In the Name of God

Journal of
Information Systems & Telecommunication
Vol. 11, No.4, October-December 2023, Serial Number 44

Research Institute for Information and Communication Technology
Iranian Association of Information and Communication Technology
Affiliated to: Academic Center for Education, Culture and Research (ACECR)

Print ISSN: 2322-1437
Online ISSN: 2345-2773
Publication License: 91/13216
Editorial Office Address: No.5, Saeedi Alley, Kalej Intersection., Enghelab Ave., Tehran, Iran,
P.O.Box: 13145-799 Tel: (+9821) 88930150 Fax: (+9821) 88930157
E-mail: info@jist.ir , infojist@gmail.com
URL: www.jist.ir

Indexed by:
- SCOPUS www. Scopus.com
- Index Copernicus International www.indexcopernicus.com
- Islamic World Science Citation Center (ISC) www.isc.gov.ir
- Directory of open Access Journals www.Doaj.org
- Scientific Information Database (SID) www.sid.ir
- Regional Information Center for Science and Technology (RICeST) www.ricest.ac.ir
- Magiran www.magiran.com

Publisher:
Iranian Academic Center for Education, Culture and Research **(ACECR)**

This Journal is published under scientific support of
Advanced Information Systems (AIS) Research Group and
Telecommunication Research Group, ICTRC

Acknowledgement

JIST Editorial-Board would like to gratefully appreciate the following distinguished referees for spending their valuable time and expertise in reviewing the manuscripts and their constructive suggestions, which had a great impact on the enhancement of this issue of the JIST Journal.

(A-Z)

- Afsharirad, Majid, Kharazmi University,Tehran, Iran
- Ahmadi, Parvin, Telecommunication Research Center, Tehran, Iran
- AleAhmad, Abolfazl, University of Tehran, Tehran , Iran
- Badie, Kambiz, Tehran University, Iran
- Fathi, Amir, Urmia University, Urmia, Iran
- Fortaki, Tarek, University of Batna, Batna, Algeria
- Fadaeieslam, Mohammad Javad, Semnan University, Iran
- Farsi, Hassan, University of Birjand, South Khorasan, Iran
- Gholami, Mohammad, Babol Noshirvani University of Technology, Mazandaran, Iran
- Gerami, Mohsen, ICT Research Institute, Tehran, Iran
- Ghaffari, Ali, Islamic Azad University, Tabriz Branch, Iran
- Izadkhah, Habib, Tabriz University, Tabriz, Iran
- Marvi, Hossein, Shahrood University of Technology, Semnan Province, Iran
- Mirzaei, Abbas, Islamic Azad University, Ardabil, Iran
- Mohammadi, Mohammad Reza, Iran University of Science and Technology, Tehran, Iran
- Mansoorizadeh, Muharram, Bu-Ali Sina University, Hamedan, Iran
- Mohammadzadeh, Sajjad, University of Birjand, South Khorasan, Iran
- Mohammadian, Ayoub, Tehran University, Tehran, Iran
- Omid Mahdi, Ebadati, Kharazmi University, Tehran, Iran
- Rasi, Habib, Shiraz University of Technology, Shiraz, Iran
- Saadatfar, Hamid, University of Birjand, Iran
- Sedghi, Shahram, Iran University of Medical Sciences, Tehran, Iran
- Sable, Nilesh, Vishwakarma Institute of Technology, Pune, Maharashtra, India
- Soleimanian Gharehchopogh, Farhad, Islamic Azad University Urmia, Iran
- Soran Saeed, Sulaimani Polytechnic University, Kurdistan Region, Iraq
- Tayefeh Mahmoodi, Maryam, Research Institute for Information and Communication Technology, Tehran, Iran
- Teymoori, Mehdi, Zanjan University, Iran
- Tanhaei, Mohammad, Ilam University, Ilam , Iran
- Tashtarian, Farzad, Islamic Azad Mashad University, Mashad, Iran
- Ziaratban, Majid, Golestan University, Gorgan, Iran
- Zayyani, Hadi, Qom University of technology, Qom, Iran

Table of Contents

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 269-281
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Implementation of Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) on Zync FPGA Device

Ahmed Belhani[1*], Hichem Semira[2], Rania. Kheddara[1], Ghada Hassis[3]

[1].Laboratoire Satellites, Intelligence Artificielle, Cryptographie, Internet des Objets « LSIACIO», Constantine 1 University, Algeria
[2].Electronics and New Technologies Laboratory (ENT), University of Oum El Bouaghi, Algeria
[3].Department of Electronics », Constantine 1 University, Algeria.

Received: 24 Jan 2022/ Revised: 04 Sep 2022/ Accepted: 02 Oct 2022

Abstract

The non-orthogonal access schemes are one of the multiple access techniques that are candidates to become an access technique for the next generation access radio. Power-domain non-orthogonal multiple-access (NOMA) is among these promising technologies. Improving the network capacity by providing massive connectivity through sharing the same spectral resources is the main advantage that this technique offers. The NOMA technique consists of exploiting the power domain which multiplex multiple users on the same resources applying a superposition coding then separating the multiplexed users at the receiver side. Due to the non-orthogonality access technique, the main disadvantage of NOMA is the presence of interferences between users. That is why this scheme is based on a successive interference cancelation (SIC) detector that separates the multiplexed signals at the receiver. In this paper, an embedded system is considered for designing and implementation of the power-NOMA For two users. The implementation is realized by employing a Zynq FPGA (Field programmable gate array) device through the Zybo-Z7 board using MATLAB/Simulink environment and Xilinx System Generator. The features offered by this device, hemps to consider the design of an uplink and a downlink scenario over Rayleigh fading channel in additive white Gaussian noise (AWGN) environment.

Keywords: Non-Orthogonal Multiple Access (NOMA); Successive Interference Cancelation (SIC); Multi-use Detection; Bit Error Rate (BER); QPSK; BPSK; Xilinx System Generator (XSG).

1- Introduction

The phenomenal development in communication systems and Internet of Things (IoT) has increased the demand for better internet connections, higher data rates, less latency, fairness, better Quality of Service (QoS), and reduced interference [1][2].

Communication networks in the next generation (5G and beyond) aspire to achieve high efficiency [2], through more users in the limited available spectrum [1]–[5] by applying a new Medium Access Control (MAC) called NOMA technique[5]. In recent years, NOMA-IoT has attracted many researchers in academia [5]–[8]. This growing interest is due to the fact that the NOMA technique is a good candidate to address some of the challenges regarding the radio spectrum scarcity. Dealing with this shortage is one of the challenges to ensure a massive machine-type communication (MMTC) (e.g., IoT framework) by attaining ultra-low latency and high reliability.

The conventional technique OMA (Orthogonal multiple access) permits multiple access by limiting the number of users to the number of available resources based on the orthogonality between users [5], Some of these techniques are, time division multiple access (TDMA), frequency division multiple access (FDMA), and code division multiple access (CDMA)[9]. These techniques are the reasons why the OMA scheme cannot meet the demands of concurrent users due to the resource shortage, especially against the increasing number of connected devices in the IoT network [10]. Contrarily, NOMA uses the same frequency block and the same time slot by assigning different power levels or by multiplexing signals in code domain [1][2], which helps to enhance the bandwidth reuse and connectivity density [6].

✉ **Ahmed BELHANI**
ahmed.belhani@umc.edu.dz

While OMA systems use orthogonal resource allocation among users to avoid interference [1], the NOMA technique intentionally introduces interference because of its principle of access and superposition of users' data. Consequently, the successive interference cancellation (SIC) technique has been widely adopted in the NOMA literature as a solution to eliminate the inter-user interference. In SIC, the first step consists of estimating the strongest signal by treating other superposed signals as noise. In the second step, the decoded signal is subtracted from the received signal to retrieve the weakest signal [8].

From an information theoretic viewpoint, the fundamental potential of NOMA over OMA has been proved. Indeed, most of the analytical studies handled in terms of information-theoretical perspectives, have revealed the efficiency of the SIC detector in both schemes uplink-NOMA and downlink-NOMA e.g. [11]–[13]. These analyzes are based on the SINR (Signal to Interference and Noise Ratio) definition and then they provide theoretical limits for throughput or outage probability. However, in throughput or outage analysis the SIC algorithm is not implemented. Therefore, the analysis could only give inferences for theoretical limits. In other words, the NOMA system with its transmitter and receiver (SIC algorithm) is not implemented or simulated while assuming perfect SIC. Hence, in real world application, to analyze the error performance of the NOMA schemes (in terms of bit error rate BER), we should implement an IQ (in-phase and quadrature inputs) modulator at the transmitters and based on the superimposed received signal, we should implement an IQ detector to decide the transmitted symbols (bits) of each user. This implementation could give us a clear idea about the imperfections of the detection technique (i.e. subtraction of the erroneous detected symbol) [13]

To study the efficiency of the SIC receiver, a practical downlink-NOMA system based on LTE specifications using an open-source software platform named Open Air Interface have been investigated in[3]. The experiment results of the work demonstrate that the NOMA scheme has a significant throughput gain compared with an orthogonal multiple access (OMA) scheme.

Currently, the use of FPGAs in research and development of embedded systems is applied to specific tasks is increasing [15], for many valuable advantages such as its great flexibility [15], which allows the re-use of any circuit as desired in different structures in a very fast time, high clock frequency, high operations per second, and parallel processing [15]. FPGAs are widely used in the communication field to benefits their capacity, function and reliability [16] such as the space-time coding and decoding algorithms for MIMO Communication system presented in [17], and the FPGA implementation of a Pseudo Chaos-signal generator for secure communication systems in [18].

In addition, the implementation of differential frequency hopping communication system is studied in [19]. In [20], NOMA technique is experimentally investigated over AWGN channel using FPGA. The authors propose an implementation based on VHDL programming for two users in the downlink scenario. In addition to the design of the transmitter and the SIC in receiver, the authors relied on a Box-Muller method for noise generation. Furthermore, the performance is verified by comparing the Monte-Carlo simulation results via MATLAB showing exact matching with that of the designed NOMA system. However, this study is only limited to the downlink scheme over AWGN environment which is not reasonable for practical situation, where the channel coefficients are consistently present to affect the received signal (path loss, fading and shadowing). In this paper, we propose the implementation on FPGA of a real-time transmission concept of the power domain NOMA for both scenarios: downlink and uplink over Raleigh fading channel. At first, derived from Monte-Carlo simulation via MATLAB, we study the performance of power NOMA technique in both schemes using QPSK and BPSK modulation. Hence, for proper implementation and avoiding overlap between the users' symbols, the power allocation principle is assured for the downlink scheme, on the other hand, we keep the uplink scheme without scheduling of power allocation as the users use their own power.

Afterwards, with the assumption of two users like the work of [20], and based on Xilinx system generator (XSG) which is a high-level design tool that allows the use of the MATLAB / Simulink environment, we design the functional blocs for an implementation of the power NOMA technique on the FPGA Zybo Z7 card for both schemes (downlink and uplink). This implementation is made simple by using Xilinx library to create elements allowing FPGA hardware design without resorting to HDL programming [21].

The rest of the paper is organized as follows. Section 2 presents some assumptions used in the work, while section 3 illustrates the system and the channel models. section 4 shows the simulation results obtained for both link: downlink and uplink over Rayleigh fading channels and AWGN. In section 5, we describe the FPGA card used in the work and the waveforms of the system generator models with the RTL schemes. Finally, a conclusion is presented in section6.

2- Assumptions

In this work, we assume the presence of two users in the network, this assumption is due to the fact that regardless

of the number of users, the remote user always considers other users signals as just additive noise as defined by NOMA theory [20]. Moreover, we assume that the power allocation factors a_N and a_F used for the two users are known, and that the path is non-selective.

In the uplink scenario, the base station is free of scheduling the power allocation as long as each user uses and manages its own battery. Hence, we assume that each user uses its own battery power to the maximum with a factor equal to 1.

3- NOMA Model

This section presents the NOMA model in power domain for both links.

3-1- Downlink NOMA

In a downlink scheme shown in Figure 1, we consider two users equipped with one antenna served by a base station (BS) equipped with a single transmitting antenna (Single Input Single Output (SISO)). The BS transmits with power P the signals of users superposed on the same resource block. The superposition is realized at the BS by adding the different symbols after power scaling both symbols are combined using superposition coding principle (SC); the more power is allocated to the far user (FU); the less power allocation ratio is allocated to the near user (NU), i.e., the information for both users is multiplexed in the BS using the same resource block (frequency/time) by allocating each user a different power coefficient by virtue of its channel conditions. The power allocation ensures fairness for all users by maintaining a specific BER thresholds for each user [22]. Hence, the signals $y_k \in \mathbb{C}$, $k \in (F, N)$, received by users are described as shown by Eq. (1):

$$y_k = h_k\left(\sqrt{a_F P}s_F + \sqrt{a_N P}s_N\right) + n_k, \quad k \in (F, N), \tag{1}$$

where h_k denotes the flat fading channel coefficients between the BS and the kth user, i.e. $k \in (F, N)$. The independent and non-identical Rayleigh fading channel coefficients h_k, $k \in (F, N)$ are modeled as complex Gaussian random variables with zero mean and variance $E[|h_k|^2]$, i.e., $h_k \sim CN(0, E[|h_k|^2])$, $k \in (F, N)$, where $|.|$ denotes the absolute value and $E[.]$ is statistical expectation. At each kth user the channel coefficients h_k, $k \in (F, N)$ is supposed to be known, such that $|h_F| < |h_N|$. As a consequence, the power allocation factors a_N and a_F are assigned such that $a_N < a_F$, where $a_F + a_N = 1$. $n_k \sim CN(0, N_0/2)$, $k \in (F, N)$, is the additive white Gaussian noise (AWGN) which is added at each node (each user's receiver) independently.

The information bits are modulated as complex symbols s_F, s_N using a phase modulated signals for both users, they can be a QPSK modulation signals which can take four possible complex values defined as $s_k = \pm 1 \pm j1$, $k \in (F, N)$, resulting from the mapping of two bits, or a BPSK modulation signals which can take one of the two possible symbols ± 1.

In the network, all users receive the same signal, which contains the information of all users. This requires a suitable detector to eliminate interference in each equipment. Since a large part of the power has been allocated to the FU, a maximum likelihood detector (MLD) is sufficient to detect the information of this user by considering the signal of the NU as interference. On the other hand, the NU must first detect the interference signal (FU) as its own symbol is noise. In the second step, it removes the detected symbol from the received signal y_N in order to detect its own symbol. This procedure is known as successive interference cancelation (SIC), and it is clear that the performance of the system depends on the channel qualities and the power allocation coefficients [13]

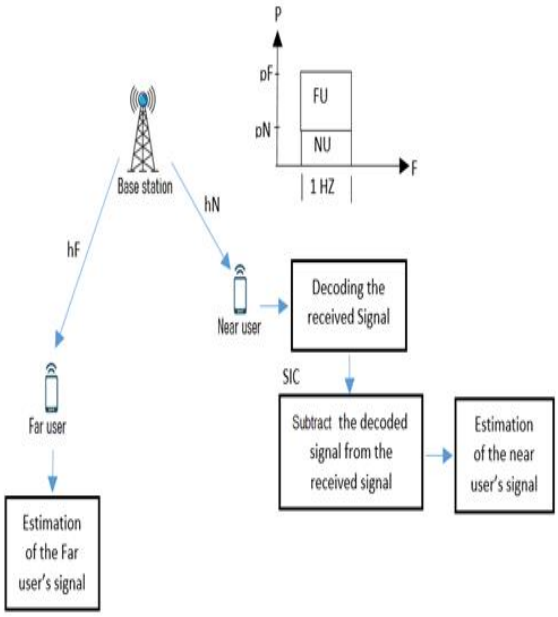


Fig. 1 Downlink Power Domain NOMA for Two Users with SIC Technique.

3-2- Uplink NOMA

As shown in Figure 2, Similar to the downlink scenario, the uplink scheme is composed of one base station and two users. Even in this scenario, we assume that each node is equipped with a single antenna (SISO-NOMA). With transmit power P_k, $k \in (F, N)$, limited by the

maximum power of the user battery, each user can independently use its own power P_k to transmit its symbols on the same frequency block within the same time slot. Therefore, the received signal $y \in \mathbb{C}$ at the BS is given by Eq. (2):

$$y = h_N\sqrt{P_N}s_N + h_F\sqrt{P_F}s_F + n \qquad (2)$$

where the Rayleigh fading channel coefficients $h_{k,}$ $k \in (F, N)$, between the users and BS are independent and identically distributed. s_N and s_F are the complex modulated information from QPSK or BPSK modulation. n denotes the AWGN at BS receiver. Like the downlink. scenario it is assumed that $\sqrt{P_F}\ |h_F| < \sqrt{P_N}|h_N|$. By virtue of the quality of the channel and the level of received power, firstly, the BS attempts to detect the symbols of the NU using MLD by considering the signal received from the FU ($h_F\sqrt{P_F}s_F$) as noise. Secondly, after having removed the detected symbol ($h_N\sqrt{P_N}\hat{s}_N$, where $\hat{s}_N$ denotes detected s_N), the BS attempts to detect the symbols $\hat{s}_F$ using MLD

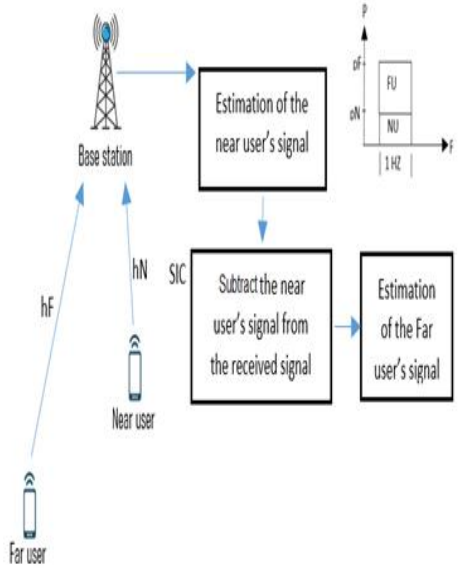


Fig. 2 Uplink Power Domain NOMA for Two Users with SIC Technique.

4- Simulation and Main Results

4-1- Downlink

Figue 3 shows the flowchart of the downlink-NOMA system. Firstly, we generate a random data for both users then we modulate them using either a QPSK modulation or a BPSK modulation, after, we multiply the two signals by the power allocation factors a_N and a_F.

In the next step, we combine the two signals to construct the NOMA signal x_{NOMA} multiplied by the Rayleigh coefficients for each user, and then we start the estimation process after adding the white noise for each user separately.

The FU estimates its data directly from the received signal via a minimum-distance criterion (MLD), whereas the NU utilizes the SIC technique to obtain its symbols.

To illustrate the performance of the NOMA system in the downlink scheme, we use the parameters presented in the table1.

