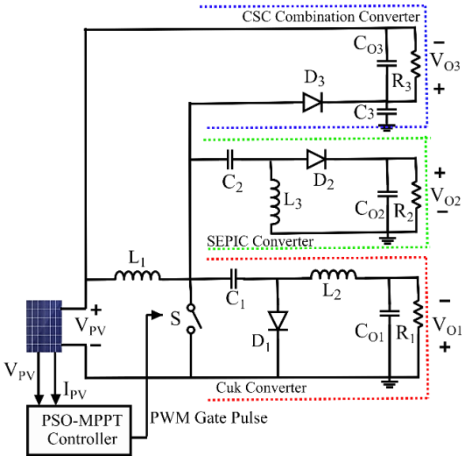


Fig. 2: Proposed SEPIC-Cuk-CSC Converter.

Operating Principal

A. PSO –MPPT Algorithm

Particle Swarm Optimization (PSO) is a population-based intelligence optimization technique, inspired by the foraging behaviour of a flock of birds and fish schooling in search of food. In PSO algorithm individual birds are referred to as individual flying particle that has their fitness value. Each particle's movement, in terms of direction and distance, ncy as calculated by the objective function and the velocity of the individual particle. Exchange of information between the particles happened based on their search process. P_{best} and G_{best} are the best position of the individual particle and the best position of all the particle, comparing all the P_{best} , respectively. All the swarm updates their direction and velocity to move towards the best position. So, convergence can be achieved [21]-[23]. The standard PSO algorithm can be represented by

$$v_i(k+1)=wv_i\left(k\right)+c_1r_1\left(P_{best}-x_i\left(k\right)\right)+c_2r_2\left(G_{best}-x_i\left(k\right)\right) \quad (1)$$

$$P_{best}=x_{ik} \quad (2)$$

$$f\left(x_{ik}\right)>f\left(P_{best,i}\right) \quad (3)$$

$$x_i(k+1)=x_i(k)+v_i(k+1) \quad (4)$$

where, $i=1,2.....N$. v_i and x_i are the velocity and the position of the particle i , the number of iteration denoted by k, w represents the inertia weight. r_1 and r_2 are the uniformly distributed random variable within [0 and 1]. Cognitive and social coefficients are denoted by c_1 and c_2 . P_{best} and G_{best} represent the individual best position of the i^th particle and the swarm best position of all the particle. If (5) is satisfied then the value of the P_{best} can be updated by (6).

$$f\left(x_{ik}\right)>f\left(P_{best,i}\right) \quad (5)$$

$$P_{best}=x_{ik} \quad (6)$$

where, f represents the objective function that should be maximized.

In Fig. 3, PSO-based MPPT topology is described. As given in the flowchart at the beginning particle swarm position and fitness value evaluation function are defined as the duty cycle and the generated output power respectively. A random initialization, within a uniform distribution, is made for the position and the velocity of each particle.

After that the fitness value of the particle is calculated, it is updated compared with the previous value. P_{best} and G_{best} of each particle are also updated against the previous values. Thereafter particle velocities and positions are updated accordingly.

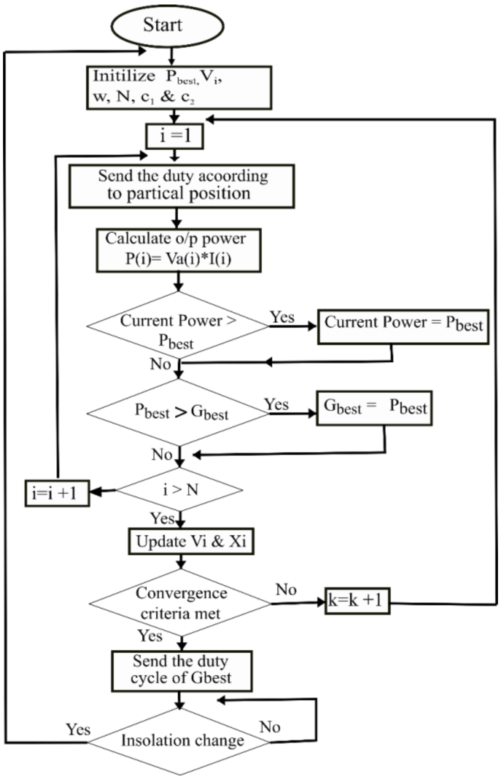


Fig. 3: PSO-based MPPT algorithm flowchart.

With the new values v_i and x_i, the convergence criteria are checked, which are either optimal solution localization or reaching the maximum number of iterations. Depending upon the weather condition and the load value, the fitness function becomes variable. So, the PSO must be reinitialized to search for a new MPP as the output of the PV module changes.

B. Single Input Multiple Output (SIMO Converter)

In this section, an interesting combination of SEPIC Cuk CSC combination converter topology is introduced. The ability to produce both positive and negative voltage simultaneously makes this converter topology suitable for renewable energy-based dc bipolar network applications. As given in Fig. 2 the CSC converter and Cuk converter produce a negative voltage whereas the SEPIC converter produces a positive voltage at the load output terminal.

C. SEPIC Converter

The Single-Ended Primary Inductance Converter or SEPIC converter is a modification of a non-isolated DC-DC converter. Some of the features, which makes this converter suitable for the PV application, are given by [24], [25] non-inverted output, the input inductor provides a low input ripple and noise, multiple inductors can be a couple in the same core, galvanic isolation can be easily obtained by replacing one of the inductors by a high-frequency transformer. The conventional SEPIC converter is shown in Fig. 4 where Vg is termed as an input dc voltage source. A MOSFET can be used as switch S, which is having a duty cycle of D.

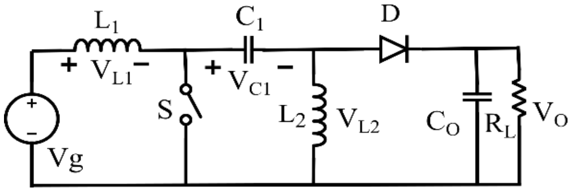
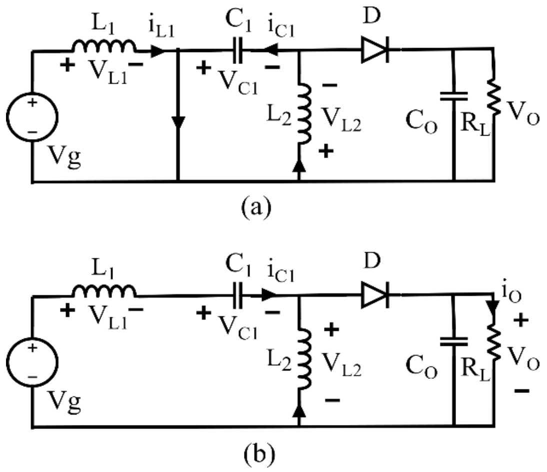


Fig. 4: SEPIC Converter topology.

In continuous conduction mode, the SEPIC converter operates in two different modes shown in Fig. 5a and Fig. 5b. In mode (a) when the switch S is turned on (duration is given by $0 \le t \le DT$, Where T represents the time period of the gate pulse), both the inductor current (I_{L1} and I_{L2}) is increasing because of charging and no energy is transferred to the load as D became reversed biased. In mode (b), when the switch S is turned off (duration given by $DT \le t \le T$), the D becomes forward biased and the energy is transferred to the load as both the inductor (I_{L1} and I_{L2}) are now discharging.



(a)

(b)

Fig. 5: Operation of SEPIC converter. (a) When S is turned on. (b) When S is turned off.

The volt-second balance across the inductor L_1 and L_2 given by

$$V_g DT + \left(V_g - V_{C1} - V_D - V_O\right)(1 - D)T = 0 \tag{7}$$

$$V_{C1} DT + \left(- V_O - V_D\right)(1 - D)T = 0 \tag{8}$$

where V_D represents the voltage drop across the diode.

The output of the SEPIC converter is represented as

$$V_O = \frac{DV_g}{(1 - D)} \tag{9}$$

The value L_1, C_1 and L_2 of the SEPIC converter can be calculated by [26]

$$L_1 = \frac{V_g D}{\Delta I_{L1}.f_s} \tag{10}$$

$$L_2 = \frac{V_g D}{\Delta I_{L2}.f_s} \tag{11}$$

$$C_1 = \frac{V_O D}{R_L \Delta V_O f_s} \tag{12}$$

D. Cuk Converter

Cuk converter is a cascaded combination of the basic boost converter and buck converter with a coupling capacitor as described in [27]. The basic structure of the Cuk converter is given in Fig. 6. Energy is transferred from the input side to the output side through the coupling capacitor.

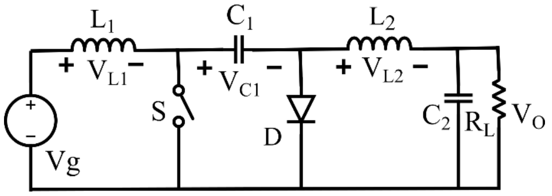
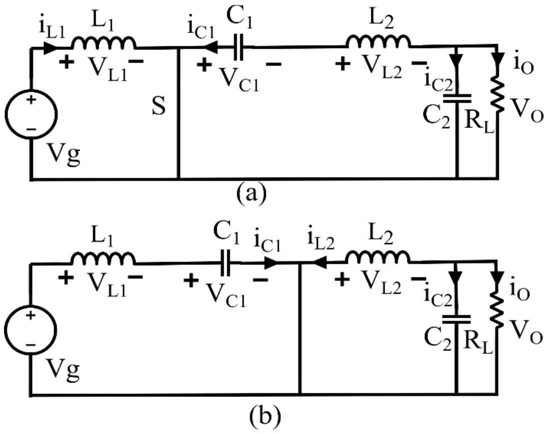


Fig. 6: Basic Cuk Converter Circuit.



(a)

(b)

Fig. 7: Operation of Cuk Converter, (a) when S is turned on. (b) when S is turned off.

The features of the Cuk converter are stated as input and output current is continuous, low switching losses and higher efficiency, have low noise generation and low electromagnetic interference. As it is a combination of buck-boost dc-dc converter, it can able to deliver output voltage both greater and less than the input voltage. The operation of the Cuk converter can be divided into two modes (a) and (b) as shown in Fig. 7a and Fig. 7b respectively.

Mode (a) begins when the switch S is turned on (duration is given by $0 \le t \le DT$). At this mode current through the inductor L_1 increase as it is getting charged by the input voltage. On the other side C_1 is discharging through the output capacitor C_2 and the inductor L_2 by making diode D reverse biased. On the other mode (b) begins when switch S is turned off (duration is given by $DT \le t \le T$). In this mode diode, D became short-circuited which help the capacitor C_1 to get charged by the supplied voltage, and the inductor L_2 transfer the energy to the load by getting discharged. So, the coupling capacitor C_1 is transferring the energy from source to load by charging and discharging. The load voltage became negative as in both the mode the current flowing through the load is opposite in direction.

Applying volt-second balance across the inductor L_1,

$$V_gDT + \left(V_g - V_{C1}\right)(1 - D)T=0 \tag{13}$$

$$V_{C1} = \frac{V_g}{(1-D)} \tag{14}$$

Applying volt-second balance across the inductor L_2

$$\left(V_O + V_{C1}\right)DT + V_O(1 - D)T=0 \tag{15}$$

$$V_O = -\frac{DV_g}{(1 - D)} \tag{16}$$

Equation (16) represents the output of the Cuk converter. Applying the power balance, the value of the current I_{L1} given as

$$I_{L1} = \frac{D^2}{(1 - D)^2}\frac{V_g}{R_L} \tag{17}$$

Voltage ripple across the capacitor C_1 is calculated as

$$\Delta V_{C1} = \frac{D^2V_gT}{R_LC_1(1 - D)} \tag{18}$$

E. Canonical Switch Cell (CSC) Converter

CSC converter is a modification of a buck-boost converter with having fewer no of devices as shown in Fig. 8. The operation of the converter is divided into two different modes (a) and (b). At mode (a), as the switch S is turned on, the input inductor L_1 is getting charged from the source Vg. Simultaneously the capacitor C_1 discharges its stored energy to L_1 through the switch S, as the diode became reversed bias.

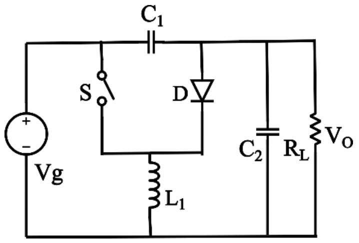


Fig. 8: CSC converter circuit.

Mode (b) begins when the switch S became turned off. Then the diode becomes forward biased and then the input inductor L_1 discharges its energy to the output capacitor C_2. Besides that, the capacitor C_1 is also getting charged by the input voltage through diode D, as shown in Fig. 9a and Fig. 9b.

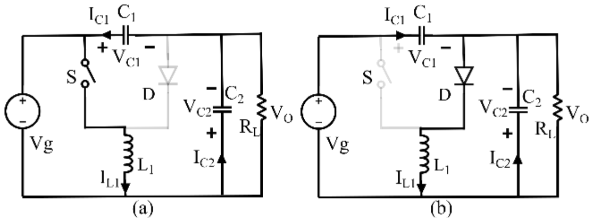


Fig. 9: Operation of CSC converter. (a) When S is turned on. (b) when S is turned off.

The expression of the capacitor C_1 and C_2 are calculated as [28]

$$C_1 = \frac{V_gD}{\Delta V_{C1}R_Lf_s} \tag{19}$$

$$C_2 = \frac{I_O}{2\omega_L\Delta V_O} \tag{20}$$

where R_L represent the equivalent DC load resistance, ω_L represent the angular frequency of the line voltage.

Designing Consideration

A. Design of the inductor (L1)

The inductor (L1) is one of the primary components of this SIMO converter as the amount of energy transfer to the different output terminals is controlled by the energy stored in the inductor during the switch turned on time. The value of the inductor can be calculated from (10). It has been observed that the value of the inductor is depending on the duty cycle (D), the magnitude of the ripple current, and the input voltage. Considering the input voltage as a constant value, the variation of the inductance for the variation of the ripple current value (15% to 20%) and duty cycle (25% to 75%) is shown in Fig. 10.

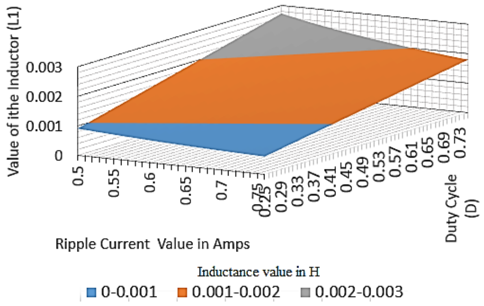


Fig. 10: Variation of the inductance (L1) depending on the duty cycle and the ripple current values (in amps).

A variation of inductance value from 0.5mH to 2.7mH can be observed in Fig. 9. Besides that, as the panel output voltage is also changed depending on the ambient condition, the variation of the inductance value, with an assumption of constant ripple current and the variable duty cycle is plotted in Fig. 11.

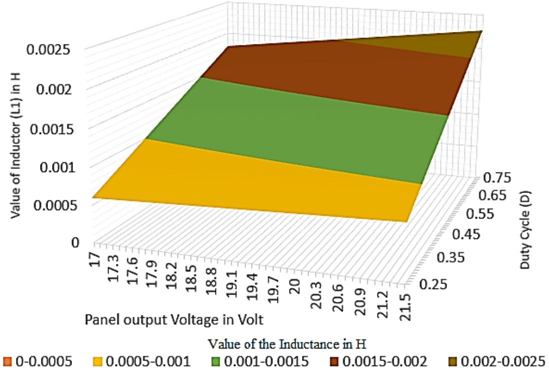


Fig. 11: Variation of the inductance (L_1) depending on the duty cycle and input voltage (in volt).

From Fig. 9 and Fig. 10, a value of 2.1mH is chosen as the value of the inductor L_1 after assuming the value of the ripple current is around 20%.

B. Design of the capacitor (C_1)

Similarly, the value of the capacitor (C_1) can also be calculated from (18).

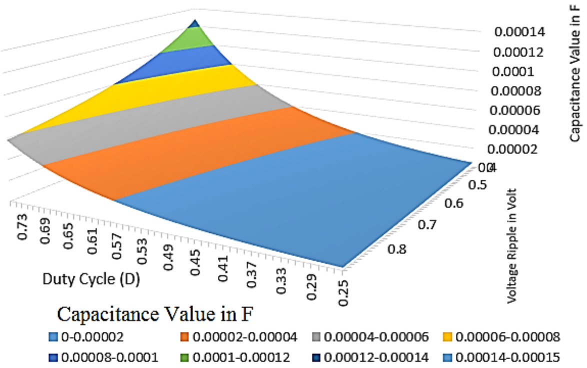


Fig. 12: Variation of the capacitance (C_1) depending on the duty cycle and ripple voltage values (in volt).

Initially considering the value of the load resistance 80Ω, the variation in the capacitance value depending on duty cycle (from 25% to 75%) and voltage ripple (.4V to .9V) is shown in Fig. 12.

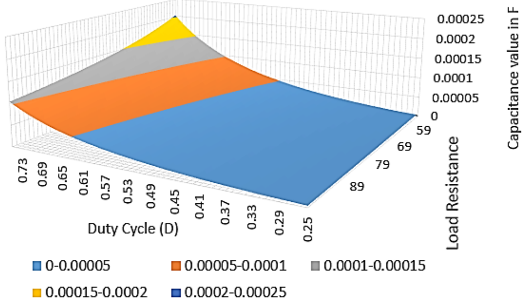


Fig. 13: Variation of the capacitance (C_1) depending on the duty cycle and Load resistance values (in ohm).

Besides, considering a constant ripple voltage of .4V, the variation of the capacitance value depending on the duty cycle (25% to 75%) and the load resistance (50 to 100) ohm is shown in Fig 13. So, after analyzing both Fig. 12 and Fig. 13, a capacitance value of 220μF is chosen as capacitance.

Stability Analysis and effect of Controller

As the SIMO converter is fed from the SPV supply, the primary purpose of the controller is to maintain a maximum power extraction from the SPV throughout the operation. Fig. 14 shows the block diagram of the proposed converter control system.

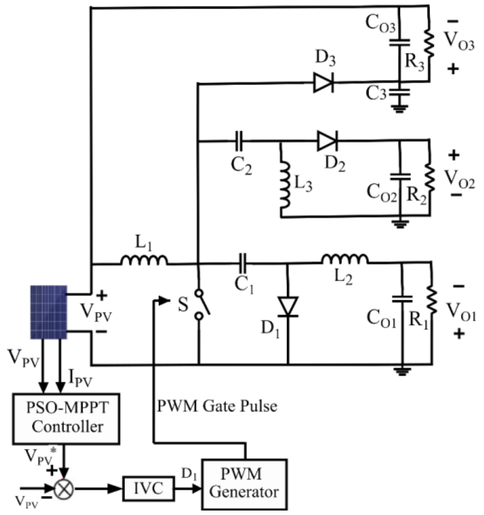


Fig. 14: Overall block diagram of total system.

A feed-forward controller consisting of an MPPT controller and an Input Voltage Controller (IVC) has been considered a total control system. After taking inputs (V_{PV} and I_{PV}) from the SPV panel, the MPPT controller develop a reference voltage V_{PV}^* with the help of the PSO algorithm. Then an error signal is generated after comparing the reference signal with the SPV output voltage. Thereafter this error signal is given as an input to the IVC and The signal D_1 is developed. An equivalent 10kHz PWM signal is generated by the PWM generator by taking the D_1 as an input.

The transfer function of the proposed converter is calculated with the help of small-signal modelling. After replacing the different component value the transfer function of the proposed converter is given by,

$$\mathrm{TF(s)} = \frac{-6.845s^3 - 2.252e^{04}s^2 - 4.33e^{05}s - 1.07e^{06}}{s^3 + 281.6s^2 + 4361s + 1.277e^{04}} \qquad (21)$$

The position of the poles and zeros, as shown in Fig. 15, depicts that the stability of the system is marginal. The stability is further enhanced by shifting the position of the pole away from the imaginary axes by inserting a PID controller.

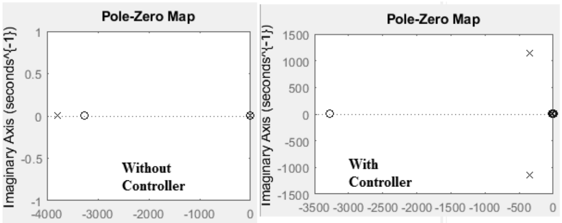


Fig. 15: Pole Zero Plot of the system with and without controller.

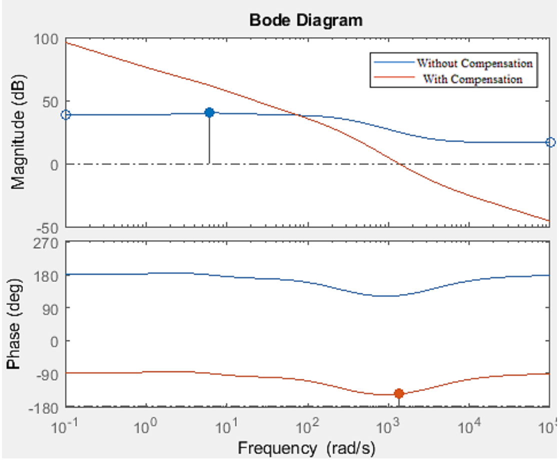


Fig. 16: Bode plot of the system with and without controller.

Similarly, improvement of the gain margin and the reduction in the gain cross-over frequency (GCF) is also achieved as shown in Fig. 16. Reduction of GCF defines a significant reduction of the system noise.

Further, the improvement of the stability of the compensated system can be observed from the nyquist plot as shown in Fig. 17. The reduction in the encirclement of the point (-1+j0) is observed after insertion of the controller.

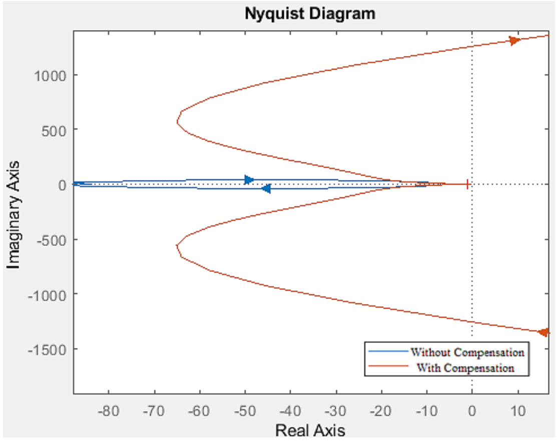


Fig. 17: Nyquist plot of the system with and without controller.

Simulation Results

To verify the PSO-MPPT based SEPIC-Cuk-CSC combinational converter characteristics, a simulation is performed in MATLAB as shown in Fig 13 the details of the SPV panel and the SIMO converter component specification, which is used in this simulation, is given in the Table 1.

Fig.18a shows the PWM gate pulse of 10 kHz developed by the PSO-MPPT. The SPV panel output voltage and the SPV panel extracted power are shown in Fig. 18b and Fig. 18c respectively. A swing of SPV voltage around the 18.2V (V_{mpp}) can be observed, which signifies a satisfactory execution of PSO-based MPPT.

Table 1: Component specification

	Name	Rating
Solar PV Panel	Input Panel Power	65W
	Open circuit Voltage (V_{oc})	22V
	The voltage at MPP (V_{mpp})	18.2V
	Short Circuit Current (I_{sc})	5.5A
	Current at MPP (I_{mpp})	3.55A
Reactance	Inductor L_1	2.1mH
	Inductor L_2 and L_3	1.35mH
	Capacitor C_1	220µF
	Capacitor C_2	470µF
	Resistance R_1-R_3	80Ω
	Switching Frequency	10kHz

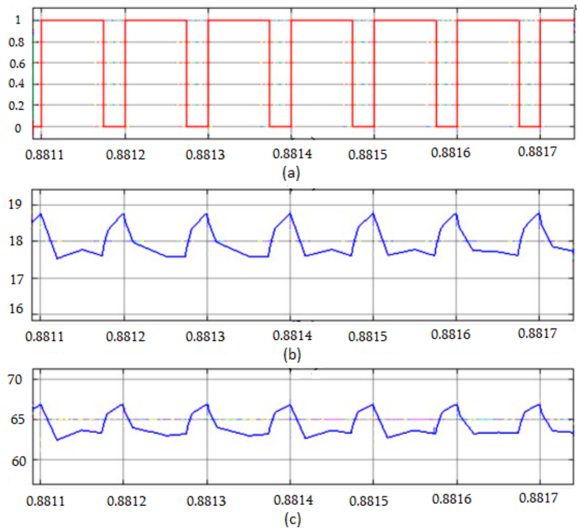


Fig. 18: (a) PWM output of the PSO-MPPT controller. (b) SPV panel output voltage (c) Solar PV panel output Power.

In Fig. 19b and Fig. 19d, the charging current of the inductor L_1 and the discharging current of the inductor L_2 are observed during the switch turn-on time. Besides that, the charging and the discharging of the capacitor C_1 are also shown in Fig. 19c.

Similarly, the charging and discharging of the capacitor C_2 along with the inductor L_3 of the SEPIC converter is shown in Fig. 20. It has been observed, in Fig. 20d, that a ripple of 0.5A is present in the inductor current. Different characteristics of the CSC converter are shown in Fig. 21. Where Fig. 21a and Fig. 21b represented the PWM gate pulse and the current characteristics of the inductor L_1.

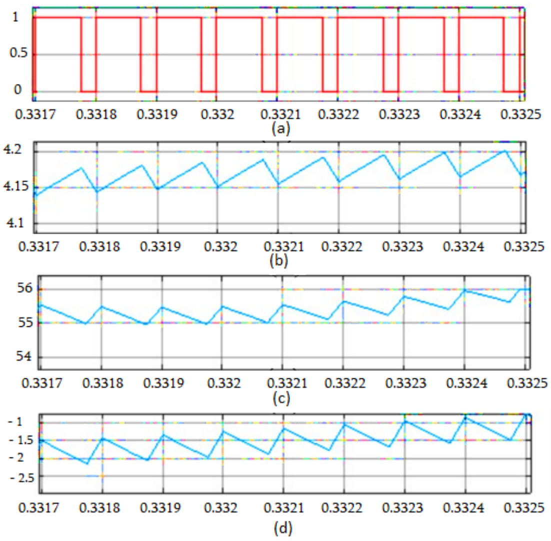


Fig. 19: Cuk converter output (a) PWM Gate pulse of switch S. (b) L1 inductor current. (c)The voltage across capacitor C1. (d) L2 inductor current.

Charging and discharging of the capacitor C_3 is verified in Fig. 21d. The voltage across the capacitor C_{03} almost remains constant as it is shown in Fig. 21c. Continuous Mode of Conduction (CCM) operation is observed in the all of the converter.

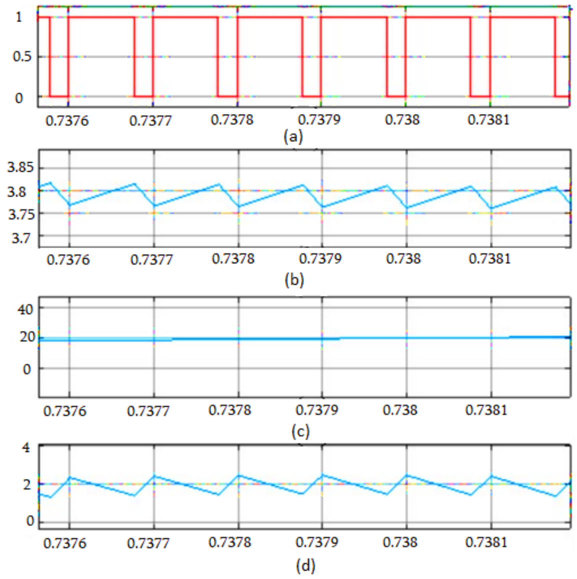


Fig. 20: SEPIC converter output (a) PWM Gate pulse of switch S. (b) L1 inductor current. (c) The voltage across capacitor C2. (d) L3 inductor current.

Three different voltage output with proper polarity is shown in Fig. 22. Where SEPIC and Cuk converter produces almost 48V and -48V with a gain of 2.6. Besides, the CSC converter is developing a voltage of around -18V.

A comparative analysis of the propose SIMO converter with some other SIMO converter is shown in Table 2.

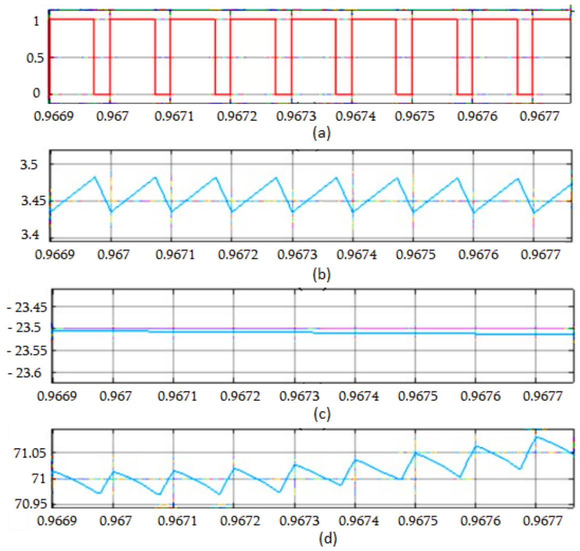


Fig. 21: CSC converter output. (a) PWM Gate pulse of switch S. (b) L1 inductor current. (c)The voltage across capacitor C_{O3}. (d) The voltage across capacitor C_3.

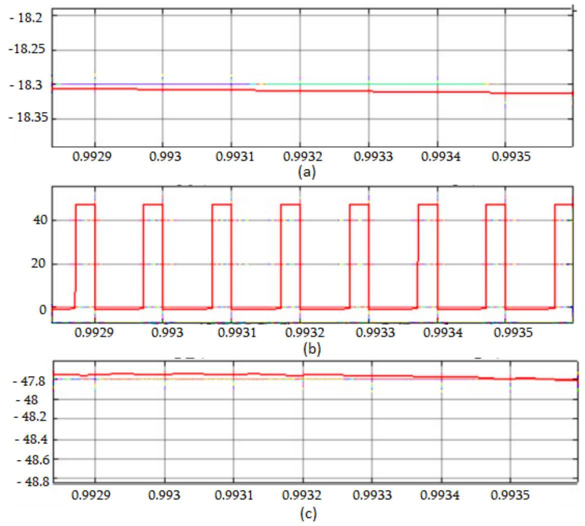


Fig. 22: Multiple output voltage of SIMO converter. (a) CSC converter output voltage. (b) SEPIC converter output voltage (c) Cuk Converter output voltage.

In [29], the no of active and passive component used is much higher than other topologies. Besides that, the topology, proposed in [14], experienced a high switching loss as hard switching technique is used. MPPT control implementation is also not proposed in this topology. Though MPPT control was implemented in [30], but only two unipolar DC voltage is generated from this converter. So, comparing the proposed topology with the other proposed converter, the advantages can be summarized as (a) Implementation of MPPT, (b) Simultaneous generation of both positive and negative regulated voltage. (c) Minimum switching loss.