Table 1: Simulation Parameters

Monte Carlo runs		10000
Power allocation factors for the near user		0.2, 0.4
Power allocation factors for the far user		0.6, 0.8
System frequency		15360000 Hz
Rayleigh Channel parameters	Path Delays	0 S
	Average Path Gains	FU: -3 dB
		NU: 0 dB
	Maximum Doppler Shift	0 HZ

Figures 4, 5 and 6 show the simulation results for the bit error rate (BER) against the Signal to Noise ratio per bit (Eb / No) for different scenarios. In figure 4, we consider a QPSK modulation for both users with different power allocation. We can easily observe the influence of the power allocation coefficients on the performance of the system. The more the symbols of the different users are well separated in power, the more we prevent the constellations overlap of different users. Furthermore, the curves show that the BER of FU is better than that of NU for $a_N = 0.2$ and $a_F = 0.8$, which shows the compensation of the effect of the channel with the enhancement of the symbols power for the FU.

In the second scenario in which the same power allocation factors are maintained, we consider two cases of modulation: Fixed QPSK modulation for both users, and adaptive modulation for the second case. In adaptive modulation, a BPSK modulation is chosen for the FU which suffers from the worst channel condition, and a QPSK modulation is used by the NU which benefits from high channel quality. For different power allocation factors, the Figures 5 and 6 show clearly the improvement of the BER for the adaptive modulation compared to the fixed modulation. In addition to avoiding overlap between the symbols of different users, the low order modulation compensates the effect of the channel by increasing the power of the symbols and also by helping the detection threshold to be improved when using the MLD for both users.

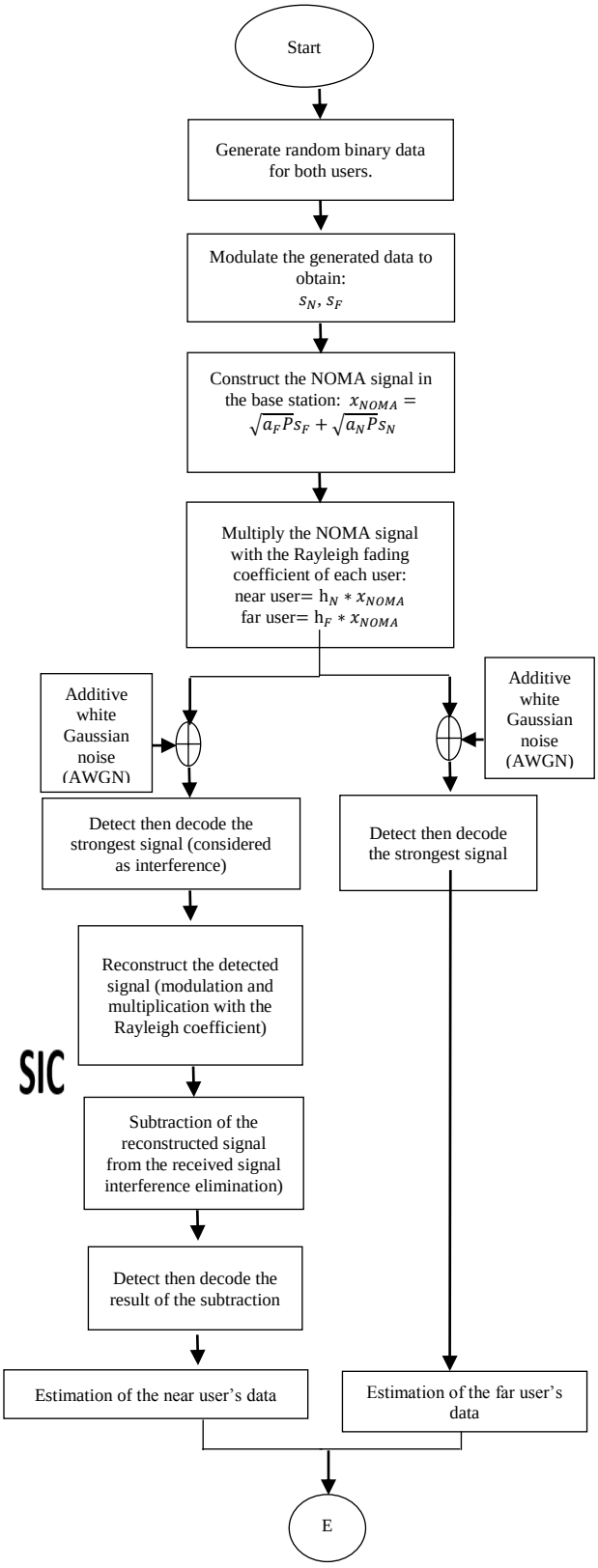


Fig. 3 Flowchart of the Downlink NOMA.

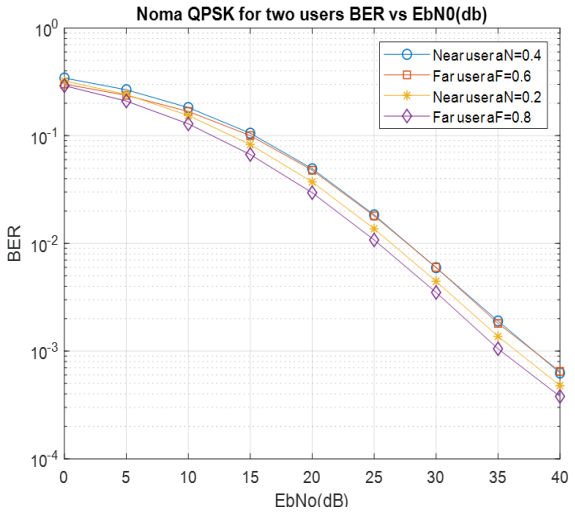


Fig. 4 BER vs E_bN_0 for the Downlink Power Domain NOMA Transmission Using the QPSK Modulation and the Power Allocation Factors of a_N = 0.2, 0.4, a_F = 0.6, 0.8.

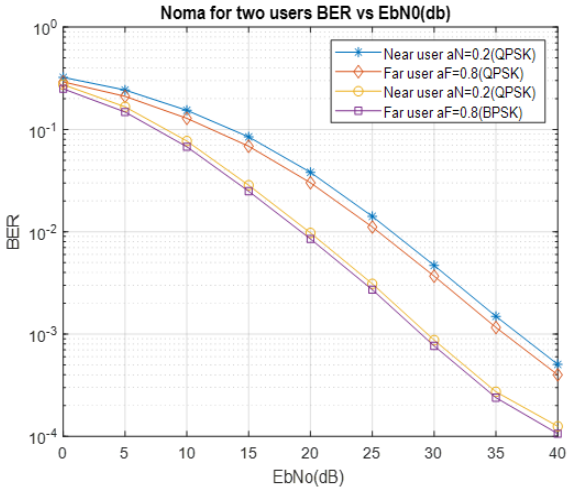


Fig. 5 BER vs E_bN_0 for the Downlink Power Domain NOMA Using the QPSK Modulation and the Power Allocation Factors of a_N = 0.2, a_F = 0.8.

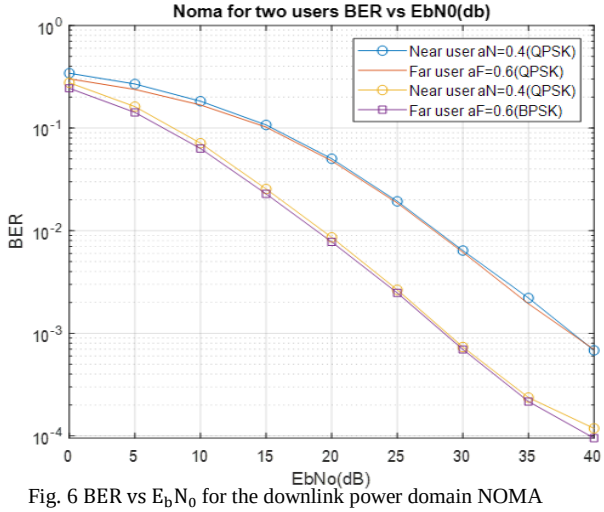


Fig. 6 BER vs E_bN_0 for the downlink power domain NOMA Transmission Using the QPSK and the BPSK Modulations and the Power Allocation Factors of a_N = 0.4, a_F= 0.6.

4-2- Uplink

Figure 7 shows the flowchart of the uplink–NOMA system. Like the previous scheme, we randomly generate the binary data for each user then we modulate them by using either a QPSK modulation or a BPSK modulation according to the channel quality. Unlike the previous diagram, in the uplink-NOMA scheme, each user sends its own data using its own power (power battery) toward the BS, which results in multiplication by independent Rayleigh coefficients and power. In consequence, the two signals are firstly multiplied by different Rayleigh coefficient (h_N for the NU and h_F for the FU) and Powers (P_N for the NU and P_F for the FU) according to their distance from the base station and their own battery. After that, the signals are multiplexed at the BS. Finally, the estimation process begins after the white noise is added.

The base station estimates the NU data directly through the ML detector by considering the FU data as noise. In the next step, the BS tries to detect the far user's symbols via MLD after subtracting the detected signal (previous step) from the received signal.

To illustrate the behavior of the SIC algorithm with the performance of the uplink-NOMA scheme, we adopt the same simulation parameters of the downlink-NOMA scheme. The figures 8 and 9 show the performance of the BER versus EbN0 in the uplink-NOMA system for two types of modulations QPSK and BPSK. To show the performance, we present two scenarios: fixed modulation and adaptive modulation in proportion to the quality of the channel. In all the simulation, we assume that $P_N = P_F$, which reflects a practical situation where the two users transmit with the same power like the IoT devices.

It can be seen from the two figures that the performance of the FU is better than the performance of Nu in terms of BER. This superiority is due to the use of SIC detector at the base station to estimate the FU's data, which suppresses the strong signal assumed to be interference. On the other hand, the base station only uses the maximum likelihood detector to extract the NU's data, which suffers from the presence of FU'S symbols considered as noise. Despite these results, we can notice from all the curves in uplink-NOMA scheme that the SIC presents a poorer performance in high SNR regime and clearly suffers from the error floor [13][23][24]. According to the figures 8 and 9, we observe that increasing the transmit power (high SNR) or using low order modulation (BPSK) at the receiver cannot remove the error floor. Indeed, whatever the improvements, the SIC detector remains unable to detect the symbols of the different users in a high SNR regime, which is an inevitable result because other signals are treated in a sequential manner on the basis that they are noise.

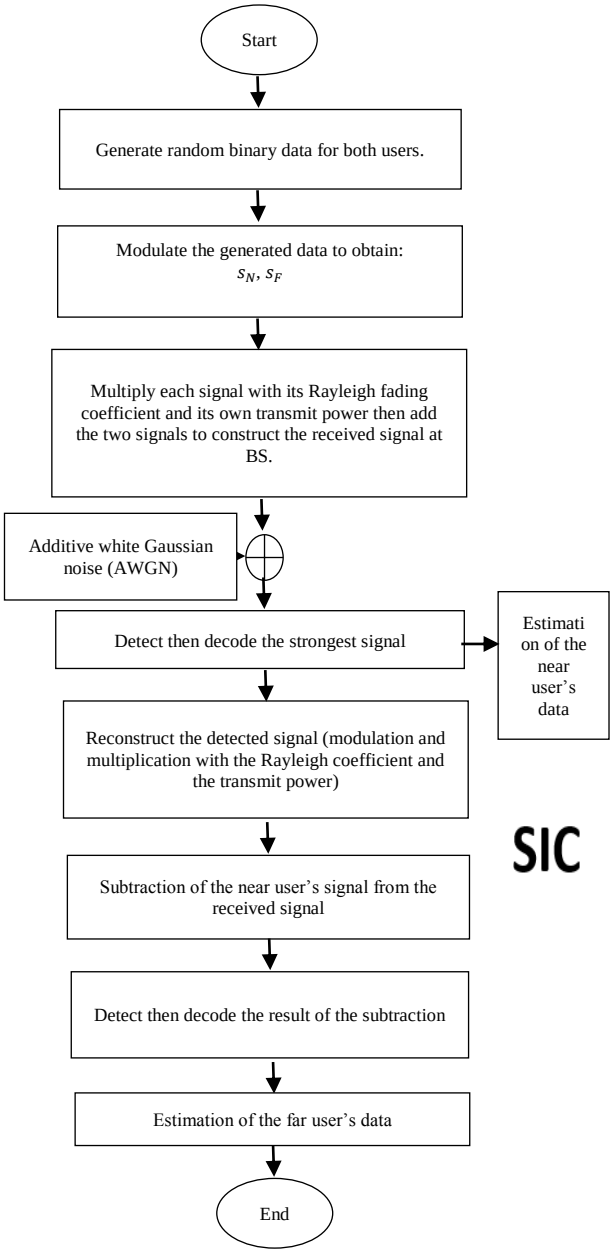


Fig. 7 Flowchart of the Uplink NOMA.

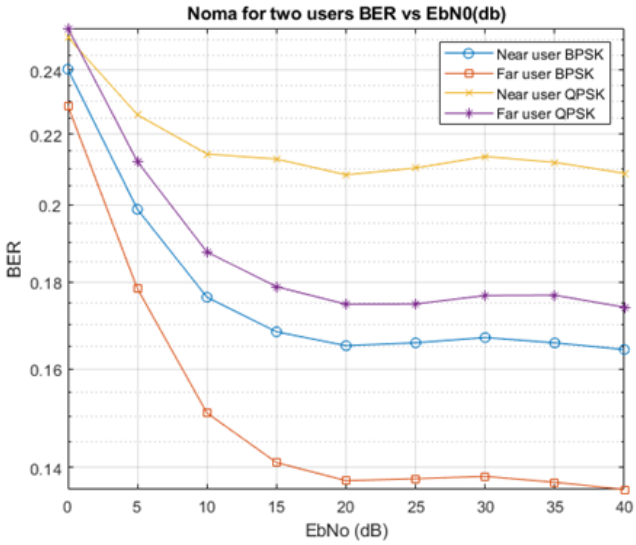


Fig. 8 BER vs E_bN_0 for the Uplink power domain. NOMA Transmission With the Using the Same Modulation for Both Users QPSK or BPSK Modulation.

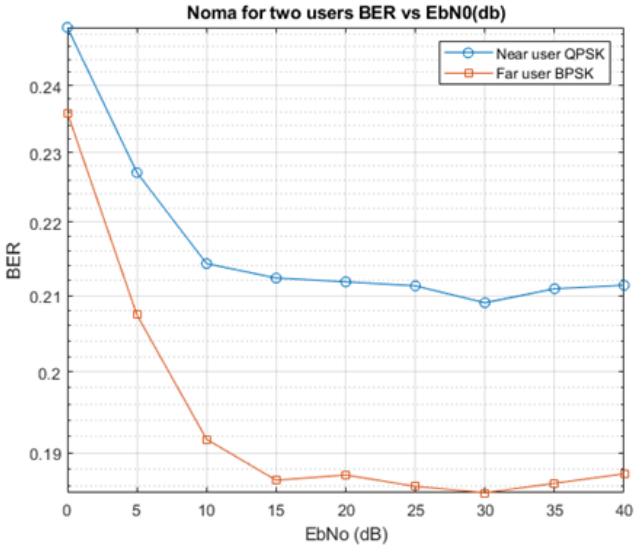


Fig. 9 BER vs E_bN_0 for the Uplink Power Domain. NOMA Transmission Using the QPSK Modulation for the NU and the BPSK Modulation for the FU.

5- Implementation of NOMA Models Using System Generator and ZYBO-Z7 Xilinx Card

In this section a Zybo Z7 card based on Zynq-7020 device is used to implement the power domain NOMA technique for both schemes: downlink and uplink. This choice is based on its features and connectivity [15] illustrated by table 2. The power tool system generator of Xilinx is combined with Matlab/Simulink to design the NOMA signal and build VHDL project, which is transferred to VIVDAO suite for waveform generation through test bench

Table 2: Zybo-Z7 Characteristics

ZYBO Z7 features	ZYBO Z7 specifications	ZYBO Z7 connectivity
667MHz dual-core Cortex-A9 processor with tightly integrated Xilinx FPGA	Clock Resources Zynq PLL with 4 outputs 4 PLLs (2 PLLs*) 4 MMCMs (2 MMCMs*) 125 MHz external clock	Wide range of USB, Ethernet
High-bandwidth peripheral controllers: 1G Ethernet, USB 2.0, SDIO)	Block RAM:630 KB (270 KB*)	Pcam camera connector
Programmable from JTAG, Quad-SPI flash, and microSD card	Internal ADC Dual-channel	HDMI sink port (input)
DDR3L memory controller with 8 DMA channels and 4 High Performance AXI3 Slave ports	DSP Slices :220 (80*)	6 Pmod ports (5*) :8 Total Processor I/O ,40 Total FPGA I/O (32*) , 4 Analog capable - 1.0V differential pairs to XADC

5-1- Downlink Implementation

Figure 10 shows the NOMA design using Xilinx system generator tools for the downlink, contains: The BPSK modulator, the BSPK demodulator, the transmitted NOMA signal, the received Far user signal and the received near user signal. We have also maintained the same parameters shown in table 1

5-2- BPSK Modulator/Demodulator

The logic elements are used to build all components, like multiplexer, multiplier, adder and delay element.
Figure 11 shows the BPSK modulator designed through XSC tools [25].
For the BPSK demodulator, we use the structure defined in [26] and accomplished it by adding a low pass filter design with eleven coefficients presented in table 3 to get more precision. Figure 12 shows the demodulator released through the DSP's FDA tool from XSG toolbox

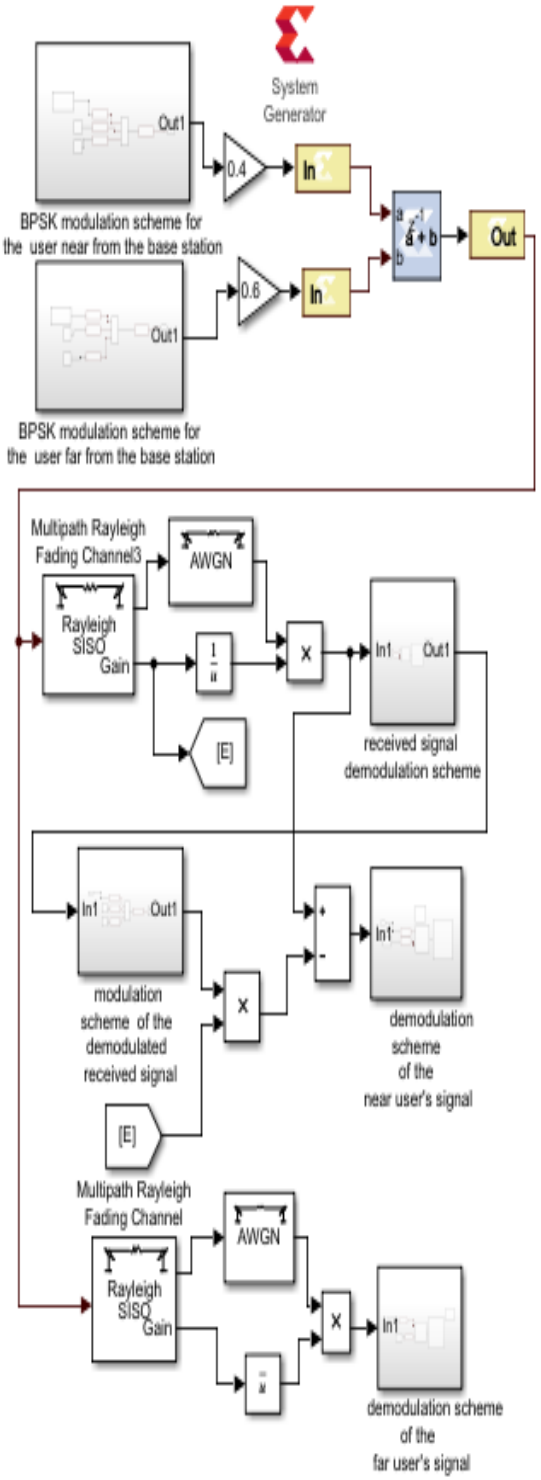


Fig. 10 NOMA Design for the Downlink Using Systems Generator and Xilinx Tools.

Table 3: Low Pass Filter CoefficientS

Coefficient number	coefficient value	Coefficient number	coefficient value
1	-0.0732	7	0.1763
2	0.1712	8	0.1601
3	0.1469	9	0.1469
4	0.1601	10	0.1712
5	0.1763	11	-0.0732

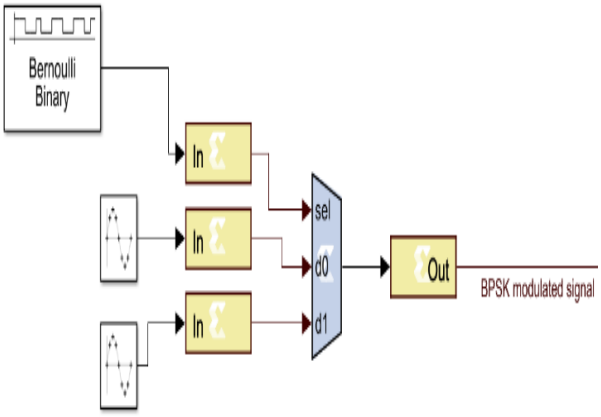


Fig. 11 BPSK Modulator.

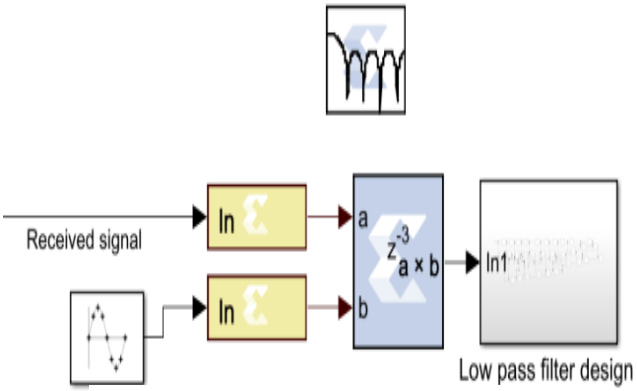


Fig. 12 BPSK Demodulator

Figures 13 and 14 show clearly the BPSK modulator and demodulator waveforms for the designed models in the downlink. The choice of using a BPSK modulation in the XSG model is only to simplify the design and the results; we can use any type of modulation, for example QPSK modulation[20].

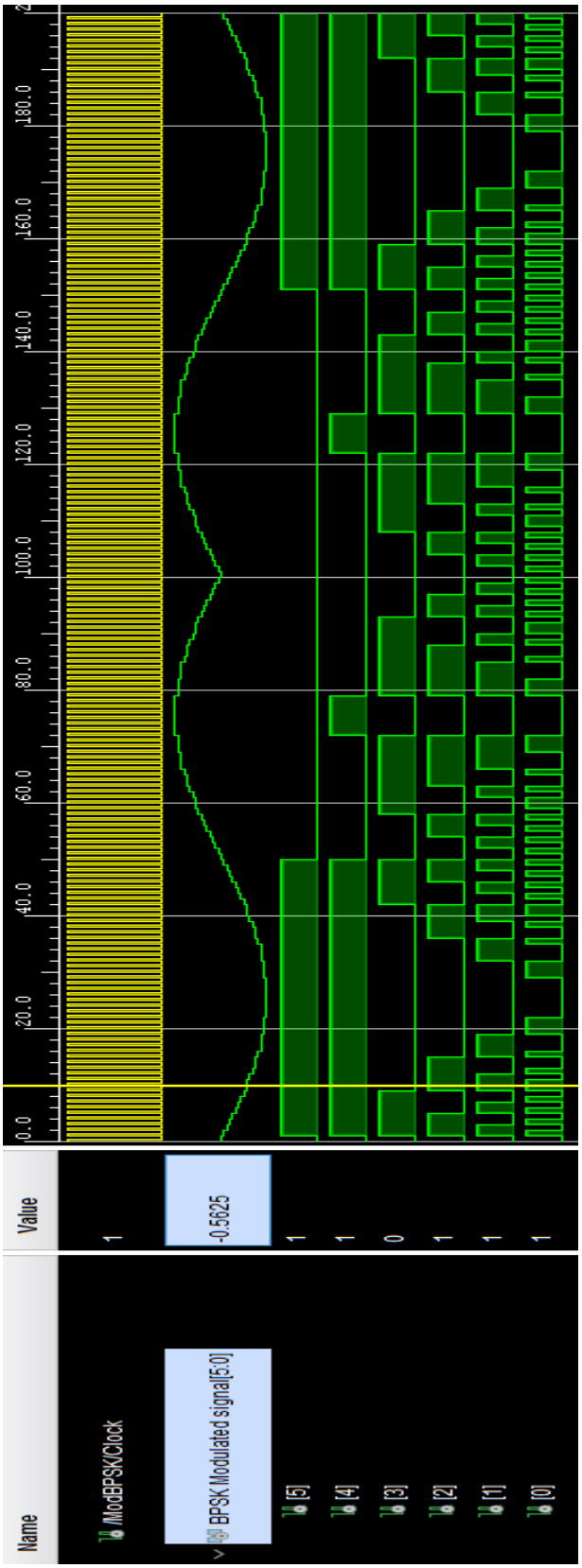


Fig. 13 BPSK Modulator Waveform

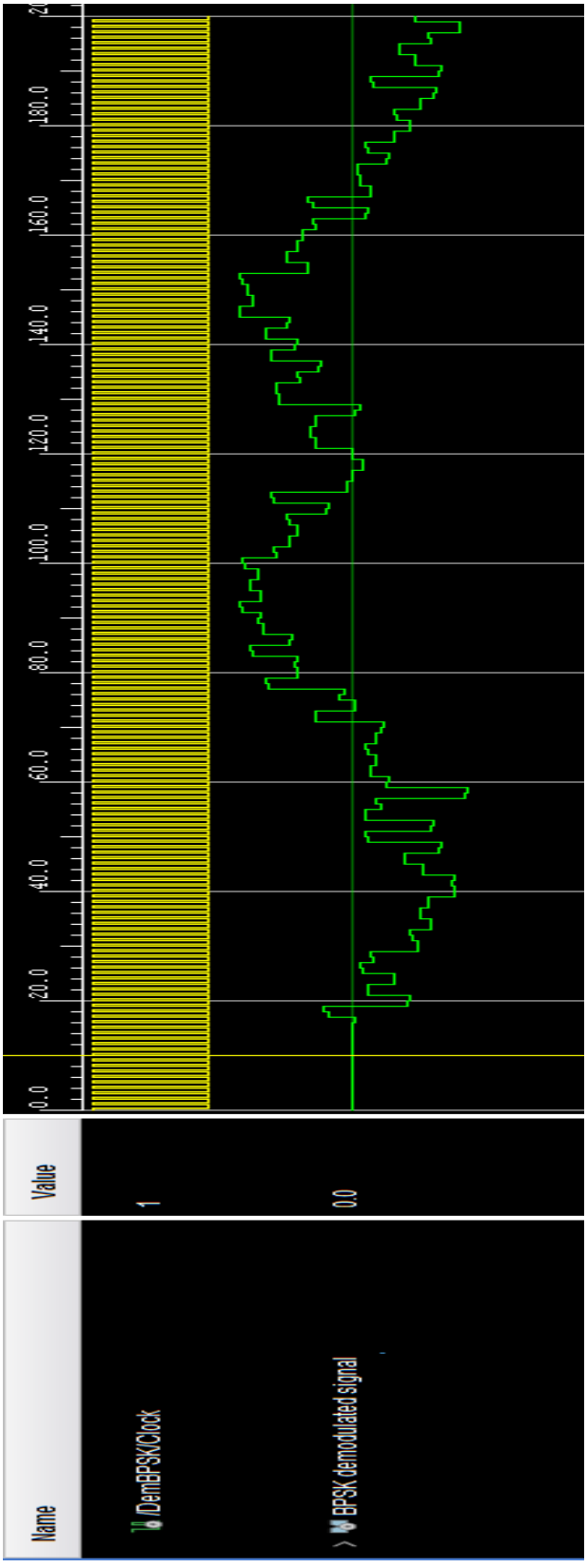


Fig. 14 BPSK Demodulator Waveform

5.2.1 NOMA Signal

Figure 15 represents the NOMA signal design for the downlink, where we multiply the BPSK modulated signal of each user with the power allocation factors, and we combine the two signals to obtain the NOMA signal, the result of the waveform is illustrated in figure 18.

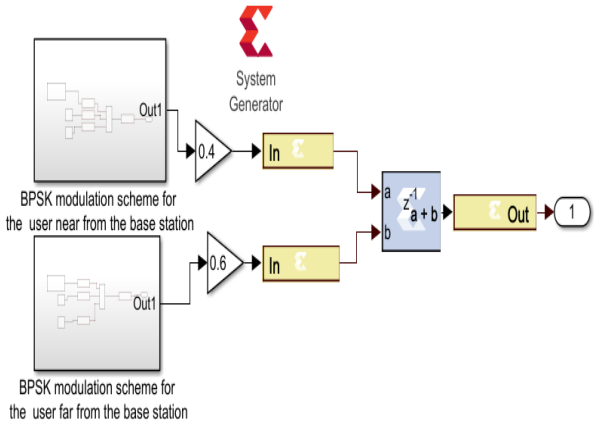


Fig. 15 NOMA Bloc Design for Downlink.

5.2.2 Far User and Near User

Figures 16 and 17 represent the two designs for the estimation process mentioned above and explained in the flowchart of the downlink. In the other hand, for the far user we implement the SIC technique to estimate the near user's signal (figure17).
In figure 18, we directly demodulate the far user's signal after passing through Rayleigh fading channel and AWGN.

5-3- Uplink Implementation

Figure 19 shows the NOMA design using Xilinx system generator tools for the uplink, which contains the same blocks used in the downlink.

5.3.1 NOMA Signal

Figure 20 represents the NOMA signal design for the uplink, which contains the two users BPSK modulated signals and the Rayleigh fading channels for each user combined and passed through Rayleigh and AWGN channel.

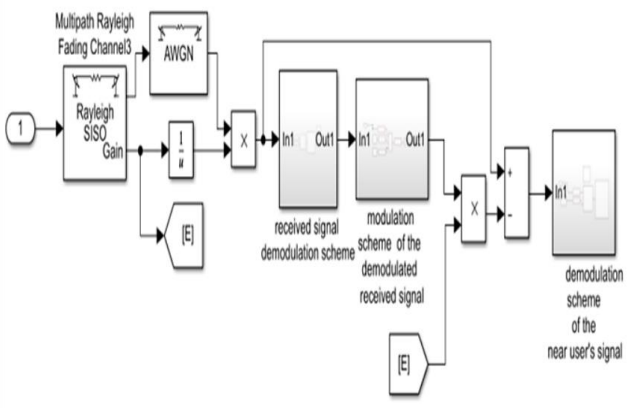


Fig. 17 Near User Reception Design for Downlink.

5.3.2 Far User and Near User

Figures 21 and 22 represent the two designs for the estimation process mentioned above and explained in the flowchart of the uplink; we first estimate the near user's signal in figure 21, afterwards, the far user's symbols are detected using SIC procedure using the same received signal as the SIC is implemented in BS.

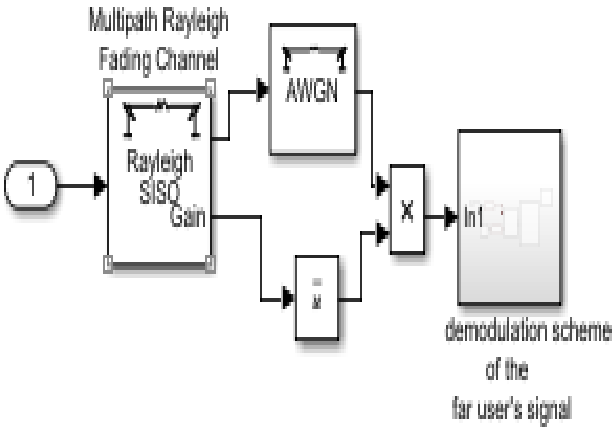


Fig. 16 Far User Reception Design for Downlink.

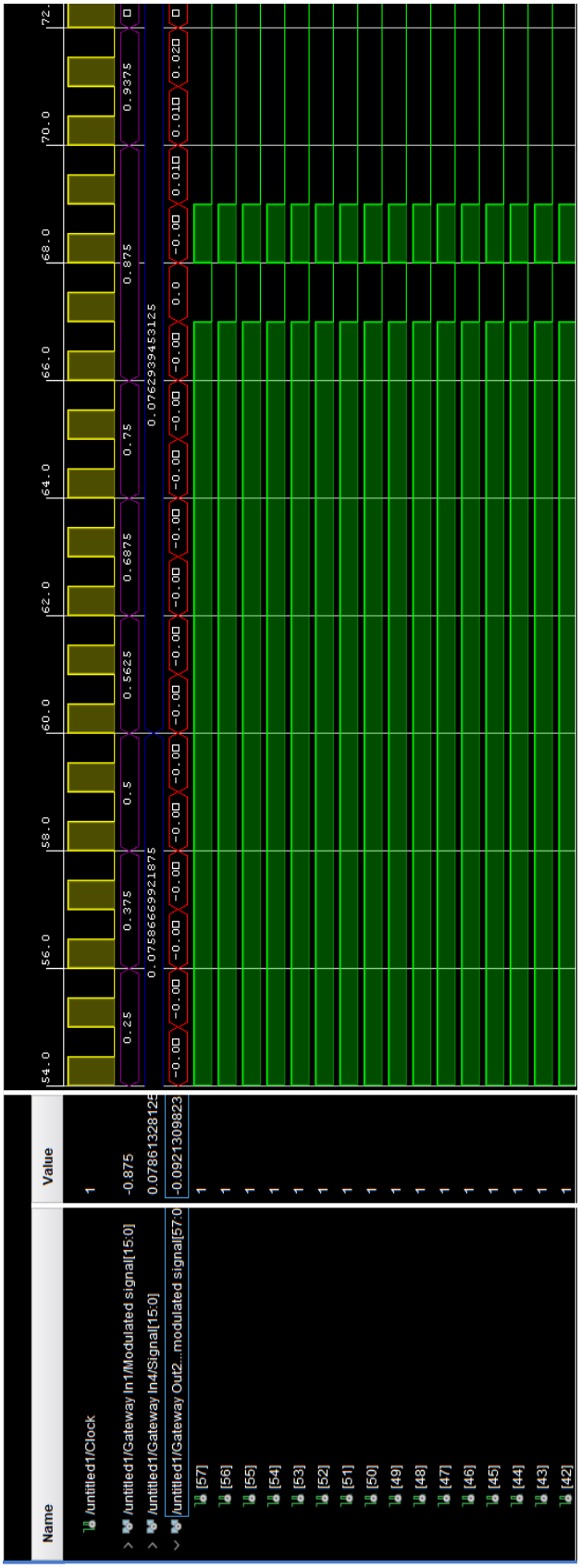


Fig. 18 NOMA Signal Waveform for the Downlink

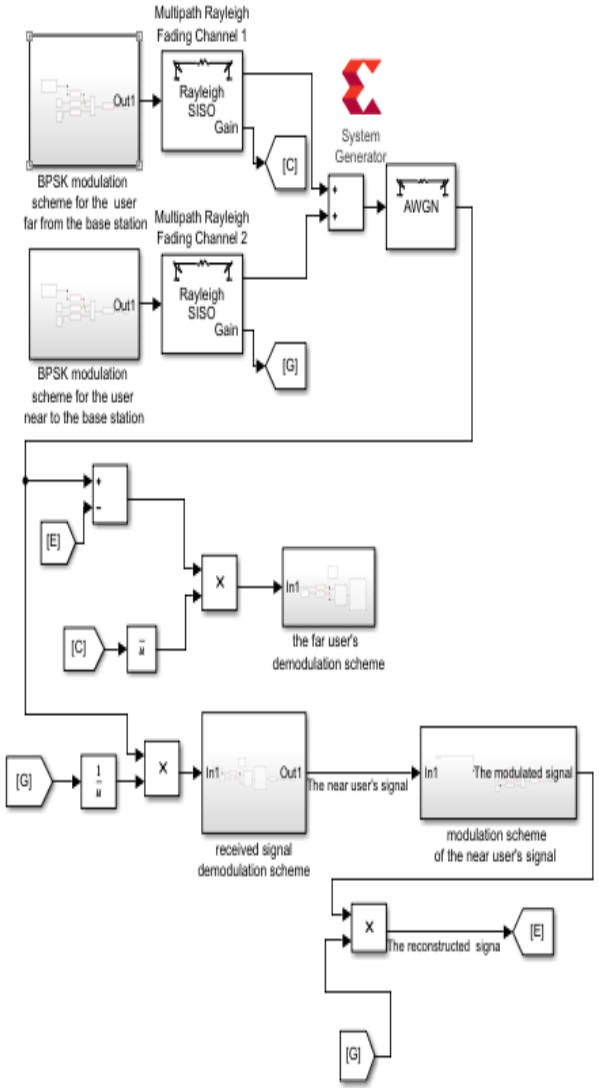


Fig. 19 NOMA Design for the Uplink Using Systems Generator and Xilinx Tools.

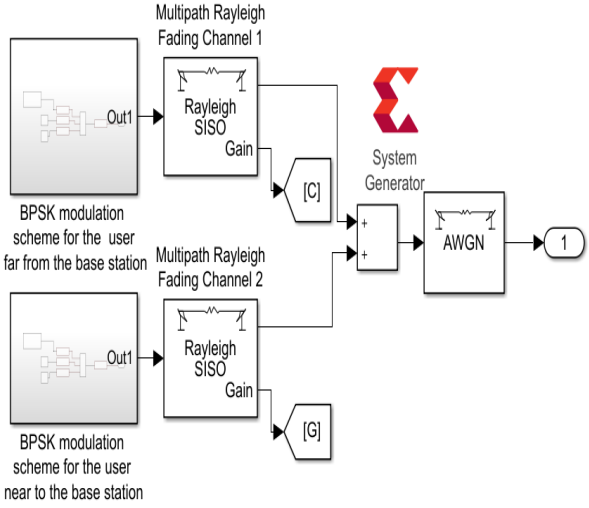


Fig. 20 NOMA Design for Uplink.