Table 2: Comparison with other SIMO topologies

Converter reference	[30]	[29]	[14]	Proposed Topology
Use of MPPT	Yes	No	No	Yes
No of Output	2	2	3	3
Output voltage polarity	Only positive	Only positive	Only Positive	Both Positive and Negative
Total No of Component	7	21	9	13
Inductor	1	1	1	3
Capacitor	2	8	3	6
Diode	2	7	0	3
Active Switch	2	2	5	1
Switching Loss	Medium	Medium	High	Low

Conclusion

This paper presents a design of the SEPIC-Cuk-CSC combination converter used for renewable energy applications. A PSO-based MPPT method is applied to extract maximum power. According to the simulation results, it is observed that the PSO method is successfully able to track the MPP in all the conditions. Reduction of design cost and loss is achieved by reducing the component requirement for developing multiple output voltage levels.

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

SISO	Single Input Single Output
SIMO	Single Input Multiple Output
MIMO	Multiple input multiple Output
PSO	Particle Swarm Optimization.
MPPT	Maximum Power Point Tracking
P_{best}	Best position of the individual Particle
G_{best}	Best position among all the particles
v_i	The velocity of the particle
x_i	Position of the particle

References

[1] R.J. Wai, W.H. Wang, C.Y. Lin, "High-performance stand-alone photovoltaic generation system," IEEE Trans. Ind. Electron., 55(1): 240–250, 2008.

[2] M.M.A. Salama, M.Z. Shams El-Dein, M. Kazerani, "Optimal photovoltaic array recon fi guration to reduce partial shading losses," IEEE Trans. Sustain. Energy, 4(1): 145–153, 2013.

[3] A. Gupta, Y.K. Chauhan, R.K. Pachauri, "A comparative investigation of maximum power point tracking methods for solar PV system," Sol. Energy, 136: 236–253, 2016.

[4] B. Subudhi, R. Pradhan, "A comparative study on maximum power point tracking techniques for photovoltaic power systems," IEEE Trans. Sustain. Energy, 4(1): 89–98, 2013.

[5] M.A. Husain, Z.A. Khan, A. Tariq, "A novel solar PV MPPT scheme utilizing the difference between panel and atmospheric temperature," Renew. Energy Focus, 19–20(00): 11–22, 2017.

[6] A.K. Podder, N.K. Roy, H.R. Pota, "MPPT methods for solar PV systems: A critical review based on tracking nature," IET Renew. Power Gener., 13(10): 1615–1632, 2019.

[7] V. Phimmasone, Y. Kondo, T. Kamejima, M. Miyatake, "Verification of efficacy of the improved PSO-based MPPT controlling multiple photovoltaic arrays," in Proc. Int. Conf. Power Electron. Drive Syst.: 1015–1019, 201.

[8] K.K. Jha and M.N. Anwar, "Solar photovoltaic based brushless DC motor driven water pumping system using PSO-MPPT algorithm," in Proc. 2019 54th Int. Univ. Power Eng. Conf. (UPEC 2019): 1–6, 2019.

[9] E. Babaei, O. Abbasi, S. Sakhavati, "An overview of different topologies of multi-port dc/dc converters for dc renewable energy source applications," in Proc. 2016 13th Int. Conf. Electr. Eng. Comput. Telecommun. Inf. Technol. (ECTI-CON 2016), 2: 0–5, 2016.

[10] O. Ray, A.P. Josyula, S. Mishra, A. Joshi, "Integrated dual-output converter," IEEE Trans. Ind. Electron., vol. 62, no. 1, pp. 371–382, 2015, doi: 10.1109/TIE.2014.2327599.

[11] R. Lin, C.R. Pan, K.H. Liu "Family of single-inductor multi-output DC-DC converters," in Proc. 2009 International Conference on Power Electronics and Drive Systems (PEDS): 1216–1221, 2009.

[12] L. Cheng, F. Wang, "A new transformerless four quadrant DC-DC converter with wide conversion ratio," in Proc. 2021 IEEE Int. Conf. Power Electron. Comput. Appl. (ICPECA 2021): 486–491, 2021.

[13] H.F. Chinchero, J.M. Alonso, "Review on DC-DC SIMO converters with parallel configuration for LED lighting control," in Proc. 2020 IEEE Andescon, Andescon 2020, Quito, Ecuador, 2020.

[14] E. Agustianno, F.D. Murdianto, H.E.H. Suharyanto, "Design and simulation of single phase controlled full wave rectifier and simo buck converter for multi output power supply," in Proc. 2nd Int. Conf. Technol. Policy Electr. Power Energy, ICT-PEP 2020, 3: 312–316, 2020.

[15] F.D. Murdianto, M.Z. Efendi, R.E. Setiawan, A.S.L. Hermawan, "Comparison method of MPSO, FPA, and GWO algorithm in MPPT SEPIC converter under dynamic partial shading condition," in Proc. ICAMIMIA 2017 Int. Conf. Adv. Mechatronics, Intell. Manuf. Ind. Autom: 315–320, 2017.

[16] B.K. Panigrahi, P.R. Thakura, "Implementation of Cuk converter with MPPT," in Proc. 3rd IEEE Int. Conf. Adv. Electr. Electron. Information, Commun. Bio-Informatics, AEEICB 2017: 105–110, 2017.

[17] J. Chen, D. Maksimović, R. Erickson, "Buck-boost PWM converters having two independently controlled switches," in Proc. PESC Rec. - IEEE Annu. Power Electron. Spec. Conf., 2: 736–741, 2001.

[18] B. Lin, S. Member, F. Hsieh, "Soft-switching zeta – flyback converter with a buck – boost type of active clamp," IEEE Trans. Ind. Electron., 54(5): 2813–2822, 2007.

[19] K. Suresh et al., "A multifunctional non-isolated dual input-dual output converter for electric vehicle applications," IEEE Access, 9: 64445–64460, 2021.

[20] H. Aboreada, A.V.J.S. Praneeth, N. Vamanan, V. Sood, S.S. Williamson, "Design and control of non-isolated, multi-input dc/dc converter for effective energy management," IEEE Int.

Symp. Ind. Electron., 2019-June: 810–815, 2019.

[21] H. Renaudineau et al., “A PSO-based global MPPT technique for distributed PV power generation,” IEEE Trans. Ind. Electron., 62(2): 1047–1058, 2015.

[22] K. Ishaque, Z. Salam, M. Amjad, S. Mekhilef, “An improved particle swarm optimization (PSO)-based MPPT for PV with reduced steady-state oscillation,” IEEE Trans. Power Electron., 27(8): 3627–3638, 2012.

[23] M.A. Abdullah, T. Al-Hadhrami, C.W. Tan, A.H. Yatim, “Towards green energy for smart cities: Particle swarm optimization based MPPT approach,” IEEE Access, 6: 58427–58438, 2018.

[24] E. Durán, J. Galán, M. Sidrach-de-Cardona, J.M. Andújar, “A new application of the buck-boost-derived converters to obtain the I-V curve of photovoltaic modules,” in Proc. 2007 IEEE Power Electronics Specialists Conference: 413–417, 2007.

[25] J.J. Chacon, R.A. Ortiz, J.E. Archila, M.A. Mantilla, M.A. Botero, J.F. Petit, “Prototype for the characterization of photovoltaic panels based on a SEPIC converter,” in Proc. 2017 IEEE Workshop on Power Electronics and Power Quality Applications (PEPQA): 1–6, 2017.

[26] M.O. Ali, A.H. Ahmad, “Design, modelling and simulation of controlled sepic dc-dc converter-based genetic algorithm,” Int. J. Power Electron. Drive Syst., 11(4): 2116–2125, 2020.

[27] S. Chakrabarti, A.K. Panja, A. Mukherjee, A.K. Bar, Intelligent Electrical Systems: A Step towards Smarter Earth. CRC Press, 2021.

[28] V. Bist, B. Singh, “A PFC-Based BLDC motor drive using a canonical switching cell converter,” IEEE Trans. Ind. Informatics, 10(2): 1207–1215, 2014.

[29] M.Y. Hassani, M. Maalandish, S.H. Hosseini, “A new single-input multioutput interleaved high step-up DC-DC converter for sustainable energy applications,” IEEE Trans. Power Electron., 36(2): 1544–1552, 2021.

[30] V.S.K. Prasadarao, S. Peddapati, S.V.K. Naresh, “A single phase seven-level MLI with reduced number of switches employing a PV Fed SIMO DC-DC converter,” in Proc. 2020 IEEE Students’ Conf. Eng. Syst. SCES 2020, 2020.

Biographies

Sarbojit Mukherjee received his M.Tech degree from Calcutta University, Kolkata in 2011 and B.Tech degree in 2009, from the West Bengal University of Technology, Kolkata. Currently, he is working as an Assistant Professor in the Department of Electrical Engineering RCC Institute of Information Technology, Kolkata, India. He is currently pursuing her PhD at the Calcutta University, Kolkata, India. His research interest includes Power Electronics DC-DC Converter, Soft-Switched Converter and Electrical Drives.

- Email: sarbo.1234@gmail.com
- ORCID: 0000-0002-8893-3010
- Web of Science Researcher ID: ABA-7292-2022
- Scopus Author ID: 57219295564
- Homepage: NA

How to cite this paper:

S. Mukherjee, “A SEPIC-Cuk-CSCCC based SIMO converter design using PSO-MPPT for renewable energy application,” J. Electr. Comput. Eng. Innovations, 10(2): 437-446, 2022.

DOI: 10.22061/JECEI.2022.8379.509

URL: https://jecei.sru.ac.ir/article_1692.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Application of Harris Hawks Optimization Algorithm and APSO-CLUSTERING in Predicting the Stock Market

I. Behravan, S.M. Razavi*

Department of Electrical Engineering, University of Birjand, Birjand, Iran.

Article Info

Article History:
Received 01 December 2021
Reviewed 18 January 2022
Revised 10 February 2022
Accepted 11 April 2022

Keywords:
Tehran Stock market
Harris Hawks Optimization (HHO)
Support Vector Regression (SVR)
APSO-Clustering
Metaheuristics

*Corresponding Author's Email Address:
smrazavi13550@gmail.com

Abstract

Background and Objectives: Stock markets have a key role in the economic situation of the countries. Thus one of the major methods of flourishing the economy can be getting people to invest their money in the stock market. For this purpose, reducing the risk of investment can persuade people to trust the market and invest. Hence, Productive tools for predicting the future of the stock market have an undeniable effect on investors and traders' profit.
Methods: In this research, a two-stage method has been introduced to predict the next week's index value of the market, and the Tehran Stock Exchange Market has been selected as a case study. In the first stage of the proposed method, a novel clustering method has been used to divide the data points of the training dataset into different groups and in the second phase for each cluster's data, a hybrid regression method (HHO-SVR) has been trained to detect the patterns hidden in each group. For unknown samples, after determining their cluster, the corresponding trained regression model estimates the target value. In the hybrid regression method, HHO is hired to select the best feature subset and also to tune the parameters of SVR.
Results: The experimental results show the high accuracy of the proposed method in predicting the market index value of the next week. Also, the comparisons made with other metaheuristics indicate the superiority of HHO over other metaheuristics in solving such a hard and complex optimization problem. Using the historical information of the last 20 days, our method has achieved 99% accuracy in predicting the market index of the next 7 days while PSO, MVO, GSA, IPO, linear regression and fine-tuned SVR has achieved 67%, 98%, 38%, 4%, 5.6% and 98 % accuracy respectively.
Conclusion: in this research we have tried to forecast the market index of the next m (from 1 to 7) days using the historical data of the past n (from 10 to 100) days. The experiments showed that increasing the number of days (n), used to create the dataset, will not necessarily improve the performance of the method.

Introduction

Stock markets have key roles in the economic situation of the countries. In order to keep or increase capital, traders are buying and selling different companies' shares. Also, the companies can increase their fund by selling their shares in the stock market. So, investing in the stock market can result in growing the economic situation and, as a result of that, expanding the industry

Doi: 10.22061/JECEI.2022.8659.541

section. Accordingly, persuading people to invest their money in the stock market is the first step to flourish the economy. On the other hand, it is necessary to reduce the risk of investment. Since the stock market can be affected by several internal and external factors, promising investment relies heavily on dependable prediction methods.

Thus, forecasting the stock markets has become one of the hot topics among traders and researchers in the last years. Numerous researches have been conducted in recent years based on the theory of the repetitive nature of stock market behavior. In these researches, several mathematical methods have been developed to forecast the future price of a stock or to forecast the future of a stock market index. Due to the nonlinearity, complexity, and noisy time-series data of the stock market, mathematical approaches are not reliable for prediction. but on the other hand, Machine learning methods, such as Support Vector Machine (SVM) and Artificial Neural Networks (ANN), have proven themselves as powerful and reliable forecasting methods. Metaheuristics are strong optimization algorithms that have gained much attention in the last years. These algorithms have been used frequently to solve complex optimization problems such as [1]-[6].

Many real-world problems in machine learning and artificial intelligence are hard to be tackled using conventional mathematical approaches such as conjugate gradient, sequential quadratic programming, fast steepest, and quasi-Newton methods [7], [8] due to the continuous, discrete, constrained, or unconstrained nature.

Thus these methods are not efficient in solving many large-scale real-world multimodal, non-continuous, and non-differentiable problems. Accordingly, metaheuristic algorithms have been invented and utilized to tackle these kinds of problems.

These algorithms have become very popular among researchers because of their simplicity, effectiveness, and ease of implementation process.

In this research, a two-stage prediction method is introduced to forecast the future of the Tehran Stock Exchange market index value.

In the first stage, a novel automatic clustering algorithm, called APSO-Clustering, is utilized to extract different clusters of the data points and in the second stage a hybrid regression method, combination of Harris Hawks Optimization algorithm (HHO) and Support Vector Regression (SVR), is used for each cluster's data points to detect the hidden patterns of them. Numerous experiments have been conducted to evaluate the efficiency of the method in predicting the stock market index value in the next m (1 to 7) days using the historical information of the past n (10 to 100) days and technical indicators.

The results of the experiments showed the effectiveness and high potential of the proposed method in forecasting the future. The main contributions of our work are:

1. Using a novel automatic clustering algorithm to divide the dataset in to smaller clusters of data to improve the performance of the model.
2. Using Harris Hawks optimization algorithm to tune the parameters of SVR and feature selection simultaneously.
3. Investigating the impact of historical information on the performance of the model.
4. Predicting the market index value for the next seven days.

There are two important issues should be mentioned as the research limitations. First, the dollar exchange rate has impact on the market but it is neglected in this research due to the lack of a structured dataset containing the daily price of dollar.

Second, some important economical indexes, such as inflation rate, have intense effect on the market index, but unfortunately, there is no structured and useful dataset containing these kinds of information for each working day. The paper is organized in the following manner:

- **Background**: a complete review on related works
- **Data preparation**: the procedure of preparing training data.
- **Proposed method:** a complete explanation about the prediction method.
- **Experimental results.**
- **Conclusion**.

Background

In 2016 Usmani et al. studied the performance of different machine learning techniques on predicting the market performance of the Karachi Stock Exchange (KSE) [9]. They have trained the classifiers on a dataset containing different attributes such as gold and oil rates, political news, historical data of the market, etc. for a binary classification problem (positive or negative market index).

Their results indicate that machine learning techniques have a great capability in predicting the stock market. On the other hand, investigating the effect of different factors on the stock market, they proved that Petrol price is the most related factor while the foreign exchange rate does not affect the performance of the stock market.

In 2017 Pyo et al. have investigated the predictability of machine learning techniques to predict the trends of the Korean stock market index (KOSPI 200) [10]. They have analyzed the performance of three nonparametric machine learning models: artificial neural network, support vector machine with polynomial, and radial basis function kernels. Their experiments revealed that the prediction of the KOSPI 200 using technical indicators (as inputs of the machine learning models) does not result in a good performance for market investments. On the other hand, they show that the google trend is not a suitable input factor in predicting the KOSPI 200 index prices.

In 2018 Senapati and his colleagues presented a hybrid method called, PSO-ANN, to predict the open price of a stock for 1 day ahead [11]. In this method, PSO is utilized to tune the weights of the Adaline neural network.

The time-series data of the Bombay stock market is used to evaluate the performance of the proposed method.

According to the reported results, PSO-ANN has shown better performance than ANN which indicates the efficiency of PSO in training ANN.

In 2018 Hu et al. introduced a method based on an improved sine cosine optimization algorithm (ISCA) and backpropagation neural network (BPNN) to predict the direction of stock markets (ISCA-BPNN) [12]. They have used ISCA to find the best possible values of the weights and biases of the neural network for maximum accuracy. In fact, ISCA is utilized to train the BPNN to predict the opening price of the next day with maximum accuracy. Evaluating the performance of the proposed method on “S&P 500” and “Dow Jones” datasets, they have demonstrated the superiority of the ISCA over GWO, WOA, and PSO.

In 2019 a new approach based on deep neural networks is introduced by Pang and his teammates [13]. In this project, two types of deep learning methods are used to forecast the Shanghai A-shares composite index: 1- LSTM with embedded hidden layer. 2- LSTM with automatic encoder.

Furthermore, in this research, a new concept, called “Stock Vector” is introduced. In fact, the input is not a single index or a single stock index, but multi-high dimensional historical data.

Gozalpour and Teshnehlab have proposed a stock price prediction method using deep neural networks in 2019 [14]. In their method dimensional reduction algorithms (PCA and autoencoder) are used to map the data points into a new feature space.

Besides, their method is designed to predict the close price of the next day using the stock price information (open price, lowest price, highest price and volume transaction) of the past 30 days. The method is tested on three NASDAQ symbols.

Ghanbari and Arian have introduced a hybrid regression method, briefly called BOA-SVR, in 2019 [15]. In this method BOA is used for parameter tuning of SVR. Also, phase space reconstruction method is used for data preparation.

In 2020 Vijh et al. have used ANN and Random Forest for the prediction of the close price [16]. The methods are trained using 6 new features extracted from the historical close prices of different stocks including 1- stock High minus Low price. 2- stock Close minus Open price. 3- Stock price’s seven days’ moving average. 4- Stock price’s fourteen days’ moving average. 5- stock price’s twenty-one day’s moving average. 6- stock price’s standard deviation for the past seven days. After testing the methods on different companies of the NASDAQ stock market, the results showed the superiority of ANN over Random Forest.

Ecer and his colleagues have introduced a hybrid stock index forecasting method in 2020 [17]. In their research, they have utilized evolutionary algorithms to train MLP to estimate the direction of the Borsa Istanbul (BIST) index using 9 technical indicators. Their experiments have shown that using $Tanh(x)$ as the output function of MLP results in better accuracy in compare to Gaussian function.

In a research project, conducted by Nabipour and his teammates in 2020, the performance of nine machine learning methods (Decision tree, Random forest, Adaptive Boosting (Adaboost), eXtreme Gradient Boosting (XGboost), Support Vector Classifier (SVC), Naïve Bayes, K-Nearest Neighbors (KNN), Logistic Regression and Artificial Neural Network (ANN)) and two powerful deep learning models (Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM)) are compared over the stock data of four different groups of stocks in Tehran stock exchange market [18]. They have used ten technical indicators as input variables for the learning models in two ways: continuous and binary. Their results showed the superiority of deep learning models over conventional machine learning models in both cases.

In a paper published recently, Awan et al. have used several machine learning methods for stock price prediction including linear regression, generalized linear regression, random forest, decision tree, naive Bayes and logistic regression [19]. They have used the historical data of the 15 famous companies and also the data in news, twitter, blogs and etc. to train the predictive models. Based on their experiments, linear regression, random forest and decision tree have shown the best performance in predicting the close value of the next day.

Kofi and his colleagues have tried to address the problem of stock price prediction in a different perspective [20]. They have proposed a fusion framework, based on convolutional neural network and a Long-Short term memory, to create a structured dataset by fusion of different heterogeneous datasets. To evaluate their framework, they have created a dataset of Ghana stock exchange market using different type of datasets ad trained a CNN on the created dataset. Their prediction accuracy showed that their framework has a positive effect on the final accuracy of the prediction model.

Tuarob et al. have proposed an end-to-end framework for stock market prediction [21]. Their framework, called DAViS, has different capabilities including data collection, analyzation and visualization. This framework can process the related heterogeneous stock data.

Also, it uses ensemble learning to predict the close price of a stock for one day ahead. Their simulation results, showed an improvement in the accuracy in compare to other baseline methods.

In [22] Muhammad Ali and his colleagues, have used the resilient back-propagation neural network to predict the direction movement of the stock market index. They have tested their method on KSE-100 index, KOSPI index, Nikkei 225 index and SZSE composite index. Comparing the results to SVM, showed the superiority of the neural network over SVM.

In most of the researches conducted recently, the researchers have used deep learning models to uncover the patterns hidden in the train data since the stock markets data contain a huge amount of information. Actually, analyzing the massive time-series data of the stock markets requires powerful learning models. Although deep learning models are strong methods that can be used to build accurate prediction models, the most important issue is the high cost of using these methods. In other words, to use deep learning methods, a powerful processing machine is needed. Another group of researchers has tried hybrid learning methods for this purpose.

Usually, in these methods, a metaheuristic algorithm is utilized to train a machine learning algorithm on the data, more effectively. Despite the effectiveness of hybrid learning methods, massive and high-dimensional datasets in addition to the iterative nature of metaheuristics result in a time-consuming training phase and decreasing the efficiency of the learning method in uncovering the hidden patterns of the data. So in this research first, the huge and high-dimensional training dataset is divided into smaller datasets based on their similarity using a novel automatic clustering algorithm.

In the next step, instead of training a single machine learning method on the whole massive data, an independent regression method in trained for each group of dataset individually.

Training different regression methods on the smaller datasets, containing similar data points, has had a key role in the performance of the prediction method. On other hand in the second phase, the metaheuristic algorithm (HHO) is used to find the best feature subset in addition to the parameters of the machine learning method (SVR).

Removing redundant features and using informative ones to train the machine learning method, improve the final accuracy of the prediction model.

Briefly, the main difference between our method and other methods, is that we have tried to reduce the amount of data and dimensions in a heuristic way to improve the performance. The experimental results show the effect of clustering and feature selection.

Data Preparation

In this research, the historical data of the past n days and four technical indicators including Bollinger bands (upper band, the middle band, and the lower band) and RSI are used to predict the index of the Tehran stock market in the next m days. For example, for $n = 20$ and $m = 1$, the historical data of the past 20 days (close, open, high, low and volume) besides four technical indicators are used to predict the market index in the next day. In this case, the first data point of the dataset is a vector with 104 items including the historical information of the first 20 working days and the corresponding technical indicators.

The target value is the close value of the 21st day. Similarly, the second object of the dataset contains the historical data and technical indicators from the second to the 21st working day and the target is the stock market index value in the 22nd working day. In fact, the main goal of this research is to predict the stock market index in the next 7 days using a two-stage prediction method. In this way, we have investigated the effect of window size (n) in the performance of the stock market index predicting method. The method is completely explained in the next section.

Proposed Method

The proposed method includes two stages: 1- clustering. 2- regression. In the first stage, APSO-Clustering [23]-[25] is utilized to cluster the training dataset. This novel automatic clustering algorithm, which can detect the number of clusters in addition to the centroids, divides the whole training dataset into different clusters. In the second stage, for each detected cluster a regression method, which is a combination of Harris Hawks Optimization algorithm (HHO) and Support Vector Regression, is hired to uncover the hidden patterns of each cluster.

In this hybrid regression method, briefly called HHO-SVR, HHO is utilized for feature selection and parameter tuning of SVM. Actually, in the training phase, HHO searches the solution space to find the best subset of features and the optimal value for the SVM's parameter. To estimate the target value of a test sample, after distinguishing its cluster, the corresponding trained regression model determines the target value of the test sample.

A. APSO-CLUSTERING

APSO-Clustering, designed based on Particle Swarm Optimization algorithm, can detect the proper number of clusters in addition to the position of centroids. This clustering method works in two phases.

Detecting the number of cluster is the main goal of the first phase while, finding the exact position of the centroids is the main goal of the second phase. Thus the main superiority of APSO-Clustering over traditional clustering methods such as K-means and fuzzy C-means is its high capability in detecting the number of clusters. This capability is more valuable when dealing with big datasets. In both phases, PSO-Clustering, a non-automatic clustering method, is used. In this clustering method, Particle Swarm Optimization algorithm is hired to find k centroids, while k should be predetermined by the user.

In fact, k is the input of PSO-Clustering. In the first phase of APSO-Clustering, PSO-Clustering is run several times sequentially with different values of k to detect the best-fitted number of clusters. Each time, the best solution found by PSO is compared with the previously found solutions to distinguish the best value of k. In each step of the first phase, the number of population and iteration numbers are set to 5 and 150 respectively, and also Calinski-Harabasz index is used for fitness evaluation. In the second phase, again PSO-Clustering finds out the exact position of k centroids. In this phase, to explore and exploit the search space completely, the iteration number is set to 600.

The pseudo-code of APSO-Clustering is shown in Fig. 1. In the next sections, HHO and HHO-SVR are described respectively.

```
Stage 1:
Inputs: number of sequences, length of sequence, number of population
For i=1 to number of sequences do:
        1.k = Generate a random integer number
        2. beginning population = Generate a random population
        3.Current state.fitness = inf
        4. Current state.k=Ø
        for j=1 to length of sequence do:
                if (j==1) do:
                        current state.k=k
                end
                best fitness = PSO-Clustering(k , beginning population)
                if best fitness < current state.fitness
                        current state.fitness=best fitness
                        curent state.k=k
                end
                k = current state.k ± ε
        end
Bestk[i] = current state.k
End
Output: find the best solution which has the best fitness amount, its corresponding k and
corresponding beginning population
Stage 2:
Final output = PSO-Clustering(best k, beginning population)
```

Fig. 1: the pseudo-code of APSO-Clustering [23].

B. Harris Hawks Optimization algorithm (HHO)

This optimization algorithm is invented by the inspiration of Harris' Hawks hunting mechanism. The Harris' Hawk is a well-known bird of prey that survives in somewhat steady groups found in the southern half of Arizona, USA [26].

These birds are known as truly cooperative predators in the raptor realm. The main tactic of Harris' Hawks to capture prey is "Surprise pounce" which is also known as the "seven kills" strategy. In this strategy, several Hawks try to cooperatively attack from different directions and converge on a detected escaping rabbit at the same time. HHO algorithm is created by mimicking the behavior of the Harris Hawks in hunting and also the behavior of the prey (rabbit) in escaping mechanisms. In fact, in this algorithm, the Hawks are the search agents and the prey is the optimum solution supposed to be found (hunted) by the search agents (Harris Hawks). Generally, HHO consists of two phases: 1- exploration. 2- exploitation. In each phase, the search agents move in the solution space using a specified criterion [27].

Exploration phase: In the exploration phase, the search agents try to discover different areas of the solution space which is also a common strategy among the Harris Hawks in nature. This strategy (exploration) is modeled in HHO by the following equation:

$$X(t+1) = \begin{cases} X_{rand}(t) - r_1 \left|X_{rand}(t) - 2r_2X(t)\right| & q \geq 0.5 \\ \left(X_{rabbit}(t) - X_m(t)\right) - r_3\left(LB + r_4(UB - LB)\right) & q < 0.5 \end{cases} \quad (1)$$

According to (1), the search agents explore the solution space using two strategies:

1- Perching based on the position of the search agents and the position of the rabbit (best search agent) for the condition of $q < 0.5$.

2- Perching on random locations inside a specified range for the condition of $q \geq 0.5$.

In this equation, $X(t+1)$ indicates the location of the search agents in the next iteration, $X_{rabbit}(t)$ is the position of the best solution found from the beginning of the optimization process, $X(t)$ is the current position of the search agents, r_1, r_2, r_3, r_4 and q are random numbers inside (0.1), LB and UB are lower and upper bounds of the variables, $X_{rand}(t)$ is a randomly selected search agent and $X_m(t)$ is the average position of the search agents.

Exploitation phase: In the exploitation phase, the Hawks perform the “surprise pounce” strategy to catch the rabbit. On the other hand, the rabbit also tries to escape from the dangerous situation. Thus based on the probability of escaping, which is defined by a random number (r) and the energy of the rabbit, shown in (2), four strategies are defined for the exploitation phase.

$$E = 2E_0(1 - \frac{t}{T}) \tag{2}$$

Soft besiege: If $r \geq 0.5$ and $|E| \geq 0.5$ the rabbit has enough energy to escape.

The following equations show the movement of search agents in this situation:

$$X(t+1) = \Delta X(t) - E|JX_{rabbit}(t) - X(\tag{3}$$

$$\Delta X(t) = X_{rabbit}(t) - X(t) \tag{4}$$

In these equations, $\Delta X(t)$ presents the difference between the position of the rabbit and the current position. r_5 is a random number and $J = 2(1 - r_5)$ is the random jump strength of the rabbit in the escaping procedure.

Hard besiege: In this situation ($r \geq 0.5$, $|E| < 0.5$) the rabbit is so exhausted and it has low escaping energy. Hence the position of the search agents is updated through the following equation:

$$X(t+1) = X_{rabbit}(t) - E|\Delta X(t)| \tag{5}$$

Soft besiege with progressive rapid dives: In this case, the prey has enough energy to escape ($|E| > 0.5$) but a soft besiege in constructed by the Hawks ($r < 0.5$). In this situation to simulate the real zigzag deceptive movements of the prey (especially rabbits) and rapid dives of Hawks around the escaping prey, the Levy Flight concept [28], [29] is utilized in the HHO algorithm. Based on this strategy the position of the search agents is updated using the following equations:

$$X(t+1) = \begin{cases} Y = X_{rabbit}(t) - E|JX_{rabbit}(t) - X(t)| & if\ F(Y) < F(X(t)) \\ Z = Y + S \times LF(D) & if\ F(Z) < F(X(t)) \end{cases} \tag{6}$$

According to this equation, it is supposed that the Hawks can evaluate their next possible move and then decide to choose the better one. In other words, in each iteration, the better position (Y or Z) is selected as the next position of the search agent.

In this equation, D is the dimension of the problem, S is a random vector by size $1 \times D$ and LF is the Levy Flight function.

Hard besiege with progressive rapid dives: When $|E| < 0.5$ and $r < 0.5$ the prey is exhausted and also a hard besiege is constructed by the Hawks. The following equations indicate how search agents update their position in this circumstance:

$$X(t+1) = \begin{cases} Y = X_{rabbit}(t) - E|JX_{rabbit}(t) - X_m(t)| & if\ F(Y) < F(X(t)) \\ Z = Y + S \times LF(D) & if\ F(Z) < F(X(t)) \end{cases} \tag{7}$$

The pseudocode of HHO is shown in Fig. 2.

```
Inputs: The population size N and maximum number of
iterations T
Outputs: The location of rabbit and its fitness value
Initialize the random population Xi(i = 1, 2, . . . , N)
while (stopping condition is not met) do
    Calculate the fitness values of hawks
    Set Xrabbit as the location of rabbit (best location)
    for (each hawk (Xi)) do
        Update the initial energy E0 and jump strength J
        Update the E using Eq. (2)
        if (|E|≥ 1) then
            Update the location vector using Eq. (1)
        if (|E|< 1) then
            if (r ≥0.5 and |E|≥ 0.5 ) then
                Update the location vector using Eq. (4)
            else if (r ≥0.5 and |E|< 0.5 ) then
                Update the location vector using Eq. (6)
            else if (r < 0.5 and |E|≥ 0.5 ) then
                Update the location vector using Eq. (7)
            else if (r < 0.5 and |E|< 0.5 ) then
                Update the location vector using Eq. (9)
Return Xrabbit
```

Fig. 2: the pseudocode of HHO [27].