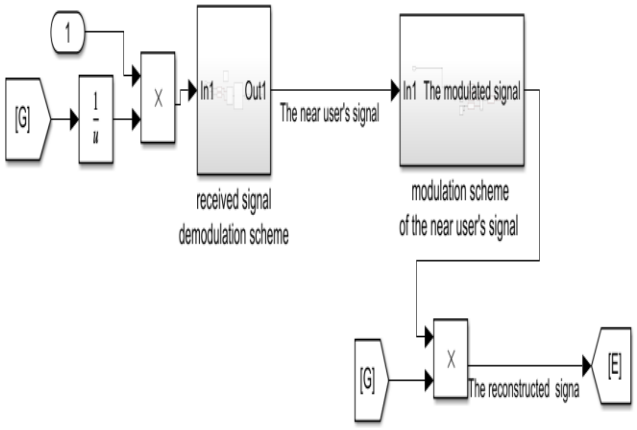


Fig. 22 Near User Reception Design for Uplink.

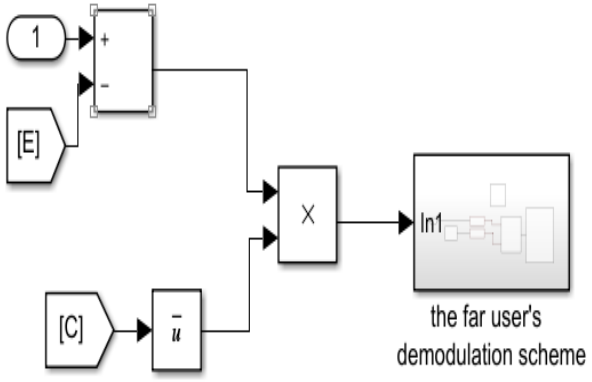


Fig. 21 Far User Reception Design for Uplink.

6- Conclusion

In this paper, the power NOMA technique in uplink and downlink links are studied and implemented on an FPGA device. The implementation is realized for two users supported by one base station over Rayleigh fading channel. In both schemes a successive interference cancelation (SIC) detector is considered in the receiver side to separate the superimposed symbols of users. The Conclusion
simulation results reveal the effectiveness of the SIC algorithm to detect the user's symbols in downlink link by using both modulation: QPSK or BPSK. By contrast, in the uplink scenario where there is no scheduling of power allocation, the SIC fails to detect any symbols in high SNR regime (suffers from the error floor). Besides, to benefit from the advantages offered by the features of the FPGA device, a design of both schemes is realized using the Xilinx system generator. The results of this work may be extended to address the implementation of this promising technique to deal with a massive connection (arbitrary number of users with higher order adaptive modulation) for IoT applications

Acknowledgement.

The authors thank Dr Cherif chibane from MIT for his help.

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, et Z. Wang, « Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends », IEEE Commun. Mag., vol. 53, no 9, p. 74‑81, sept. 2015.

[2] M. Moltafet, N. Mokari, M. R. Javan, H. Saeedi, et H. Pishro-Nik, « A New Multiple Access Technique for 5G: Power Domain Sparse Code Multiple Access (PSMA) », IEEE Access, vol. 6, p. 747‑759, 2018.

[3] X. Wei et al., « Software Defined Radio Implementation of a Non-Orthogonal Multiple Access System Towards 5G », IEEE Access, vol. 4, p. 9604‑9613, 2016.

[4] Z. Ding et al., « Application of Non-Orthogonal Multiple Access in LTE and 5G Networks », IEEE Commun. Mag., vol. 55, no 2, p. 185‑191, févr. 2017.

[5] M. Vaezi, Z. Ding, et H. V. Poor, Éd., Multiple Access Techniques for 5G Wireless Networks and Beyond. Cham: Springer International Publishing, 2019.

[6] A. E. Mostafa, Y. Zhou, et V. W. S. Wong, « Connection Density Maximization of Narrowband IoT Systems With NOMA », IEEE Trans. Wireless Commun., vol. 18, no 10, p. 4708‑4722, oct. 2019.

[7] Y. Yuan et al., « NOMA for Next-Generation Massive IoT: Performance Potential and Technology Directions », IEEE Commun. Mag., vol. 59, no 7, p. 115‑121, juill. 2021.

[8] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, et S. J. Johnson, « Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey », IEEE Commun. Surv. Tutorials, vol. 22, no 3, p. 1805‑1838, 2020.

[9] F. A. Khales et G. A. Hodtani, « An evaluation of the coverage region for downlink Non-Orthogonal Multiple Access (NOMA) based on Power Allocation Factor », in 2017 Iran Workshop on Communication and Information Theory (IWCIT), Tehran, Iran, mai 2017, p. 1‑5.

[10] Q. C. Li, H. Niu, A. T. Papathanassiou, et G. Wu, « 5G Network Capacity: Key Elements and Technologies », IEEE Veh. Technol. Mag., vol. 9, no 1, p. 71‑78, mars 2014.

[11] X. Liang, X. Gong, Y. Wu, D. W. K. Ng, et T. Hong, « Analysis of Outage Probabilities for Cooperative NOMA Users with Imperfect CSI », in 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, déc. 2018, p. 1617‑1623.

[12] A. Agarwal, R. Chaurasiya, S. Rai, et A. K. Jagannatham, « Outage Probability Analysis for NOMA Downlink and Uplink Communication Systems With Generalized Fading Channels », IEEE Access, vol. 8, p. 220461‑220481, 2020.

[13] N. Tutunchi, A. Haghbin, et B. Mahboobi, « Complexity Reduction in Massive-MIMO-NOMA SIC Receiver in Presence of Imperfect CSI », Journal of Information Systems and Telecommunication (JIST), vol. 2, no 30, p. 113, août 2020.

[14] F. Kara et H. Kaya, « BER performances of downlink and uplink NOMA in the presence of SIC errors over fading channels », IET Communications, vol. 12, no 15, p. 1834‑1844, sept. 2018.

[15] C. A. Ramos-Arregu'n et al., « FPGA Open Architecture Design for a VGA Driver », Procedia Technology, vol. 3, p. 324‑333, 2012.

[16] « Zybo Z7 Reference Manual - Digilent Reference ». accessed on july 04, 2021. [Online]. available on: https://digilent.com/reference/programmable-logic/zybo-z7/reference-manual

[17] T. Tami, T. Messaoudene, A. Ferdjouni, et O. Benzineb, « Chaos secure communication' implementation in FPGA », in 2018 International Conference on Applied Smart Systems (ICASS), Medea, Algeria, nov. 2018, p. 1‑6.

[18] W. Tang, S. Yang, et X. Li, « Implementation of Space-time Coding and Decoding Algorithms for MIMO Communication System Based on DSP and FPGA », in 2019 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC), Dalian, China, sept. 2019, p. 1‑5.

[19] H. Sreenath et G. Narayanan, « FPGA Implementation of Pseudo Chaos-signal Generator for Secure Communication Systems », in 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, sept. 2018, p. 804‑807.

[20] Q. Yingchao et Y. Feng, « Design and Implementation of Differential Frequency Hopping Communication System Based on FPGA », in 2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, déc. 2018, p. 1006‑1010.

[21] M. A. Ahmed, K. F. Mahmmod, et M. M. Azeez, « On the performance of non-orthogonal multiple access (NOMA) using FPGA », IJECE, vol. 10, no 2, p. 2151, avr. 2020.

[22] M. Mekhfioui, A. Satif, O. Mouhib, R. Elgouri, A. Hadjoudja, et L. Hlou, « Hardware Implementation of Blind Source Separation Algorithm Using ZYBO Z7& Xilinx System Generator », in 2020 5th International Conference on Renewable Energies for Developing Countries (REDEC), Marrakech, Morocco, Morocco, juin 2020, p. 1‑5.

[23] T. Assaf, A. Al-Dweik, M. S. E. Moursi, H. Zeineldin, et M. Al-Jarrah, « NOMA Receiver Design for Delay-Sensitive Systems », IEEE Systems Journal, vol. 15, no 4, p. 5606‑5617.

[24] H. Semira et F. Kara, « Error Performance of Uplink SIMO-NOMA with Joint Maximum-Likelihood and Adaptive M-PSK », in 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Bucharest, Romania, mai 2021, p. 1‑6.

[25] H. Semira, F. Kara, H. Kaya, et H. Yanikomeroglu, « Multi-User Joint Maximum-Likelihood Detection in Uplink NOMA-IoT Networks: Removing the Error Floor », IEEE Wireless Commun. Lett., vol. 10, no 11, p. 2459‑2463, nov. 2021.

[26] « Digital Modulation in FPGAs Xilinx using system generator (ASK, BPSK, FSK, OOK, QPSK) - File Exchange - MATLAB Central ». accessed on july 04, 2021. [online]. available on: https://www.mathworks.com/matlabcentral/fileexchange/14650-digital-modulation-in-fpgas-xilinx-using-system-generator-ask-bpsk-fsk-ook-qpsk?s_tid=FX_rc3_behav

[27] H. CJ, S. D. Hanwate, et A. S. Mali, « Hardware Implementation of BPSK System on Virtex2-Pro FPGA Using Xilinx System Generator », IRJES, vol2, issue1,p 18-24, jan 2013

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 282-293
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A Recommender System for Scientific Resources Based on Recurrent Neural Networks

Hadis Ahmadian Yazdi[1], Seyyed Javad Seyyed Mahdavi[2*], Maryam Kheirabadi[1]

[1].Department of Computer Engineering, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran
[2].Department of Electrical Engineering, Mashhad Branch, Islamic Azad University, Mashhad, Iran

Received: 23 Sep 2022/ Revised: 04 Mar 2023/ Accepted: 08 Apr 2023

Abstract

Over the last few years, online training courses have had a significant increase in the number of participants. However, most web-based educational systems have drawbacks compared to traditional classrooms. On the one hand, the structure and nature of the courses directly affect the number of active participants; on the other hand, it becomes difficult for teachers to guide students in choosing the appropriate learning resource due to the abundance of online learning resources. Students also find it challenging to decide which educational resources to choose according to their condition. The resource recommender system can be used as a Guide tool for educational resource recommendations to students so that these suggestions are tailored to the preferences and needs of each student. In this paper, it was presented a resource recommender system with the help of Bi-LSTM networks. Utilizing this type of structure involves both long-term and short-term interests of the user and, due to the gradual learning property of the system, supports the learners' behavioral changes. It has more appropriate recommendations with a mean accuracy of 0.95 and a loss of 0.19 compared to a similar article.

Keywords: Deep Learning Networks; Recurrent Methods; Educational Resource; Recommender System.

1- Introduction

Nowadays, the purpose of online courses is to provide students with the opportunity to learn quickly. However, designing this course is not always the most effective method for all learners and has led to high dropout rates and low academic effectiveness. This model may be as effective as face-to-face training by enhancing artificial intelligence. The global corona epidemic has forced students to rely on technology in unprecedented ways to teach themselves. These technologies can be significantly improved with artificial intelligence and machine learning algorithms. Virtual educators can help students find the right curriculum from the available curriculum through the referral system [1,2,3].

Internet services today offer various content. This high diversity is both an opportunity and a threat. The opportunity is to present the customer's favorite content with more probability, and the customer can use it. The danger is that the customer may not find the content they need in the vast amount of content. It can be concluded that the rapid increase of the information volume, the limitation of search engines to search for information, and the increasing number of visitors to websites in recent years, are critical challenges in the recommender systems. A Student must find the required Educational Resources from the appropriate sources, which would be time-consuming and costly without information filtering and recommender systems [4].

Recommender systems can discover the users' interests and predict their preferences, and among the high volume of data, refine the items which are likely to be of interest to the user and save time by suggesting them. Of course, the only efficiency of a recommendation system (which is done by search engines) is not only searching the items in less time with less energy, but its primary purpose is to discover items. These systems, with the ability to store and analyze the user's past behaviors, also infer services and information that users have not noticed but are probably

✉ **Seyyed Javad Seyyed Mahdavi**
Mahdavi@Mashdiau.ac.ir

interested in and provide exciting results to users. Recommender systems are one of the main tools to overcome information overload and are an intelligent complement to the retrieval concepts and information filtering to analyze the users' behaviors.

In addition, students have individual differences such as educational background, study method, age, etc., which emphasizes the need to take feedback from students to guide them in the educational process better [5]. In e-learning systems, students are eager for personalized services to be automatically trained, monitored, supported, and evaluated. With such a personalized service, student loyalty increases [6].

The Recommender system can help teachers personalize the curriculum and resources for each student based on their unique skills and weaknesses. It is used to develop more complex and attractive methods of student assessment that are very time-consuming for teachers today. Artificial intelligence and machine learning have the potential to address many of the problems that have arisen in the transfer of teaching methods to online learning. Including students' resistance to changing their education, increasing curriculum planning, and addressing the loss of personal interaction between students and teachers [4].

This paper aims to design a time-based educational recommender system to suggest new resources to users based on the features that include a person's pre-clicked or downloaded educational resources and the rating that the user has given to each resource. If there is an inherent structure that the model can exploit, deep neural networks are very efficient for this issue.

Because the nature of the problem of recommending textbooks depends on the time and long-term review of student performance, the sequential structure of sessions or report clicks is very appropriate for inferential errors in conventional or recursive models. In many methods, only the user's past information is used in learning. While in the present article, having a network that looks both backward and forward can also cover changes in learner behavior and offer more up-to-date recommendations. In the following, we will review the work done on educational recommenders. In the next section, we present the proposed method, a hybrid architecture of BI-LSTM and MLP networks. In the fourth section, we review the results of implementing the proposed algorithm based on accuracy and efficiency, and finally, in the fifth section, we will present future suggestions.

2- Related Works

Session-based recommender systems are an excellent example of recommender systems. They are primarily researched, although not a new research topic [7, 8, 9]. Compared to traditional recommender systems, a session-based referral system is more suitable for learning dynamic and sequential user behaviors.

The purpose of recommender systems is to generate search results close to the user's needs and make predictions based on their priorities. In virtual learning environments, educational recommender systems have learning objects based on students' characteristics, priorities, and learning needs. A Learning Object (LO) is a unit of educational content that can assist students in their learning process [10]. A learning object is defined by the IEEE [11] as a digital or non-digital entity with educational design features that can be reused or referenced during the computer-assisted learning process.

Recently, deep learning has dramatically changed the architecture of recommenders and provided more opportunities to improve the performance of recommenders. Deep learning can capture nonlinear and meaningful user-item relationships and result in abstract data representations at higher levels. Besides, it can obtain complex relationships within data from other sources such as conceptual, textual, and visual information [12]. In the traditional technology method, the recommender system faces problems such as a large amount of data, shared filtering, poorly given information, and a cold start, which are also addressed in this study.

A new generation of algorithms is required for recommender systems due to the importance of recommender systems in online servers and the dynamics, privacy, and bulk of data and problems in these systems (such as cold start). Deep learning is offered as a solution to the problem of recommender systems.

In recent years, artificial neural networks have attracted much attention due to their increased computing power and big data storage capabilities. Many new methods in image processing, object recognition, natural language processing, and voice recognition now use deep neural networks as a primary tool [13]. The remarkable capabilities of deep learning methods encourage researchers to apply deep architectures in "recommendation". Deep learning-based recommender models can be categorized in Fig1. [14].

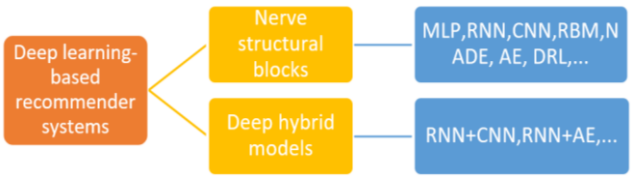


Fig. 1. General Models of Deep learning for the Recommender.

In the following, the work that has been done specifically on educational advisors will be reviewed.

2-1- Data Mining Methods

In [15], the proposed method integrates the features of online learning style, including participatory filtering (CF) features, association rules criteria, and online learning style (OLS) in the recommendation algorithm. The output of the proposed method has improved by 25% compared to the technique without students' characteristics.
In [16], attention-focused neural networks (CNN) have been used to obtain predictions of user ratings and user profiles and to recommend superior courses. Then, they integrated a participatory filter to enable real-time recommendations and reduce server workload. The model ultimately recommends courses to students. However, the proposed system may continue to suffer from the problem of recommending similar courses as MOOCs develop and the number of courses increases.

2-2- Development on Traditional Methods

Most traditional online learning systems based on refining methods depend on user's behavior towards different sources. For users who behave similarly, the results of resource recommendations are often unsatisfactory [17].
The aim is to help students make informed decisions about their learning paths using a hybrid counseling system. By combining content-based similarity and dispersion, based on structural information about module space, the detectability of long-term choices that are consistent with students' preferences and goals can be improved.
One of the advantages of this is that you can add scatter to the set of recommendations. The goal of [15] is to provide a personal reference system that leads to better recommendations in the shortest possible time. The proposed system uses user profiles to create neighborhoods and predicts weights. To overcome the problem of cold start and scattered data, student profiles are created using the learning method. Resources that are of interest to the user are suggested through calculations calculated with new features and participatory refinement method.

2-3- Machine learning-based Methods

In [18], the aim is to provide an advisory system based on reasoning theory that combines content-based, participatory, and knowledge-based recommendation methods. This method recommends training resources so that the system can generate further arguments to justify its competence.
In [19], the AROLS method is proposed. This method is an advanced recommendation integrated with a comprehensive learning style model for online students. This method considers the learning method as prior knowledge and provides recommendations. First, it creates clusters of different learning styles. Then the behavioral patterns presented by the matrix of similarity of learning resources and communication rules of each group are extracted using students' review history. Finally, it creates a set of personal recommendations based on the data mining results of the previous steps. This method presents the recommendation results more accurately while maintaining the computational advantage than the traditional participatory refining (CF) recommendation.

2-4- Artificial Intelligence-based Methods

[20] A multi-factor technology-based referral system has been developed that helps e-learning referral systems offer students the most appropriate learning resources. This work utilizes the capabilities of multi-agent technology to create a plan that combines web use and extraction algorithms such as content-based methods and collaborative refinement to find the most appropriate training resources. The performance of this combined method is better than other algorithms in it. Advances have also been made in building models for searching and retrieving learning objects stored in heterogeneous repositories.
In [21], the aggregation of two multifactorial models is introduced that can carry a specific LO corresponding to the characteristics of a student and carry the LO to the instructors to help them in creating lessons. The aim is to create an integrated multi-factor model that meets the needs of students and educators and thus improves the learning and teaching process.

2-5- Methods based on Neural Networks and Deep learning Networks

The paper [11] introduces a high-precision resource recommender model (MOOCRC) based on deep belief networks (DBNs) to increase the efficiency and enthusiasm of learners in MOOC environments. This model extracts the characteristics of learners and their curriculum content. User-lesson vector vectors are constructed as model input. Instructors' grades are processed into lessons as supervised labels. The MOOCRC model is taught without supervisor pre-training and is fine-tuned using supervisor feedback. The model's performance has been evaluated using selective data from educators obtained from the starC MOOC platform of Central China Normal University. The results show that MOOCRC has higher recommendation accuracy and faster convergence than other traditional recommending methods.
The article [22], with the aim of recommending educational resources, tries to help learners achieve better academic results. The proposed model consists of deep learning recursive layers that have been improved with the attention technique. After testing the model's performance

with OULAD data, 95% accuracy was obtained, which is a better result than similar works.

3- The Proposed Method

In this paper, the purpose of the first phase is to obtain the users' database, including their interest in the study resources and the amount of use and click on these resources and related features, and then select the practical items among them. In the second phase, the recommender system is trained with acceptable accuracy using deep neural networks. The recommendation algorithm includes data extraction from OULAD information resource files, data preprocessing, building an add-on deep learning network from MLP, Bi-LSTM, initial parameterization, training, and predicting scores. Finally, it is offered resources to users using the trained network.

Two-way short-term memory (Bi-LSTM) is a type of recurrent neural network. This process processes data in two directions because it works with two hidden layers. It is the main point of divergence with LSTM. This method has proven promising results in natural language processing. One advantage of two-way LSTM over one-way LSTM is that two-way LSTM looks to both the past and the future to make predictions. Still, one-way LSTM only looks to the past, so two-way LSTM can be sensitive to diversity in the short-term interests of the user in both directions and thus cover learners' behavioral changes.

In the proposed architecture, as presented in Fig.2, a Bi-LSTM cell is provided in each layer for each feature in the database. The cells focus on one feature of each record, and each cell represents only one feature pattern. Consider and ultimately, the combination of these patterns will lead to better results.

Among the papers in this field, we found the works based on Bi-Lstm, which required short-term and long-term memory; or for new mechanisms that, in practice, create shortcuts and ignore several time steps. These shortcuts also allow the production error to be easily transferred to the post-diffusion phase without quickly losing. It can significantly handle the vanishing gradient problems.

3-1- Predicting Course Resource Scores

In the model training process, the data labeled class is used as a training set for the model. Then, based on the user-class feature vector, the recommendation problem becomes the category prediction problem. In this paper, using the label of rating classes, error information is published to each layer from top to bottom with fine-tuning of the parameters to the observer. After training the model to achieve a certain error, the test set can be used to test the performance of the recommender model. The data in the test set is divided into two categories: user feature - lesson vector and lesson evaluation. Each user-lesson feature vector corresponds to a category level, and each level corresponds to a point. All lessons that correspond to a user are sorted according to the expected score, and then lesson recommendations are generated

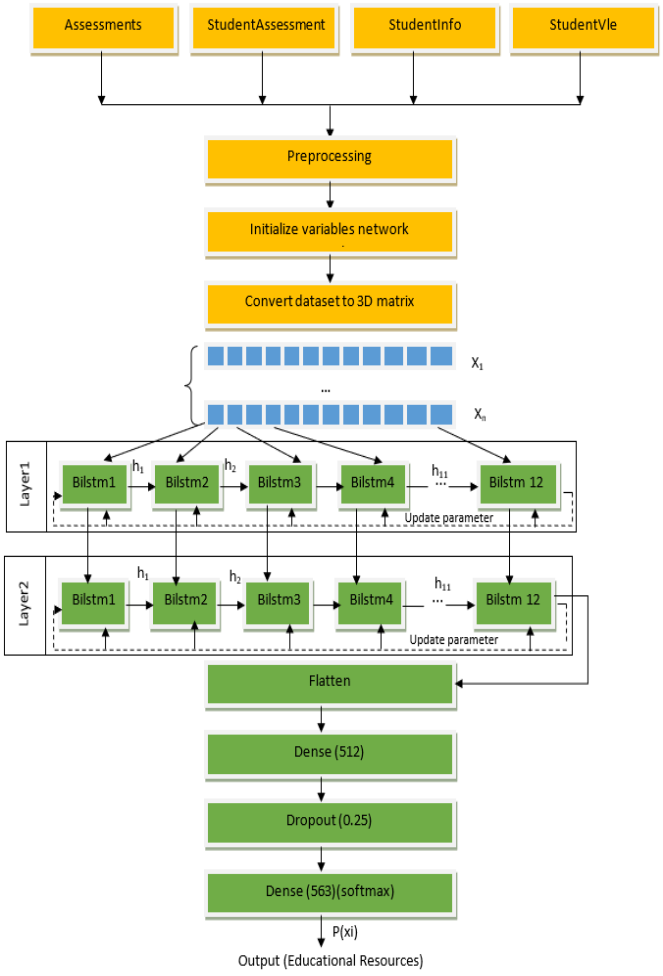


Fig.2 The Structure of Proposed Methods.

3-2- Implementation

The selected data as input to the database implementation are divided into three sections: students, teacher, and course and includes information about 22 courses offered, 32593 students, their evaluation results, and their mutual reports with the virtual learning environment (VLE), which is provided by summaries of students' daily clicks on various "resources" (10,655,280 entries).

- **Database**[1]

Students generate various behavioral data by learning in an online learning environment. This behavioral data is collected and stored through data collection methods (OULAD). The reference database provides data sources for this platform. This curriculum database can extract content features that reflect students' interest in reference. Students' feature vectors are constructed by combining students' characteristics and lesson features, and then hybrid behavioral characteristics and user-lesson feature vectors are generated (Jacob Kozilk et al., 2007).
The frequency of each student's recorded activities is presented in Fig3. The database is anonymous using the ARX [PK15] data encryption tool. The data was reviewed for error detection and verified and published by Open Data Institute1. The frequency of each student's recorded activities is presented in Fig3. The database is anonymously using the ARX [PK15] data encryption tool. The data were reviewed for error detection and verified and published by Open Data Institute1.
The main table contains the student files attached to the courses (A student can have more than a one registered course). Each course has several assessments, which are related to students and include the history of student assessment results. There are three types of assessment: Teacher Assessment (TMA), Computer Assessment (CMA), and Final Exam (Exam).

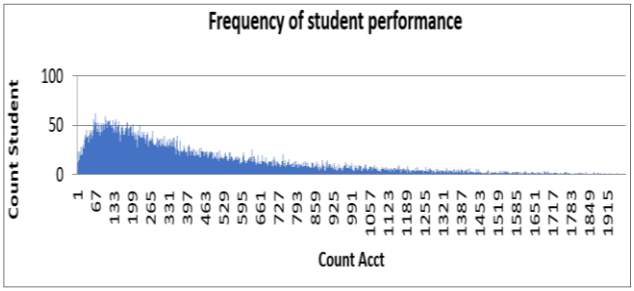


Fig.3 Frequency of Student Performance.

- **Data Preprocessing**

As presented in Fig4. In the pre-processing stage, the input first includes four sections: resources provided, students' characteristics, courses held, and student performance and assessment history in each course. These four sections are combined, and further analysis, categorization, feature mapping, clean blank or incorrect data, and feature normalization are performed on them.
The normalization of features is done in the range (0, 1).
The data must be placed at an equal distance so that the data that contains a larger range of numbers does not have a more significant effect on the algorithm than the others. This prevents network weights from fluctuating too much, and the amount of network loss fluctuates slightly when modeling data. In other words, the higher the convergence speed, the better and smoother the network model [23].

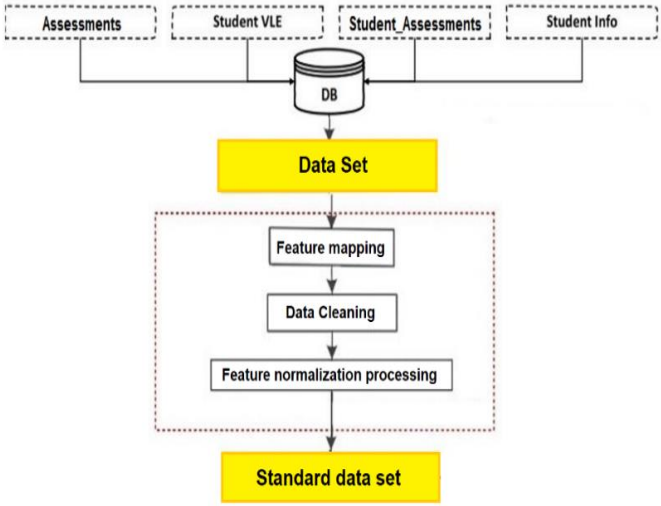


Fig.4 Data Preprocessing Steps.

Empirical evidence shows that data standardization is useful in terms of accuracy. It may be related to the descent of the gradient. It is easy to understand why normalization improves training time. Large input values saturate activation functions such as sigmoid or ReLu (negative input). This type of feedback activation function has little or no gradient in the saturated region and thus reduces the Training speed [23]. Eq1 is used to perform normalization.

$$X^{*} = \frac{(X - Xmin)}{(Xmax - Xmin)} \qquad (1)$$

Where X min represents the lowest eigenvalue as X min {X1, X2, ..., Xn}= Xmin.Xmax represents the maximum eigenvalue as X max = max {X1, X2, ..., Xn}; X*indicates the normalized value, X represents the original data. Another step in the preprocessing step is to convert the string values in the database to numeric values.
Table1 shows an example of the values in the database that are mapped to numerical values in Table2.
After the initial steps of preprocessing, the datasheet is tagged in two following ways:

- **Maximum Click in Maximum Point Average:**

Among the set of activities registered for the standard courses for each student, the source with the most clicks can indicate the student's taste and interest) in the course that had the highest GPA (showing the practical resources studied in that course) was selected as a label. Five hundred sixty-two labels were created, mapped from 0 to 561. The frequency of count labels is presented in Fig5.

[1] [1] https://analyse.kmi.open.ac.uk/open_dataset

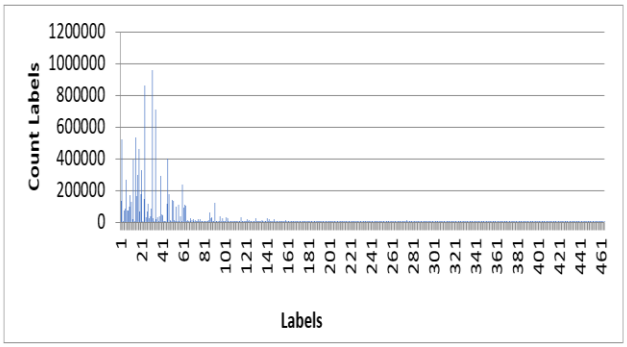


Fig. 5. Frequency of Count labels (Method 1: The Most Click in the Maximum Average Score

- **Most Recent Clicks:**

From the set of each student's activities, the source with the most clicks (indicating the student's taste and interest) were selected as the label on the last day of the student's activity, assuming that the current study subject is essential to him. One thousand five hundred ninety-four labels were created, mapped from 0 to 1593. The frequency of count labels is presented in Fig6. After preprocessing, the data is divided into a training set and a test set. Finally, the training set enters the proposed network as input.

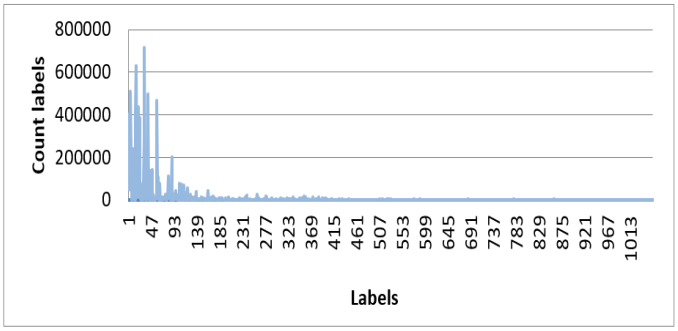


Fig. 6. Frequency of Count labels (Method 2: The Most Click)

Table 1. A Sample of Database Data before Mapping

Code module	Code presentation	Id student	gender	Age band	Sum click
AAA	2013J	11391	M	55<=	16
AAA	2013J	11391	M	55<=	44
AAA	2013J	11391	M	55<=	1
AAA	2013J	11391	M	55<=	2
AAA	2013J	11391	M	55<=	1
AAA	2013J	11391	M	55<=	2
AAA	2013J	11391	M	55<=	2
AAA	2013J	11391	M	55<=	16
AAA	2013J	11391	M	55<=	44

Highest education	final_result	score_ mean	id_site	date
HE Qualification	Pass	82	546669	-5
HE Qualification	Pass	82	546662	-5
HE Qualification	Pass	82	546652	-5
HE Qualification	Pass	82	546668	-5
HE Qualification	Pass	82	546652	-5
HE Qualification	Pass	82	546670	-7
HE Qualification	Pass	82	546671	-7
HE Qualification	Pass	82	546669	-5
HE Qualification	Pass	82	546662	-5

Table 2. The Values of Features Mapped to the Number

Code module		Code presentation		age_band		gender		highest_education		final_result	
AAA	0.1	2013B	540	0-35	0.1	F	0.1	A Level or Equivalent	0.1	Distinction	0.1
BBB	0.2	2013J	720	35-55	0.2	M	0.2	HE Qualification	0.2	Fail	0.2
CCC	0.3	2014B	180	55<=	0.3	-	-	Lower Than A Level	0.3	Pass	0.3
DDD	0.4	2014J	360	-	-	-	-	No Formal quals	0.4	Withdrawn	0.4
EEE	0.5	-	-	-	-	-	-	Post Graduate Qualification	0.5	-	-
F F F	0.6	-	-	-	-	-	-	-	-	-	-
GGG	0.7	-	-	-	-	-	-	-		-	-

- **Correlation Between Variables and Labels**

In this research, the method of the maximum average method has been used for labeling. We examined the possible correlation between the label and the existing variables that were used as input for teaching the model by correlation test. As you can see in Fig7, there is no significant correlation between the variables and their labels.

- **Network Construction**

The Bi-Lstm library of KERAS has been used to implement the idea of this paper. The data is entered into a

two-layer Bi-Lstm architecture with 512 neurons, and finally, the output of this layer enters the MLP network.
In the model training process, the model parameters must repeatedly be adjusted to achieve better results in feature extraction. During the learning process, the 512 minibatch processing method is used to solve the problem of large data volumes. Also, the learning rate parameter 0.0001, the number of counts Epoch = 100, and the SoftMax activator function are initialized.

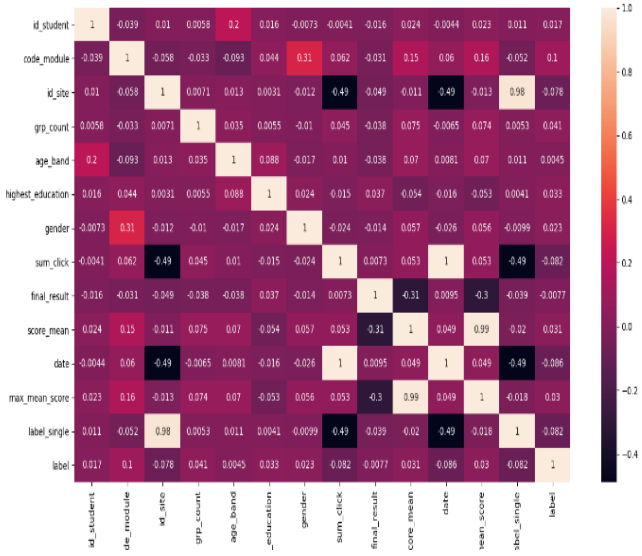


Fig7. The Result of the Correlation Test.

The number of records used in the testing process of the networks implemented in this article is Train = 8434945, test = 2108737, and validation split = 0.2. After completing Epochs, the diagrams and the results of implementations show that the accuracy and loss of the work are far better than the results of the implementation of the proposed network [24].

3-3- Methods and Tools of Data Analysis

The primary purpose of the proposed system presented in this article is to predict the best sequence of educational resources. There are many criteria for measuring different aspects of bid performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Indicates what percentage of experimental records are properly categorized.

$$Precision = \frac{\sum_{x\in X}|R(x)\cap H(x)|}{\sum_{x\in X}|R(x)|} \qquad (3)$$

$$Recall = \frac{\sum_{x\in X}|R(x)\cap H(x)|}{\sum_{x\in X}|H(x)|} \qquad (4)$$

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \qquad (5)$$

Here x is a student from the set of all students X, R (x) represents the learning resources recommended for student x, H (x) represents the learning resources observed by learner x [25].

4- Ablation Study

4-1- Investigating the Effect of the Number of Cells in Each Layer

As observed in Table 3, the results of the implementation of three types of LSTM, GRU, and Bi-LSTM networks were implemented and examined as single-celled structures in three single-layer architectures, two layers, and three layers are not desirable. In nine architectures, one cell in each layer could not find the pattern of different features, the relationship of features to other features in combination, permutation, and different models. Increasing the number of layers has not been able to play an influential role in improving the results.
As observed in Table4, The implementation of multicellular single-layer architectures implemented and investigated in three single-layer architectures of LSTM, Bi-LSTM and GRU had more favorable results than the single-cell architectures.
In fact, by changing the layout cells in the proposed models, it is possible to use and extract features in both single and multiple forms. When entering the first feature of the model, it examines and extracts the information contained in the same feature individually. After entering the second feature of the model and extracting the information contained in this feature, it also examines and extracts the connections and information between these two features. With the introduction of the third feature, in addition to extracting the information of the same feature individually and examining the existing communications and information with the previous features in pairs, the communications are also examined on a permutation basis. As a result, the model achieves more features and is more important than single-cell networks.
Meanwhile, the GRU network with higher generalizability power than the two types of Bi-LSTM and LSTM networks with a loss of 0.2 and an accuracy of 0.91 has had a better performance.

Table 3. Results of Training and Testing of one to Three-layer Single-Cell Networks: LSTM, Bi-Lstm and GRU

LSTM	1Layer	train	loss	val_loss	accuracy	val_ acc
			2.6907	2.6617	0.2825	0.2879
		test	accuracy			
			0.28			
	2Layer	train	loss	val_loss	accuracy	val_ acc
			2.8305	2.8172	0.2607	0.2577
		test	accuracy			
			0.18			
	3Layer	train	loss	val_loss	accuracy	val_ acc
			2.8633	2.8369	0.266	0.2548
		test	accuracy			
			0.25			
GRU	1Layer	train	loss	val_loss	accuracy	val_ acc
			2.9989	2.9215	0.2389	0.2539
		test	accuracy			
			0.25			
	2Layer	train	loss	val_loss	accuracy	val_ acc
			2.9429	2.9185	0.2439	0.2562
		test	accuracy			
			0.25			
	3Layer	train	loss	val_loss	accuracy	val_ acc
			2.9811	2.9577	0.2317	0.2388
		test	accuracy			
			0.23			
Bi-Lstm	1Layer	train	loss	val_loss	accuracy	val_ acc
			1.9011	1.3059	0.4481	0.6298
		test	accuracy			
			0.62			
	2Layer	train	loss	val_loss	accuracy	val_ acc
			2.4977	2.4328	0.3133	0.3231
		test	accuracy			
			0.31			
	3Layer	train	loss	val_loss	accuracy	val_ acc
			2.5016	2.466	0.3057	0.3132
		test	Accuracy			
			0.31			

Table 4. Results of Training and Testing Single layer Multi-Cellular Networks of LSTM, GRU and BI-LSTM

LSTM	train	loss	val_loss	accuracy	val_ acc
		0.6685	0.6067	0.7573	0.7704
	test	accuracy			
		0.77			
Bi-Lstm	train	loss	val_loss	accuracy	val_ acc
		0.3268	0.2366	0.8757	0.9036
	test	accuracy			
		0.9			
GRU	train	loss	val_loss	accuracy	val_ acc
		0.2319	0.2021	0.898	0.9099
	test	accuracy			
		0.91			

4-2- Investigating the Effect of the Number of Layers of Network Architecture

In the architecture of deep learning networks, one issue is the relationship between the cells in each layer and themselves, which was studied in detail. The second issue will be the relationship between the different layers in the implemented architecture. This part of the architecture does an overview of the features in the database. Some previous work has suggested that multilayer Bi-Lstm in neural networks can further improve classification or regression performance [26]. In addition, some related theoretical supports have shown that a deep hierarchical model is more efficient in delivering some functions than the shallow type.