C. HHO-SVR

Support vector regression (SVR) is the developed version of the support vector machine classifier (SVM) which is suitable for the regression problems [30], [31]. Considering the smallest risk minimization principle in high-dimensional feature space, this well-known regression method finds the best regression hyperplane. This non-linear method maps the data points from vector space to high-dimensional feature space using a kernel function to facilitate the process of distinguishing different objects [32]. Several kernel functions have been introduced up to now. The Gaussian function is one of the most popular functions which has been used frequently in different researches [33], [34]. This function maps the data points into feature space using the following equation:

$$K(x_i.x_j) = exp\left(-\gamma\|x_i - x_j\|^2\right) \tag{8}$$

The amount of γ (Gaussian kernel's parameter. See (10)) has a significant effect on the performance of SVR. In Support Vector Regression machine, the user should select a kernel function and set the kernel parameter in order to achieve better generalization performance.

One of the most important parameters is the kernel parameter which implicitly defines the structure of the high dimensional feature space where the maximal margin hyperplane is found.

Too rich a feature space would cause the system to overfit the data.

The Gaussian kernel function is the most common used kernel function. Therefore, its parameter, γ, needs to be determined before the SVR is trained. It has been proved that γ, in Gaussian kernel function, dramatically affects the generalization performance of SVR. When γ is very small, all the training data will be regarded as support vector, and therefore they can be classified correctly.

However, for any unseen data, the SVM may not give right distinction due to "over-fitting" training. On the other hand, when γ is very large, all the training data are regarded as one point in feature space, the SVM cannot recognize any unseen data due to "under-fitting" training. Obviously, these two extreme situations should be avoided. Deeper analysis on this important topic is provided in [35], [36].

So finding the optimal value of γ is an important task that is done by HHO in this research. In other words, removing redundant features and building a regression model based on informative features usually results in better prediction performance.

So HHO should search the solution space to find the best feature subset and the best value of γ simultaneosly. Thus, each search agent contains f cells for feature selection ($f = number\ of\ features$), which are encoded binary (1 for selecting the feature and 0 for removing the feature), and one cell for the value of γ. For fitness evaluation Mean Squared Error (*MSE*) function is used.

Results and Discussion

The performance of the proposed method is evaluated on the historical data of the Tehran Stock Exchange market index from 6/12/2008 to 1/11/2020. To evaluate the performance of the method after training on 70% of the dataset, the following evaluation metrics have been calculated on the rest of the data:

$$MSE = \frac{1}{n}(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2) \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}(\sum_{i=1}^{n}(y_i - \hat{y}_i)^2)} \tag{10}$$

$$MAE = \frac{1}{n}(\sum_{i=1}^{n}(y_i - \hat{y}_i)) \tag{11}$$

$$R^2 = 1 - (\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \mu)^2}) \tag{12}$$

The first three metrics (mean squared error (*MSE*), root mean squared error (*RMSE*), and mean absolute error (*MAE*)) show the error rate. In other words, these metrics show the difference between the actual and estimated target values. On the other hand, R^2 shows the similarity.

Bigger value of R^2 means better performance of the method.

In the next subsections, the values of performance metrics achieved for different amounts of n (window size) and m (target day) are presented. The simulation parameters and their assigned values are shown in Table 1.

Table 1: Simulation parameters

Parameter	Value
PSO_Iter_1	150
PSO_Pop_1	5
PSO_Iter_2	600
PSO_Pop_2	5
HHO_Iter	50
HHO_Pop	15

The description of these parameters are as follows:

- *PSO_Iter_1:* iteration number of PSO, in PSO_Clustering in the first stage of APSO-Clustering.
- *PSO_Pop_1:* population number of PSO_Clustering in the first stage of APSO-Clustering
- *PSO_Iter_2:* the iteration number of PSO_Clustering in the second stage of APSO-Clustering.
- *PSO_Pop_2:* population number of PSO_Clustering in the second stage of APSO-Clustering.
- *HHO_Iter:* iteration number of HHO in the second phase of the proposed method.
- *HHO_Pop:* population number of HHO in the second stage of the proposed method.

A. Predicting one day ahead (m=1)

In this case, the target day is tomorrow. In other words, the main goal is to predict the stock market index value of tomorrow (close value) using the last n days. The evaluation metrics of the method for different values of n are presented in Table 2.

According to this table, the proposed method gives the best performance when $n = 20$. This means that, to predict the index value of the next day, considering the historical information of the last 20 days is enough. The details of the results of this experiment ($n = 20$) are presented in Table 3.

This table indicates that APSO-Clustering has detected 2 clusters.

In the second step, two hybrid regression methods are trained on each cluster’s data. According to this table, HHO has detected 21 and (only) 3 features for the first and second clusters respectively.

Among these features, none of the technical indicators (RSI, upper Bollinger band, lower Bollinger band, and mid-Bollinger band) is selected.

Table 2: Performance evaluation of the proposed method for m=1

n	20	40	60	80	100
MSE	$\mathbf{1.55\times10^{-4}}$	4.07×10^{-4}	1.907×10^{-4}	2.97×10^{-4}	4.51×10^{-4}
RMSE	**0.0125**	0.0202	0.0138	0.017	0.0212
MAE	**0.0051**	0.0069	0.0052	0.0066	0.0084
R^2	**0.9929**	0.9816	0.99	0.986	0.9809

Table 3: Details of the achieved results for m=1 and n=20

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	21	1	none
HHO-SVR-2	3	1	none

In Fig. 3, the curves of real and predicted index values are demonstrated for $n = 20$. According to Fig. 3, the predicted values are very close to the real values which shows the fabulous performance of the presented method in predicting the market index value of 1 day ahead.

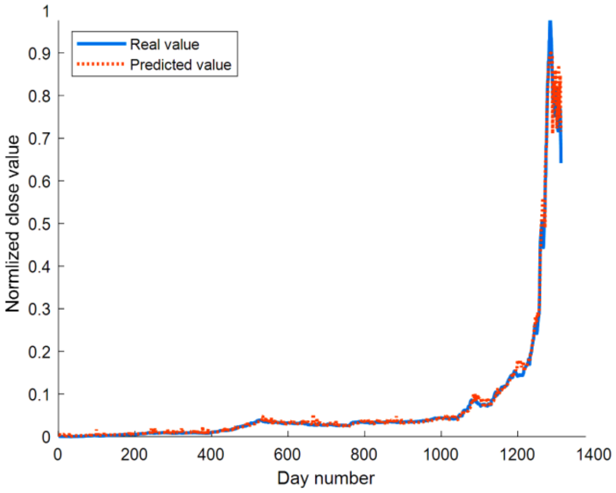


Fig. 3: Real and predicted close values for m=1 and n=20.

B. Predicting two days ahead (m=2)

In these experiments, the performance of the method in predicting the index value of the next two days, is evaluated for different values of n.

In Table 4, the values of different evaluation metrics are shown. According to Table 4, the proposed method gives the best performance when $n = 100$. Also, when $n = 60$ the R^2 index of the method is 99.38% which is very promising. In Table 5 more details of the results, for $n = 100$, are shown.

Table 5 indicates that SVR can predict the market situation for the next 2 days using only 4 and 2 features for the first and second clusters respectively. Furthermore, no technical indicators are needed for the prediction.

In other words, having the historical information of the past 100 days, the presented method can predict the stock market index for the next two days accurately.

Also, the best detected value for γ for each of the trained models is 1.

In Fig. 4, the curves of real and estimated values are presented.

Table 4: Performance evaluation of the proposed method for m=2

n	20	40	60	80	100
MSE	3.4×10^{-4}	5.03×10^{-4}	1.56×10^{-4}	8.43×10^{-4}	**7.82×10^{-5}**
RMSE	0.0186	0.0224	0.0125	0.029	**0.0088**
MAE	0.0069	0.0075	0.0064	0.0071	**0.0058**
R^2	0.986	0.9771	0.9938	0.9614	**0.9958**

Table 5: Details of the achieved results for m=2 and n=100

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	4	1	None
HHO-SVR-2	2	1	None

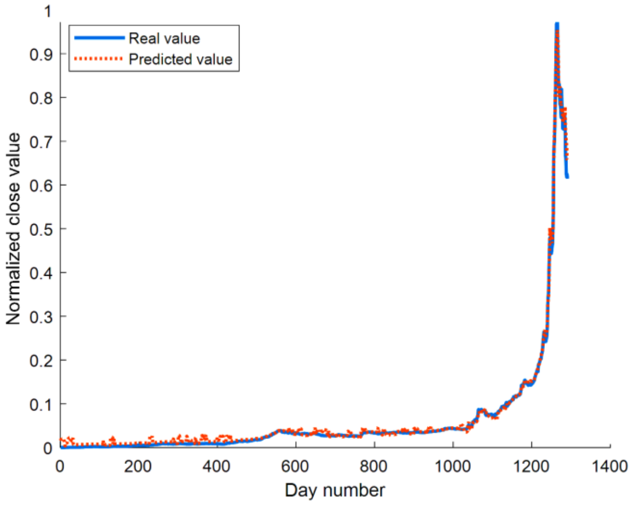


Fig. 4: Real and predicted close values for m=2 and n=100.

C. Predicting three days ahead (m=3)

The values of the evaluation metrics and the details of the best result are shown in Tables 6 and 7 respectively.

Table 6: Performance evaluation of the proposed method for m=3

n	20	40	60	80	100
MSE	1.27×10^{-4}	6.13×10^{-4}	3.004×10^{-4}	9.18×10^{-5}	**9.16×10^{-5}**
RMSE	0.0113	0.0248	0.0173	0.0096	**0.0096**
MAE	**0.0045**	0.0085	0.0057	0.0059	0.0068
R^2	0.9938	0.9723	0.9883	**0.9955**	0.9952

Table 7: Details of the achieved results for m=3 and n=80

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	9	1	None
HHO-SVR-2	178	5	None

According to Table 6, the least mean squared error is achieved when $n = 100$ while the best R^2 is 0.9955 for $n = 80$.

Also, when $n = 20$, the proposed method predicts the next three days fairly accurately (99.38%). Fig. 5 shows the real and estimated prices for this experiment (m=3 and n=80).

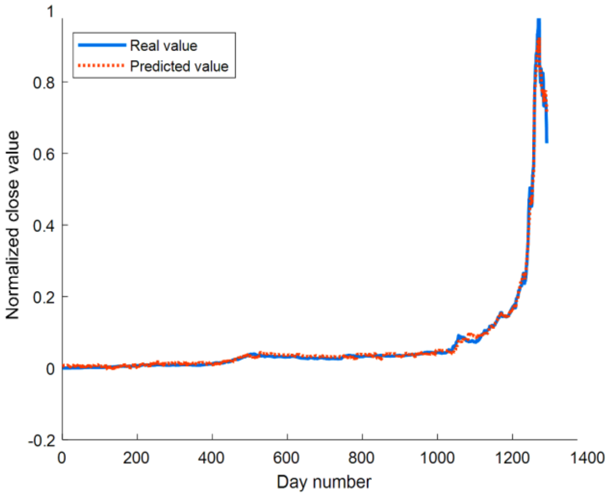


Fig. 5: Real and predicted close values for m=3 and n=80.

D. Predicting four days ahead (m=4)

In Table 8, the performance evaluation of the method is presented for $m = 4$.

According to this table, although the best performance (R^2 index) has been achieved for $n = 100$, the best performance in terms of mean absolute error (MAE) belongs to $n = 20$ and $n = 40$.

Table 8: Performance evaluation of the proposed method for m=4

n	20	40	60	80	100
MSE	2.45×10^{-4}	5.12×10^{-4}	0.0013	0.0011	**1.35×10^{-4}**
RMSE	0.0157	0.0226	0.036	0.0331	**0.0116**
MAE	**0.0068**	0.0068	0.0082	0.0093	0.0079
R^2	0.9894	0.9766	0.9392	0.9604	**0.994**

In this case, the method has shown great performance for $n = 20$. This indicates that to predict the stock market index value of the next 4 days, using the historical information of the past 20 days will result in good performance although it is better to use the past 100 days. Besides that, giving the weakest performance for $n = 60$ reveals the fact that increasing the number of past days (n) will not always result in better performance.

It can add ambiguity to the data and thus degrade the performance of the method.

Table 9 shows that in the first phase of the method, two clusters have been detected by APSO-Clustering. For the first cluster, HHO has selected 218 features including the lower Bollinger band, while for the second cluster 276 features are selected without any of the technical indicators.

The best-detected values for γ are 5 and 18 for the first and second clusters respectively.

Fig. 6 demonstrates the curves of real and predicted values for $n = 100$. Fig. 6 shows the high capability of the introduced method in detecting and tracking the market trend.

Table 9: Details of the achieved results for m=4 and n=100

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	218	5	Lower Bollinger band
HHO-SVR-2	276	18	None

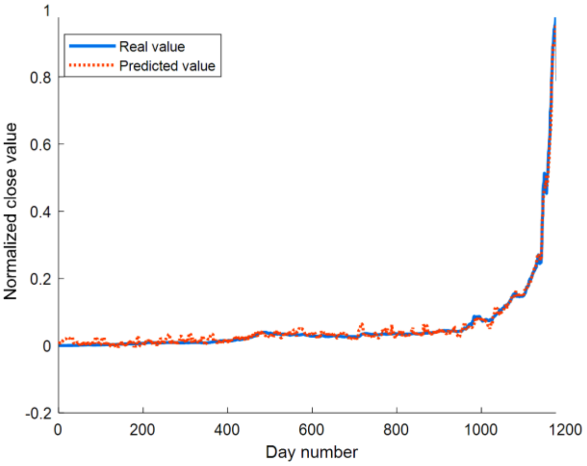


Fig. 6: Real and predicted close values for m=4 and n=100.

E. Predicting five days ahead (m=5)

Table 10 and Table 11 present the results provided by the introduced method for $m = 5$. These tables show that including the historical information of the last 40 days, APSO-Clustering has divided the data points of the created dataset into 2 clusters in the first phase. In the second phase, training two regression methods using 13 and 3 features of the first and second cluster's data respectively, have given the best performance. Furthermore, according to Table 10, using the historical data of the past 100 days has given the same result as using the historical data of the past 20 days which confirms that more number of days will not always result in better performance in decision making. Fig. 7 shows the real and estimated curves for $n = 40$.

Table 10: Performance evaluation of the proposed method for m=5

n	20	40	60	80	100
MSE	6.31×10^{-4}	**2.1×10^{-4}**	0.0018	7.6×10^{-4}	4.63×10^{-4}
RMSE	0.0251	**0.0145**	0.0425	0.0276	0.0215
MAE	0.008	**0.0058**	0.0107	0.0079	0.008
R^2	0.9713	**0.9908**	0.9223	0.9702	0.9782

Table 11: Details of the achieved results for m=5 and n=40

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	13	1	None
HHO-SVR-2	3	1	None

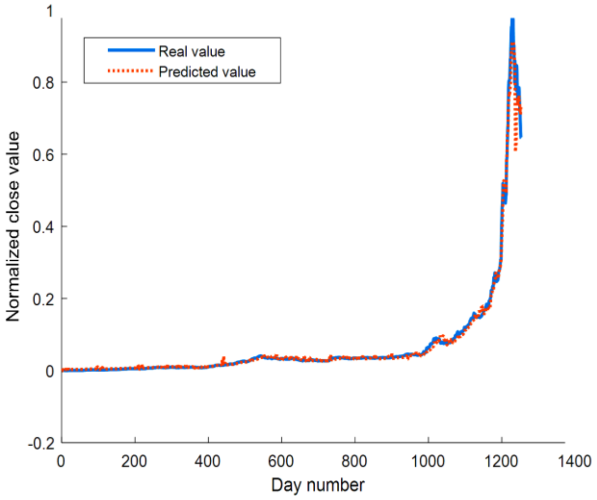


Fig. 7. Real and predicted close values for m=5 and n=40.

F. Predicting six days ahead (m=6)

In this case, according to Table 12, forecasting the market situation of the next 6 days using the historical data of the past 100 days has brought the best performance while the method has shown a very close performance for $n = 40$ ($R^2 = 0.9848$), $n = 60$ ($R^2 = 0.9934$) and $n = 80$ ($R^2 = 0.9946$). Table 13 shows that the second regression method has been trained on the data of the second cluster with 185 features including upper and lower Bollinger bands. In Fig. 8, the curves of the real and predicted values are shown.

Table 12: Performance evaluation of the proposed method for m=6

n	20	40	60	80	100
MSE	7.98×10^{-4}	3.61×10^{-4}	1.39×10^{-4}	1.42×10^{-4}	**9.11×10^{-5}**
RMSE	0.0282	0.019	0.0118	0.012	**0.0095**
MAE	0.0078	0.0065	**0.006**	0.0068	0.0068
R^2	0.9668	0.9848	0.9934	0.9946	**0.9958**

Table 13: Details of the achieved results for m=6 and n=100

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	2	1	None
HHO-SVR-2	185	4	Upper and lower Bollinger bands

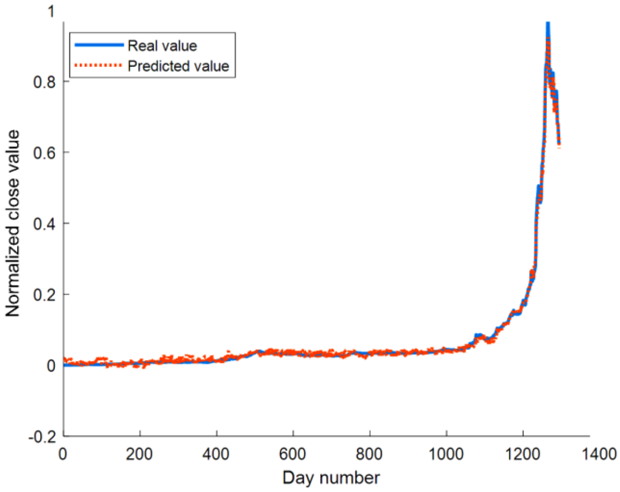


Fig. 8: Real and predicted close values for m=6 and n=100.

G. Predicting seven days ahead (m=7)

According to Table 14, like the previous case ($m = 6$), the best performance is achieved when $n = 100$, while for $n = 20$ and $n = 40$ the method has shown promising performance in forecasting the next 7 days.

Table 15 shows that including RSI, mid and lower Bollinger bands in the data points of the second cluster have a key role in reducing the error rate of the method. The great performance of the method is shown in Fig. 9 where the real and predicted curves are very close to each other.

Table 14: Performance evaluation of the proposed method for m=7

N	20	40	60	80	100
MSE	2.12×10^{-4}	2.19×10^{-4}	2.54×10^{-4}	4.25×10^{-4}	**1.28×10^{-4}**
RMSE	0.0146	0.0148	0.160	0.0206	**0.0113**
MAE	**0.0055**	0.0063	0.0071	0.0067	0.0081
R^2	0.9903	0.9905	0.9872	0.9808	**0.9945**

Table 15: Details of the achieved results for m=7 and n=100

	Number of selected features	γ	Selected technical indicators
HHO-SVR-1	2	1	None
HHO-SVR-2	347	5	RSI, mid and lower Bollinger band

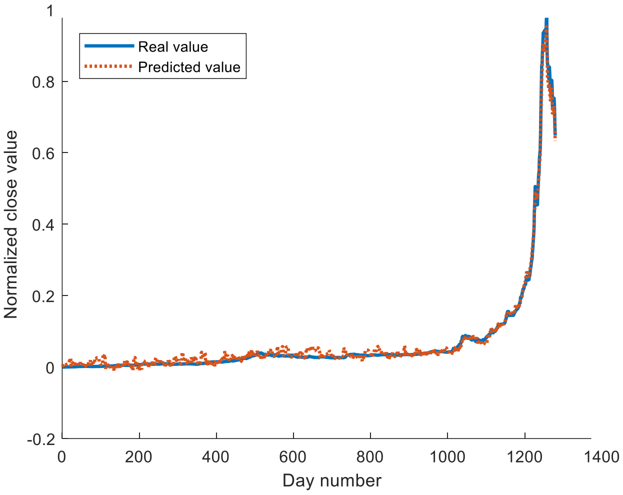


Fig. 9: Real and predicted close values for m=7 and n=100.

Comparing With Other Metaheuristics, Self-Tuned SVR and Linear Regression

In order to reach a fair judgment about the efficiency of HHO-SVR, the second stage of the proposed method is implemented using other well-known metaheuristics (PSO, MVO [37], GSA [38], and IPO [39]). Just like HHO, the number of iteration and population for all of the metaheuristic optimization methods are set to 50 and 15 respectively. In Tables 16 to 19 the performance evaluation of PSO-SVR, MVO-SVR, GSA-SVR, and IPO-SVR for $m = 7$, are demonstrated respectively. To compare the results easily, the graphs of R^2 values of each method are shown in Fig. 10.

Table 16: Performance evaluation of PSO-SVR for m=7

n	20	40	60	80	100
MSE	0.007	0.0154	0.0175	0.0017	**1.14×10^{-4}**
RMSE	0.0838	0.1243	0.1324	0.0414	**0.0107**
MAE	0.0179	0.0332	0.0405	0.0099	**0.0077**
R^2	0.6776	0.3307	0.1220	0.9225	**0.9951**

Table 17: Performance evaluation of MVO-SVR for m=7

n	20	40	60	80	100
MSE	**4.13×10^{-4}**	0.0083	0.0034	0.0016	0.0249
RMSE	**0.0203**	0.0912	0.0582	0.0401	0.1578
MAE	**0.0074**	0.022	0.0156	0.0094	0.0551
R^2	**0.9810**	0.6396	0.8304	0.9273	0.0592

Table 18: Performance evaluation of GSA-SVR for m=7

n	20	40	60	80	100
MSE	0.0135	0.0244	0.208	0.0027	**0.0014**
RMSE	0.1161	0.1561	0.1443	0.0516	**0.0380**
MAE	0.0366	0.0555	0.0511	0.0145	**0.0125**
R^2	0.3812	0.0556	0.043	0.8798	**0.9386**

Table 19: Performance evaluation of IPO-SVR for m=7

n	20	40	60	80	100
MSE	0.0228	0.0244	0.0184	**0.0049**	0.016
RMSE	0.1511	0.1561	0.1355	**0.0698**	0.1265
MAE	0.0547	0.0555	0.0452	**0.0225**	0.0843
R^2	0.0487	0.0556	0.0799	**0.7794**	0.3188

Tables 13 to 18 and Fig. 10, show that HHO-SVR has a high potential in solving such a hard and complex optimization problem while the other methods failed in detecting the global optimum point successfully.

Besides that, according to Tables 13 to 18, all of the methods have shown a promising performance in predicting the index value of the next week using the historical information of the past 80 days.

In other words, using the information of the past 80 days reduces the complexity of the problem for the optimization algorithms which in turn increases the likelihood of finding the near-global optimum point by the optimization algorithm.

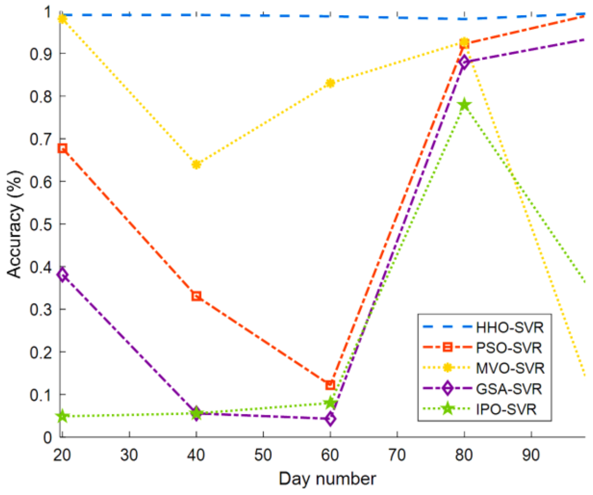


Fig. 10: R^2 values' graphs of different methods for m=7.

Also, to compare the performance of the whole method with other regression methods, the values of evaluation metrics of linear regression and the fine-tuned SVR, introduced in [41], for m=7, are presented in Tables 20 and 21 respectively. These two well-known regression methods are trained on the whole training data.

Table 20: Performance evaluation of linear regression SVR for m=7

n	20	40	60	80	100
MSE	0.0218	0.0231	0.02	0.0221	0.0235
RMSE	0.1476	0.1519	0.1415	0.1486	0.1533
MAE	0.0746	0.0745	0.0721	0.0747	0.0753
R^2	0.053	0.048	0.056	0.047	0.04

According to these two tables, although fine-tuned SVR has shown good performance but it cannot overcome the proposed method in predicting the stock market index value of the next 7 days.

Table 21: Performance evaluation of fine-tuned for m=7.

n	20	40	60	80	100
MSE	2.7×10^{-4}	7.9×10^{-4}	0.0025	5.4×10^{-4}	0.0055
RMSE	0.0165	0.0282	0.05	0.0233	0.0743
MAE	0.0055	0.0095	0.0203	0.0125	0.0282
R^2	0.9876	0.9655	0.87	0.9755	0.7652

Conclusion and Future Works

In this paper, an effective method for forecasting the future of the stock market is proposed which works in two stages. In the first stage, the training dataset is clustered using a novel automatic clustering method, called APSO-Clustering that can detect the proper number of clusters and the position of the centroids simultaneously.

This capability is very valuable when dealing with massive and high-dimensional datasets. In the second stage a hybrid regression method, called HHO-SVR, is trained for each cluster's data points.

In this regression method, HHO is utilized for feature selection and parameter tuning of SVR. To estimate the target value of an unknown sample, after determining its cluster, the corresponding regression method estimates the target value.

The main goal of this research was to predict the future of the Tehran Stock Exchange market. For this purpose, the historical data of the market index in addition to some technical indicators are used for data preparation. Several experiments have been conducted to evaluate the accuracy and effectiveness of the method.

In fact, in this research we have gone further in compare to our previous work [40]. While in our previous research we have analyzed the performance of the proposed method in predicting the price of the stocks in one day ahead, in this research we have tried to forecast the situation of the whole market in the next seven days which gives the traders a good opportunity to make a proper decision.

Besides, we have tried to forecast the market index of the next m (from 1 to 7) days using the historical data of the past n (from 10 to 100) days. The experiments show that increasing the number of days (n), used to create the dataset, will not necessarily improve the final accuracy of the method. Although in the last two experiments, the best performance has been achieved for $n = 100$, in most experiments the method has shown high accuracy in forecasting the future using the past 20 days.

On the other hand, in this research a new metaheuristic optimization algorithm is used for prediction which has shown a great accuracy in forecasting the market index value in the next week. This method has shown a great performance in predicting and tracking trends. Definitely, several unpredictable factors (political events, natural disasters, economic situation and etc.) affects the stock market that can produce sudden fluctuations.

These sudden jumps and rises in the index value are hard to be predicted since their causing factors are hardly predictable. Nevertheless, the most important thing is the prediction of the trends in the near future, that the methodology introduced in this paper performs it effectively.

For future works, the effect of different information (such as dollar exchange rate and inflation rate) on the performance of the proposed method, can be investigated.

Also, the method can be extended in order to process other types of data such as news and twits. Furthermore, the proposed method can be used to predict the future of other markets such as cryptocurrencies, which are very popular and interested nowadays.

Besides, other regression methods such as logistic regression and also deep learning methods can be used in the second phase, instead of HHO-SVR, to investigate their performance. Generally, our method is useful in solving different important regression problems such as electricity forecasting and etc.

Author Contributions

The introduced method is designed by Iman Behravan, while Dr. Seyed Mohammad Razavi has interpreted the final results.

Acknowledgements

We would like to appreciate Dr.Roberto Trasarti and Prof. Seyyed Hamid Zahiri for their helps and advices in the road of designing APSO-CLUSTERING.

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

PSO	Particle Swarm Optimization
APSO-Clustering	Automatic Particle Swarm Optimization-Clustering
HHO	Harris's Hawks Optimization
MVO	Multi-Verse Optimization
GSA	Gravitational Search Algorithm
IPO	Inclined Planes Optimization

References

[1] H. Chen, D.L. Fan, L. Fang, W. Huang, J. Huang, C. Cao, et al., "Particle swarm optimization algorithm with mutation operator for particle filter noise reduction in mechanical fault diagnosis," Int. J. Pattern Recognit. Artif. Intell., 34: 2058012, 2020.

[2] M. Janga Reddy, D. Nagesh Kumar, "Evolutionary algorithms, swarm intelligence methods, and their applications in water resources engineering: a state-of-the-art review," H2Open J., 3: 135-188, 2021.

[3] Q.V. Pham, D.C. Nguyen, S. Mirjalili, D.T. Hoang, D.N. Nguyen, P.N. Pathirana, et al., "Swarm intelligence for next-generation wireless networks: Recent advances and applications," arXiv preprint arXiv:2007.15221, 2020.

[4] M. Schranz, G.A. Di Caro, T. Schmickl, W. Elmenreich, F. Arvin, A. Şekercioğlu, et al., "Swarm intelligence and cyber-physical systems: concepts, challenges and future trends," Swarm Evol. Comput., 60: 100762, 2021.

[5] A. Kaveh, A.D. Eslamlou, Metaheuristic optimization algorithms in civil engineering: New applications: Springer, 2020.

[6] B. Yang, J. Wang, X. Zhang, T. Yu, W. Yao, H. Shu, et al., "Comprehensive overview of meta-heuristic algorithm applications on PV cell parameter identification," Energy Convers. Manage., 208: 112595, 2020.

[7] J. Nocedal, S. Wright, Numerical optimization: Springer Science & Business Media, 2006.

[8] G. Wu, "Across neighborhood search for numerical optimization," Inf. Sci., 329: 597-618, 2016.

[9] M. Usmani, S.H. Adil, K. Raza, S.S.A. Ali, "Stock market prediction using machine learning techniques," in Proc. 2016 3rd international conference on computer and information sciences (ICCOINS): 322-327, 2016.

[10] S. Pyo, J. Lee, M. Cha, H. Jang, "Predictability of machine learning techniques to forecast the trends of market index prices: Hypothesis testing for the Korean stock markets," PloS one, 12: e0188107, 2017.

[11] M.R. Senapati, S. Das, S. Mishra, "A novel model for stock price prediction using hybrid neural network," J. Inst. Eng. (India): Ser. B, 99: 555-563, 2018

[12] H. Hu, L. Tang, S. Zhang, H. Wang, "Predicting the direction of stock markets using optimized neural networks with Google Trends," Neurocomputing, 285: 188-195, 2018.

[13] X. Pang, Y. Zhou, P. Wang, W. Lin, V. Chang, "An innovative neural network approach for stock market prediction," J. Supercomputing, 76: 2098-2118, 2020.

[14] N. Gozalpour, M. Teshnehlab, "Forecasting stock market price using deep neural networks," in Proc. 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS): 1-4: 2019.

[15] M. Ghanbari, H. Arian, "Forecasting stock market with support vector regression and butterfly optimization algorithm," arXiv preprint arXiv:1905.11462, 2019.

[16] M. Vijh, D. Chandola, V. A. Tikkiwal, A. Kumar, "Stock closing price prediction using machine learning techniques," Procedia Comput. Sci., 167: 599-606, 2020.