It has been implemented and trained three architectures of two-layer LSTM (Fig8, a and b), two-layer GRU (Fig8, c and d), and two-layer BI-LSTM (Fig8, e and f) to evaluate the effectiveness of the relationship between the layers. As observed in Table5, the results of two-layer architectures have been more favorable compared to the single-layer architectures.

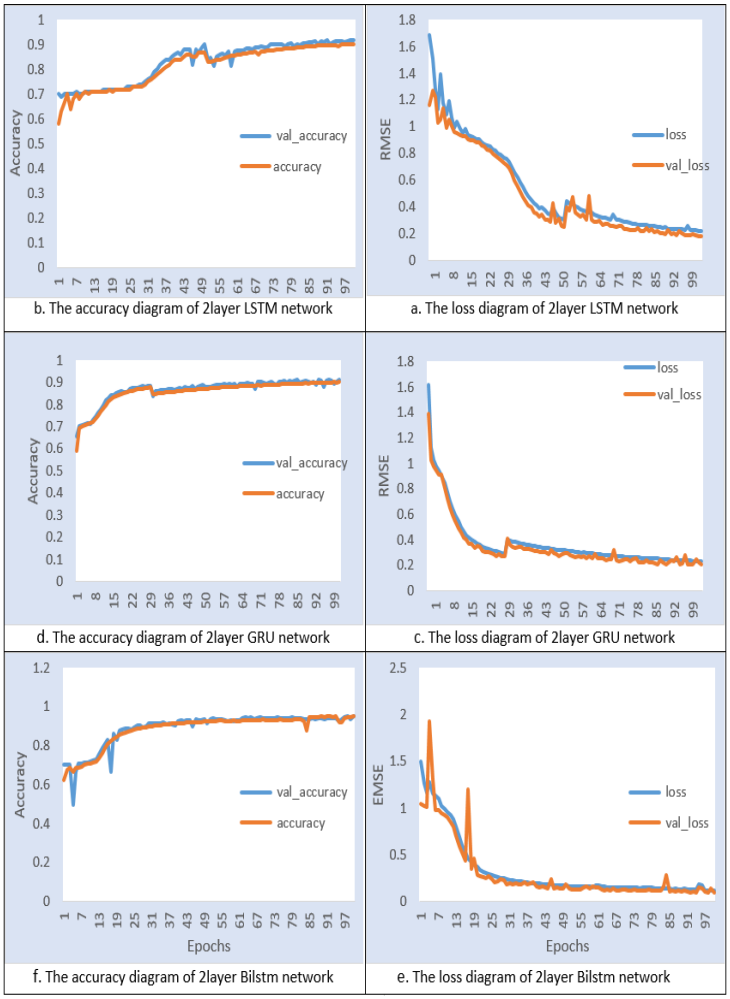
According to Table5, the proposed architecture results, based on BI-LSTM bilayer and multi-cellular, with a loss of 0.9 and accuracy of 0.95, were much more accurate and desirable than the proposed architecture [25- 28]. As observed in Fig8, the Loss reduction speed indicates that the number of selected AIPs is appropriate, and since the accuracy and validation accuracy diagrams are almost the same, over-fitting did not occur in this experiment.

As can be seen in Table 7, the results of our proposed architecture are very acceptable and desirable.

According to Figure 8, considering the loss rate, it can be seen that the number of selected epics is appropriate. Besides, according to the accuracy diagram, it can be seen that as the accuracy and validation accuracy diagrams are almost the same, overfitting does not occur in this experiment.

Table 5. Results of Training and Testing of Dual-layer Multi-Cell Networks

LSTM 2 Layer	train	loss	val_loss	accuracy	val_ acc
		0.2207	0.1836	0.9021	0.9169
	test	accuracy			
		0.91			
GRU 2 Layer	train	loss	val_loss	accuracy	val_ acc
		0.2321	0.2063	0.9013	0.9121
	test	accuracy			
		0.92			
Bi-Lstm2 Layer	train	loss	val_loss	accuracy	val_ acc
		0.1171	0.0967	0.9529	0.9539
	test	accuracy			
		0.95			



b. The accuracy diagram of 2layer LSTM network
a. The loss diagram of 2layer LSTM network
d. The accuracy diagram of 2layer GRU network
c. The loss diagram of 2layer GRU network
f. The accuracy diagram of 2layer Bilstm network
e. The loss diagram of 2layer Bilstm network

Fig. 8. Results of Training and Testing of Dual-layer Multi-Cell Networks

5- Result and Discussion

In the proposed network and the training phase, the highest Validation accuracy was 0.9529 in epoch 47, and the minimum loss was 0.0995 in epoch 45. After completing the training step, it is entered the test data was input into the network, and the final result was 0.95.

Due to the multi cellularity of the layers, the proposed network allows the study and extraction of features in both single and multiple forms. When entering the first feature of the model, it examines and extracts the information contained in the same feature individually. After entering the next features, the available communications and information with the features are checked in pairs, and the communications are also examined on a permutation basis. On the other hand, the multi-layering of multicellular structures will create a connection between the different layers in the implemented architecture. An overview of the features in the database is done by this part of the architecture.

In addition, the model consisting of Bi-LSTM layers has been more successful due to its two-way structure and the extraction of the relationship between past and future features. As a result, the model acquires more features and is more important than single-cell and single-layer networks.

5-1- Investigating the Effect of Unconventional Data on Model Accuracy

As shown in the analysis of the existing database data and the frequency charts of student activities, Figure 3 and the frequency of classes, Figures 5 and 6, we are faced with an unbalanced data set. In this case, in addition to accuracy, it is better to use call parameters and F1-score to evaluate the model. To make sure that the model does not intelligently categorize all the data presented in an iterative class in the training process to achieve high accuracy.

We have used the content of the database in 3 different sections to teach and test the model.

At the end of the test phase, the average value for all three f1-score reminder parameters is 0.95. We examined the classification accuracy of each group separately and found that the data of all classes were well categorized. For example, Table 6 presents the results of ten groups of groups with the lowest frequency of repetition and ten groups of groups with the highest frequency of repetition.

As you see, in all three sections, 30-70, 20-80, and 10-90 as a group with a support value of one, The result of most recall and f1-score is one. On the other hand, the value one for recall and f1 points is too low in the groups with the most members. This shows that our model, in addition to the data volume challenge, has also responded well to the unconventional data challenge.

5-2- Comparison of the Performance of the Proposed Model with other Models

We have compared the result of the proposed model in line 1 of Table 7 with other methods presented in previous studies or implemented by ourselves. As can be seen, the results are more favorable for different evaluation parameters of the proposed model than other implemented methods. All evaluations were performed on OULAD shared data.

The proposed method [16] has been implemented and has been trained, tested, and evaluated with OULAD data. As can be seen in Table 7, it performed worse than our proposed model in terms of both error and accuracy criteria.

In [25], the three criteria Recall, Prec, and F1 for the 3 methods itemCF, Clustering + itemCF, and AROLS are examined and show that their proposed algorithm

Table 6. Values of Mapping Properties are Given in Numbers

Train test split(test-size=0.1)					Train test split(test-size=0.2)					Train test split(test-size=0.3)					
support	f1-score	recall	precision	label	support	f1-score	recall	precision	label	support	f1-score	recall	precision	label	
1	0.4	1	0.25	530	1	0.4	1	0.25	209	1	1	1	1	326	10 groups with the lowest frequency
1	1	1	1	498	1	1	1	1	228	1	1	1	1	418	
1	1	1	1	453	1	1	1	1	547	1	1	1	1	463	
2	0.02	0.5	0.01	477	1	0.4	1	0.25	228	1	0.4	1	0.25	391	
2	1	1	1	305	1	1	1	1	90	2	1	1	1	446	
2	1	1	1	460	2	1	1	1	184	2	0.4	1	0.25	442	
3	1	1	1	213	3	1	1	1	133	2	0.01	0.5	0.01	171	
3	1	1	1	503	3	1	1	1	391	3	1	1	1	408	
3	0.02	0.5	0.01	411	3	1	1	1	326	3	1	1	1	547	
3	1	1	1	103	4	1	1	1	557	3	1	1	1	550	
82212	0.96	0.96	0.96	8	82212	0.96	0.96	0.97	1	86317	0.95	0.95	0.96	16	10 groups with the highest frequency
84858	0.99	0.99	0.99	34	84858	0.98	0.98	0.98	57	94754	0.98	0.98	0.99	20	
99036	0.99	1	0.98	5	99036	0.98	0.99	0.98	60	112033	0.98	0.99	0.98	12	
101120	0.98	0.99	0.98	28	101120	0.96	0.96	0.97	5	112418	0.99	1	0.99	46	
111173	0.96	0.95	0.98	56	111173	0.97	0.97	0.98	11	131070	0.97	0.97	0.98	17	
137559	0.98	0.99	0.98	13	137559	0.98	0.99	0.98	18	148366	0.98	0.99	0.98	1	
165625	0.97	0.98	0.97	33	165625	0.98	0.98	0.99	16	154820	0.98	0.98	0.99	14	
198928	0.98	0.98	0.99	49	198928	0.98	0.98	0.99	19	202925	0.98	0.98	0.98	34	
262277	0.97	0.97	0.97	56	262277	0.96	0.96	0.97	26	249276	0.97	0.96	0.97	23	
265671	0.98	0.98	0.98	12	265671	0.95	0.95	0.96	41	275379	0.98	0.97	0.98	30	

(AROLS) has a better Prec compared to the other two cases. At the same time, F1 and Recall remain relatively constant during the n top recommendation.

The report in [26] shows that the performance of AROLS is much better than traditional participatory filtering. In particular, the User-AROLS call and accuracy has more than tripled, and the UserCF call and accuracy are much lower than ItemCF, Probably because UserCF focuses more on the interest of learners who are more like a particular learner. ItemCF's recommendation, on the other hand, is more personal because it largely suggests similar ones based on the learner's interest. As you can see in the first row of results, our proposed model performed better than all 7 methods reviewed in these two papers.

In [27] the results show that OLS characters can make the recommendation algorithm more accurate and robust, but as you can see in the results of row one, our proposed model performed better than both methods studied.

Table 7. Comparison of the Proposed Method with the Results Obtained from other Implementations Performed by us and Studies [16,25- 27]

Model	Accuracy	Recall	Prec.	F1	ref
Our proposed model	0.9529				-
Naive Bayes(nb)	0.1981	0.4725	0.2937	0.1853	
Logistic Regression(Lr)	0	0	0	0	
Latent Dirichlet Allocation(lda)	0.5415	0.0772	0.4314	0.4579	
DBN	0.2912	-	-	-	**16**
AROLS	-	0.022	0.28	0.04	**25**
itemCF	-	0.018	0.18	0.027	
Clustering + itemCF	-	0.024	0.24	0.041	
ItemCF	-	0.026	0.1334	0.0435	**26**
Item-AROLS	-	0.0406	0.1880	0.0668	
User-AROLS	-	0.0018	0.0046	0.0026	
UserCF	-	0.0005	0.0011	0.0007	
CF with ARM	-	0.6874	0.076	0.1374	**27**
Proposed article method	-	0.8647	0.1033	0.1842	

6- Conclusion

Recommendation of Educational Resources is an essential and challenging task, especially in a course with the rapid development of the Internet, which consists of a massive variety of educational resources. The challenge is due to the massive amount of educational information in almost all academic fields and the inevitable neglect of personal needs for specific knowledge. Therefore, research on timely learning of learners' behaviors and then personal guidance of their learning process becomes more necessary. This study has analyzed the online learning behaviors to improve personal recommendations in Educational Resources. It is necessary to use different sources and design a centralized framework to combine them and thus provide superior recommendations.

Informal learning environments will also focus on teacher support. Systems must be able to participate in the teachers' tasks, especially when the continuous monitoring and assessment of student homework during the semester is needed [39].

References

[1] Konstan, J. A., |Introduction to recommender systems. In Proceedings of the 2008 ACM SIGMOD international Conference on Management of Data, Vancouver, Canada, (Jun, 2008).

[2] Resnick, P. and Varian, H. R. 1997. Recommender systems, Commun. ACM 40, 3 (Mar, 1997).

[3] Schafer, J. B., Konstan, J., and Riedi, J. Recommender systems in e-commerce. In Proceedings of the 1st ACM Conference on Electronic Commerce, Denver, Colorado, United States, (Nov, 1999).

[4] Gordan Durovic, Martina Holenko Dlab and Natasa Hoic-Bozic, Educational Recommender Systems: An Overview and Guidelines for Further Research and Development, Croatian Journal of Education Vol.20; No.2, 2018, 531-560.

[5] Paula Rodríguez et all, An educational recommender system based on argumentation theory, AI Communications 30 (2017), 19–36.

[6] learning Technology Standards Committee, IEEE Standard for Learning Object Metadata. Institute of Electrical and Electronics Engineers, New York (2002).

[7] Zhang, S.; Yao, L.; Sun, A. Deep learning based recommender system: A survey and new perspectives. arXiv 2017, arXiv:1707.07435.

[8] Ludewig,M.; Jannach, D. Evaluation of Session-based Recommendation Algorithms. arXiv 2018, arXiv:1803.09587.

[9] Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; Tikk, D. Session-based Recommendations with Recurrent Neural Networks. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2-4 May 2016; pp. 1-10.

[10] Paula Rodriguez et all, An educational recommender system based on argumentation theory, AI Communications 30 (2017), 19-36.

[11] Zhang, H., et al., MOOCRC: A highly accurate resource recommendation model for use in MOOC environments. Mobile Networks and Applications, 2019. 24(1): p. 34-46.

[12] Shuai Zhang, et. all, \Deep Learning based Recommender System: A Survey and New Perspectives", ACM Computing Surveys, Vol. 1, No. 1,2018.

[13] Zeynep Batmaz, Ali Yurekli, Alper Bilge and Cihan Kaleli, A review on deep learning for recommender systems: challenges and remedies, Springer Nature B.V. 2018.

[14] SHUAI ZHANG,LINA YAO, AIXIN SUN and YI TAY, Nanyang Technological University Deep Learning based Recommender System: A Survey and New Perspectives, 2018, ACM Computing Surveys, Vol. 1, No. 1, Article 1.

[15] Chen, H., et al., Enhanced learning resource recommendation based on online learning style model. Tsinghua Science and Technology, 2019. 25(3): p. 348-356.

[16] Li, R., et al., Online learning style modeling for course recommendation, in Recent Developments in Intelligent Computing, Communication and Devices. 2019, Springer. p. 1035-1042.

[17] Hagemann, N., M.P. O'Mahony, and B. Smyth. Visualising module dependencies in academic recommendations. in Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion. 2019.

[18] Rodríguez, P., et al., An educational recommender system based on argumentation theory. AI Communications, 2017. 30(1): p. 19-36.

[19] Dawen Liang, Rahul G Krishnan, Mathew D Ho man, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. arXiv preprint arXiv:1802.05814 (2018).

[20] Neto, J. Multi-agent web recommender system for online educational environments. in International Conference on Practical Applications of Agents and Multi-Agent Systems. 2017. Springer.

[21] Rodríguez, P., N. Duque, and S. Rodríguez, Integral Multi-agent Model Recommendation of Learning Objects, for Students and Teachers, in Management Intelligent Systems. 2013, Springer. p. 127-134.

[22] Ahmadian Yazdi, H., S.J. Seyyed Mahdavi Chabok, and M. Kheirabadi, Dynamic Educational Recommender System Based on Improved Recurrent Neural Networks Using Attention Technique. Applied Artificial Intelligence, 2021: p. 1-24.

[23] Shanker, M., M.Y. Hu, and M.S. Hung, Effect of data standardization on neural network training. Omega, 1996. 24(4): p. 385-397.

[24] Zhang, H., Huang, T., Zhihan, Lv., Liu, S., and Yang, H. MOOCRC: A Highly Accurate Resource Recommen- dation Model for Use in MOOC Environments. Springer, Mobile Networks and Applications, Springer ,2018.

[25]Charnelli, M.E., Sistemas recomendadores aplicados en Educación. 2019, Universidad Nacional de La Plata.

[26]Li, R., et al., Online learning style modeling for course recommendation, in Recent Developments in Intelligent Computing, Communication and Devices. 2019, Springer. p. 1035-1042.

[27]Yan, L., et al. Learning Resource Recommendation in E-Learning Systems Based on Online Learning Style. in International Conference on Knowledge Science, Engineering and Management. 2021. Springer.
[28] P. Resnick and H.R. Varian. Recommender systems. Communications of the ACM, 40(3):56.58, (1997).

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023,294-306
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

A New Power Allocation Optimization for One Target Tracking in Widely Separated MIMO Radar

Mohammad Akhondi Darzikolaei[1*], Mohammad Reza Karami Mollaei[1] , Maryam Najimi[2]

[1] Department of Electrical and Computer Engineering, Babol Noshirvani University of Technology,Babol, Iran
[2].Department of Electrical and Computer Engineering, University of Science and Technology of Mazandaran,Behshahr, Iran

Received: 03 Oct 2022/ Revised: 04 Dec 2022/ Accepted: 18 Jan 2023

Abstract

In this paper, a new power allocation scheme for one target tracking in MIMO radar with widely dispersed antennas is designed. This kind of radar applies multiple antennas which are deployed widely dispersed from each other. Therefore, a target is observed simultaneously from different uncorrelated angles and it offers spatial diversity. In this radar, a target's radar cross section (RCS) is different in each transmit-receive path. So, a random complex Gaussian RCS is supposed for one target. Power allocation is used to allocate the optimum power to each transmit antenna and avoid illuminating the extra power in the environment and hiding it from interception. This manuscript aims to minimize the target tracking error with constraints on total transmit power and the power of each transmit antenna. For calculation of target tracking error, the joint Cramer Rao bound for a target velocity and position is computed and this is assumed as an objective function of the problem. It should be noted that a target RCS is also considered as unknown parameter and it is estimated along with target parameters. This makes a problem more similar to real conditions. After the investigation of the problem convexity, the problem is solved by particle swarm optimization (PSO) and sequential quadratic programming (SQP) algorithms. Then, various scenarios are simulated to evaluate the proposed scheme. The simulation results validate the accuracy and the effectiveness of the power allocation structure for target tracking in MIMO radar with widely separated antennas.