[17] F. Ecer, S. Ardabili, S.S. Band, A. Mosavi, "Training multilayer perceptron with genetic algorithms and particle swarm optimization for modeling stock price index prediction," Entropy, 22: 1239, 2020.

[18] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, A. Mosavi, "Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis," IEEE Access, 8: 150199-150212, 2020.

[19] M.J. Awan, M.S.M. Rahim, H. Nobanee, A. Munawar, A. Yasin, A.M. Zain, "Social media and stock market prediction: A big data approach," Comput. Mater. Continua, 67: 2569-2583, 2021.

[20] I. K. Nti, A.F. Adekoya, B.A. Weyori, "A novel multi-source information-fusion predictive framework based on deep neural networks for accuracy enhancement in stock market prediction," J. Big Data, 8: 1-28, 2021.

[21] S. Tuarob, P. Wettayakorn, P. Phetchai, S. Traivijitkhun, S. Lim, T. Noraset, et al., "DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction," Financ. Innovation, 7: 1-32, 2021.

[22] M. Ali, D.M. Khan, M. Aamir, A. Ali, Z. Ahmad, "Predicting the direction movement of financial time series using artificial neural network and support vector machine," Complexity, 2021: 1-13, 2021.

[23] I. Behravan, S.H. Zahiri, S.M. Razavi, R. Trasarti, "Clustering a big mobility dataset using an automatic swarm intelligence-based clustering method," J. Electr. Comput. Eng. Innovations, 6(2): 251-271, 2018.

[24] I. Behravan, S.H. Zahiri, S. M. Razavi, R. Trasarti, "Finding roles of players in football using automatic particle swarm optimization-clustering algorithm," J. Big data, 7: 35-56, 2019.

[25] I. Behravan, S.M. Razavi, "A novel machine learning method for estimating football players' value in the transfer market," Soft Comput., 25: 2499-2511, 2021.

[26] J.C. Bednarz, "Cooperative hunting Harris' hawks (Parabuteo unicinctus)," Science, 239: 1525-1527, 1988.

[27] A.A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, H. Chen, "Harris hawks optimization: Algorithm and applications," Future Gener. Comput. Syst., 97: 849-872, 2019.

[28] S.M. Yusuf, C. Baber, "Multi-Agent searching adaptation using levy flight and inferential reasoning," Int. J. Electr. Inf. Eng., 14: 290-297, 2020.

[29] X.S. Yang, Nature-inspired metaheuristic algorithms: Luniver press, 2010.

[30] J. Cheng, D. Yu, Y. Yang, "Application of support vector regression machines to the processing of end effects of Hilbert–Huang transform," Mech.l Syst. Sig. Process., 21: 1197-1211, 2007.

[31] M. Sabzekar, S.M.H. Hasheminejad, "Robust regression using support vector regressions," Chaos, Solitons & Fractals, 144: 110738, 2021.

[32] H. Drucker, C.J. Burges, L. Kaufman, A. Smola, V. Vapnik, "Support vector regression machines," Adv. Neur. Inf. Process. Syst., 9: 155-161, 1997.

[33] I. Behravan, O. Dehghantanha, S. H. Zahiri, "An optimal SVM with feature selection using multi-objective PSO," in Proc. 2016 1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC): 76-81, 2016.

[34] B. Madhu, A.K. Paul, R. Roy, "Performance comparison of various kernels of support vector regression for predicting option price," Int. J. Discrete Math., 4: 21, 2019.

[35] Y. Tang, W. Guo, J. Gao, "Efficient model selection for support vector machine with Gaussian kernel function," in Proc. 2009 IEEE Symposium on Computational Intelligence and Data Mining: 40-45, 2009.

[36] W. Wang, Z. Xu, W. Lu, X. Zhang, "Determination of the spread parameter in the Gaussian kernel for classification and regression," Neurocomputing, 55: 643-663, 2003.

[37] S. Mirjalili, S. M. Mirjalili, A. Hatamlou, "Multi-verse optimizer: a nature-inspired algorithm for global optimization," Neur. Comput. Appl., 27: 495-513, 2016.

[38] E. Rashedi, H. Nezamabadi-Pour, S. Saryazdi, "GSA: a gravitational search algorithm," Inf. Sci., 179: 2232-2248, 2009.

[39] M.H. Mozaffari, H. Abdy, S.H. Zahiri, "IPO: an inclined planes system optimization algorithm," Comput. Inf., 35: 222-240, 2016.

[40] I. Behravan, S. Razavi, “Stock price prediction using machine learning and swarm intelligence,” J. Electr. Comput. Eng. Innovations, 8(1):31-40, 2020.

[41] R.K. Dash, T.N. Nguyen, K. Cengiz, A. Sharma, "Fine-tuned support vector regression model for stock predictions," Neur. Comp. Appl., 2021: 1-15, 2021.

Biographies

Iman Behravan received his B.Sc. in electronics engineering from Shahid Bahonar University of Kerman, Kerman, Iran. Also, he received his M.Sc. and Ph.D. degrees from the University of Birjand, Birjand, Iran. Now he is a post-doctoral researcher at the University of Birjand under the supervision of Professor Seyed Mohamad Razavi. His research interests include Big data analytics, pattern recognition, machine learning, swarm intelligence, and soft computing.

- Email: i.behravan@birjand.ac.ir
- ORCID ID: 0000-0003-0319-1765
- Web of science ID: J-5326-2017
- Scopus ID: 57190213695
- Home page: https://scholar.google.com/citations?user=w9GKiVcAAAAJ&hl=en

Seyyed Mohammad Razavi received the B.Sc. degree in Electrical Engineering from the Amirkabir University of Technology, Tehran, Iran, in 1994 and the M.Sc. degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 1996, and the Ph.D. degree in Electrical Engineering from the Tarbiat Modares University, Tehran, Iran, in 2006. Now, he is an Associate Professor in the Department of Electrical and Computer Engineering, the University of Birjand, Birjand, Iran. His research interests include Computer Vision, Pattern Recognition, and Artificial Intelligence Algorithm.

- Email: smrazavi@birjand.ac.ir
- ORCID ID: 0000-0002-3493-7614
- Web of science ID: AGS-8258-2022
- Scopus ID: 3861265
- Home page: https://scholar.google.com/citations?user=iSd4OusAAAAJ&hl=en

How to cite this paper:

I. Behravan, S.M. Razavi, "Application of Harris Hawks optimization algorithm and APSO-CLUSTERING in predicting the stock market," J. Electr. Comput. Eng. Innovations, 10(2): 447-462, 2022.

DOI: 10.22061/JECEI.2022.8659.541

URL: https://jecei.sru.ac.ir/article_1693.html

SRTTU

Journal of Electrical and Computer Engineering Innovations (JECEI)

Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

A Novel Analytical Approach for Time-response shaping of the PI controller in Field Oriented Control of the Permanent Magnet Synchronous Motors

H. Salimi, A. Zakipour, M. Asadi*

Department of Electrical Engineering, Arak University of Technology (AUT), Arak, Iran.

Article Info

Article History:
Received 23 December 2021
Reviewed 17 January 2022
Revised 11 April 2022
Accepted 11 April 2022

Keywords:
Permanent Magnet Synchronous Motor (PMSM)
Filed Oriented Control (FOC)
Controller tunning
Dynamic responses
Newton-Raphson numerical method

*Corresponding Author's Email Address: *zakipour@arakut.ac.ir*

Abstract

Background and Objectives: Permanent magnet synchronous motors (PMSM) have received much attention due to their high torque as well as low noise values. However, several PI blocks are needed for field, torque, and speed control of the PMSM which complicates controller design in the vector control approach. To cope with these issues, a novel analytical approach for time-response shaping of the Pi controller in the filed oriented control (FOC) of the PMSM is presented in this manuscript. In the proposed method, it is possible to design the controlling loops based on the pre-defined dynamic responses of the motor speed and currents in dq axis. It should be noted that as decoupled model of the motor is employed in the controller development, a closed loop system has a linear model and hence, designed PI controllers are able to stabilize the PMSM in a wide range of operation.

Methods: To design the controllers and choose PI gains, characteristic of the closed loop response is formulated analytically. According to pre-defined dynamic responses of the motor speed and currents in dq-axis e.g., desired maximum overshoot and rise-time values, gains of the controllers are calculated analytically. As extracted equation set of the controller tuning includes a nonlinear term, the Newton-Raphson numerical approach is employed for calculation of the nonlinear equation set. In addition, designed system is evaluated under different tests, such as step changes of the references. Finally, it should be noted that as the decoupled models are employed for the PMSM system, hence exact closed loop behavior of the closed loop system can be expressed via a linear model. As a result, stability of the proposed approach can be guaranteed in the whole operational range of the system.

Results: Controlling loops of the closed loop system are designed for speed control of the PMSM. To evaluate accuracy and effectiveness of the controllers, it has been simulated using MATLAB/Simulink software. Moreover, the TMS320F28335 digital signal processor (DSP) from Texas Instruments is used for experimental investigation of the controllers.

Conclusion: Considering the simulation and practical results, it is shown that the proposed analytical approach is able to select the controlling gains with negligible error. It has shown that the proposed approach for rise time and overshoot calculations has at most 0.01% for step response of the motor speed at 500 rpm.

Introduction

In recent years, permanent magnet synchronous motors (PMSM) have been used in a wide range of applications such as robotics, electric vehicles, aerospace, and

Doi: 10.22061/JECEI.2022.8709.545

medical equipment [1] due to higher torque and lower noise, weight, and power loss compared to the induction motors with identical power rating [2].

Two main control approaches for closed loop control of the PMSM have been proposed in the literature. First method is called scalar control and basically it can be implemented by keeping the voltage to frequency ratio constant (V/f=constant).

In [3], this technique is employed in the PMSM. Even though the implementing of the scalar controller is straightforward in this condition, however its main drawbacks are complexity of the simultaneous control of speed and torque, as well as dynamic response control. For these reasons, scalar control is not a preferred choice in most of the applications.

The second approach for closed loop control of the electrical motors is vector control. Basically, two general techniques on vector control of the PMSM are reported: direct torque control (DTC) and field-oriented control (FOC).

Application of the DTC method on PMSM has been reported in [4], [5]. Main advantages of the DTC are simplicity of the implementation and acceptable dynamic response of the motor torque. On the other hand, its main drawbacks are variable switching frequency and high torque ripple which restricts widespread application of the mentioned approach [6]-[8].

On the other hand, FOC is widely used in industrial applications for closed loop control of the DC to AC inverters. In this method, AC machine can be assumed as a separately excited DC machine [9]. In other word, torque and speed of the electrical motor can be controlled separately if FOC approach is employed. Main advantages of the FOC method such as possibility of torque control at low speeds, and fast dynamic response of the speed loop make it more attractive for closed loop control of the AC motors in the industrial applications [10].

The FOC control system is implemented based on the separation of motor model in dq-axes. Actually, if the mentioned separation be possible in a closed loop system, a variety of controllers can be employed for speed/torque control of an electrical motor. For example, in [11], [12] the sliding mode controller is used according to FOC scheme.

Although the mentioned controllers can overcome the parameter's uncertainty challenges, however in [11], [12] chattering phenomena is seen in the controllers output signal which deteriorates advantages of the sliding mode technique.

In addition, in practical applications, the controller requires a high-speed processor and monitoring circuits which can considerably increase implementation costs of the practical system. In [13], [14], model reference adaptive control system has been developed for closed loop vector control the PMSM. The controller's parameters are optimally adjusted by using the PMSM model on different operational points. Moreover, implementation of the adaptive controller is a time consuming and complicated task.

In [15], [16] hysteresis current controller is used for FOC of the electrical motor. In ideal conditions, this controller has a zero steady-state error, however its main drawback is variable switching of the converter which complicates practical implementation of the hysteresis approach. In [17]-[19], model predictive controller is used for closed-loop control of the PMSM. Using this controller and by selecting an appropriate cost function, various goals can be achieved such as optimal system dynamics, switching frequency adjustment, etc., however, requirement of an accurate system model, uncertainty in model parameters, and time-consuming online calculation limit widespread application of the mentioned approach.

On the other hand, application of the linear PI controllers in the FOC structure has been increased particularly in industry applications [20]-[22]. Simplicity of the implementation has increased widespread application of the linear controllers. Actually, if a linear controller be tuned properly, it can stabilize the closed loop system and reject disturbances satisfactorily in an operating point.

In the FOC structure, three controllers are needed; two controllers are employed in the inner loop for dq currents adjustment, and the third controller in the outer loop generates the reference current of the inner loops.

To implement the FOC approach more efficiently by using the PI controllers, the proportional and integral coefficients must be selected properly for both inner and outer loops. One of the popular methods for selection of the control gains in closed-loop control of the PMSM is offline application of the innovative algorithms e.g. genetic [23], [24], PSO [25]-[27], and fuzzy logic [28]-[30]. Despite advantages of the mentioned training algorithms, large volume of real data is needed in training phase of the mentioned techniques. In addition, they may stick on the local optimum points and in other word, there is no guarantee that the mentioned search algorithms generate the globally optimum point and the best answer. In [31], [32] adaptive PI has been used for online calculation of the controller's coefficients. This method can be useful for systems in a wide range of operation. As mentioned, adaptive approaches are not an ideal choice for motor drive due to practical implementation issues as time-consuming on-line calculations are needed.

Regarding the PI controller tunning, various performance measures such as steady-state error, integral of absolute error, integral of square error, integral of time square error, and integral of time absolute error is defined [22], [33], [34]. Actually, the controller is designed so that to minimize one/some of the mentioned measures.

In this condition, if different controllers are employed for selection of the PI gains, it is possible to compare them in terms of the performance measure values. Hence, a superior approach will result in the minimum index regarding response rise time, settling time, and maximum overshoot.

However, in this manuscript, it should be noted that minimization of the mentioned indexes isn't employed for controller tuning.

Actually, a novel approach for time-response shaping of the closed loop system is developed in FOC of the PMSM. Regarding the pre-defined values of the overshoot and rise-time in the step response of the closed loop system, the PI gains are selected according to the numerical analysis of plant exact model. Clearly, if different values are selected as a design input (overshoot and rise-time) in the proposed method of this manuscript, different PI gains which demonstrate different performance indexes will be obtained.

In the meantime, to ensure the overall functionality of the controller, the design inputs (rise time and overshoot of the response) should be selected properly. For example, in this manuscript, desired overshoot and rise time of the speed controller are selected as 10% and 150ms and PI controller gains are calculated based on these values. Regarding the controller stability analysis which is added to the revised paper (please referrer the question 5), it is shown that the employed PI controller is stable despite large changes of the model parameters and uncertainties.

Briefly, this manuscript focuses on the time-response shaping and selection of the PI gains in FOC structure of the PMSM.

Gain selection algorithm is based on stability and robustness of the speed control loop. Due to application of the PI controller, elimination of the steady-state error and removal of the parameter's uncertainty issues, as well as input disturbance rejection will also be possible in the designed controllers.

Briefly, structure of the manuscript is as follows: In section II model of the PMSM is reviewed.

The PI controllers are designed for the inner and outer loops. Then selection of the controlling gains is explained in the third section.

Considering the XML-SE09MEKE PMSM, the performance of the proposed analytical approach for controller tuning is evaluated using a case study example in the next section. Finally, simulation and practical responses of the PMSM in different scenarios is evaluated in section V and conclusions are given in section VI.

Decoupled Model of the PMSM in FOC Approach

The mathematical model of PMSM in the *dq0* reference frame is expressed by the following differential equations, where the iron saturation, magnetic flux leakage, eddy current as well as hysteresis losses are assumed to be negligible [35]:

$$\frac{di_d(t)}{dt} = \frac{1}{L_d}\Big(v_d(t) - R_s i_d(t) + \omega_e(t) L_q i_q(t)\Big) \tag{1}$$

$$\frac{di_q(t)}{dt} = \frac{1}{L_q}\big(v_q(t) - R_s i_q(t) + \omega_e(t) L_d i_d(t) - \omega_e(t)\psi_r\big) \tag{2}$$

$$\frac{d\omega_e(t)}{dt} = \frac{P}{J_m}\left(T_e - \frac{B_v}{P}\omega_e(t) - T_L\right) \tag{3}$$

$$T_e = \frac{3}{2} P \psi_r i_q \tag{4}$$

where ω_e is the electrical speed of the motor, which depends on the rotor speed based on $\omega_e = P\omega_r$ and the parameter P refers number of motor poles pairs. Moreover, v_d and v_q represent the stator voltage and i_d and i_q are the motor current in the *dq* frame. It should be mentioned that T_L and T_e are the load and electromagnetic torques. Also, the parameters of the PMSM model include *dq*-axis inductances (L_d, L_q), field flux (ψ_r), stator resistance (R_s), motor inertia (J_m), and viscous coefficient (B_v).

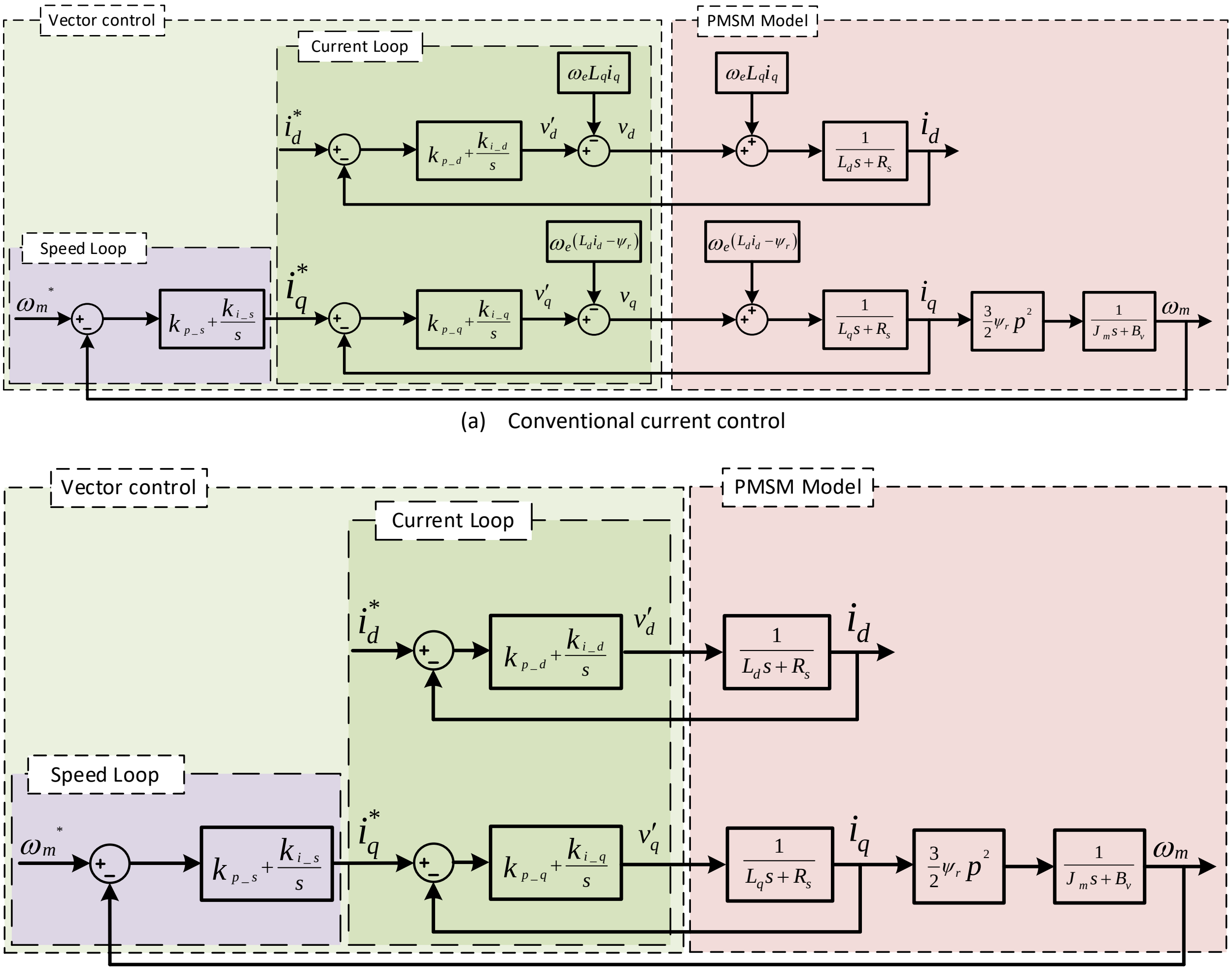
Fig. 1(a) shows PMSM closed-loop speed control block diagram in which the PI blocks are responsible for control of the *dq*-axis currents in the inner loops and the speed controller of the outer loop. According to (1) and (2), it is seen that there is a nonlinear coherence between model state-variables on the *dq*-axis.

In other word, all the variables are dependent to each other and as a result, controlling loops cannot be employed directly which complicates closed loop system design.

To overcome this problem, it is well-known that the decoupling variables can be introduced as follows:

$$\frac{1}{L_d}\hat{v}_d = \frac{1}{L_d}\left(v_d + \omega_e L_q i_q\right) \tag{5}$$

$$\frac{1}{L_q}\hat{v}_q = \frac{1}{L_q}\left(v_q - \omega_e L_d i_d - \omega_e \psi_r\right) \tag{6}$$



(a) Conventional current control

(b) Simplified decoupled FOC controller

Fig. 1: Closed-loop control of the speed controller in PMSM.

By replacing (5) and (6) in (1) and (2), the electrical equations of the motor can be rewritten as:

$$\frac{di_d}{dt} = -\frac{R_s}{L_d} i_d + \frac{1}{L_d}\hat{v}_d \tag{7}$$

$$\frac{di_q}{dt} = -\frac{R_s}{L_q} i_q + \frac{1}{L_q}\hat{v}_q \tag{8}$$

According to Fig. 1(b) and based on (7) and (8), two separate feedback control loops should be designed by adjustment of the $\hat{v}_d$ and $\hat{v}_q$to achieve control target.

Analytical Approach for Controller's Tunning

Main objective of this manuscript is selection of the PI gains in FOC of the PMSM.

To address this issue, closed loop controller is implemented through different blocks e.g., inner and outer loops which is analyzed below.

A. Current Loop Design

To design the inner loops for *dq*-axes, considering proportional and integral gains as k_{p_d}, k_{i_d} and k_{p_q}, k_{i_q}, the PI controller of the *dq*-axis current can be written as*:*

$$\begin{aligned} v_d(t) = {} & k_{p\_d}\big(i_d{}^*(t) - i_d(t)\big) \\ & + k_{i\_d}\int_0^t \big(i_d{}^*(\tau) - i_d(\tau)\big)d\tau \\ & - \omega_e(t) L_q i_q(t) \end{aligned} \tag{9}$$

$$\begin{aligned} v_q(t) = {} & k_{p\_q}\left(i_q{}^*(t) - i_q(t)\right) \\ & + k_{i\_q}\int_0^t \left(i_q{}^*(\tau) - i_q(\tau)\right)d\tau \\ & + \omega_e(t) L_d i_d(t) - \omega_e(t)\psi_r \end{aligned} \tag{10}$$

By defining the virtual controller and according to Fig. 1(b) which shows simplified block diagram of the control system, coupled mutual sentences of the model can be removed and hence, this technique facilitates controller design for PMSM. According to Fig. 1(b), to design the current PI controller, electrical term of the motor transfer functions in the frequency domain can be written as:

$$\frac{i_d(s)}{\hat{v}_d(s)} = \frac{\frac{1}{R_s}}{\frac{L_d}{R_s}s + 1} \tag{11}$$

$$\frac{i_q(s)}{\hat{v}_q(s)} = \frac{\frac{1}{R_s}}{\frac{L_q}{R_s}s + 1} \tag{12}$$

where $\hat{v}_d(s)$ and $\hat{v}_q(s)$ are auxiliary variables of the *dq*-axis voltages.

Considering the block diagram of the decoupled controller in Fig. 1(b), closed loop transfer function of the system can be calculated easily as:

$$G_{CL_C} = \frac{K_c(k_{i\_q} + k_{p\_q}s)}{\tau_C s^2 + (K_c k_{p\_q} + 1)s + K_c k_{i\_q}} \tag{13}$$

where $K_c = \frac{1}{R_s}$, $\tau_C = \frac{L_q}{R_s}$.

From (13), step response of close loop system in time domain is:

$$s_{CL\_C}(t) = \frac{(K_c k_{p\_q} - 1)\sinh(Wt)\, e^{(Yt)}}{2\tau_C W} - e^{(Yt)}\cosh(Wt) + 1 \tag{14}$$

where:

$$W = \frac{\sqrt{K_c^2 k_{p\_q}^2 + 2K_c k_{p\_q} - 4T_{sC} k_{i-_q} K_c + 1}}{2\tau_C} \tag{15}$$

$$Y = -\frac{(K_c k_{p\_q} + 1)}{2\tau_C} \tag{16}$$

According to (14), rise time can be calculated as:

$$t_{rise} = \frac{2\tau_C \log\left(-\frac{2\sqrt{\left(-K_c(k_{p\_q} - \tau_C k_{i\_q})\right)}}{H - K_c k_{p\_q} + 1}\right)}{H} \tag{17}$$

and

$$H = \sqrt{K_c^2 k_{p\_q}^2 + 2K_c k_{p\_q} - 4\tau_C k_{i-_q} K_c + 1} \tag{18}$$

It should be noted that the (14) expresses the step response of the closed loop system in time domain. According to time-derivative of the (14) and setting it into zero, time of the maximum point is obtained in (19).

This equation is calculated and simplified using the 'MATLAB symbolic analysis toolbox'. By replacing t_{MP} from (19) in (14), maximum value of the overshoot can be written as (20), where:

$$M = \sqrt{-\tau_C k_{i\_q}(k_{p\_q} - \tau_C k_{i\_q})} \tag{21}$$

$$N = k_{p\_q} - 2\tau_C k_{i-q} + K_c k_{p\_q}^2 - k_{p\_q}Q \tag{22}$$

$$Q = \sqrt{K_c^2 k_{p\_q}^2 + 2K_c k_{p\_q} - 4\tau_C K_c k_{i-q} + 1} \tag{23}$$

To calculate the controller coefficients, desired values of system rise time and maximum overshoot should be replaced in the (17) and (20).

As these equations are nonlinear, so it may be challenging task to introduce an analytical solution using the classical methods. On the other hand, numerical approaches e.g., the Newton-Raphson method can be employed to cope with the mentioned issues. The employed block diagram of the Newton-Raphson method is illustrated in Fig. 2.

According to Fig. 1(b), it is observed the current loop for *d*-axes is similar to *q*-axes. As a result, *d*-axes controller can be designed using the same equations if L_q is replaced with L_d.

$$t_{MP} = \frac{\tau_c * \log\left(-\frac{2*\left(-k_{i\_q}*\tau_c*\left(k_{p\_q} - k_{i\_q}*\tau_c\right)\right)^{\frac{1}{2}}}{k_{p\_q} - 2*k_{i\_q}*\tau_c + K_c*{k_{p\_q}}^2 - 2*k_{p\_q}*\left(\frac{K_c^2*{k_{p\_q}}^2}{4} + \frac{K_c*k_{p\_q}}{2} - k_{i\_q}*\tau_c*K_c + \right.}\right.}{\left(\frac{K_c^2*{k_{p\_q}}^2}{4} + \frac{K_c*k_{p\_q}}{2} - k_{i\_q}*\tau_c*K_c + \frac{1}{4}\right)^{\frac{1}{2}}} \tag{19}$$

$$MP = \frac{\frac{1}{2^{\frac{K_c k_{p-q}+1}{Q}}}\left(\tau_C k_{i-q} - k_{p-q} + 2^{\frac{K_c k_{p-q}+1}{Q}}\left(\frac{M}{N}\right)^{\frac{K_c k_{p-q}+1}{Q}} M\right)}{\left(\frac{M}{N}\right)^{\frac{K_c k_{p-q}+1}{Q}} M} - 1 \tag{20}$$

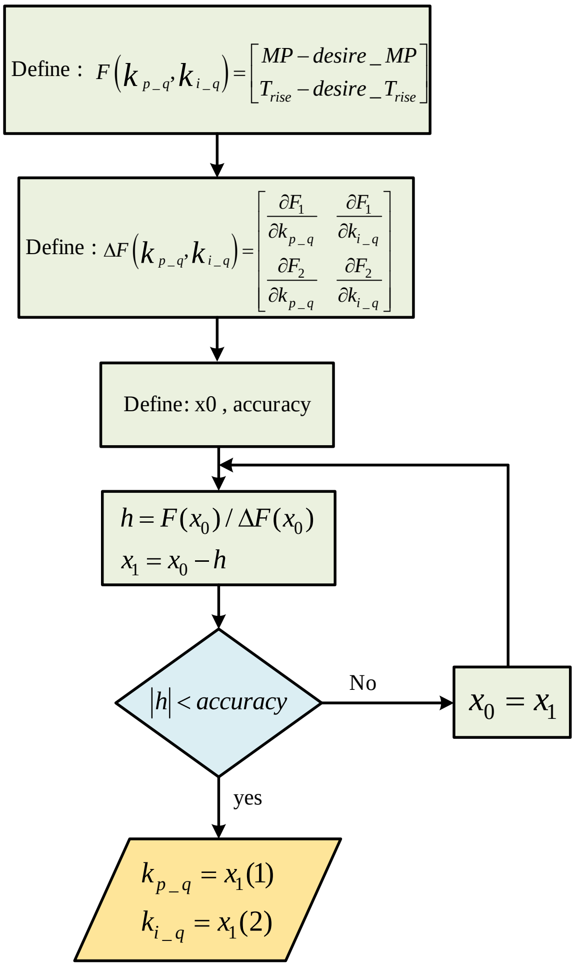

Define : $F(k_{p\_q}, k_{i\_q}) = \begin{bmatrix} MP - desire\_MP \\ T_{rise} - desire\_T_{rise} \end{bmatrix}$
Define : $\Delta F(k_{p\_q}, k_{i\_q}) = \begin{bmatrix} \frac{\partial F_1}{\partial k_{p\_q}} & \frac{\partial F_1}{\partial k_{i\_q}} \\ \frac{\partial F_2}{\partial k_{p\_q}} & \frac{\partial F_2}{\partial k_{i\_q}} \end{bmatrix}$
Define: x0 , accuracy
$h = F(x_0) / \Delta F(x_0)$
$x_1 = x_0 - h$
$|h| < accuracy$
No
$x_0 = x_1$
yes
$k_{p\_q} = x_1(1)$
$k_{i\_q} = x_1(2)$

Fig. 2: Block diagram of the Newton-Raphson numerical method for calculation of the q-axes controlling gains.

B. Speed-Loop Design

Outer-loop of the introduced closed-loop system can be designed using the (3) and (4). These equations are used to describe mechanical behavior of the model as well as coherence between mechanical and electrical equations. By substituting (4) in (3), mechanical dynamic behavior of the PMSM can be described as:

$$\frac{d\omega_e(t)}{dt} = \frac{3}{2}\frac{P^2\psi_r}{J_m} i_q(t) + \frac{3}{2}\frac{P^2}{J_m}(L_d - L_q) i_d(t) i_q(t) - \frac{B_v}{J_m}\omega_e(t) - \frac{P}{J_m}T_L \quad (24)$$

The equation includes $(L_d - L_q)$ term. In non-salient pole PMSM, it is well known that $L_d = L_q$.

However, in salient motors where $L_d \neq L_q$, then $i_d(t)$ should be set to zero in the control system.

So, in both conditions, the second term of the equation $(\frac{3}{2}\frac{P^2}{J_m}(L_d - L_q) i_d(t) i_q(t))$ will be zero.

Also, $\frac{B_v}{J_m}\omega_e(t) - \frac{P}{J_m}T_L$ term in the (24) is directly related to the load torque on the motor shaft. This parameter can be considered as a disturbance during the controller operation.

Hence, if an integrator is employed in the outer speed controller, steady-state error can be eliminated. In this condition, (24) can be rewritten as in frequency domain:

$$\left(s + \frac{B_v}{J_m}\right)\omega_e(s) = \frac{P^2\psi_r}{J_m} i_q(s) \quad (25)$$

According to Fig. 1(b), by substituting (15) into (25), speed transfer function to the reference current signal of the *q*-axis can be written as:

$$\frac{\omega_e(s)}{i_q^*(s)} = \left(\frac{\frac{3}{2}\frac{P^2\psi_r}{J_m}}{s + \frac{B_v}{J_m}}\right) G_{CL\_C} \quad (26)$$

In (26), it is seen that $s = -\frac{B_v}{J_m}$ is related to mechanical behavior of the system and hence it can be assumed as a dominant pole of the closed loop system. So, for outer speed loop, closed loop transfer function of the system can be written as follow:

$$G_{CL_S} = \frac{K_M(k_{i\_s} + k_{p\_s}s)}{\tau_M s^2 + (K_M k_{p\_s} + 1)s + K_M k_{i\_s}} \quad (27)$$

where:

$$K_M = \frac{3}{2}\frac{P^2\psi_r}{J_m} \quad (28)$$

$$\tau_M = \frac{J_m}{B_v} \quad (29)$$

As transfer function of the current loop in (13) is similar to (27), hence, similar approach which is presented for numerical solution of the inner loop in Fig. 2, can be employed for outer speed controller design based on the desired rise time and maximum overshoot values.

Case Study for XML-SE09MEKE PMSM

In this section, developed approach is employed for XML-SE09MEKE PMSM from LS Electric Company. The performance of the proposed analytical approach for controller tuning is evaluated using a case study example.

The nominal parameters of the tested closed loop system including motor and inverter parameters are listed in Table 1.

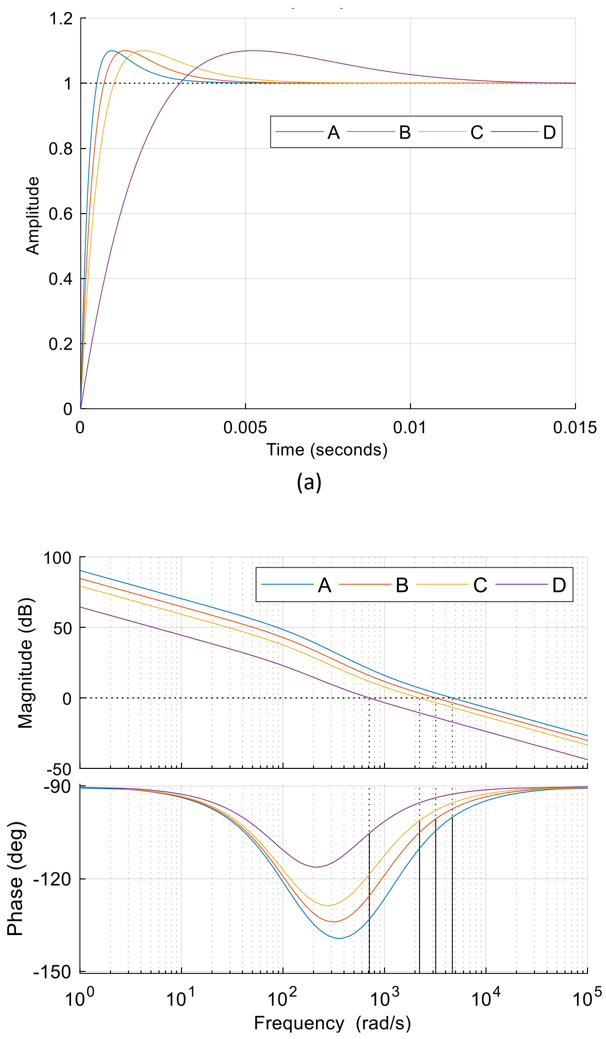
A. Case-1- Selection of the controller gains for a constant $\boldsymbol{MP}$ and several different $\boldsymbol{t_r}$ values in the inner current loop

Assuming a fixed overshoot in the controller's response, step response and bode diagram of the closed loop system is shown in Fig. 3 for several rise time values. Obtained controller gain for each rise time value is listed in Table 2.

It is observed that the controller gain tuning algorithm has an acceptable error. Maximum value of the error in this test is less than 7.5%. Also, according to gain and phase margins of the inner-loop in Table 2 as well as in Fig. 3, it can be concluded that the control system is able to stabilize the closed-loop plant in the mentioned operating point.

Table 1: Inverter and XML-SE09MEKE PMSM parameters

Parameter	Symbol	**Unit**	Value
Rated Power	P_n	W	900
Rated Torque	T_{en}	N.m.	8.59
Rated Speed	w_n	RPM	1000
Voltage DC Link	Vdc	V	310
Pole pairs	P	-	4
Stator Resistance	R_s	Ω	1
Stator inductance	L_s	mH	7.5
Electrical Back EMF Constant	ψ_m	wb	0.3108
Rotor Inertia	J_m	$kg.m^2$	20.6e-4
viscous coefficient	B_v	$kg.m$^2 /s	0.0001
Maximum Current	I_{max}	A	5.2
Sampling Time	T_s	μs	25
Switching Frequency	f_{sw}	KHz	40
IGBT Driver	UCC2154a Texas Instrument		
IGBT	40N120FL2		



(a)

(b)

Fig. 3: Step response (a) and bode plot (b) of the inner loop for several different rise time values using a constant overshoot.

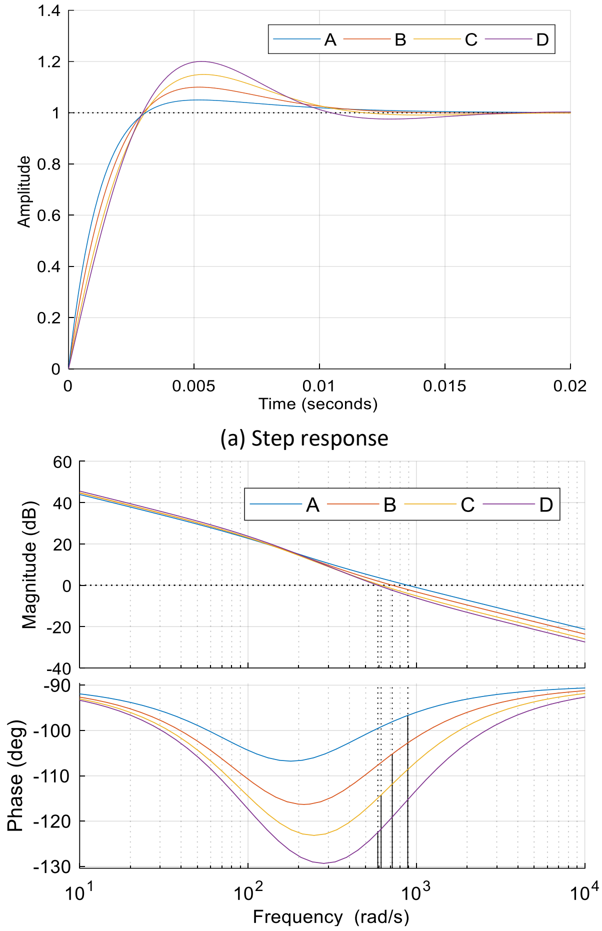
Table 2: Selected controller gains for several different rise time values and a constant overshoot

		Rise time (ms)*			Margins Value**		Current loop PI parameters	
No.	MP (%)	t_r^d	t_r^m	Err. (%)	GM (dB)	PM (deg)	k_{i_c}	k_{p_c}
A	10	0.5	0.48	4	Inf	79.91	33209	34.8
B	9.998	0.8	0.74	7.5	Inf	79.47	17287	23.4
C	10	1	0.95	5	Inf	78.83	9337	16.14
D	9.98	3	3.1	3.33	Inf	74.76	1682	4.88

*t_r^d:desired rise time t_r^m: measured rise time
** GM: Gain Margin PM: Phase Margin

B. Case-2- Selection of the controller gains for a constant t_r and several different MP values in the inner current loop

In accordance with the previous test in case-1, controller gains are tuned based on a constant rise-time value. Step response and bode diagram of the closed loop system in case-2 is shown in Fig. 4 for different MP values. Obtained controller gain for each condition is listed in Table 3. Maximum error value in this case is less than 0.2%.



(a) Step response

(b)Bode diagram

Fig. 4: Response of the inner-current loop in case-2.

Table 3: Selected controller gains in case-2

No.	t_r (ms)	overshoot (%)*			Margins Value**		Current loop PI parameters	
		MP^d	MP^m	Err. (%)	GM (dB)	PM (deg)	k_{i_c}	k_{p_c}
A	3.06	5	4.99	0.2	Inf	83.32	1569	6.53
B	2.98	10	9.99	0.1	Inf	74.82	1707	4.94
C	3.07	15	14.99	0.06	Inf	65.69	1733	3.81
D	2.95	20	19.99	0.05	Inf	57.55	1897	3.19

*MP^d:desired overshoot MP^m: measured overshoot
** GM: Gain Margin PM: Phase Margin

In order to study the outer speed loop, step response of the current loop for selected controller gains is shown in Fig. 5.

In this condition, desire values of maximum overshoot and rise time are selected equal to 10% and 3 milliseconds, respectively.

As a result, according to the proposed tuning algorithm, 9.99% overshoot and 2.9 milliseconds rise-time are achieved respectively.

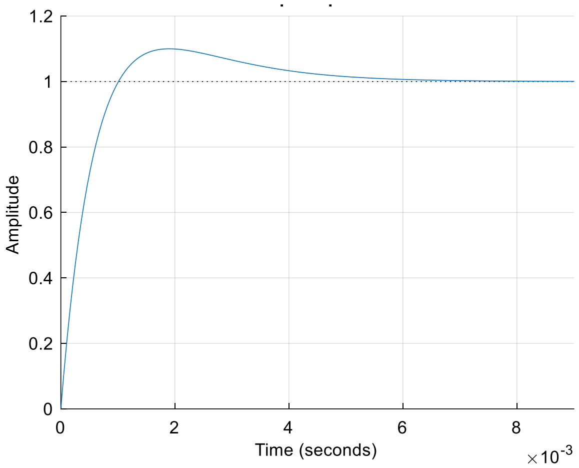


Fig.5: step response of the inner loop based on the selected controller gains

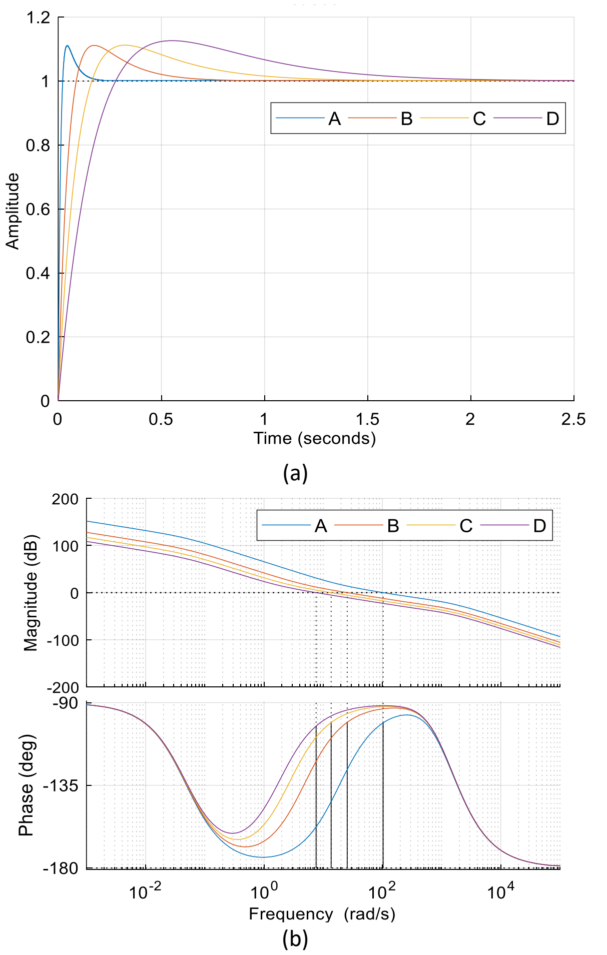
C. Case-3- Selection of the controller gains for a constant MP and several different t_r values in the outer speed loop

Fig. 6 shows step response and bode diagram of designed controller for a constant overshoot (10%) and variable rise time. Results of this test is shown in Table 4. According to the Table 4, the maximum value of the desired rise time is close to the real values. Maximum error in this test is less than 2.3%. Also, values of gain and phase margins in Table 4 and Fig. 6 illustrate stability of control system in these conditions.

Table 4: Controller parameters in case-3

No.	MP (%)	Rise time (ms)*			Margins Value**		Current loop PI parameters	
		t_r^d	t_r^m	Err. (%)	GM (dB)	PM (deg)	k_{i_s}	k_{p_s}
A	10.02	20	19.78	1.1	25	79.9	0.536	0.0286
B	10.01	80	79.81	0.2	37.4	80.5	0.034	0.0071
C	9.99	150	150.1	0.1	43	80.6	0.01	0.0038
D	10.5	250	255.8	2.3	47.9	80.5	0.003	0.0021

*t_r^d:desired rise time t_r^m: measured rise time
** GM: Gain Margin PM: Phase Margin



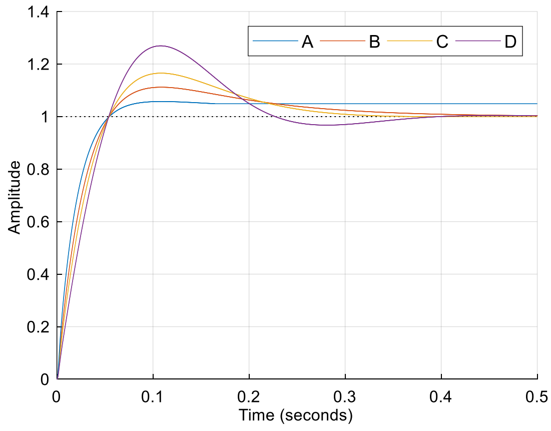
(a)

(b)

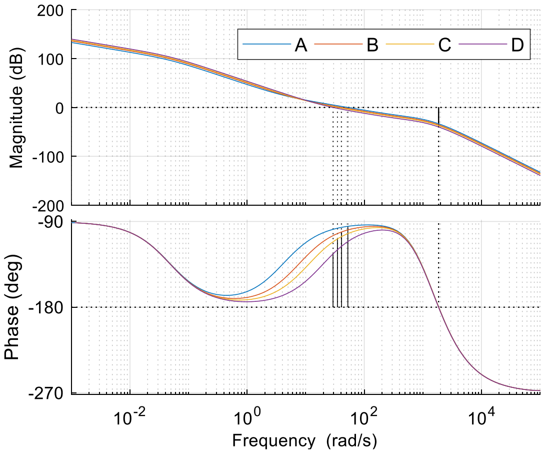
Fig. 6: Step response (a) and bode diagram (b) of the outer loop in case-3.

D. *Case-4- Selection of the controller gains for a constant t_r and several different MP values in the outer speed loop*

Similar to the previous test in case-3, controller gains are tuned based on a constant rise-time (50ms) value in case-4 and step response and bode diagram of the closed loop system are shown in Fig. 7 for different MP values. Obtained controller gain for each condition is listed in Table 5. Maximum error value in this case is less than 0.45%.



(a)



(b)

Fig. 7: step response (a) and bode diagram (b) of the outer loop in case 4.

Table 5: controller parameters in case-4

No.	t_r (ms)	overshoot (%)*			Margins Value**		Current loop PI parameters	
		MP^d	MP^m	Err. (%)	GM (dB)	PM (deg)	k_{i_s}	k_{p_s}
A	49.7	5	5	0	31	85.7	0.0609	0.0148
B	49.7	10	10.02	0.2	33.2	80.4	0.0869	0.0114
C	49.8	15	15.05	0.4	34.8	73.8	0.1061	0.0094
D	49.8	25	25.11	0.44	37.4	58.6	0.1352	0.0069

*MP^d:desired overshoot MP^m: measured overshoot
** GM: Gain Margin PM: Phase Margin

Results and Discussion

In this section, validation of the vector control approach is studied based on the selected controller gains under different scenarios. Experiments are conducted out using the XML-SE09MEKE three-phase PMSM from *LS Electric*.

A. *Simulation Result*

In this study, currents of the *dq*-axes, control signal and speed of the machine are shown in Fig. 8 based on simulation of the designed closed loop system in MATLAB/Simulink software.

Desired values of maximum overshoot and rise time are selected as 15% and 100ms respectively. From the simulation results in Fig. 8, it is seen that these variables are equal to 16% and 98ms which are properly compatible.

Furthermore, it is shown that the control signal ($\mathbf{U}_{\text{abc}}$) is quite stable.

Also, currents of the *dq*-axes can follow their reference values with an acceptable transient response and zero steady-state error.

Moreover, in the Fig. 8, the three-phase motor input voltages ($\mathbf{V}_{\mathrm{Line-abc}}$) which are supplied through the inverter is illustrated.

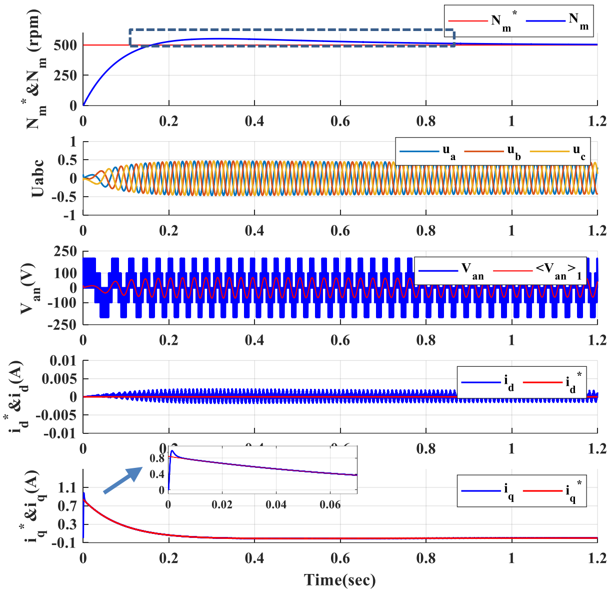


Fig. 8: Simulation of the designed closed loop system in MATLAB/Simulink.

In this paper, a proportional-integral linear controller is designed for FOC control of the PMSG.

It is well known that integral term can suppress the steady-state error of the closed loop control system despite uncertainty of the model parameters which has been demonstrated in the following simulation results. Also, regarding the robustness analysis, it can be considered as the stability of the closed-loop controller in case of uncertainties.

Actually, model of the plant, which is employed for controller design, may include some differences respect to the practical system due to inherent tolerance of the parameters, un-modelled dynamics, and nonlinearities. Hence, robustness analysis is a vital issue during the controller design.

Time-response of the designed controllers for step changes of the motor intertie, friction (viscous) coefficient, load torque, stator series resistance and inductance are shown in Fig. 9.

All the motor parameters are stepped up from the normal operating point to the two-times of nominal values.

Despite changes of the system operating point in a wide range, it can be concluded that the designed controller is stable and robust with respect to parameters changes.

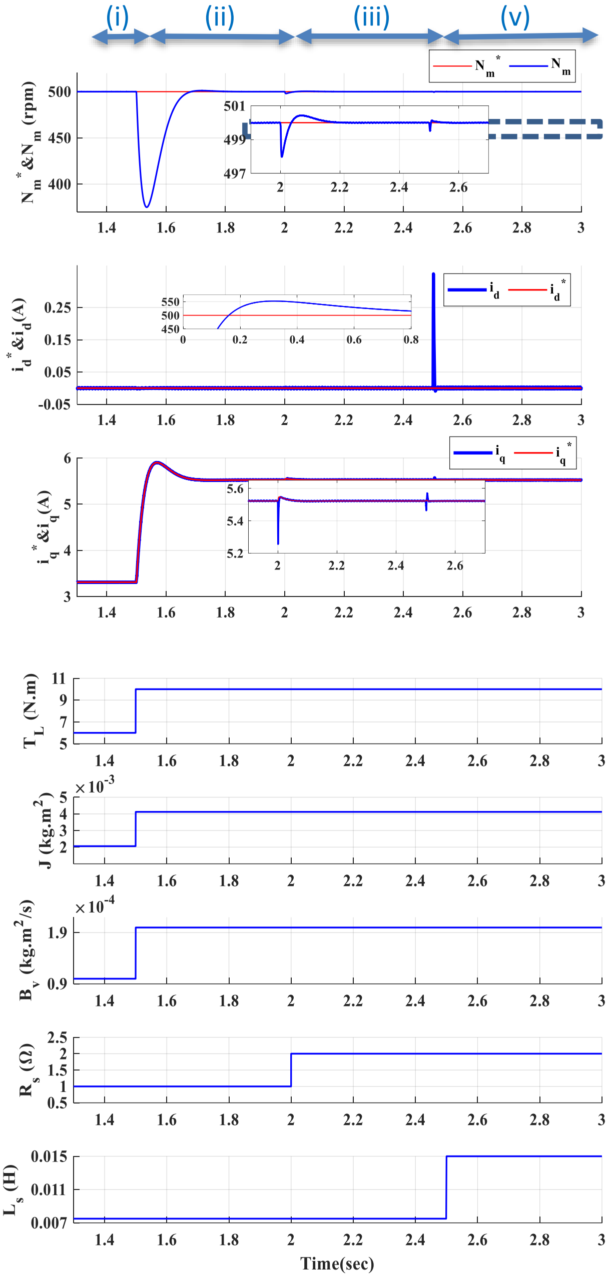


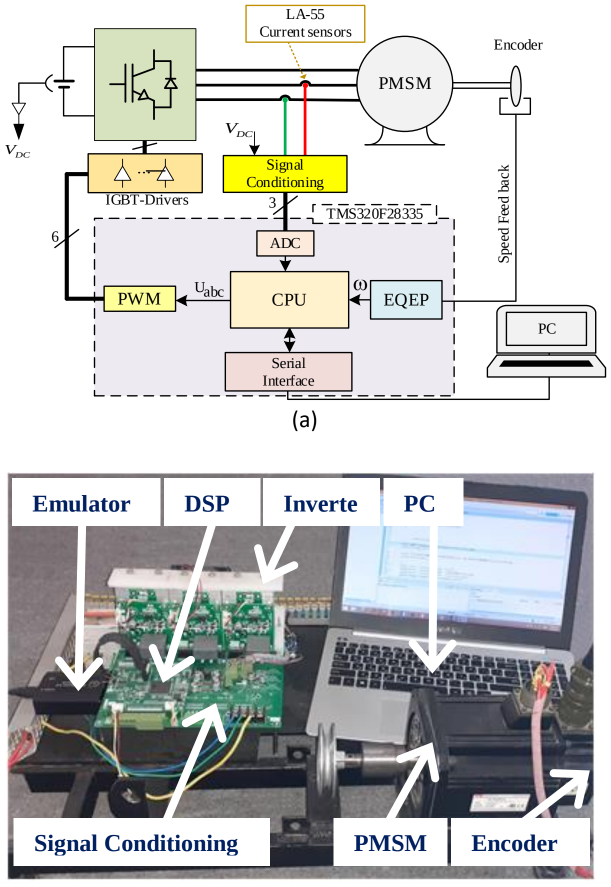
Fig. 9: Time response of the designed controllers for step changes of the motor intertie, load torque, stator series resistance and inductance.

B. *Experimental results*

The experimental test bench which is employed for evaluation of the designed controller based on Table 1, is shown in Fig. 10.

The proposed FOC algorithm is implemented using the TMS320F28335 DSP board from *Texas Instruments*. A 3000 pulse per round embedded encoder is adopted for measurement of the rotor position and speed feedback signals.

In order to plot the measured waveforms, a serial interface is employed for data transfer between DSP and computer.



(a)

(b)

Fig. 10: Block diagram of the experimental setup (a) and photo of the test bench (b).

Test 1- Experimental response of the closed loop system for different gains

According to different design conditions in Table. 6, experimental step response of the developed closed loop system for outer speed loop is shown in Fig. 11. Accuracy of the developed algorithm is evaluated in three different cases (A, B and C).

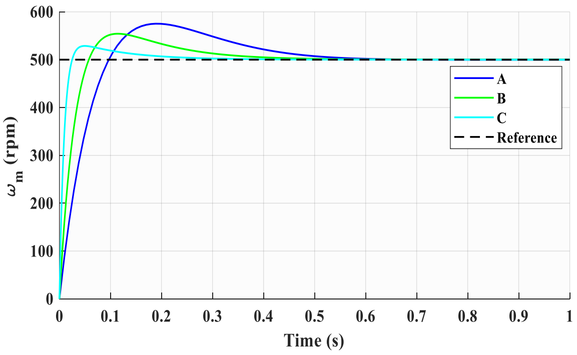


Fig. 11: Experimental response of the outer-control loop for different design conditions

Test 2- Experimental start-up response

In Fig. 12, experimental response of the designed closed loop system is illustrated during start-up and steady-state conditions. In this test, desired values of the maximum overshoot and rise time are assumed as 15% and 100ms respectively. So, as it is written in Table. 6, controller gains will be equal to $k_{p\_s} = 0.0346$ and $k_{i\_s} = 0.2163$ for the speed loop. According to the experimental response, it is seen that measured values for the mentioned parameters are 14% and 97ms, respectively which is compatible with design criteria. Also, controller has a negligible error during the steady-state condition.

Table 6: Summary of the experimental results in test-1

No.	overshoot (%)*			Rise time (ms)**			Current loop PI parameters	
	MP^d	MP^m	Err. (%)	t_r^d	t_r^m	Err. (%)	k_{i_s}	k_{p_s}
A	15	15.07	0.47	100	96	4	0.0346	0.2163
B	10	10.85	8.5	50	57	14	0.0687	0.5092
C	5	5.74	14.8	25	24.7	1.2	0.2048	1.9086

*MP^d:desired overshoot MP^m: measured overshoot
**t_r^d:desired rise time t_r^m: measured rise time

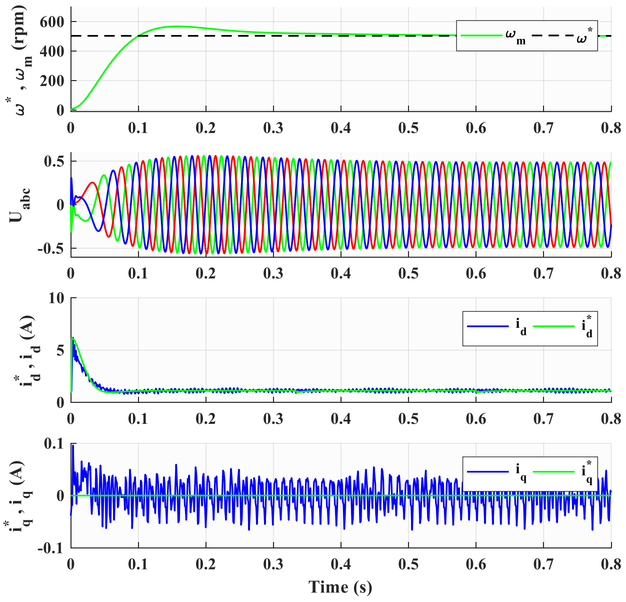


Fig. 12: Experimental response of the controller during start-up.

Test 3- Dynamic response of the controllers

To verify stability of the proposed system in different operating points, speed reference of the closed-loop control system is stepped between 350 and 500 rpm in test-3.

In spite of step changes of speed reference in a wide range in Fig. 13, it is seen that both inner and outer loops are stable during transients with zero steady-state error. In this test, similar gains are used for both inner and outer loops of the controller.

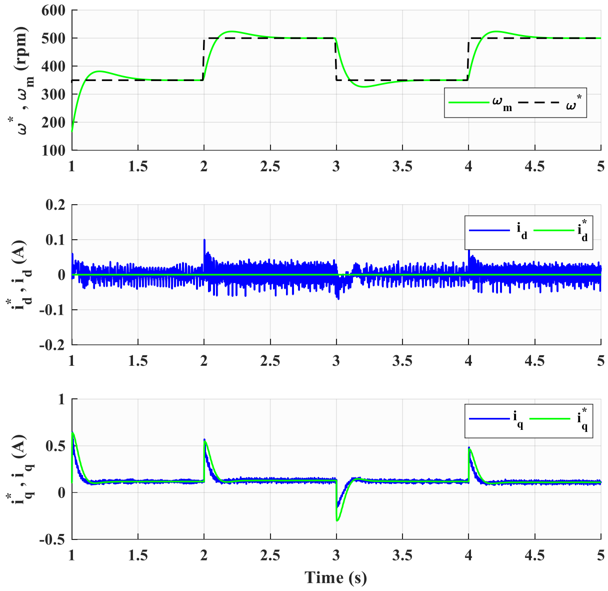


Fig. 13: Experimental dynamic response of the designed closed loop system.

Conclusion

In this manuscript, a novel method for tuning of the PI gains in FOC structure of the PMSM drivers is presented. The proposed method enjoys fast calculation time, so it can be employed in wide ranges of application *e.g*. CNC machine and electric vehicles.

In this method, the rise time and maximum overshoot values can be employed separately for tuning of the controller. Designed technique employs the Newton Raphson method to solve the nonlinear equations of the model.

The stability of the proposed method has also been evaluated by using the bode plot analysis. According to simulation and experimental results in different operational conditions, the proposed inner-current and outer-speed loops have stable and robust responses with zero steady-state error.

Author Contributions

This paper is the result of H. Salimi's M.Sc. thesis supervised by A. Zakipour and advised by A. Asadi. All of the authors have the same contribution in different parts of the manuscript including system modeling, controller design, simulation and experimental.

Acknowledgment

The author gratefully acknowledges the Power Quality Lab of Arak University of Technology.

Conflict of Interest

The author declares that there is no conflict of interests 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

PMSM	Permanent Magnet Synchronous Motor
PWM	Pulse Width Modulation
PI	Proportional Integral
FOC	Field Oriented Control

References

[1] S. Khodakaramzadeh, M. Ayati, M. Haeri Yazdi, "Fault diagnosis of a permanent magnet synchronous generator wind turbine," J. Electr. Comput. Eng. Innovations, 9(2): 143-152, 2021.

[2] M.E. Gerlach, M. Zajonc, B. Ponick, "Mechanical stress and deformation in the rotors of a high-speed PMSM and IM," e & i Elektrotech. Informationstech., 138: 96-109, 2021.

[3] W.J. Kim, S.H. Kim, "MTPA operation scheme with current feedback in V/f control for PMSM drives," Journal Power Electron., 20: 524-537, 2020.