Keywords: unknown RCS; Target tracking; Power allocation; MIMO radar; Joint Cramer Rao bound; PSO; SQP

1- Introduction

Radar system applies eletromagnetic waves to assign target position, velocity, and other features [1-2]. In the last decades, MIMO radars become an important and attractive issue in radar research [3]. A MIMO radar uses multiple receiver and transmitter antennas to illuminate the specific waveform [4]. The superiority of MIMO radars over conventional radars has been recently proved in many aspects. These radars include of many transmitters and receivers located far from each other. In this scenario, the MIMO radar can observe the targets from different directions. One of the advantages of these radars is exploitation of Doppler frequencies from different transmitter-target-receiver paths. The extracted Doppler frequencies can be used for estimation of target parameters so that, the radar can track the targets with reasonable accuracy[5]. Collocated and widely separated antennas are the common types of MIMO radar. In the collocated type, the antennas are deployed so near, similar to Phased Array. In the widely separated MIMO radar, all antennas are deployed in a great geographical environment and target is observed from various uncorrelated aspect angles.

Power consumption is an essential challenge in wireless networks in UAV communications [6], underwater communications, cooperative cognitive radio netwrok [7], and radar systems. Power allocation is usually applied to allocate the optimum power value between the transmit antennas. to minimize the tracking error with power constraints in transmitter or its converse is a common strategy for power allocation scheme in MIMO radar system [8]. Power allocation is also essential to hiding the radar from other LPI radars [9]. Power allocation technique in MIMO radar systems is investigated in recent research. Using power allocation technique in MIMO radar with widely separated antennas is investigated in [10]. In target tracking cycle just target range is

✉ **Mohammad Reza karami Mollaei**
mkarami@nit.ac.ir

considered. The problem is constructed by aiming to maximize B-FIM1. it is derived and then the problem is formed as one cooperative game. Then, the problem is solved to distribute the total power between transmitting antennas. the selection of antennas and power allocation technique for localization in distributed type of MIMO radar are proposed in [11]. A constrained problem by aiming to minimize the estimation error of target position is solved. The transmit antenna number and power budget were the constraints of this problem. [12] Introduces a joint method in antenna selection for target tracking problem in distributed MIMO radar. Resource restrictions in radars make it essential to choose radars at per time cycle and maintain the performance in the high condition. Therefore, the PCRLB 2 is applied as an optimization criterion for this problem. [13] Proposes a new resource allocation technique for the multi-target tracking in widely separated MIMO radar and it considers just a velocity as an unknown parameter. The authors selected one key target. They applied the MSE of that target velocity estimation as an optimization problem criterion. The choosing of receive and transmit antennas and assigning of transmit power and signal time are the parameters that are obtained in this problem. [14] Considers a netted Collocated MIMO radar and suggests joint beam and power schemes for multi-target tracking. a distributed fusion is also used to reduce the communication requirements while keeping the system robustness. The distributed fusion schemes use covariance intersection fusion. [15] Designs a joint antenna placement and power allocation technique in MIMO radar with widely separated antennas to increase target detection performance. First, a problem for the Neyman-Pearson detector by using the Lagrange power allocation scheme and the antenna deployment optimization is considered. Then with the iterative method, the problem is solved. The power allocation scheme based on PSO for one target tracking strategy is introduced in [16]. The problem is formed for MIMO radar with widely dispersed antennas. The power allocation technique with aiming to minimize the tracking error with constraints on the total power and each transmit antenna power is constructed. Then with the PSO algorithm, that problem is solved. In this reference, the joint target position and velocity are considered unknown parameters and then the CRB of estimation error is calculated and it is used as an objective function. In [17], a solution for joint beam and power scheduling in the netted Collocated MIMO radar systems for distributed multi-target tracking is suggested. This solution contains a distributed fusion architecture that decreases the communication requirements while maintaining the overall robustness of the system. The distributed fusion architecture employs the covariance intersection fusion to address the unknown information correlations among radar nodes. An adaptive sensor scheduling integrated with power and bandwidth allocation is presented for centralized multiple target tracking in the netted collocated MIMO radar in [18].

By reviewing the above research, in this paper, we slove some challenges including using joint target velocity and position in the calculation of target tracking error, and considering random complex Gaussian target RCS. Besides considering these two challenges, we consider random complex Gaussian RCS as an unknown parameter and it is estimated along with the target position and velocity parameter. This is similar to real conditions. Because in other research, they consider RCS known but is obvious that we usually do not have any information from the target RCS. Therefore, the estimation of RCS is a very essential issue that should be performed in the estimation cycle. To our knowledge, this is the first time performed for power allocation problem for target tracking in MIMO radar with widely separated antennas.

The scope of this manuscript are including:

1. First, The system model and the antenna deployment model for widely separated MIMO radar are determined. Then target motion model is chosen. random model with complex Gaussian distribution is selected for target RCS and it is used in the computation of Cramer Rao bound for target parameters estimation error. And also, besides the target parameters, the target RCS is considered an unknown parameter. (This is the first time considered for MIMO radar with widely separated antennas). It is the essential assumption because target RCS depends on many target factors and it cannot be known and should be achieved in the estimation process.
2. CRB for unknown target parameters and the variance of random RCS are computed and then Joint CRB for target velocity and position estimation is obtained. The joint CRB has used an objective function for the power allocation problem.
3. The power allocation scheme is designed. The minimizing one target tracking errors by considering the transmit power of each transmit antenna and total transmit power limitations is the power allocation problem of this

[1] Bayesian fisher information matrix
[2] Posterior Cramer Rao lower bound

manuscript. We aim to minimize the tracking errors with the above constraints.
4.The PSO and SQP algorithm are utilized to slove this problem. These algorithms are formed to assign optimal value to each transmit antenna and satisfy the constraints in the problem.
The remainder of the paper is structured as follows:
the system model is mentioned in section 2. In the next, target parameters error is calculated. Part 4 constructs a power allocation problem and applies two SQP and PSO algorithms for solving it. Section 5 presents the simulation results, the conclusion comments are mentioned in section 6, and in the final part, the appendices are presented.

2- System Model

In this model, nth receive antenna is located in (x_n, y_n), where $n = 1,2, \ldots\ldots, N$. The position of the target is in (x_q, y_q) and the target velocity equals $(\dot{x}_q, \dot{y}_q)$. A set of orthogonal signals,$s_m(t)$, is illuminated. ($\int_{T_m} |s_m(t)|^2 dt = 1$). period, effective bandwidth and transmit power of mth transmit waveform are shown as T_m,β_m, P_m. RCS of mnth path is expressed as a zero-mean complex Gaussian random variable $\xi_{mn} \sim \mathcal{CN}(0, \sigma_{mn}^2)$. Where σ_{mn}^2 is the mnth path variance and we conisder it unknown in this paper.
The assumptions of this paper are as follow:
1. w_{mn} (Noise of mnth transmit-receive path) and ξ_{mn} in mn th paths are mutually independent.
2. the transmit signals are orthogonal. This also true for time delays and Doppler shifts [19]:
3. we consider $\sigma_w^2 = 1$.
4. The antennas are adequately spaced far [20]. Therefore, the observation of the target in each path is independent and RCS, ξ_{mn}, is independent.
The time delay of (m, n)th transmit-receive path in kth time slot is:

$$\tau_{mn,k} = \frac{d_{m,k} + d_{n,k}}{c} \tag{1}$$

And,

$$d_{m,k} \triangleq \sqrt{\left(x_{q,k} - x_m\right)^2 + \left(y_{q,k} - y_m\right)^2}$$
$$d_{n,k} \triangleq \sqrt{\left(x_{q,k} - x_n\right)^2 + \left(y_{q,k} - y_n\right)^2} \tag{2}$$

Where, c is light velocity. $d_{m,k}$ and $d_{n,k}$show the distance from mth transmitter and nth receiver from the target.
The received signal from mth transmit antenna at nth receive antenna at time k is:

$$r_{mn,k}(t) = \sqrt{\alpha_{mn,k} P_m}\ \xi_{mn,k}\ s_m\left(t - \tau_{mn,k}\right) e^{j2\pi f_{mn,k} t} + w_{mn,k}(t) \tag{3}$$

where, $w_{mn,k} \sim \mathcal{CN}(0, \sigma_w^2)$. The loss of pass is illustrated as $\alpha_{mn,k} = \frac{1}{(4\pi)^3} \frac{1}{f_c^2} \frac{1}{d_{m,k}^2\, d_{n,k}^2}$. Where f_c is the carrier frequency.
Doppler frequency in mn path and time-slot k can be expressed as:

$$f_{mn,k} = \frac{\dot{x}_{q,k}\left(x_m - x_{q,k}\right) + \dot{y}_{q,k}\left(y_m - y_{q,k}\right)}{\lambda d_{m,k}} + \frac{\dot{x}_{q,k}\left(x_n - x_{q,k}\right) + \dot{y}_{q,k}\left(y_n - y_{q,k}\right)}{\lambda d_{n,k}} \tag{4}$$

λ shows the wavelength.

2-1- Motion Model

The constant velocity (CV) is considered for the target motion model of this paper. This model expressed as [10]:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{F}\boldsymbol{\theta}_k + \boldsymbol{w}'_k \tag{5}$$

$\boldsymbol{\theta}_k = \left[x_{q,k}, \dot{x}_{q,k}, y_{q,k}, \dot{y}_{q,k}\right]^T$ is unknown target position and velocity vector. That it will be estimated in tracking cycle. $\boldsymbol{w}'_k$ is a Gaussian vector and represent noise and it is modeled as $\mathcal{N}(0, \boldsymbol{\Sigma}_k)$. Where $\boldsymbol{\Sigma}$ illustrates the covariance matrix, and **F** shows the state transition matrix [20]:

$$\boldsymbol{F} = \begin{bmatrix} 1 & \mathrm{T} & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \mathrm{T} \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{6}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \frac{\mathrm{T}^2}{3} & \frac{\mathrm{T}^2}{2} & 0 & 0 \\ \frac{\mathrm{T}^2}{2} & \mathrm{T} & 0 & 0 \\ 0 & 0 & \frac{\mathrm{T}^2}{3} & \frac{\mathrm{T}^2}{2} \\ 0 & 0 & \frac{\mathrm{T}^2}{2} & \mathrm{T} \end{bmatrix} \mathrm{l} \tag{7}$$

sample intervals and the process noise density are shown in T and l.
In this paper, besides$\boldsymbol{\theta}_k$, the target RCS in each transmit-receive path, $\xi_{mn,k}$, is also supposed to be unknown. Therefore, the unknown parameter vector is changed as:

$$\boldsymbol{\theta'}_k = \left[x_{q,k}, \dot{x}_{q,k}, y_{q,k}, \dot{y}_{q,k}, \boldsymbol{\xi}_k^T\right]^T \tag{8}$$

Where, $\boldsymbol{\xi}_k = \left[\xi_{11,k}, \xi_{12,k}, \ldots\ldots, \xi_{NM,k}\right]^T$. In this case, the state transition matrix is changed as $'$:

$$\boldsymbol{F'} = \begin{bmatrix} \boldsymbol{F} & \mathbf{0}_{\mathrm{n}_\theta \times \mathrm{NM}} \\ \mathbf{0}_{\mathrm{NM} \times \mathrm{n}_\theta} & \boldsymbol{I}_{NM} \end{bmatrix} \tag{9}$$

It is noted that the RCS transition model is like first-order Markov process and it is obtained as [21]:

$$\boldsymbol{\xi}_k = \boldsymbol{\xi}_{k-1} + \boldsymbol{\mu}_{k-1} \tag{10}$$

Where $\boldsymbol{\mu}_{k-1}$ white Gaussian noise with $\boldsymbol{Q}_{\xi,k-1}$ covariance. Therefore, according to (9), the unknown parameter transition model is achieved as:

$$\boldsymbol{\theta'}_k = \boldsymbol{F'}\boldsymbol{\theta'}_{k-1} + \boldsymbol{\eta}_{k-1} \quad (11)$$

Where in above equations, n_θ shows the dimension of the unknown parameter vector and $\boldsymbol{\eta}_{k-1}$ is Gaussian noise with covariance equals to $\boldsymbol{Q}_{\boldsymbol{\theta'}} = blkdiag\{\boldsymbol{\Sigma}, \boldsymbol{Q}_{\xi,k}\}$.
Since $\xi_{mn,k}$ is random variable with zero mean and $\sigma^2_{mn,k}$ variance, therefore, the $\sigma^2_{mn,k}$ is used in unknown parameter vector instead of $\xi_{mn,k}$:

$$\boldsymbol{\theta'}_{\mathrm{k}} = \left[x_{q,k}, \dot{x}_{q,k}, y_{q,k}, \dot{y}_{q,k}, , \sigma^2_{11,k}, \sigma^2_{12,k}, \dots., \sigma^2_{NM,k}\right]^T \quad (12)$$

In the next part, the joint CRB for target tracking error is calculated.

3- Joint CRB for Target Tracking Error

Log-likelihood ratio of the unknown parameter ($\boldsymbol{\theta}_{\mathrm{k}}$) is obtained as [13]:

$$\mathrm{L}_{mn}\left(\boldsymbol{\theta}_k; r_{mn}(t)\right) = Ln\,\Lambda_{mn}\left(\boldsymbol{\theta}_k; r_{mn,k}(t)\right) = \frac{\sigma^2_{mn}P_m}{\sigma^2_{mn}P_m + 1}\left|\int_{-\infty}^{+\infty} r_{mn,k}(t)\, s^*_m\big(t - \tau_{mn,k}\big)e^{-j2\pi f_{mn,k}t}\, dt\right|^2 + C_{mn} \quad (13)$$

Where $C_{mn} = -\ln(\sigma^2_{mn}P_m + 1)$, and $r_{mn,k}(t)$ shows the observation signal in nth receiver from mth transmitter
According to our assumptions, RCS and noise are independent, the joint likelihood ratio term are achieved by [13]:

$$\Lambda_J\left(\boldsymbol{\theta}_k; \boldsymbol{r_k}(t)\right) = \prod_{m=1}^{M}\prod_{n=1}^{N}\Lambda_{mn}\left(\boldsymbol{\theta}_k; r_{mn,k}(t)\right) \quad (14)$$

Where $\boldsymbol{r_k}(t)$ is expressed as [16]:

$$\boldsymbol{r_k}(t) = [r_{11,k}(t), r_{12,k}(t), \dots, r_{NM,k}(t)] \quad (15)$$

According to [19], BIM for unknown parameter vector $\boldsymbol{\theta}_k$ is as:

$$\mathbf{J}_{\mathrm{B}}(\boldsymbol{\theta}_{\mathrm{k}}) = [\boldsymbol{\Sigma} + \mathbf{F}\,\mathbf{J}_{\mathrm{B}}^{-1}(\boldsymbol{\theta}_{\mathrm{k-1}})\,\mathbf{F}^{\mathrm{T}}]^{-1} + \mathbb{E}[\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{k}})] \quad (16)$$

Where $\mathbf{J}_{\mathrm{D}}$ represents the Fisher Information Matrix (FIM). For the CRB calculation, first, FIM is calculated [23]:

$$\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{k}}) = \mathbb{E}_{\boldsymbol{r_k}(t);\boldsymbol{\theta}_k}\{\nabla_{\boldsymbol{\theta}_k}\ln\Lambda_J(\boldsymbol{r_k}(t);\boldsymbol{\theta}_k)[\nabla_{\boldsymbol{\theta}_k}\ln\Lambda_J(\boldsymbol{r_k}(t);\boldsymbol{\theta}_k)]^T\} = -\mathbb{E}_{\boldsymbol{r_k}(t);\boldsymbol{\theta}_k}\{\nabla_{\boldsymbol{\theta}_k}[\nabla_{\boldsymbol{\theta}_k}\ln\Lambda_J(\mathbf{r_k}(t);\boldsymbol{\theta}_k)]^T\} \quad (17)$$

Since (13) is the function of $\tau_{mn,k}$ and $f_{mn,k}$, a new unknown parameter is defined as :

$$\boldsymbol{\vartheta'}_k = \left[\tau_{11,k}, \tau_{12,k}, \dots, \tau_{NM,k}, f_{11,k}, f_{12,k}, \dots, f_{NM,k}, \sigma^2_{11,k}, \sigma^2_{12,k}, \dots., \sigma^2_{NM,k}\right]^T \quad (18)$$

According to Chain rule, a new FIM is as:

$$\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta'}_{\mathrm{k}}) = \left(\nabla_{\boldsymbol{\theta'}_k}\boldsymbol{\vartheta'}_k{}^T\right)\mathbf{J}_{\mathrm{D}}(\boldsymbol{\vartheta'}_k)\left(\nabla_{\boldsymbol{\theta'}_{\mathrm{k}}}\boldsymbol{\vartheta'}_k{}^T\right)^T \quad (19)$$

$\nabla_{\boldsymbol{\theta'}_k}\boldsymbol{\vartheta'}_k{}^T$ Is calculated as (20).

$$\nabla_{\boldsymbol{\theta'}_k}\boldsymbol{\vartheta'}_k{}^T = \begin{bmatrix} \frac{\partial\tau_{11,k}}{\partial x_{q,k}} & \frac{\partial\tau_{12,k}}{\partial x_{q,k}} & \dots & \frac{\partial\tau_{NM,k}}{\partial x_{q,k}} & \frac{\partial f_{11,k}}{\partial x_{q,k}} & \frac{\partial f_{12,k}}{\partial x_{q,k}} & \dots & \frac{\partial f_{NM,k}}{\partial x_{q,k}} & 0 & \dots & 0 \\ \frac{\partial\tau_{11,k}}{\partial y_{q,k}} & \frac{\partial\tau_{12,k}}{\partial y_{q,k}} & \dots & \frac{\partial\tau_{NM,k}}{\partial y_{q,k}} & \frac{\partial f_{11,k}}{\partial y_{q,k}} & \frac{\partial f_{12,k}}{\partial y_{q,k}} & \dots & \frac{\partial f_{NM,k}}{\partial y_{q,k}} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \frac{\partial f_{11,k}}{\partial \dot{x}_{q,k}} & \frac{\partial f_{12,k}}{\partial \dot{x}_{q,k}} & \dots & \frac{\partial f_{NM,k}}{\partial \dot{x}_{q,k}} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \frac{\partial f_{11,k}}{\partial \dot{y}_{q,k}} & \frac{\partial f_{12,k}}{\partial \dot{y}_{q,k}} & \dots & \frac{\partial f_{NM,k}}{\partial \dot{y}_{q,k}} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & \dots & 1 \end{bmatrix}_{5\times 3NM} \quad (20)$$

The number of 0 and 1 in the right side of above matrix is NM. The above matrix parameters are calculated in [16]. By defining $\mathbf{J}_{\mathrm{D}}(\boldsymbol{\vartheta'}_k)$ as:

$$\mathbf{J}_{\mathrm{D}}(\boldsymbol{\vartheta'}_k) = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} & \boldsymbol{G}_{13} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} & \boldsymbol{G}_{23} \\ \boldsymbol{G}_{31} & \boldsymbol{G}_{32} & \boldsymbol{G}_{33} \end{bmatrix} \quad (21)$$