[4] H. Mesloub, R. Boumaaraf, M.T. Benchouia, A. Goléa, N. Goléa, K. Srairi, "Comparative study of conventional DTC and DTC_SVM based control of PMSM motor—Simulation and experimental results," Math. Comput. Simul., 167: 296-307, 2020.

[5] S.J. Kim, J. Park, D.H. Lee, "Zero voltage vector-based predictive direct torque control for PMSM," in Proc. 2019 IEEE Student Conference on Electric Machines and Systems (SCEMS 2019): 1-6, 2019.

[6] D. Ahmed, B. Mokhtar, B. Aek, "DTC hybrid control by different methods of observation with artificial intelligence for induction machine drives," Int. J. Power Syst., 10(4), 2019.

[7] S. Khateri-Abri, F.Y. Notash, S. Tohidi, "A reduced-switch 3-level vsi based direct torque control of PMSM," in Proc. 2019 27th Iranian Conference on Electrical Engineering (ICEE): 565-569, 2019.

[8] M. Alizadeh Pahlavani, H. Damroodi, "LPV Control for speed of permanent magnet synchronous motor (PMSM) with PWM Inverter," J. Electr. Comput. Eng. Innovations, 4: 185-193, 2016.

[9] A. Kushwaha, M. Gopal, "Reinforcement learning-based controller for field-oriented control of induction machine," Soft Computing for Problem Solving, ed: Springer, 737-749, 2019.

[10] R. Marouane, Z. Malika, "Particle swarm optimization for tuning PI controller in FOC chain of induction motors," in Proc. 2018 4th International Conference on Optimization and Applications (ICOA): 1-5, 2018.

[11] X. Wang, M. Reitz, E.E. Yaz, "Field oriented sliding mode control of surface-mounted permanent magnet AC motors: Theory and applications to electrified vehicles," IEEE Trans. Veh. Technol., 67: 10343-10356, 2018.

[12] A. Zakipour, N. Ghaffari, M. Salimi, "State space modeling and sliding mode current control of the grid connected multi-level flying capacitor inverters," J. Electr. Comput. Eng. Innovations, 9(2): 215-228, 2021.

[13] A.T. Nguyen, M.S. Rafaq, H.H. Choi, J.W. Jung, "A model reference adaptive control based speed controller for a surface-mounted permanent magnet synchronous motor drive," IEEE Trans. Ind. Electron., 65: 9399-9409, 2018.

[14] A. Chouya, M. Chenafa, A. Mansouri, "Adaptive field-oriented control with MRAC regulator for the permanent magnet synchronous motor," Int. J. Control Syst. Rob., 4: 52-57, 2019.

[15] J. Zhang, H. Yang, T. Wang, L. Li, D.G. Dorrell, D.D.C. Lu, "Field-oriented control based on hysteresis band current controller for a permanent magnet synchronous motor driven by a direct matrix converter," IET Power Electron., 11: 1277-1285, 2018.

[16] M. Sreejeth, M. Singh, "Performance analysis of PMSM drive using hysteresis current controller and PWM current controller," in Proc. 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS): 1-5, 2018.

[17] S. Wang, D.D. Xu, C. Li, "Dynamic control set-model predictive control for field-oriented control of VSI-PMSM," in Proc. 2018 IEEE Applied Power Electronics Conference and Exposition (APEC): 2630-2636, 2018.

[18] W. Zhang, W. Yang, W. Yan, D. Xu, "Improved finite control set model predictive control for permanent magnet synchronous motor drives," in Proc. 2019 Chinese Control Conference (CCC): 2977-2982, 2019.

[19] Y. Ahmed, A. Hoballah, E. Hendawi, S. Al Otaibi, S.K. Elsayed, N.I. Elkalashy, "Fractional order PID controller adaptation for PMSM drive using hybrid grey wolf optimization," Int. J. Power Electron. Drive Syst. (IJPEDS), 12: 745-756, 2021.

[20] H. Celik, T. Yigit, "Field-oriented control of the PMSM with 2-DOF PI controller tuned by using PSO," in Proc. 2018 International Conference on Artificial Intelligence and Data Processing (IDAP): 1-4, 2018.

[21] O. EROL, M. AKTAŞ, Y. ALTUN, "Obtaining of PI control parameters for vector controlled PMSM," in Proc. Int. Conf. Hydraul. Pneum. Tools, Seal. Elem. Fine Mech. Specif. Electron. Equip. Mechatronics, 2017.

[22] F. Jamshidi, M. Vaghefi, "WOA-based interval type II fuzzy fractional-order controller design for a two-link robot arm," J. Electr. Comput. Eng. Innovations, 7: 69-82, 2018.

[23] G. Demir, R.A. Vural, "Speed control method using genetic algorithm for permanent magnet synchronous motors," in Proc. 2018 6th International Conference on Control Engineering & Information Technology (CEIT): 1-6, 2018.

[24] H. Chaoui, M. Khayamy, O. Okoye, H. Gualous, "Simplified speed control of permanent magnet synchronous motors using genetic algorithms," IEEE Trans. Power Electron., 34: 3563-3574, 2018.

[25] A.A. Abd Samat, M. Zainal, L. Ismail, W.S. Saidon, A.I. Tajudin, "Current PI-gain determination for permanent magnet synchronous motor by using particle swarm optimization," Ind. J. Electric. Eng. Comput. Science, 6: 412-421, 2017.

[26] T. M. Reda, K.H. Youssef, I.F. Elarabawy, T.H. Abdelhamid, "Comparison between optimization of PI parameters for speed controller of PMSM by using particle swarm and cuttlefish optimization," in Proc. 2018 Twentieth International Middle East Power Systems Conference (MEPCON): 986-991, 2018.

[27] R. Pilla, T.S. Gorripotu, A.T. Azar, "Tuning of extended Kalman filter using grey wolf optimisation for speed control of permanent magnet synchronous motor drive," Int. J. Autom. Control, 15: 563-584, 2021.

[28] D. Gu, Y. Yao, D.M. Zhang, Y.B. Cui, F.Q. Zeng, "Matlab/simulink based modeling and simulation of fuzzy PI control for PMSM," Procedia Computer Science, 166: 195-199, 2020.

[29] S. Sakunthala, R. Kiranmayi, P.N. Mandadi, "Investigation of PI and fuzzy controllers for speed control of PMSM motor drive," in Proc. 2018 International Conference on Recent Trends in Electrical, Control and Communication (RTECC): 133-136, 2018.

[30] W.A.A. Salem, G. F. Osman, S.H. Arfa, "Adaptive neuro-fuzzy inference system based field oriented control of PMSM & speed estimation," in Proc. 2018 Twentieth International Middle East Power Systems Conference (MEPCON): 626-631, 2018.

[31] S.C. Chen, H.K. Hoai, "Studying an adaptive fuzzy PID controller for PMSM with FOC based on MATLAB embedded coder," in Proc. 2019 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW): 1-2, 2019.

[32] P.Q. Khanh, H.P. H. Anh, C.V. Kien, "Advanced sensor-less control of IPMSM motor using adaptive neural FOC approach," Applied Mechanics and Materials: 149-157, 2019.

[33] B. Verma, P.K. Padhy, "Robust fine tuning of optimal PID controller with guaranteed robustness," IEEE Trans. Ind. Electron., 67: 4911-4920, 2019.

[34] Q. Xu, C. Zhang, L. Zhang, C. Wang, "Multiobjective optimization of PID controller of PMSM," J. Control Scie. Eng., 2014: 1-10, 2014.

[35] R. Krishnan, Permanent magnet synchronous and brushless DC motor drives: CRC press, 2017.

Biographies

Hesam Salimi was born in Iran, 1995. He received his B.S. and M.S. degrees in power Electrical Engineering from Arak University of Technology (ArakUT), Arak, Iran, in 2018 and 2021, respectively. His research interests include design and control of the DC/DC and DC/AC converter, switching power supplies and Electrical Machine drive.

- Email: h.salimi1995@gmail.com
- ORCID: 0000-0002-7937-9222
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: NA

Adel Zakipour was born in Iran, in 1981. He received his Ph.D. degrees in Electrical Engineering from K.N.Toosi University of technology, Tehran, Iran, in 2017. Currently, he is an assistant professor in power electronics at department of electrical engineering, Arak university of technology. His research interests include design and control of the DC/DC and DC/AC converter, grid connected inverters and variable speed drive.

- Email: zakipour@arakut.ac.ir
- ORCID: 0000-0003-4900-5841
- Web of Science Researcher ID: 4236434
- Scopus Author ID: 55613351600
- Homepage: http://arakut.ac.ir/fa/zakipur.html

Mehdi Asadi was born in Iran, 1979. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Iran University of Science and Technology (IUST), Tehran, Iran, in 2002, 2004, and 2013, respectively. Since 2013, he has jointed to the Arak University of Technology, Arak, Iran. His research interests include high-power and high-frequency converters, electrical machines drives, power quality, battery chargers, and FACTS devices.

- Email: m.asadi@arakut.ac.ir
- ORCID: 0000-0001-7342-1584
- Web of Science Researcher ID: NA
- Scopus Author ID: NA
- Homepage: http://arakut.ac.ir/fa/asadi.html

How to cite this paper:

H. Salimi, A. Zakipour, M. Asadi, "A novel analytical approach for time-response shaping of the pi controller in field oriented control of the permanent magnet synchronous motors," J. Electr. Comput. Eng. Innovations, 10(2): 463-476, 2022.

DOI: 10.22061/JECEI.2022.8709.545

URL: https://jecei.sru.ac.ir/article_1694.html

SRTTU

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

JECEI

Research paper

Adaptive Energy-Efficient Variation-Aware Dynamic Frequency Management

H. Dorosti[*]

Department of Computer Systems Architecture, Faculty of Computer Engineering, Shahid Rajaee Teacher Training University, Tehran, Iran.

Article Info

Article History:
Received 18 September 2021
Reviewed 27 October 2021
Revised 12 January 2022
Accepted 13 January 2022

Keywords:
Internet of things
Timing slack monitoring
Negative slack measurement
Clock stretching
Frequency scaling
Ultra-Low-Energy

[*]Corresponding Author's Email Address: *hdorosti@sru.ac.ir*

Abstract

Background and Objectives: Considering the fast growing low-power internet of things, the power/energy and performance constraints have become more challenging in design and operation time. Static and dynamic variations make the situation worse in terms of reliability, performance, and energy consumption. In this work, a novel slack measurement circuit is proposed to have precise frequency management based on timing violation measurement.

Methods: the Proposed slack measurement circuit is based on measuring the delay difference between the edge clock pulse and possible transition on path end-points (primary outputs of design). The output of the proposed slack monitoring circuits is a digital code related to the current state of target critical path delay. In order to convert this digital code to equivalent delay difference, the delay of a reference gate is mandatory which is the basic unit in the proposed monitor. This monitor enables the design to have more precise and efficient frequency management, while maintaining the correct functionality regarding low-power mode.

Results: Applying this method on a MIPS processor reduces the amount of performance penalty and recovery energy overhead up to 30% with only 2% additional hardware. Results for benchmark applications in low-power mode, show 7-30% power improvement in normal execution mode. If the application is resilient against occurred errors duo to timing violations, the proposed method achieves 20-60% power reduction considering approximate computation as long as application is showing resilience. The performance of the proposed method depends on the degree of application resilience against the timing errors. In order to keep generality of the proposed monitor for different applications, the resilience threshold is user programmable to configure according to the requirements of each application.

Conclusion: The results show that precise frequency scheduling is more energy/power efficient in static and dynamic variation management. Utilizing a proper monitor capable of measuring the amount of violation will help to have finer frequency management. On the other hand, this method will help to use the resilience of application according to estimation about the possible error value based on measured violation amount.

Introduction

Internet of Things (IoT) is a fast emerging technology which enables continuous sensing data flow and actuation controls through everything and involves different applications from health, industry, automation, military and etc. The demands for these applications are different in terms of performance, power/energy, reliability, and lifetime.

Energy efficiency is the common objective for applications ranging from energy-constrained with low performance and high lifetime requirements to high-performance maintainable needs.

Ultra-low-energy processors providing performance

Doi: 10.22061/JECEI.2022.8331.507

demands for IoT applications must be kept at a reasonable lifetime for network operation.

Feature size scaling makes the design of those processors more cost effective, but in contrast, due to leakage current and process variation, energy efficiency is dropped down and reliability issues are imposed. On the other hand, wear-out mechanisms will shorten the lifetime of the network and impose energy overhead. Clearly, any variation related to design timing, due to Process-Voltage- Temperature variations (PVT) & Aging, will result in reliability and performance issues. Timing-error if occurs, will impose energy overhead during the operation of the processor. Therefore, timing errors in design time as well as frequency management during the operation of the design should be considered.

The rest of this paper is organized as follows: In the next section, related works are explored and their achievements are presented. After literature review, we discuss about the proposed method and design details as another section. The experimental results and analysis of the proposed method will come afterward. Finally we conclude the paper in the last section.

Related Works

In order to address reliability issues due to PVT & Aging variations, timing error detection is a need in recent technologies. To date studies have mostly focused on timing error detection or slack measurement to prevent the design functionality failure instead of inserting longer guard-bands (20% margins in [1]) between the path delays and clock period.

There are a class of timing error detection methods namely called In-situ timing error detection consist of RAZOR [2], [3], RAZOR Lite [4]. There exists also another method called Replica circuits [5] which are configured with the latency of the design paths to forecast the error probability due to temporal and spatial dependency of circuit elements. Measurements of these methods are used to capture the statistical state of the timing variations inside the design and to manage the voltage and frequency or even to utilize the error correction [7] and recovery [2], [3], [7] methods in order to compensate the variation effects and to retrieve the processor against failure state.

Another class of timing error forecasting is utilizing positive slack monitoring circuits in conjunction with timing guard-band. The authors of [8] have proposed slack monitoring circuit at end-points of the design to measure the available guard-band between the path delays and the period of the clock cycle. The measurement results are available through the scan chain and the slack monitors are inserted using ATPG toolset. In case the design timing is altered, the signal cannot cross the whole elements of the delay line and the embedded register will capture the state of the delay margin at rising edge of the clock cycle. When timings margin is reached to critical state, the processor is configured to increase the margin and to ensure the correct functionality.

The authors of [9] used activation probability and correlated detection window to maximize the efficiency of measurements of positive slack monitors. Using this method, slack monitors are placed in proper nodes while more precise measurements are achieved.

A. Benhassain et al. have used a voltage regulation feedback based on positive slack monitor to perform a better voltage selection according to variation in path delays or circuit operation modes. Using this method, the proper voltage and frequency selection is simpler and sign-off difficulties for different operating conditions are reduced [10].

SlackProbe in [11] is a method to increase the efficiency of positive slack monitor insertion as well as to increase the observability and as a result, to reduce the number of required slack monitors. In this work, the monitors are inserted at intermediate nodes instead of the end-points for early detection and smaller area overhead. This method could be used jointly with different compensation techniques.

The authors of [12] have proposed a novel slack monitor circuit in transistor level to improve its accuracy, power, area, and aging resistance. The designed monitor circuit is placed in master-slave flip-flop to add the monitor capability into the registers. These registers are used with a high-resolution TDC circuit to measure the remaining positive slack of the selected paths. TDC [13] circuit is similar to ring oscillator [14] design based on basic gate elements which is configured to measure the path delay with higher timing resolution.

Timing speculation [15] is another proposed method to prevent the effects of timing violation within the same clock cycle. In this work and similar publications [16] the monitors are placed in mid-point of candidate path and detect the failure event before the end of clock period. Global clock stretching is a short time method to reduce the power and performance penalty and could not provide the lifetime efficiency. Authors of [17] used machine learning methods to estimate current aging state and remaining lifetime by 97% average precision. This work achieves an acceptable estimation, but not sufficient for precise real-time frequency scaling.

There are another class of works such as [18] which try to improve design lifetime by utilizing aggressive voltage and frequency scaling without any precise measurements. These methods rely on planning voltage and frequency of the design to reduce the stress and improve the lifetime by 40%.

A major group of recent works have been exploited positive slack monitoring for timing margin measurement to scale the working frequency in order to prevent timing error occurrence. These methods act precautionary to ensure correct operation by placement of in-situ timing circuits in mid or end points of candidate paths. Another group of studies utilize the error detection, state retrieval, and re-compute with new configuration. These methods are known as recovery methods. They provide however limited power/energy and lifetime improvements and have lower flexibility considering different application requirements due to lack of the exact timing behavior consideration. In order to have energy efficient variation management, both of likelihood and severity of timing violation should be considered, while recent studies only consider threshold-based monitoring of delay growth (as severity of possible violation). Precise monitoring of delay growth by tracking the occurrence of timing violation over a period of time will result in more energy efficient voltage and frequency scaling method.

In this work, we propose another intermediate method to monitor the amount of negative slack at the end-points. This method reduces the performance and energy overhead significantly and provides more precise frequency management capability. The key contributions to this work include:

1) A novel negative slack monitoring circuit capable of measuring the amount of timing violations up to ½ clock period.

2) A novel monitor insertion method based on timing analysis of the design.

3) A novel frequency management scheme with the introduction of the forecasting pipe stage based on the proposed slack monitor. This method provides more efficient voltage and frequency management in terms of performance and energy consumption while preventing functional failure of the design.

Negative Slack Monitoring

In this section, we present a new architecture for negative slack monitor and explain its functionality and design in details. Careful timing considerations are required which will be discussed in this section. The proposed slack monitor insertion method is used based on the statistical static timing analysis of the design. The architecture level utilization of the monitors together with the added benefits based on the timing analysis of the design will be presented as follows.

Negative Slack Monitor Architecture

In order to measure the positive slack, designers [8]-[12] used a circuit which generates a pulse according to difference between the input and output of flip-flops which are placed at the circuit end-points. The length of this pulse is measured using another circuit known as time-to-digital converter (TDC) and the measurement result is captured in rising edge of the clock signal. The story behind the negative slack monitor is quite different with the one with positive slack, in which the negative slack measurement starts at the rising edge of the clock signal and continues until a transition detector detects the difference between the input and output of the end-point flip-flop. Fig. 1 shows the architecture of the negative slack monitor. As shown in this figure, the clock signal is fed to the TDC circuit and the generated signal from the transition detector is used to capture the measurement results. While in positive slack monitor, the output of the design path crosses the TDC circuit and the clock signal captures the measurement result as available timing margin.

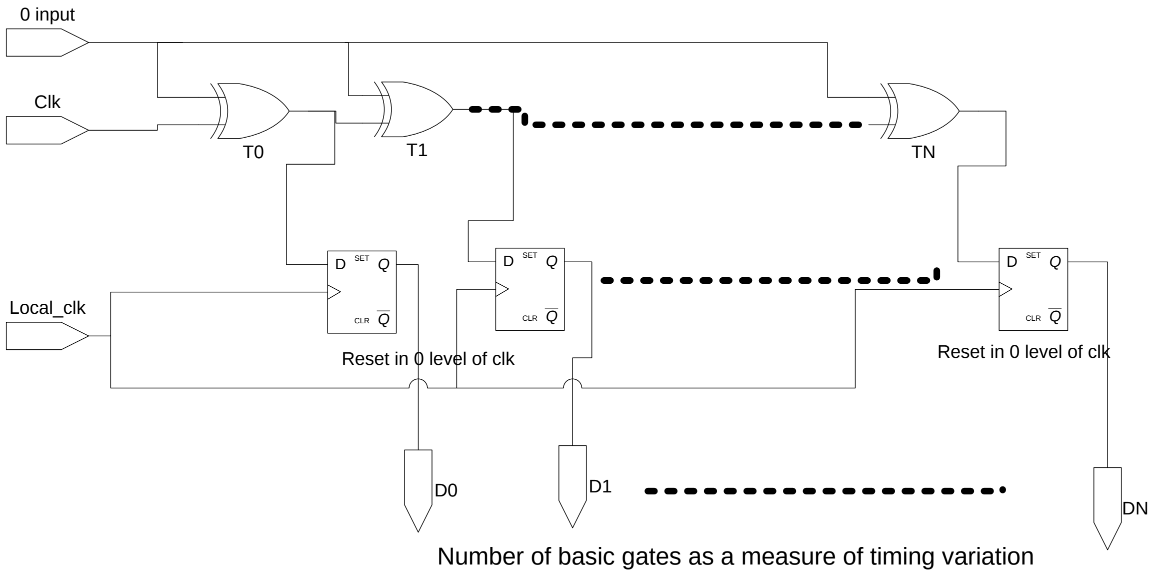


Fig. 1: The architecture of negative slack monitoring sensor.

Fig. 2 shows the structure of transition detector which is used jointly with the circuit above. The signal naming in both figures are the same to follow the dependency between these figures.

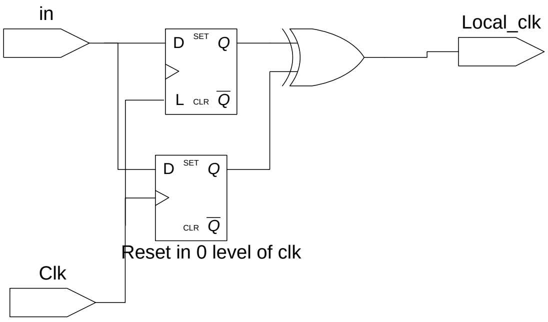


Fig. 2: The architecture of transition detector which is used with slack monitor circuit.

In order to use this monitor in the design, careful timing considerations are required. Minimum delay of the paths ending to the transition detector should be more than half of the clock period. The mentioned constraint means that the negative slack should be less than half of the clock period and the TDC should have capability to measure the latency with higher precision. If the delay of activated path ending to the slack monitor has no timing violation, there is no transition and meanwhile the register captures zero value meaning no negative slack occurred. If the signal violates the safety margin or even crosses the rising edge of the clock cycle, the transition detector will activate the error signal (Local_clk) as a result. At the rising edge of the error signal, the state of the clock signal crossing the delay line is captured in the existing flip-flops (Fig. 1). The captured digital data is a code that represents the amount of negative slack for the expected end-point, and will count the amount of '1's in this code to return the negative slack in unit delay. In this structure the unit delay means the delay of a basic XOR gate which shapes a buffer with a specific delay. In this scheme, the transition detector has a delay that is equal to two-unit delay measurement and should be considered in the final result.

It is important to know that the existence of a latch in transition detector is mandatory and the propagation delay of that latch should be greater than or equal to that of the flip-flop. Removing the latch from the design creates instability in measurements due to detector delay discrepancy. There is a hidden exception in the operation of the proposed monitor circuit, when the amount of negative slack is smaller than the propagation delay of the main flip-flop. In this case, the flip-flop delay will be counted on the measured negative slack, while it is not counted for greater negative slacks. Here, consider a condition in which the value of the end-point remains the same for a set of two clock cycles. After the rising edge of the second pulse, the transition detector activates the error signal due to the late signal arrival. In this situation when the logic state of the arriving signal remains the same in comparison with the previous clock edge, the measured value for the same negative slack will vary time to time and the delay of the main flip-flop in transition detector will not be counted on delay measurement accordingly. This event can therefore affect the accuracy of the sensor and should be compensated in order to resolve the problem; we need to reset the transition detector flip-flop during zero level of the clock signal to remove the consecutive '1's. Fig. 3 presents the proposed flip-flop structure using two latches to enable the flip-flop reset at zero level of the clock signal. The two consecutive zero values problem is resolved using additional flip-flop that will capture the comparison between the two consecutive clocked signals. When the output of this flip-flop is activated and the measured delay is less than or equal to the three unit delays, then the delay of the flip-flop should be taken into account for measurement correction. In order to reduce the compensation area and power overhead, a latch is placed between the arriving signal and comparator to make the problem to be uniformed for all instability conditions.

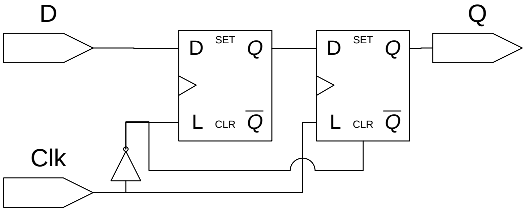


Fig. 3: The flip-flop structure enabling reset on clock zero level for negative slack monitor.

Our proposed slack monitoring circuit is capable of detecting the delayed transition from the beginning of safety margin to the next half clock period.

A. Slack Monitor Insertion

To have an efficient monitor insertion, statistical timing analysis is required. Timing violation and safety margin impose that the most vulnerable endpoints are the best candidates to insert the proposed monitor. Since any timing variation, due to a path entering the safety margin, may cause that path to violate the timing, the ending point to that path can be a candidate to place the monitor. The insertion method for a standard synthesis flow, according to Fig. 4, could be extended to every netlist generated from physical design and layout by applying the timing analysis to the netlist. Statistical static timing analysis removes unnecessary timing margins and reduces the amount of critical nodes which results in smaller area and energy overhead due to slack monitors. The delay distribution of the paths is extracted from the timing analysis. Critical end points and critical paths will be identified with regards to timing margin requirement and different variation sources (such as 10% of clock period). These nodes are identified as critical nodes in which the slack monitors are inserted to measure the amount of timing violation.

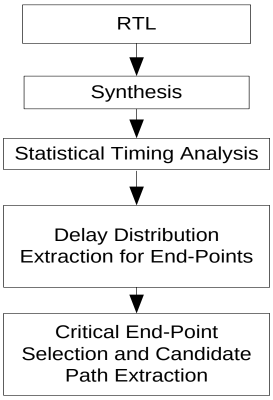

RTL
Synthesis
Statistical Timing Analysis
Delay Distribution Extraction for End-Points
Critical End-Point Selection and Candidate Path Extraction

Fig. 4: The monitor insertion flow diagram.

First of all, we need to analyze the path delay distribution for the design to evaluate the nature of the delay distribution with respect to the clock period. Considering Tacc as an acceptable path delay,

Tacc = Clock Period – Safety Margin (1)

It is obvious that major parts of the design paths are shorter than the Tacc. These paths do not cause the safety margin to enter the critical state except some minor parts (known as critical path) that could be the main source of timing violation but are rarely activated. Figs. 5, 6, and 7 are shown the delay distribution of the paths in combination with the activity probability for three ISCAS'85 benchmarks which are achieved applying random test vectors and confirm the above sentence. The delay distribution combination of the design paths with activity probability is used to determine the candidate end-points.

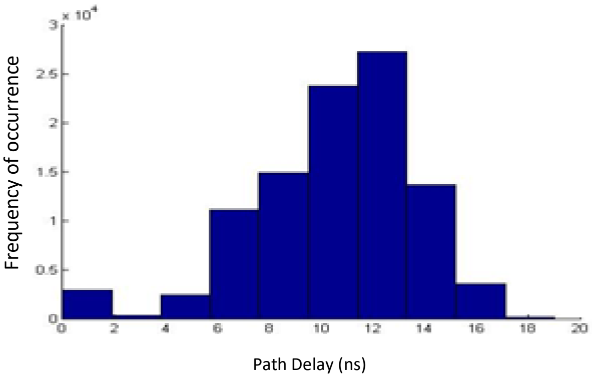


Fig. 5: The delay distribution (histogram) of paths for C3540 ISCAS benchmark.

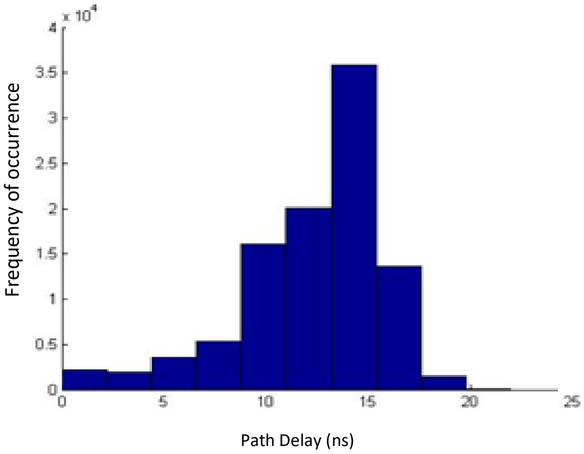


Fig. 6: The delay distribution (histogram) of paths for C6288 ISCAS benchmark.

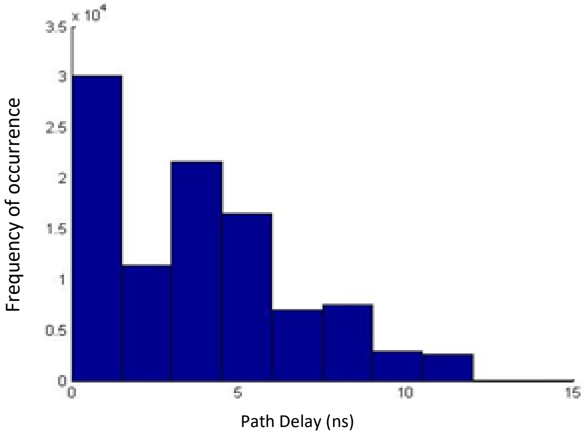


Fig. 7: The delay distribution (histogram) of paths for C7552 ISCAS benchmark.

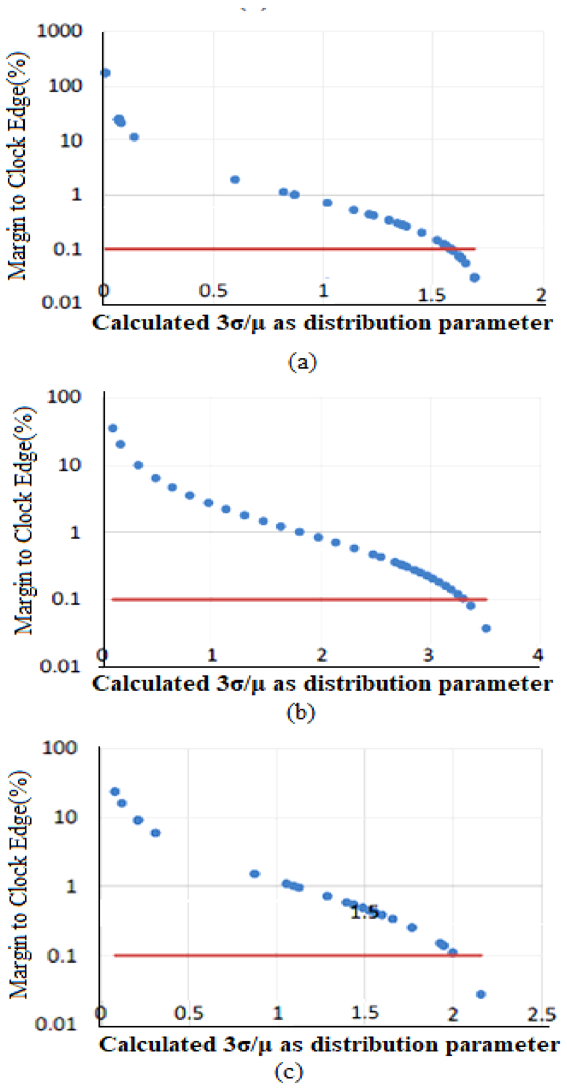
These candidates have the capability to grow and violate the timing constraints which result in timing errors and possible functional failures. We have used Tacc in our previously published work [21] as a measure of criticality to prevent the failure due to process variations.

The usage of this parameter for every static and dynamic variation which modifies the timing specifications is proposed accordingly.

The candidate paths are then extracted for the design, according to delay distributions of the paths terminating each end-point, and the maximum variability due to different variation sources and performance requirements (clock frequency).

Any point, at the end of those paths violating the safety margin of the clock cycle, remains candidate for timing error and the others are considered to be error-prone end-points.

Fig. 8 presents the candidate end-points for the benchmark circuits with 10% safety margin assigned (Red Line).



(a)

(b)

(c)

Fig. 8: Number of candidate end-points for a) c3540, b) c6288, and c) c7552 ISCAS benchmarks. Horizontal axis is calculated using the average and standard deviation of path delays for each benchmark.

Utilization Method

After the end-point selection, it is time to measure slacks to be used in dynamic management method and to compensate the variation effects.

As mentioned in the second section, recent studies use the slack monitor to measure the positive slack and to keep the safety guard-band in order to ensure the correct functionality. When the safety guard-band could not be established, the use of dynamic frequency scaling is inevitable as a result. Considering the lower probability of timing error at the end-points and the severity of delay growth with respect to different variation sources, there is no need to scale the frequency (and probably the voltage) of the design. Frequency scaling burdens more performance and energy overhead while clock stretching will resolve the problem with smaller overhead. In order to clarify the situation, one could consider a design working with 2GHz frequency and 10% activation probability for timing errors.

According to recent studies, anytime the path delays enter the safety margin, the clock frequency will be scaled to 1.5GHz. Reconfiguration reduces the performance of the design by 25% which makes the design to consume almost 25% more energy in order to finish the same task. The more the frequency steps are further away, the more energy overhead will be imposed to the circuit.

While, clock stretching to prevent timing error, will imposes only 10% performance and energy overhead. In order to reduce the supply voltage to achieve higher energy efficient design, this method makes the design to perform more reliable functionality and to achieve a reasonable performance.

Fig. 9 shows the timing diagram of the negative slack monitor and the clock stretch mechanism.

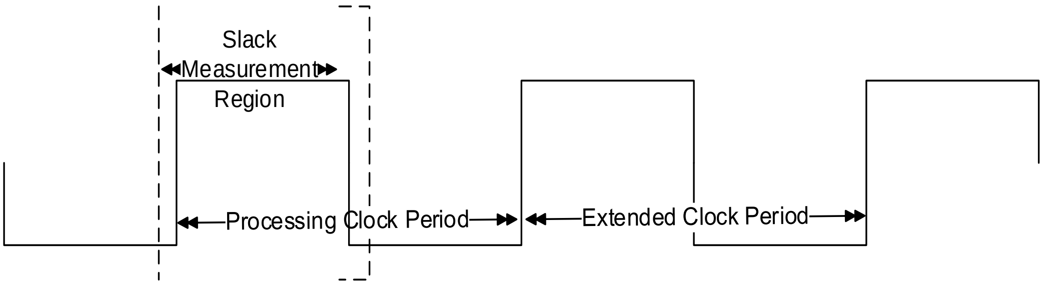


Fig. 9: The timing diagram for negative slack monitoring and clock stretching.

Since the continuous adjustment of the clock frequency is not possible in reality and the designers use discrete frequency steps to manage working frequency, there is a cross point between the model of performance overhead for the proposed stretching method and the methods presented in literature review (2) shows the mathematical description for this cross point in which the left hand side is for the proposed method and the right hand side is for the frequency scaling.

$$(N_{clk} + p \times N_{clk}) \times T_{clk} = N_{clk} \times (1+k) \times T_{clk} \qquad (2)$$

where N_{clk} is the number of cycles required to execute the application, p is the activation probability of timing error, k is the steps of frequency scaling and T_{clk} is period of baseline clock signal.

According to (2), the proposed method is demonstrated to have higher performance and lower energy (neglecting the energy consumption for additional hardware for complex design) consumption compared to the others for k > p and vice versa. The imposed overhead is the same for k = p. Therefore, the proposed method will act as a finer frequency management method in comparison with the recent studies when k > p.

It is important to know that, when the paths exceed beyond the two consecutive cycles, this method cannot prevent timing error and frequency scaling is therefore mandatory. Then, we have to measure the amount of the negative slack rather than relying only to the timing error detection. Using the negative slack monitor, any latent transition is detected and measured to be considered in stretching and scaling of the clock cycle. In order to have efficient frequency management method, we need to introduce replica pipe stage as the forecasting stage. Candidate end-points and critical paths ending to these points are copied to a prior pipeline stage which is known as replica stage, and the same input vector is applied to both stages. When the monitor asserts the timing error prediction signal, due to similarity in both of replica and original end-points and identical input vector, it is obvious to occur in original stage. Then, the next clock is stretched or scaled to extend the design lifetime. Fig. 10 presents the position of both replica (gray blocks) and original stages (white blocks).

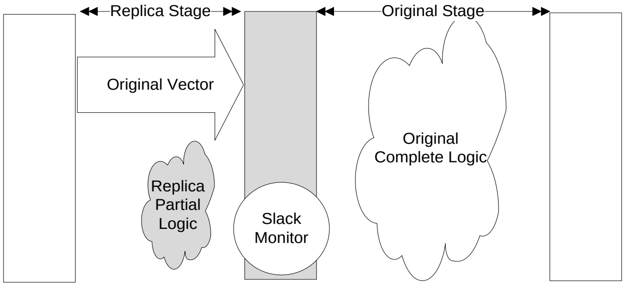


Fig. 10: The schematic of replica pipe stage which is used to measure negative slack monitoring in dynamic frequency management.

This method gives another chance to use approximate computing without any additional overhead in terms of area, performance and energy consumption. The amount of computation error has a direct relation to the severity of timing variation (or the amount of measured negative slack). According to the application and its resiliency against the computation errors, we have measured the tolerable error threshold and mapped the threshold to the slack measurements. When the negative slack is shorter than the tolerable value, the timing error is ignored and masked by the application and there is no need to further compensation in architecture level. This is an outstanding feature of the proposed method for multimedia applications which have a degree of resilience against the errors.

Results and Discussion

The negative slack monitor using the mentioned insertion and utilization method is applied to ISCAS benchmarks and MIPS processor. The benchmark circuits are synthesized in 45nm technology using standard synthesis tools, and the energy consumption and performance of the design are achieved using power compiler and timing analyzer tools. The variation effects on path delays are modeled using proper models for different sources in gate level simulation utilizing PLI interface. The delay distribution of the basic gates is used to model the effect of process variation. The effects of voltage and temperature variations are modeled as random fluctuation in candidate basic elements. Wear-out effects are modeled as progressive timing growth with respect to the effective parameters such as voltage, frequency, activity factor, operating time, and etc.

Table 1 presents the specifications of the proposed slack monitor. According to the results, the monitor has negligible area and energy overhead and it measures the negative slack with higher accuracy.

Table 1: Specifications of the proposed slack monitor in 45nm

Area (nm^2)	Power (uW)
19	25

Appling the monitor to the candidate ISCAS benchmarks shows that the proposed method provides higher performance and lower energy consumption compare to pure guard-banding and frequency scaling methods. The proposed method is compared to a baseline end-point monitor insertion method [8] as baseline method. Table 2 presents the performance and energy consumption of both methods on candidate benchmark circuits. The last two columns of this table present the performance overhead in comparison with the state with no timing variation.

Table 2: Comparison of ISCAS benchmarks using the baseline and the proposed slack monitors in terms of energy and performance in 45nm technology

Benchmark	Power (mW)		Performance Reduction	
	Proposed	[8]	Proposed	[8]
C3540	0.65	0.8	5%	30%
C6288	1.4	1.7	6%	30%
C7552	1.15	1.3	10%	30%

Considering the execution unit of the MIPS processor to be vulnerable to the timing errors, the candidate end-points for this stage is extracted and the replica stage is added to the architecture. Table 3 presents the area and energy overhead in comparison to the baseline method. In most cases, the proposed method results in better performance and lower energy consumption. In some cases, both methods have the same efficiency due to severity of the delay growth in design paths.

Table 3: Efficiency of the baseline and the proposed slack monitor and imposed overhead on MIPS processor in 45nm

Error Rate	Power (mW)		Performance Reduction	
	Proposed	[8]	Proposed	[8]
10%	3.32	3.91	10%	30%
30%	3.95	3.94	30%	30%
50%	3.95	3.96	30%	30%
80%	3.95	3.96	30%	30%

Using the capability of the proposed method to reduce the voltage deliberately to achieve lower power/energy consumption while keeping the application functionality, 7-30% power improvement is achieved (Table 4). In this case, we have used lower supply voltages which results in timing violation to improve power/energy consumption, and the proposed method tolerates the effect of supply voltage reduction by stretching the clock period to keep the correct functionality.

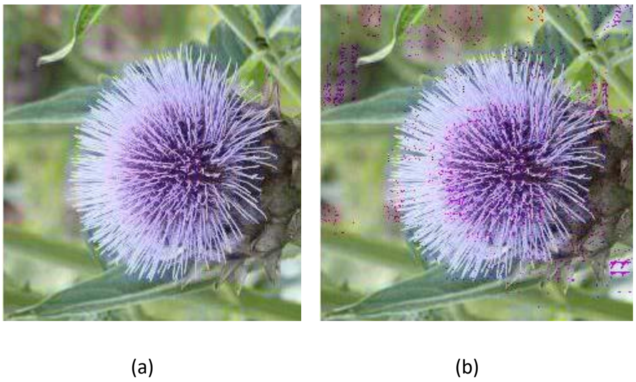
Table 4: Power improvement of the processor when the studied benchmarks have been run under different operating voltage level without performance overhead

Benchmark	V_{DD} (V)	Power (mW)	Improvement
adpcm	0.9	2.8	25%
	1.1	3.7	
blowfish	0.9	3.2	30%
	1.1	4.6	
Dfmul	0.9	2.2	19%
	1.1	2.8	
Gsm	0.9	2.6	25%
	1.1	3.5	
Gaussian blur	0.9	1.6	8%
	1.1	1.8	
JPEG	0.9	2.4	12%
	1.1	2.7	
Sobel	0.9	1.7	7%
	1.1	1.8	

Also, considering inherent error resiliency of the benchmark applications, remarkable results are achieved. Table 5 shows the power/voltage improvements for both the normal and low-power mode. Fig. 11 shows an execution result for JPEG benchmark application in both operating modes. The acceptable condition for error resilience is considered as MSSIM [22] to be greater than 0.9 for all benchmark applications and the negative slack threshold is then determined according to this parameter. Enabling low power mode in comparison to the sense of error resilience capability in the proposed method results in 20-60% power improvement without significant performance overhead.

Table 5: Power improvement of the processor when the studied benchmarks have been run under different operating voltage level achieving MSSIM [20] greater than 0.9 without performance overhead

Benchmark	V_{DD} (V)	Power (mW)	Improvement
adpcm	0.9	1.7	54%
	1.1	3.7	
blowfish	0.9	2	57%
	1.1	4.6	
dfmul	0.9	1.7	39%
	1.1	2.8	
gsm	0.9	1.6	54%
	1.1	3.5	
Gaussian blur	0.9	1.4	22%
	1.1	1.8	
JPEG	0.9	2	26%
	1.1	2.7	
Sobel	0.9	1.4	22%
	1.1	1.8	


(a)

(b)

Fig. 11: JPEG benchmark output, (a) original image, (b) degraded image of voltage 0.9.

In order to have better comparison between performance of the proposed work and the state of the art literature, we have provided comparison data with two recent works in terms of area and power overhead. Table 6 summarizes the comparison data. The reported area and power overhead is for all required extra hardware and power consumption after applying the method compared to initial design. Power overhead is a total value extracted from individual evaluation process of each work and reported as an average of experiments. In this table, it is clear that the proposed work operates with only 2% extra hardware (smallest area overhead) and consumes 3% extra power.

Table 6: Efficiency of the proposed monitor compared to literature

Overhead	Proposed	[8]	[11]	[15]
Area overhead (%)	2	5	9	3
Power overhead (%)	3	10	16	3

Conclusion

Considering the growth rate of low-power internet of things, the power/energy and performance constraints for these applications are so tight. Static and timing variations make the situation worse in terms of reliability, performance and power/energy consumption. In this work, a new slack monitoring circuit and monitor insertion method is proposed. A novel frequency management scheme is introduced based on the proposed slack monitor at the candidate end-points. We have used negative slack monitor instead of positive one to enable more precise frequency management by measuring the actual delay growth of design paths. In order to have in situ error correction by clock stretching, the measurement on the selected end-points was performed one cycle before. A *replica pipe stage* is required to measure the probability and severity of timing error to decide the clock cycle stretch or to scale the frequency. This method provided another low-power mode which enabled the design to execute the applications in lower supply voltages with correct functionality with 7-30% power improvement. Also, in this method there is another option to sense the resilience of application against the timing error which reduces the amount of performance and energy overhead (20-60% power improvement) to high extent.

Author Contributions

H. Dorosti designed the experiments and collected the data through proper simulations, and data analysis is carried out by him. Finally H. Dorosti interpreted the results and wrote the manuscript.

Acknowledgment

The author would like to acknowledge the scientific support of Professor Sied Mehdi Fakhraie whose experience in the field and insightful guidelines was very helpful for the project. The authors would like also to thank Dr. Fakhraie for providing the computing equipments at the Silicon Intelligence laboratory of the University of Tehran.

Conflict of Interest

Author declare that there is no conflict of interests 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

IoT	Internet of Things
PVT	Process, Voltage and Temperature
TDC	Time to Digital Converter
ATPG	Automatic Test Pattern Generation

References

[1] W. Abadeer, W. Ellis, “Behavior of NBTI under AC dynamic circuit conditions,” presented at the 41’st International Reliability Physics Symposium, Dallas, TX, USA, 2003.

[2] D. Ernst, N. S. Kim, S. Das, S. Pant, R. Rao, T. Pham, C. Ziesler, D. Blaauw, T. Austin, K. Flautner, T. Mudge, “Razor: A low-power pipeline based on circuit-level timing speculation,” presented at the IEEE/ACM Int. Symposium Microarchitecture, 2003.

[3] S. Das, C. Tokunaga, S. Pant, W.H. Ma, S. Kalaiselvan, K. Lai, D.M. Bull, D.T. Blaauw, “RazorII: In situ error detection and correction for PVT and SER tolerance,” IEEE J. Solid-State Circuits, 44(1): 32–48, 2009.

[4] S. Kim, I. Kwon, D. Fick, M. Kim, Y.P. Chen, D. Sylvester, “Razor-lite: A side-channel error-detection register for timing-margin recovery in 45nm SOI CMOS,” in Proc. 2013 IEEE ISSCC: 264–265.

[5] F. Firouzi, F. Ye, K. Chakrabarty, M.B. Tahoori, “Representative critical-path selection for aging-induced delay monitoring,” in Proc. 2013 IEEE International Test Conference (ITC): 1-10, 2013.

[6] M. Fojtik, D. Fick, Y. Kim, N. Pinckney, D.M. Harris, D. Blaauw, D. Sylverster, “Bubble razor: Eliminating timing margins in an ARM cortex-M3 processor in 45 nm CMOS using architecturally independent error detection and correction,” IEEE J. Solid-State Circuits, 48(1): 66–81, 2013.

[7] K. Bowman, J. Tschanz, C. Wilkerson, S. L. Lu, T. Karnik, V. De, S. Borkar, “Circuit techniques for dynamic variation tolerance,” in proc. Design Automation Conference (DAC), 2009.

[8] M. Sadi, L. Winemberg, M. Tehranipoor, “A robust digital sensor IP and sensor flow for in-situ path timing slack monitoring in SoCs,” in Proc. IEEE 33rd VLSI Test Symposium (VTS), 2015.

[9] S. Sarrazin, S. Evain, I. Miro-Panades, L.A.D. B. Naviner, V. Gherman, “Flip-flop selection for in-situ slack-time monitoring based on activation probability of timing-critical paths,” in Proc. IEEE 20th Int. On-Line Testing Symposium (IOLTS), 2014.

[10] A. Benhassain, F. Cacho, V. Huard, M. Saliva, L. Anghel, C. Parthasarathy, A. Jain, F. Giner, “Timing in-situ monitors: implementation strategy and applications results,” in Proc. IEEE Custom Integrated Circuits Conference (CICC), 2015.

[11] L. Lai, V. Chandra, R. C. Aitken, P. Gupta, “SlackProbe: A flexible and efficient in situ timing slack monitoring methodology,” IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 33(8): 1168-1179, 2014.

[12] N. Pour Aryan, G. Georgakos, D. Schmitt-Landsiedel, “Reliability Monitoring of Digital Circuits by in situ Timing Measurement,” in

Proc. 23rd International Workshop on Power and Timing Modeling, Optimization and Simulation (PATMOS), 2013.

[13] D. Fick, N. Liu, Z. Foo, M. Fojtik, J. S. Seo, D. Sylvester, D. Blaauw, “In situ delay-slack monitor for high-performance processors using an all-digital self-calibrating 5ps resolution time-to-digital converter,” presented at the IEEE ISSCC, San Francisco, CA, USA, 2010.

[14] X. Wang, M. Tehranipoor, R. Datta, “Path-RO: A novel on-chip critical path delay measurement under process variations,” presented at the IEEE/ACM ICCAD, San Jose, CA, USA, 2008.

[15] H.A. Balef, H. Fatemi, K. Goossens, J.P.D. Gyvez, “Timing speculatoin with optimal In-Situ monitoring placement whithing-cycle error prevention,” IEEE Trans. Very Large Scale Integr. VLSI Syst., 27(5): 1206-1217, 2019.

[16] H.A. Balef, K. Goossens, J.P.D. Gyvez, “Chip health tracking using dynamic In-Situ delay monitoring,” presented at the 2019 Design, Automation & Test in Europe Conference & Exhibition (DATE), Florence, Italy, May, 2019.

[17] Y.G. Chen, I.C. Lin, Y.C. Wei, “A novel NBTI-Aware chip remaining lifetime prediction framework using machine learning,” presented at ISQED, Santa Clara, CA, USA, April, 2021.

[18] F. Nakhaee, M. Kamal, A. Afzali-Kusha, M. Pedram, S.M. Fakhraie, H. Dorosti, “Lifetime improvement by exploiting aggressive voltage scaling during runtime of error-resilient applications,” Integr. VLSI J., 61: 29-38, 2018.

[19] H. Reyserhove, W. Dehaene, “Design margin elimination through robust timing error detection at ultra-low voltage,” in Proc. IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2017.

[20] Y. Sazeides, A. Bramnik, R. Gabor, C. Nicopoulos, R. Canal, D. Konstantinou, G. Dimitrakopoulos, “2D error correction for F/F based arrays using In-Situ Real-Time Error Detection (RTD),” in Proc. IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), 2020.

[21] M. Faryabi, H. Dorosti, M. Modarresi, S.M. Fakhraie, “Process variation-aware approximation for efficient timing management of digital circuits,” in Proc. IEEE East-West Design and Test Symposium (EWDTS), 2015.

[22] Z. Wang, A.C. Bovik, H.R. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., 13: 600-612, 2004.

[23] J. Rabaey, “Low Power Design Essentials,” Springer Science & Business Media, USA, 2009.

[24] E. Karl, D. Blaauw, D. Sylvester, T. Mudge, “Reliability modeling and management in dynamic microprocessor-based systems,” in Proc. the 43rd ACM/IEEE in Design Automation Conference: 1057–1060, 2006.

[25] F. Oboril, M.B. Tahoori, “Reducing wearout in embedded processors using proactive fine-grain dynamic runtime adaptation,” in Proc. the 17th IEEE European Test Symposium (ETS): 1–6, 2012.

[26] B. Soltani, H. Dorosti, M. E. Salehi, S. M. Fakhraie, "Ultra-low-energy DSP processor design for many-core parallel applications," JECEI, 8 (1): 71-84, 2019.

[27] H. Dorosti, et al., "Ultralow-energy variation-aware design: adder architecture study," IEEE Trans. Very Large Scale Integr. VLSI Syst., 24(3): 1165-1168, 2016.

[28] A. Teymouri, H. Dorosti, M.E. Salehi, S.M. Fakhraie, “Energy-efficient variation-resilient high-throughput processor design,” JECEI, doi: 10.22061/JECEI.2021.8253.499, 2022.

[29] M.B. Taylor, "A landscape of the new dark silicon design regime," Micro, IEEE Micro, 33(5): 8-19, 2013.

[30] H. Esmaeilzadeh, et al., "Dark silicon and the end of multicore scaling," in Proc. 2011 38th Annual International Symposium on Computer Architecture (ISCA): 365-376, 2011.

Biographies

Hamed Dorosti was born in Khoy, in 1986 and received the B.S. and M.S. degree in computer engineering from University of Tehran, Tehran, Iran, in 2009 and 2011, respectively. He received his Ph.D. in computer engineering (computer architecture) from University of Tehran in 2017. Since 2009, he was member of Silicon Intelligence and VLSI Signal Processing Lab., University of Tehran and co-operated in low-power ASIP project from 2010 to 2012. His research interest includes VLSI design, digital signal processing, adaptive timing error detection and correction and low-power high-throughput/performance processor architecture design considering static and dynamic variations. He is now an assistant professor of Shahid Rajaee Teacher Training University.

- Email: hdorosti@sru.ac.ir
- ORCID: 0000-0001-6554-1607
- Web of Science Researcher ID: L-5928-2019
- Scopus Author ID: 36662029700
- Homepage: https://www.sru.ac.ir/dorosti

How to cite this paper:

H. Dorosti, “Adaptive energy-efficient variation-aware dynamic frequency management,” J. Electr. Comput. Eng. Innovations, 10(2): 477-486, 2022.

DOI: 10.22061/JECEI.2022.8331.507

URL: https://jecei.sru.ac.ir/article_1701.html

Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

An Adaptive Routing Algorithm for Wireless Network on Chips

A. Tajary[*]*, E. Tahanian*
Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.

Article Info

Article History:
Received 16 December 2021
Reviewed 24 January 2022
Revised 16 February 2022
Accepted 20 April 2022

Keywords:
Wireless network on chip
Routing
NoC

[*]Corresponding Author's Email Address:
tajary@shahroodut.ac.ir

Abstract

Background and Objectives: Wireless Network on Chip (WNoC) is one of the promising interconnection architectures for future many-core processors. Besides the architectures and topologies of these WNoCs, designing an efficient routing algorithm that uses the provided frequency band to achieve better network latency is one of the challenges.

Methods: Using wireless connections reduces the number of hops for sending data in a network, which can lead to lower latency for data delivery and higher throughput in WNoCs. On the other hand, since using wireless links reduces the number of hops for data transfer; this can result in congestion around the wireless nodes. The congestion may result in more delay in data transfer which reduces the network throughput of WNoCs. Although there are some good routing algorithms that balance traffic using wired and wireless connections for synthetic traffic patterns, they cannot deal with dynamic traffic patterns that existed in real-world applications. In this paper, we propose a new routing algorithm that uses the wireless connections as much as possible, and in the case of congestion, it uses the wired connection instead.

Results: We investigated the proposed method using eight applications from the Parsec benchmark suite. Simulation results show that the proposed method can achieve up to 13.9% higher network throughput with a power consumption reduction up to 1.4%.

Conclusion: In this paper, we proposed an adaptive routing algorithm that uses wireless links to deliver data over the network on chip. We investigated the proposed method on real-work applications. Simulation results show that the proposed method can achieve higher network throughput and lower power consumption.

Introduction

Demand for more computation power within a chip resulted in multicore systems in which the bus technology was used to connect processing elements and cache modules on a single chip [1], [2]. Due to the poor scalability of bus base systems, a new paradigm is formed for future many-core systems: Network on Chip (NoC) [3]-[6]. In the NoC, many processing elements and cache systems are connected through an on-chip interconnection network. Each interconnection node contains a processing element and a router [7]. The nodes can connect in 2D (like mesh) or 3D structures.

Although this paradigm is promising, some issues should be considered. One of the main issues in the traditional NoCs is the large latency of packets for traveling between the distant source and destination nodes. Sending packets to a neighbor node can be done in one hop, while sending a packet to a distant node,

Doi: 10.22061/JECEI.2022.8464.518

may take tens of hops, which results in long latency and probably makes congestion in the network [8], [9]. To alleviate this problem, several new technologies like 3D NoC [10]-[12], Photonic NoC [13]-[15] and Wireless NoC [16]-[18] has been proposed.

Wireless NoC (WNoC) is one of the main options for future NoCs [16], [17]. There are many kinds of research conducted to investigate different aspects of this technology such as the possible architectures and routing algorithms in these new architectures [18]-[22]. In the WNoCs, some nodes have a wireless antenna which makes them the wireless nodes. To reduce the number of hops, sending packets through the wireless nodes is preferred over the traditional wired nodes [9]. However, this makes congestion around the wireless nodes and leads to more latency for packets and more congestion in the network.

A very famous solution is to consider an additional cost for the wireless connections [23], [9], [24]. The number of hops is a common metric to make the selection between the wired and wireless paths. Considering the additional cost for the wireless paths makes them be less selected by the packets. Therefore, the congestion around the wireless nodes will be decreased.

Although this method is effective in some situations, it has the main drawback, especially, when the network experiences the traffic of real applications. In real-world applications, the traffic pattern is not static and changes over time. Therefore, when a constant cost is considered, it is possible to have no congestion at one time and surprisingly observe huge congestion at another time.

To reduce the noise effect of the wireless links on wired links, several methods have been proposed [34], [9]. For example, the method in [9] uses parallel-plate waveguide as a dedicated structure for transferring wireless signal. Since we consider the architectural view of the NoC, we did not consider these effects in our simulations. It is also important to note that, hierarchical NoC is one of the promising architectures for wireless NoCs. Of course, if the wired path is shorter than the wireless one, the routing algorithm forces the source of flit to send it through wire path.

In this paper, we present a method to balance the traffic for avoiding congestion in the WNoC. In this method, the congestion at wireless nodes is continually checked. Whenever the congestion is detected, it uses a wired connection. To avoid congestion, average flit latency for wireless and wired connections is compared and consequently, the routing will be done. However, we consider some level of randomness in the path selection which makes it possible to react to the changes in the traffic pattern.

For wireless communications, we used the radio-hub component in [25]. It can support one or more wireless channels. We used multiple channels for sending and receiving of the data and acknowledge signals. It is important to note that each wireless node transmits the flits by only a specific channel but can receive all the channels. According to the destination wireless node ID, it can accept or reject the received flits. So, since the frequencies of the transmitters are different, the effect of collision has been neglected.

Many kinds of research on the WNoC used synthetic traffic patterns available in NoC simulators such as Noxim [25] and booksim [26]. In this paper, we ran the experimental results on the real-world applications from the Parsec benchmark suite [27].

Simulation results showed that the proposed routing algorithm can improve the network throughput by 13.9% while reducing the power consumption by 1.4% over related works.

The rest of this paper is organized as follows. In Section 2, the related works are discussed. In Section 3, the proposed method is presented. The experimental results are reported in Section 4, and the conclusions are given in section 5.

Related Works

The XY routing algorithm is one of the traditional deadlock-free deterministic routing algorithms in NoCs [7]. The wireless-XY is an extension of the XY routing algorithm for the WNoCs. Several variants for this algorithm are proposed. The main idea is to use the wireless equipment in the WNoC as much as possible. Since the excessive use of wireless links causes congestion in the wireless nodes, researchers suggest considering a cost for the wireless connections. The value of this cost is based on the packet injection rate of the WNoC. For low packet injection rates, lower values of the cost are needed. But, as the packet injection rate increases, higher values for the cost are required. However, the researchers have considered only a predetermined value for the cost in the previous studies. It is important to note that the wireless-XY algorithm is also deterministic and deadlock-free. The deadlock-freedom is achieved by utilizing virtual channels.

In [28] a Q-tabled-based adaptive routing algorithm is proposed. In this method, each node contains its Q-table, which actions are the selection of the output port and the agent is the packet to be transmitted. The number of rows of the table is equal to the number of nodes in the network. When the current node wants to send a packet to another node, it checks the action values from the corresponding row in the table and selects the lowest value. Then the packet will go through that output port. The values of the table are updated based on the local observation of the packet latency and the Q values of the neighbor nodes. Although the paper gives a very interesting idea, there are some issues with the algorithm. Since this routing algorithm is adaptive and the path of the packets changes over time, it is possible to encounter live lock or deadlock. To alleviate the deadlock problem, the authors used packet-based routing and did not use the well-known wormhole switching. Another problem of this method is the use of information from the adjacent nodes which needs dedicated buses for communication and therefore, leads to more power and area overhead.

In [29], Q-learning is used to propose a congestion-aware routing algorithm for NoCs. In this reference, the network is divided into some clusters and each cluster has a Q-table. Selecting an output channel is based on the density values extracted from the Q-table. Although this method uses wormhole switching, it does not consider wireless network-on-chip.

In [31] authors proposed a wireless NoC architecture that uses on-chip antennas for wireless communication between the long-distance cores. They synthesized and implemented the architecture in Altera Quartus || tool. They proposed a custom routing algorithm for the proposed architecture that uses the vertical and horizontal distance of the source and the destination node as a parameter for selecting the routing path. They used synthetic traffic in evaluation of the proposed routing algorithm.

In [32] four different routing algorithms are compared based on their performance in the wireless NoCs. The evaluations are based on the synthetic traffic patterns. They compared latency, throughput, and wireless utilization of the routing algorithms. The simulation results showed that the XY algorithm achieved the best latency and throughput. It is important to note that they did not consider extra cost on using wireless connections.

In [33] authors proposed an arbitration mechanism for crossbar switches in wireless NoCs. The arbitration mechanism tries to eliminate the port contention in wireless routers. For this new architecture, a routing algorithm is proposed that considers λ as an extra cost for wireless links. They also used synthetic traffic pattern for the evaluation of their proposed method.

Proposed Method

The architecture of a small world NoC with 64 nodes is shown in Fig. 1. The gray squares indicate the position of the wireless nodes. This WNoC uses the mesh topology and has 8 rows and 8 columns. The data in traditional NoCs are transferred based on a router to router mechanism. For example, suppose that the green node wants to send data to the red node in Fig. 1. In each step, the data will be delivered to an adjacent router until it reaches its destination. In the XY routing algorithm, this takes four columns to the right and seven rows to the bottom which takes 11 hops. The wireless technology can be used to reduce the number of hops for delivery of the data for distant node. In this example, using wireless connections reduces the number of hops to three. If a node decides that its packet should pass through the wireless links, it sends the packet to the closest wireless router. The wireless router sends the packet to the closest wireless router to the destination, and that wireless router sends the packet to the destination node.

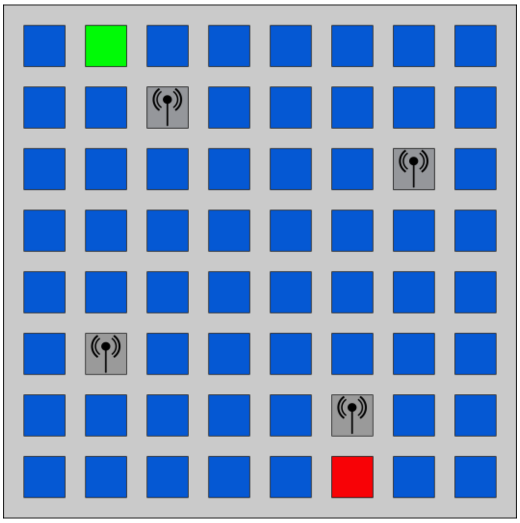


Fig. 1: The schematic view of a 64 node WNoC.

In the small-world WNoC, the wireless nodes are used to minimize the distance between distant nodes. As they can provide faster communication between nodes, the first insight is to send as much traffic as possible through the wireless nodes. This makes congestion around wireless nodes. For two nodes to communicate there are many options, wired connection is one of them, while there are many wireless variants. In wired connection, there are many deadlock-free routing algorithms, XY is one of them. For XY routing, the cost of using it can be computed as the following equation:

$$cost = |sx - dx| + |sy - dy| \tag{1}$$

where sx is the row number of the source node, dx is the row number of the destination node, sy is the column number of the source node and dy is the column number of the destination node. When there are some wireless nodes, another possible path would be to send the packet to a wireless node (usually the nearest one) and then the wireless node, sends the packet to another wireless node (usually the closest to the destination node), and finally, the second wireless node, delivers the packet to the destination node. In this scenario, the cost can be computed as (2):

$$cost = |sx - wnsx| + |sy - wnsy| + |wndx - dx| + |wndy - dy| + \delta \tag{2}$$

In this formula, wns is the wireless node near the source node and wnd is the wireless node near the destination node. The δ is a factor for considering a cost value for using wireless. The first idea is to have no δ value, but in this case, the congestion will happen around the wireless nodes. Many nodes prefer to send their packets through the wireless links and there will be congestion for the source wireless nodes. On the other hand, in comparison with the wireless connections, the wired network bandwidth is so limited. Therefore, the destination wireless node cannot send the packet to the destination immediately. Consequently, this causes the source wireless node to wait until the destination wireless node accepts the packet. The source wireless node keeps sending the same packet until the destination node accepts it. So having a δ value will enforce a cost on using wireless links. Now, consider two nodes that want to communicate together. There are two options for them, use wireless or just use wired connections. To find the optimum option, the costs of the two different paths are computed and the lowest value can be accepted. Having a low value for δ makes congestion around the wireless nodes and considering a high value for the cost prevent using the wireless medium as much as possible. On the other hand, the cost value considered in the previous works is constant and therefore it cannot react to the traffic type. For example, for a specific type of traffic, the cost value 2 may be proper, but if the traffic gets congested, it is not good anymore.

In this paper, we use a congestion-aware routing method that can detect congestion around a wireless node and react to it. For each node, we use a table to determine the routing algorithm. There are two routing algorithms, wired and wireless options. It should be noted that wired routing can be used for all communications, but wireless one cannot be used always. For example, if the destination node is adjacent to the source node, the wireless routing cannot be used. The routing table contains two columns and 63 rows, where the number of rows is equal to the number of the possible destinations for each node. The columns are the average latency of staying in the output buffer of the current router for each routing algorithm. As an example, the table of node 14 is shown in Fig. 2.

Node Id	Wired Latency	Wireless Latency
0	6.3	8.4
1	7.2	10.1
⋮	⋮	⋮
63	14.8	9.2

Fig. 2: Routing table for node 14.

As can be seen in this figure, the average latency of being in the output buffer for the wired routing is less than the latency for the wireless routing for node 0. It means that we should use wired routing for sending packets to node 0. On the other hand, consider the last row of the table which is for node 63. For this node, the wireless routing has lower latency than the wired one. So we should use wireless routing for sending packets to this node. It should be noted that we do not use this greedy idea for selecting the routing algorithm for all the packets. Instead, we use an ϵ-greedy algorithm for selection. In this selection algorithm, we consider ϵ as a non-zero value and with the probability of ϵ we select the routing algorithm with the higher latency. So, with the probability of $1-\epsilon$, we use the routing algorithm with less latency, and with the probability of ϵ, we select the routing algorithm with the greatest latency. Since the traffic pattern of real applications are not uniform, and they change over time, we give a chance to the non-optimum routing algorithm to refresh its latency. For example, as previously mentioned, wireless routing is better for sending packets to node 63. After a while, it may be congestion around the wireless nodes and wired routing becomes the optimum algorithm for sending packets to this node. Usingϵ, we give this chance to the wired routing algorithm to refresh its latency and therefore be selected for the next packets.

Algorithm 1 describes the above-mentioned selection algorithm. In this algorithm, *r* is a random variable created from a uniform distribution between 0 and 1. If *r* is less than or equal to ϵ (line 7) then the non-optimum selection would be done, otherwise, the normal routing will be selected (line 12). The id of the destination node comes from the arrived packet, and the values of *wiredLatency(id)* and *wirelessLatency(id)* can be derived from the routing table.

Algorithm 1 Selection of the routing algorithm

```
1   Inputs:
2         r: the random value between 0 and 1
3         id: the id of the destination node
4   Output:
5         The selected routing algorithm
6   Algorithm:
7     if r <=  ε then
8         if wiredLatency(id) < wirelessLatency(id) then
9            return wireless_algorithm
10        else
11           return wired_algorithm
12    else
13        if wiredLatency(id) < wirelessLatency(id) then
14           return wired_algorithm
15        else
16           return wireless_algorithm
```

To send a packet, we use wormhole scheduling which divides each packet into some flits. Delivering a packet means delivering all of its flits in order. To send a packet, each flit is sent to the output port of the source router. If the flit can be received by the next router, flit sending is done; otherwise, the flit should stay until the next router can pick it. We compute the latency of this waiting time for each flit and accordingly, we calculate the entries of the routing table. When a new latency is observed, we update the corresponding entry in the routing table according to the weighted sum of the old values and the new ones. The new value of the entry can be calculated as (3):

$$newValue = (1-\alpha) * oldValue + \alpha * observedLatency \qquad (3)$$

where the value of α is between 0 and one. If wired routing is selected for the current packet, and then the value of the wiredLatency column will be updated, otherwise, the value of the *wirelessLatency* column will be updated. In the next section, the effect of ϵ and α on the proposed algorithm will be checked.

Deadlock freedom is the most important property of a routing algorithm. If a routing algorithm is not deadlock-free, it is not applicable. It is important to note that the proposed algorithm works just like a normal wireless routing algorithm which is deadlock-free [9]. The wireless algorithm needs a virtual channel to be deadlock-free. One virtual channel is used for regular wired routing and routing to the first wireless nodes. The other virtual channel is used for the flits that used wireless communication.

Therefore, the virtual channel of flit changes during wireless communication. Before wireless communication, its virtual channel is 0 and after that, its virtual channel is one. It should be noted that the routing from the source node to the wireless node and the routing from the wireless node to the destination node, is done using the XY routing algorithm. Therefore, the wireless node acts as a temporary destination node for the source node and a temporary source node for the destination node.

A. *Motivation and Innovation*

As stated by the literature, using wireless links in NoC can lead to having better performance results. These performance results can be achieved by a proper routing algorithm. For example, if all nodes in the NoC try to use wireless links to deliver their data, there will be congestion around the wireless routers, which leads to high latency and lower throughput in the NoC.

To lower the wireless usage, related works use static methods like adding an extra cost to the wireless connections, for selecting between wired and wireless routes. These static methods achieve acceptable results in synthetic traffic patterns which the traffic pattern does not change over time. On the other hand, in real world applications, the network traffic pattern will change over time, so using a static method is not suitable for these applications.

In this paper, we proposed an adaptive routing algorithm in wireless NoCs that responds to the network congestion and improves the network performance. We evaluated the proposed method on real world applications from the Parsec benchmark suite.

The detail design of the routing algorithm is shown in Fig. 3. As can be seen in this figure, the routing table is the main component of the routing algorithm. Each router in the NoC maintains its own routing table. The routing table will be updated when the trail flit of the packet exits from the router.

The *ws* and *w* in Fig. 3 are the estimated wireless latency and wired latency corresponding to the packet's destination (*p.d*), respectively.

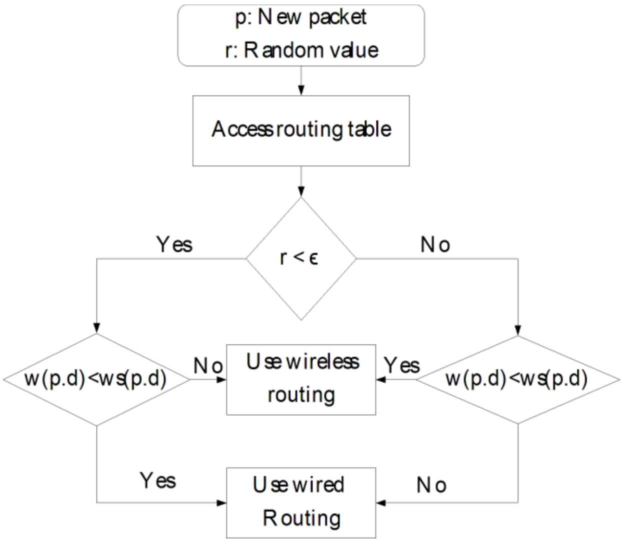


Fig. 3: the flowchart of the proposed routing algorithm.

Results and Discussion

A. *Real-World Benchmark Application*

We use real-world applications which are from the Parsec benchmark suite. It contains scientific benchmark programs that use multi-threaded programming and target the many-core and multicore architectures. Since the NoC architectures act as the communication infrastructure for the future many-core processors, we selected this benchmark suite for the performance evaluation of the proposed method. It is important to note that, many related works on the NoC just use the synthetic traffics generated for simple scenarios like uniform and hotspot traffics which may not occur in real-world applications.

B. *Simulation Environment*

We use a modified version of the Noxim cycle-accurate NoC simulator for experimental results. We modified it to have the virtual channels and to be proper for implementing our routing algorithm. The traffic patterns of the applications are extracted from the Netrace [30] tool and the packets are injected into the Noxim simulator from an implemented interface. Since the Netrace tool is generating traffic for a 64 node NoC, we use a WNoC with 64 nodes to investigate the proposed algorithm.

The parameters of the NoC simulator are shown in Table 1.

Table 1: Simulation parameters

parameter	Value
number of rows	8
number of columns	8
number of the wireless nodes	4
flit width	32

C. *Comparison Metrics*

We used three comparison metrics for comparing the routing algorithms.

The first metric is the network throughput computed as the ratio of the flits injected to the network per core in each cycle. Throughput is the main metric for comparing the routing algorithms. The second metric is the average flit latency.

It is the average value for the latency of delivering a flit. The latency of a flit is computed as the difference between the times of the injection and delivery of a flit. The third metric is the power consumption of the routing algorithm.

Routing algorithms with low power consumption are desirable.

D. *The effect of Cost (δ)*

For the first set of experiments, we investigate the effect of δ on the above-mentioned metrics.

The effect of δ on network throughput is shown in Fig. 4.

As can be seen in this figure, having greater values of δ improves the network latency.

The main reason for this effect is that in such cases, fewer packets are delivered through the wireless connection and therefore there is less congestion around wireless nodes.

It is important to note that having greater values for δ wastes the wireless equipment because very low ratio of the packets go through the wireless connections.

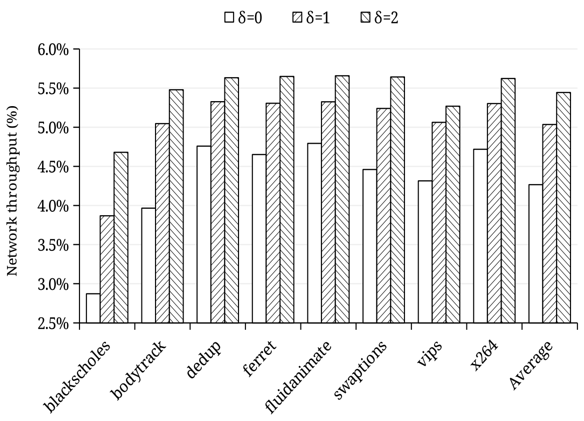


Fig. 4: The effect of δ on the throughput of the NoC.

The effect of δ on the average flit latency is shown in Fig. 5. As shown in this figure, having greater values for δ leads to lower values for average flit latency. The main reason for this effect is having less congestion around wireless nodes. Finally, Fig. 6 shows the effect of δ on the power consumption of the routing algorithm. As can be seen, the average power is almost the same for all routing algorithms which means having δ does not incur additional power consumption on the routing algorithm.

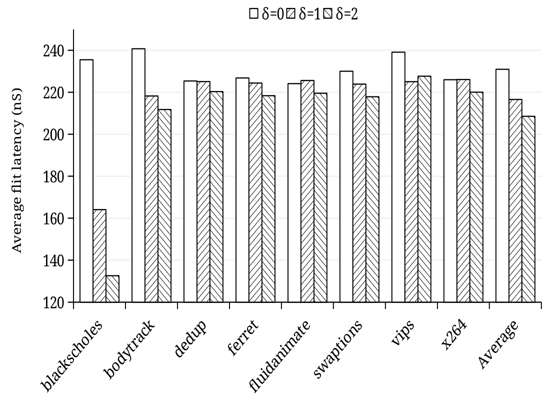


Fig. 5: The effect of δ on the average flit latency of the NoC.

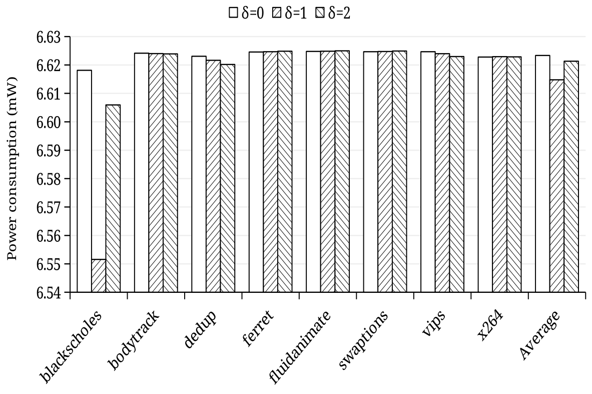


Fig. 6: The effect of δ on the power consumption of the NoC.

E. The effect of ϵ

To investigate the effect of the ϵ, we consider five different values from 0.01 to 0.5 for it. The obtained results have been shown in Fig. 7, Fig. 8, and Fig. 9. Although the different programs have different behavior againstϵ, two facts should be considered: 1) ϵ affects the network throughput of the routing algorithm, 2) the results show that $\epsilon = 0.05$ is the best value from the point of view of the network throughput.

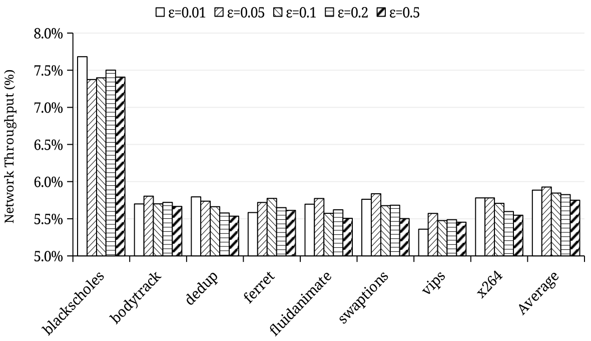


Fig. 7: The effect of ϵ on the throughput of the NoC.

The effect of ϵ on the average flit latency and the power consumption of the proposed algorithm have been illustrated in Fig. 8 and Fig. 9 respectively. As shown in these figures, on average, having a higher value for ϵleads to the greatest latency and more power consumption. Although the average flit latency for ϵ=0.01 and ϵ=0.05 are almost equal, the power consumption of the routing algorithm for ϵ=0.05 is 2.37% higher than the power consumption of the routing algorithm with ϵ=0.01.

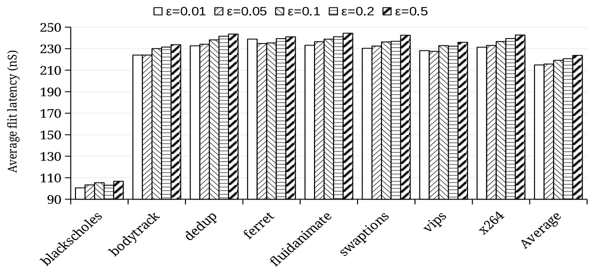


Fig. 8: The effect of ϵ on the average flit latency of the NoC.

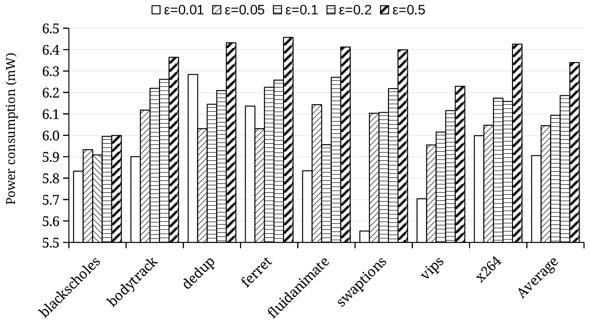


Fig. 9: The effect of ϵ on the power consumption of the NoC.

F. The effect of α

The α value is used to regulate the contributions of the flit latency history and the new flit latency in (3). Higher values for α means that we consider more weight for the new flit latency, and lower values mean that we pay more attention to the history of the flit latency. To investigate the effect of the α on the performance of the proposed routing algorithm, six values of α ranging from 0.1 to 1.0 are considered. Fig. 10 shows the effect of α on the network throughput of the proposed algorithm. As shown in this figure, lower values of α result in higher throughput. An interesting case is for α=0.9 which has better network throughput than α=1.0.

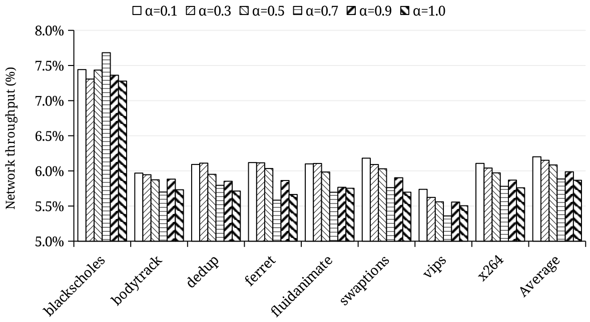


Fig. 10: The effect of α on the throughput of the NoC.

The effect of α on the average flit latency and the power consumption of the proposed method is shown in Fig. 11 and Fig. 12. As shown in Fig. 11, α=0.5 has the lowest average flit latency and α=1.0 has the highest average flit latency. It is important to note that the difference between these two values is less than 4.88%. As can be seen in this figure, the α has regular effect on the average flit latency of the proposed method. On the contrary, the power consumption of the proposed method declines with higher values of α. The power consumption of the proposed method with α=0.1 is the highest, while the power consumption with α=0.9 has the lowest value. Again the difference between the highest and the lowest values is 35.8%. Having low power consumption for α=0.9 gives the idea of a low power routing algorithm alongside a high performance routing algorithm (As in α=0.1).

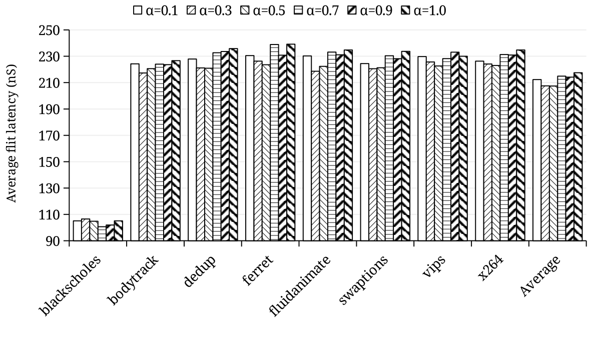


Fig. 11: The effect of α on the average flit latency of the NoC

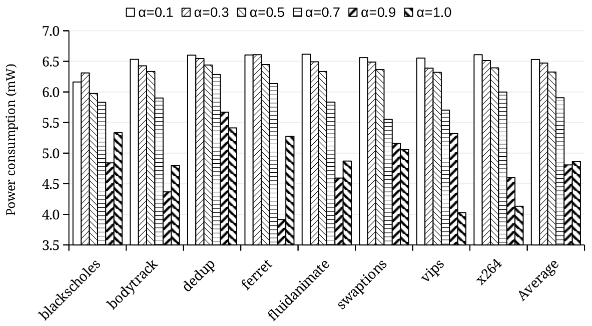


Fig. 12: The effect of α on the power consumption of the NoC.

G. *Comparison to the Related Works*

To compare the proposed method with the related works, we considered two configurations: 1) a high-performance configuration with the best network throughput, and 2) a low power configuration with acceptable network throughput. We compared the proposed configurations to the routing algorithm used in [9], [31], [32], and [33]. It is important to note that the simulation results obtained by the Parsec benchmark for all the methods. The comparison of the network throughput, average flit latency, and power consumption with the related works has been shown in Table 2. As shown in this table, the first configuration has the best network throughput compared to the related works. On the other hand, the second configuration has the best power consumption. It is important to note that the method in [9] achieved the lowest flit latency.

Table 2: Comparison of the proposed method to the related works

Method	Throughput (%)	Latency (nS)	Power (mW)
Proposed Method (α =0.1)	**6.20**	212.32	6.53
Proposed Method (α =0.9)	5.98	214.17	**4.81**
[9]	5.44	**208.51**	6.62
[31]	4.53	229.80	6.61
[32]	4.27	230.96	6.62
[33]	5.20	242.39	6.54

Conclusions

Having congestion around the wireless nodes is the main drawback of the conventional routing algorithms for the WNoC architectures. In this paper, we proposed a learning-based algorithm that uses a history of the flit latency to select the proper trajectory of the flits. To evaluate the proposed method, three metrics were considered: 1) the network throughput, 2) the average flit latency, and 3) the power consumption. We thoroughly investigated the parameters of the proposed routing algorithm and finally, achieved two configurations. The first one is a high-performance configuration that has the best network throughput and the second one has the lower power consumption while having an acceptable network throughput. We investigated the proposed algorithm on eight real-world scientific applications from the Parsec benchmark suite. The simulation results showed that the proposed algorithm can achieve 13.9% more throughput and consume 1.4% less power in comparison with the previous works.

Author Contributions

The main idea of the paper is proposed by A. Tajary and E. Tahanian. The simulation and data analysis are carried out by A. Tajary. A. Tajary wrote the first draft of the manuscript and E. Tahanian corrected it.

Acknowledgment

The authors would like to thank the editor and reviewers for their valuable comments on the 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

NoC	Network on Chip
WNoC	Wireless Network on Chip
sx	The row number of the source node
dx	The row number of the destination node
sy	The column number of the source node
dy	The column number of the destination node
wnsx	The row number of the wireless node near the source node
wnsy	The column number of the wireless node near the source node
wndx	The row number of the wireless node near the destination node
wndy	The column number of the wireless node near the destination node
δ	The cost value for using wireless connection
ϵ	The value of epsilon in ϵ-greedy algorithm
α	The update rate of the algorithm

	based on the current and previous values of the latency
r	A random value with uniform distribution between 0 and 1
ws	The cost of using wireless links
w	The cost of using wired links
p.d	The destination of the packet

References

[1] H. Jang et al., "Developing a multicore platform utilizing open risc-v cores," IEEE Access, 9: 120010–120023, 2021.

[2] P. Kansakar, A. Munir, "A design space exploration methodology for parameter optimization in multicore processors," IEEE Trans. Parallel Distrib. Syst., 29(1): 2–15, 2018.

[3] Y. Liu, S. Kato, M. Edahiro, "Analysis of memory system of tiled many-core processors," IEEE Access, 7: 18964–18977, 2019.

[4] A. Vijaya Bhaskar, T. Venkatesh, "Performance analysis of network-on-chip in many-core processors," J. Parallel Distrib. Comput., 147: 196–208, 2021.

[5] F.N. Sibai, "A two-dimensional low-diameter scalable on-chip network for interconnecting thousands of cores," IEEE Trans. Parallel Distrib. Syst., 23(2): 193–201, 2012.

[6] A. Balakrishnan, A. Naeemi, "Optimal global interconnects for networks-on-chip in many-core architectures," IEEE Electron Device Lett., 31(4): 290–292, 2010.

[7] S.D. Chawade, M.A. Gaikwad, R.M. Patrikar, "Review of xy routing algorithm for network-on-chip architecture," Int. J. Comput. Appl., 43(21): 975–8887, 2012.

[8] B. Bahrami, M.A. Jabraeil Jamali, S. Saeidi, "A novel hierarchical architecture for wireless network-on-chip," J. Parallel Distrib. Comput., 120: 307–321, 2018.

[9] E. Tahanian, A. Tajary, M. Rezvani, M. Fateh, "Scalable thz network-on-chip architecture for multichip systems," J. Comput. Netw. Commun., 2020: 8823938, Dec. 2020.

[10] B. Halavar, B. Talawar, "Power and performance analysis of 3D network-on-chip architectures," Comput. Electr. Eng., 83: 106592, 2020

[11] Z. Wang, H. Gu, Y. Chen, Y. Yang, K. Wang, "3D network-on-chip design for embedded ubiquitous computing systems," J. Syst. Archit., 76: 39–46, 2017.

[12] Z. Liu, G. Song, Y. Zhou, Z. Zhang, F. Cheng, "Layout exploration for three-dimensional networks-on-chip: Chemical equilibrium simulation approach," Microelectron. J., 115: 105195, 2021.

[13] D.A. Hamdi, S. Ghoniemy, Y. Dakroury, M.A. Sobh, "A scalable software defined network orchestrator for photonic network on chips," IEEE Access, 9: 35371–35381, 2021.

[14] K. Sharma, V.K. Sehgal, "Energy-efficient and sustainable communication in optical networks on chip," Sustainable Comput: Inform. Syst., 28: 100426, 2020.

[15] A.B. Ahmed, T. Yoshinaga, A.B. Abdallah, "Scalable photonic networks-on-chip architecture based on a novel wavelength-shifting mechanism," IEEE Trans. Emerg. Top. Comput., 8(2): 533–544, 2020.

[16] D. DiTomaso, A. Kodi, D. Matolak, S. Kaya, S. Laha, W. Rayess, "A-winoc: Adaptive wireless network-on-chip architecture for chip multiprocessors," IEEE Trans. Parallel Distrib. Syst., 26(12): 3289–3302, 2015.

[17] H.K. Mondal, S.H. Gade, M.S. Shamim, S. Deb, A. Ganguly, "Interference-aware wireless network-on-chip architecture using directional antennas," IEEE Trans. Multi-Scale Comput. Syst., 3(3): 193–205, 2017.

[18] N. Mansoor, P.J.S. Iruthayaraj, A. Ganguly, "Design methodology for a robust and energy-efficient millimeter-wave wireless network-on-chip," IEEE Trans. Multi-Scale Comput. Syst., 1(1): 33–45, 2015.

[19] S. Abadal, J. Torrellas, E. Alarcón, A. Cabellos-Aparicio, "OrthoNoC: A broadcast-oriented dual-plane wireless network-on-chip architecture," IEEE Trans. Parallel Distrib. Syst., 29(3): 628–641, 2018.

[20] R.S. Narde, J. Venkataraman, A. Ganguly, I. Puchades, "Intra- and inter-chip transmission of millimeter-wave interconnects in noc-based multi-chip systems," IEEE Access, 7: 112200–112215, 2019.

[21] K. Duraisamy, Y. Xue, P. Bogdan, P.P. Pande, "Multicast-aware high-performance wireless network-on-chip architectures," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., 25(3): 1126–1139, 2017.

[22] A. Tajary, E. Tahanian, "A routing-aware simulated annealing-based placement method in wireless network on chips," J. AI Data Min., 8(3): 409–415, 2020.

[23] W.H. Hu, C. Wang, N. Bagherzadeh, "Design and analysis of a mesh-based wireless network-on-chip," J. Supercomput., 71(8): 2830–2846, 2015.

[24] B. Bahrami, M.A. Jabraeil Jamali, S. Saeidi, "A hierarchical architecture based on traveling salesman problem for hybrid wireless network-on-chip," Wireless Networks, 25(5): 2187–2200, 2019.

[25] V. Catania, A. Mineo, S. Monteleone, M. Palesi, D. Patti, "Improving the energy efficiency of wireless network on chip architectures through online selective buffers and receivers shutdown," in Proc. 2016 13th IEEE Annual Consumer Communications Networking Conference (CCNC): 668–673, 2016.

[26] N. Jiang et al., "A detailed and flexible cycle-accurate network-on-chip simulator," in Proc. 2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS): 86–96, 2013.

[27] C. Bienia, "Benchmarking modern multiprocessors," PhD thesis, Princeton University, 2011.

[28] F. Farahnakian, M. Ebrahimi, M. Daneshtalab, P. Liljeberg, J. Plosila, "Q-learning based congestion-aware routing algorithm for on-chip network," in Proc. 2011 IEEE 2nd International Conference on Networked Embedded Systems for Enterprise Applications: 1–7, 2011.

[29] F. Farahnakian, M. Ebrahimi, M. Daneshtalab, J. Plosila, P. Liljeberg, "Adaptive reinforcement learning method for networks-on-chip," in Proc. 2012 International Conference on Embedded Computer Systems (SAMOS): 236–243, 2012.

[30] J. Hestness, S.W. Keckler, "Netrace: Dependency-tracking traces for efficient network-on-chip experimentation," The University of Texas at Austin, Department of Computer Science, 2011.

[31] M. Devanathan, V. Ranganathan, P. Sivakumar, "Congestion-aware wireless network-on-chip for high-speed communication," Automatika, 61(1): 92–98, 2020.

[32] A.I. Fasiku, B.O. Ojedayo, O.E. Oyinloye, "Effect of routing algorithm on wireless network-on-chip performance," in Proc. 2020 second international sustainability and resilience conference: Technology and innovation in building designs (51154): 1–5, 2020.

[33] F. Rad, M. Reshadi, A. Khademzadeh, "A novel arbitration mechanism for crossbar switch in wireless network-on-chip," Cluster Comput., 24(2): 1185–1198, 2021.

[34] M.S. Shamim, N. Mansoor, R.S. Narde, V. Kothandapani, A. Ganguly, J. Venkataraman, "A wireless interconnection framework for seamless inter and intra-chip communication in multichip systems," IEEE Trans. Comput., 66(3): 389–402, 2017.

Biographies

Alireza Tajary received his B.Sc., M.Sc., and Ph.D. from Amirkabir University of Technology in 2008, 2011, and 2018 respectively. Since 2018, he is with the Faculty of Computer Engineering, Shahrood University of Technology as an Assistant Professor. His main research interests include computer architecture, network on chip, and fault tolerant computing.

- Email: tajary@shahroodut.ac.ir
- ORCID: 0000-0003-0530-9827
- Web of Science Researcher ID: AGY-1402-2022
- Scopus Author ID: 35811245700
- Homepage: https://shahroodut.ac.ir/en/as/?id=S908

Esmaeel Tahanian received the B.Sc. and M.Sc. degrees in communication engineering from K.N.T university of technology, Tehran, Iran in 2008 and 2010, respectively. He received the Ph.D. degree in communication engineering from Shahed university, Tehran, Iran in 2016. His main research interests include computer network especially wireless network and Network on Chip (NoC).

- Email: e.tahanian@shahroodut.ac.ir
- ORCID: 0000-0002-5418-2490
- Web of Science Researcher ID: AGY-1846-2022
- Scopus Author ID: 57221046099
- Homepage: https://shahroodut.ac.ir/en/as/?id=S887

How to cite this paper:

A. Tajary, E. Tahanian, “An adaptive routing algorithm for wireless network on chips,” J. Electr. Comput. Eng. Innovations, 10(2): 487-496, 2022.

DOI: 10.22061/JECEI.2022.8464.518

URL: https://jecei.sru.ac.ir/article_1702.html