Where, $\boldsymbol{G}_{11}$ contains second-order derivatives with respect to $\tau_{mn,k}$, $\boldsymbol{G}_{12}, \boldsymbol{G}_{21}$ are second-order derivatives with respect to $\tau_{mn,k}$ and $f_{mn,k}$, $\boldsymbol{G}_{22}$ is second-order derivatives with respect to $f_{mn,k}$, $\boldsymbol{G}_{13}, \boldsymbol{G}_{31}$ are second-order derivatives with respect to $\tau_{mn,k}$ and $\sigma^2_{mn,k}$, $\boldsymbol{G}_{23}, \boldsymbol{G}_{32}$ includes second-order derivatives with respect to $f_{mn,k}$ and $\sigma^2_{mn,k}$, and $\boldsymbol{G}_{33}$ contains second-order derivatives with respect to $\sigma^2_{mn,k}$ for all m and n in time slot k.. Therefore, (we show the proof procedure in **Appendix I**):

$$\boldsymbol{G}_{11} = \boldsymbol{C}\odot(\boldsymbol{I}_N \otimes diag\,\{\epsilon_1, \epsilon_2, \dots, \epsilon_M\}) \quad (22)$$
$$\boldsymbol{G}_{12} = \boldsymbol{G}_{21} = \boldsymbol{C}\odot diag\left\{\gamma_{11,k}, \gamma_{12,k}, \dots, \gamma_{NM,k}\right\} \quad (23)$$
$$\boldsymbol{G}_{22} = \boldsymbol{C}\odot diag\left\{\eta_{11,k}, \eta_{12,k}, \dots, \eta_{NM,k}\right\} \quad (24)$$
$$\boldsymbol{G}_{13} = \boldsymbol{G}_{31} = \boldsymbol{C_1}\odot(\boldsymbol{I}_N \otimes diag\,\{\chi_1, \chi_2, \dots, \chi_m\}) \quad (25)$$
$$\boldsymbol{G}_{23} = \boldsymbol{G}_{32} = \boldsymbol{C_1}\odot diag\left\{\varpi_{11,k}, \varpi_{12,k}, \dots, \varpi_{NM,k}\right\} \quad (26)$$
$$\boldsymbol{G}_{33} = \boldsymbol{C_2} \quad (27)$$

Where,

$$\boldsymbol{C} = 8\pi^2\, diag\,\{\frac{\sigma^4_{11}P^2_1}{\sigma^2_{11}P_1 + 1}, \dots., \frac{\sigma^4_{MN}P^2_M}{\sigma^2_{MN}P_M + 1}\} \quad (28)$$

$$\boldsymbol{C_1} = 4\pi\, diag\,\{\frac{\sigma^2_{11}P^2_1}{\sigma^2_{11}P_1 + 1}, \dots., \frac{\sigma^2_{MN}P^2_M}{\sigma^2_{MN}P_M + 1}\} \quad (29)$$

$$\boldsymbol{C_2} = diag\{\frac{2P_1^2}{\sigma_{11}^2 P_1 + 1}, \ldots, \frac{2P_M^2}{\sigma_{MN}^2 P_M + 1}\} \quad (30)$$

And also:

$$\chi_{m,k} = -j(\int f|S_m(f)|^2 df) \quad (31)$$

$$\varpi_{mn,k} = j\left(\int_{-\infty}^{+\infty} t \mid s_m^*(t-\tau_{mn,k})\right|^2 dt\Big) \quad (32)$$

$$\epsilon_m = \int_{-\infty}^{+\infty} f^2 |S_m(f)|^2 df - \left|\int_{-\infty}^{+\infty} f |S_m(f)|^2 df\right|^2 \quad (33)$$

$$\gamma_{mn,k} = \frac{1}{2\pi}\Im\{\int_{-\infty}^{+\infty} t\, s_m^*(t - \tau_{mn,k})\frac{\partial s_m(t-\tau_{mn,k})}{\partial \tau_{mn,k}} dt\} - \int_{-\infty}^{+\infty} f |S_m(f)|^2 df \cdot \int_{-\infty}^{+\infty} t |s_m(t-\tau_{mn,k})|^2 dt \quad (34)$$

$$\eta_{mn,k} = \int_{-\infty}^{+\infty} t^2 |s_m(t-\tau_{mn,k})|^2 dt - \left|\int_{-\infty}^{+\infty} t |s_m(t-\tau_{mn,k})|^2 dt\right|^2 \quad (35)$$

Where, $S_m(f)$ denotes a Fourier transform of $s_m(t)$.
If we divide the $\nabla_{\boldsymbol{\theta'}_k}\boldsymbol{\vartheta'}_k{}^T$ to matrix block as:

$$\nabla_{\boldsymbol{\theta'}_k}\boldsymbol{\vartheta'}_k{}^T = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0_{NM}} & \boldsymbol{0_{NM}} & \boldsymbol{1_{NM}} \end{bmatrix} \quad (36)$$

Where, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{D}, \boldsymbol{0}$ are the $2 \times NM$ matrices and $\boldsymbol{0_{NM}}$ and $\boldsymbol{1_{NM}}$ are zero and one vectors with $1 \times NM$ dimension.

Since $\mathbf{J}_{\mathrm{D}}(\boldsymbol{\vartheta'}_k) = \begin{bmatrix} \boldsymbol{G_{11}} & \boldsymbol{G_{12}} & \boldsymbol{G_{13}} \\ \boldsymbol{G_{21}} & \boldsymbol{G_{22}} & \boldsymbol{G_{23}} \\ \boldsymbol{G_{31}} & \boldsymbol{G_{32}} & \boldsymbol{G_{33}} \end{bmatrix}$, therefore:

$$\boldsymbol{J}_D(\boldsymbol{\theta'}_k) = \begin{bmatrix} \mathbf{A} & \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \mathbf{0} \\ \mathbf{0_{NM}} & \mathbf{0_{NM}} & \mathbf{1_{NM}} \end{bmatrix}\begin{bmatrix} \mathbf{G_{11}} & \mathbf{G_{12}} & \mathbf{G_{13}} \\ \mathbf{G_{21}} & \mathbf{G_{22}} & \mathbf{G_{23}} \\ \mathbf{G_{31}} & \mathbf{G_{32}} & \mathbf{G_{33}} \end{bmatrix}\begin{bmatrix} \mathbf{A} & \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} & \mathbf{0} \\ \mathbf{0_{NM}} & \mathbf{0_{NM}} & \mathbf{1_{NM}} \end{bmatrix}^{\mathrm{T}} \quad (37)$$

And

$$\boldsymbol{J}_D(\boldsymbol{\theta'}_k) = \begin{bmatrix} \mathbf{AG_{11}A^T + BG_{21}A^T + AG_{12}B^T + BG_{22}B^T} & \mathbf{AG_{12}D^T + BG_{22}D^T} & \mathbf{AG_{13}1_{NM}^T + BG_{23}1_{NM}^T} \\ \mathbf{DG_{21}A^T + DG_{22}B^T} & \mathbf{DG_{22}D^T} & \mathbf{DG_{23}1_{NM}^T} \\ \mathbf{1_{NM}G_{31}A^T + 1_{NM}^T G_{32}B^T} & \mathbf{1_{NM}G_{32}D^T} & \mathbf{1_{NM}G_{33}1_{NM}^T} \end{bmatrix} \quad (38)$$

After making block matrix:

$$\boldsymbol{J}_D(\boldsymbol{\theta'}_k) = 8\pi^2 \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \frac{\sigma_{mn}^4 P_m^2}{\sigma_{mn}^2 P_m + 1}\begin{bmatrix} \boldsymbol{J}_{Dnew}^{\boldsymbol{UL}} & \boldsymbol{J}_{Dnew}^{\boldsymbol{UR}} \\ \boldsymbol{J}_{Dnew}^{\boldsymbol{LL}} & \boldsymbol{J}_{Dnew}^{\boldsymbol{LR}} \end{bmatrix} \quad (39)$$

Therefore, $\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{UL}}$, $\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{UR}}$, $\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{LL}}$, and $\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{LR}}$ are obtained in (40), (41), (42), and (43).

$$\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{UR}} = (\frac{1}{2\pi\sigma_{mn}^2})\begin{bmatrix} \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}(a_{mn,k}\chi_{m,k} - e_{mn,k}\varpi_{mn,k}) \\ \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}(b_{mn,k}\chi_{m,k} - g_{mn,k}\varpi_{mn,k}) \\ -V_{mn,k}\varpi_{mn,k} \\ -q_{mn,k}\varpi_{mn,k} \end{bmatrix} \quad (41)$$

$$\boldsymbol{J}_{Dnew}^{\boldsymbol{LL}} = \left(\frac{1}{2\pi\sigma_{mn}^2}\right) \times \begin{bmatrix} \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}(a_{mn,k}\chi_{m,k} - e_{mn,k}\varpi_{mn,k}) & \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}}\sum_{\boldsymbol{n=1}}^{\boldsymbol{N}}(b_{mn,k}\chi_{m,k} - g_{mn,k}\varpi_{mn,k}) & -V_{mn,k}\varpi_{mn,k} & -q_{mn,k}\varpi_{mn,k} \end{bmatrix} \quad (42)$$

$$\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{LR}} = \left(\frac{MN}{4\pi^2\sigma_{mn}^4}\right) \quad (43)$$

We can reform the (39) as :

$$\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta'}_{\mathrm{k}}) = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2 \frac{\sigma_{mn}^4 P_m^2}{\sigma_{mn}^2 P_m + 1} \cdot (\mathrm{G_{mn}})_{5\times 5} \quad (44)$$

Where $\mathrm{G_{mn}}$ is function of target velocity and position and target RCS. $\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta'}_{\mathrm{k}})$ is 5×5 block matrix which is function transmit power of transmit antenna.
In the next part, the power allocation problem is constructed.

4- Power Allocation

In the construction of the power allocation problem, trace of Cramer Rao Bound matrix is considered as tracking error and it is obtained as:

$$\mathbb{F}_{\mathrm{B}}(\boldsymbol{\theta'}_k, \boldsymbol{P}) = \mathrm{trace}(\boldsymbol{\Upsilon}_{\mathrm{k}}([\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{1,1} + [\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{2,2} + [\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{3,3} + [\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{4,4}\boldsymbol{\Upsilon}_{\mathrm{k}}^{\mathrm{T}}) \quad (45)$$

Where, $[\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{1,1}$, $[\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{2,2}$, $[\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{3,3}$, and $[\boldsymbol{J}_D^{-1}(\boldsymbol{\theta'}_k)]_{4,4}$ are the CRB (lower bound) of the variance

$$\mathbf{J}_{\mathrm{Dnew}}^{\boldsymbol{UL}} = \begin{bmatrix} \epsilon_m a_{mn,k}^2 + 2\gamma_{mn,k}\, a_{mn,k}\, e_{mn,k} + \eta_{mn,k} e_{mn,k}^2 & (\epsilon_m a_{mn,k} + \gamma_{mn,k} e_{mn,k}) b_{mn,k} + (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) g_{mn,k} & (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) V_{mn,k} & (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) q_{mn,k} \\ (\epsilon_m a_{mn,k} + \gamma_{mn,k} e_{mn,k}) b_{mn,k} + (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) g_{mn,k} & \epsilon_m b_{mn,k}^2 + 2\gamma_{mn,k}\, b_{mn,k}\, g_{mn,k} + \eta_{mn,k} g_{mn,k}^2 & (\gamma_{mn,k} b_{mn,k} + \eta_{mn,k} g_{mn,k}) V_{mn,k} & (\gamma_{mn,k} b_{mn,k} + \eta_{mn,k} g_{mn,k}) q_{mn,k} \\ (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) V_{mn,k} & (\gamma_{mn,k} b_{mn,k} + \eta_{mn,k} g_{mn,k}) V_{mn,k} & \eta_{mn,k} V_{mn,k}^2 & \eta_{mn,k} V_{mn,k} q_{mn,k} \\ (\gamma_{mn,k} a_{mn,k} + \eta_{mn,k} e_{mn,k}) q_{mn,k} & (\gamma_{mn,k} b_{mn,k} + \eta_{mn,k} g_{mn,k}) q_{mn,k} & \eta_{mn,k} V_{mn,k} q_{mn,k} & \eta_{mn,k} q_{mn,k}^2 \end{bmatrix} \quad (40)$$

of target position and also target velocity in axis of x and y in kth frame. and $\mathbf{P} = [P_1, \dots, P_M]$ is a transmit antenna power vector, $\boldsymbol{\Upsilon}_k$ shows the normalization matrix and the target tracking error is illustrated in $\mathbb{F}_B$. $\boldsymbol{\Upsilon}_k$ is introduced as $\boldsymbol{\Upsilon}_k = I_2 \otimes \begin{bmatrix} 1 & 0 \\ 0 & T \end{bmatrix}$, Where I_2 is 2×2 identity matrix and $\otimes$ denotes the Kronecker product operator.

Our power allocation problem for target tracking in widely separated MIMO Radar by applying random Gaussian RCS and also considering target RCS as unknown parameter is expressed as:

$$\min_{P_m} \mathbb{F}_B(\boldsymbol{\theta'}_k, \boldsymbol{P}) \tag{46}$$

$$s.t. \sum_{m=1}^{M} P_m \le P_T \tag{47}$$

$$P_{min} \le P_m \le P_{max} \quad , m = 1,2, \dots, M \tag{48}$$

the first constraint of this problem shows that the sum of transmit powers is lower than a predetermined value, P_T. since MIMO radar wants to utilize the least power for one target tracking and it avoids to being intercepted. another constraint illustrates that each transmit antenna also has power limitation. To solve the (46) subject to (47) and (48), we use PSO and SQP algorithms. In **appendix III**, we prove that the problem is convex. The overall structure of solving the power allocation problem of this paper is illustrated in Fig.1.

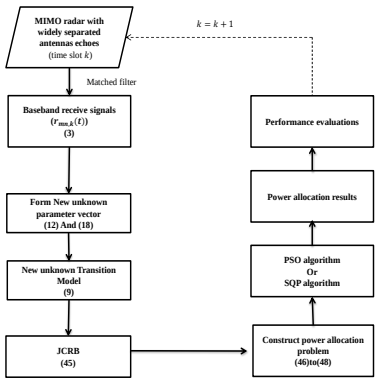


Fig.1. The Overall Structure of Solving the Power Allocation Problem of This Paper

4-1- The Problem Solution Based on PSO

The PSO algorithm [25] is based on animal's social behavior. The swarms construct a cooperative approach to find food and each member keeps varying the search pattern based on the learning experiences of its own and others. The pseudo code of algorithm to solve the problem of this paper is described in Algorithm1.

Algorithm1 The pseudo-code of PSO algorithm for this problem solution

1.Initialize parameters $(M, N, K, dimension(nvar), popsize, Maxiter, \omega, c_1, c_2, \omega_{damp}, P_{min}, P_{max})$

2. Uniformly randomly initialize each P_m in the population

$$par.var = P_{min} + (P_{max} - P_{min}). * rand(1, nvar) \tag{49}$$

3. define initial velocities randomly for each particle

$$par.vel = 0 \tag{50}$$

4. the fitness evaluation of each particle with the Cost function (51)

$$Cost(P_m) = \mathbb{F}_B(\boldsymbol{\theta'}_k, \mathbf{P}) + \gamma * (\sum_{m=1}^{M} P_m - P_T) \tag{51}$$

(Note that γ is a great value coefficient, to illustrate better constraint (47) in (46)).

$$par.cost = Cost(par.var) \tag{52}$$

5. define p_{best} and g_{best} in population and time slot (k) (in this paper is the minimum case)

6. while $iter < Maxiter$ do

7. for $k = 1{:}K$

8. for $i = 1{:}popsize$

9. Update velocity

$$par(i).vel = w * par(i).vel + \dots c1 * rand * (bpar(i).var - par(i).var) + \dots c2 * rand * (gpar(k).var - par(i).var); \tag{53}$$

10. Update variable

$$par(i).var = par(i).var + par(i).vel; \tag{54}$$

11. over merge checking

$$par(i).var = min(par(i).var, P_{max}) \tag{55}$$

$$par(i).var = max(par(i).var, P_{min}) \tag{56}$$

12. Compute the fitness values of new particle P_m with the Cost function (51)

$$par(i).cost = Cost(par(i).var) \tag{57}$$

13. If new particle value from (57) is better than p_{best} and p_{best} is better than g_{best} ,

Define new P_m as an optimal variable.

```
if par(i).cost < bpar(i).cost
    bpar(i) = par(i)
    if bpar(i).cost < gpar(k).cost
        gpar(k) = bpar(i);
    End if
End if
```
(58)

14. update decreasing coefficient ω

$$\omega = \omega * \omega_{damp} \tag{59}$$

15. End (for k)

16. End (for i)

17. $iter = iter + 1$;

18. End (while)

4-2- The Problem Solution Based on SQP

To prove the performance of first algorithm, and also because the problem is convex, we use another algorithm named SQP. In addition, the time complexity of PSO is

usually high and the SQP has less time consumption, we prefer to utilize SQP and compare these two algorithm results for our power allocation problem.

The structure of SQP algorithm for nonlinear problem is described as [24]:

$$\text{Min: } f(\mathbf{x}) \quad (60)$$

$$\text{s.t: } g_i(\mathbf{x}) \le 0, \quad i = 1, \dots, n_g \quad (61)$$

$$h_i(\mathbf{x})=0, \quad i = 1, \dots, n_h \quad (62)$$

$$x^l < \mathbf{x} < x^u \quad (63)$$

Where, f denotes an objective function, g and h show the inequality and equality function and $f, g,$ and h are twice continuously differentiable. $\boldsymbol{x}$ is the favorable variable matrix and it is limited by upper and lower bound x^u and x^l.

$\boldsymbol{\lambda}^k$ and $\boldsymbol{\nu}^k \ge 0$ are the Lagrange coefficients. Consider the below QP sub-problem as a direct extension of QP($\boldsymbol{x}^k, \boldsymbol{\lambda}^k$):

$$min_{\mathbf{d}^k} f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T \mathbf{d}^k + \frac{1}{2}\mathbf{d}^{kT}\nabla^2_{xx}\mathcal{L}(\mathbf{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k) \quad (64)$$

$$\mathbf{d}^k$$

$$s.t.\ g_i(\mathbf{x}) + \nabla g_i(\mathbf{x}^k)^T \mathbf{d}^k = 0, \quad i = 1, \dots, n_g \quad (65)$$

$$h_i(\mathbf{x}) + \nabla h_i(\mathbf{x}^k)^T \mathbf{d}^k = 0, \quad i = 1, \dots, n_h \quad (66)$$

(53) is named as QP($\boldsymbol{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k$) and $\mathcal{L}(\boldsymbol{x}^k, \boldsymbol{\lambda}^k, \boldsymbol{\nu}^k) = f(\boldsymbol{x}) + \boldsymbol{\nu}^T \boldsymbol{g}(\boldsymbol{x}) + \boldsymbol{\lambda}^T \boldsymbol{h}(\boldsymbol{x})$.

In this paper, problem (46) is the objective function and P_m is our favorable variable. (46) is f in (60) and $\mathbf{P}$ in (46) is $\boldsymbol{x}$ in (60). By supposing $\sum_{m=1}^{M} P_m - P_T \le 0$, we can say that $\boldsymbol{g}(\boldsymbol{x})$ in (61) is equal to $\sum_{m=1}^{M} P_m - P_T$ in our problem. Fig.2 shows the flowchart of SQP algorithm which is used in this paper.

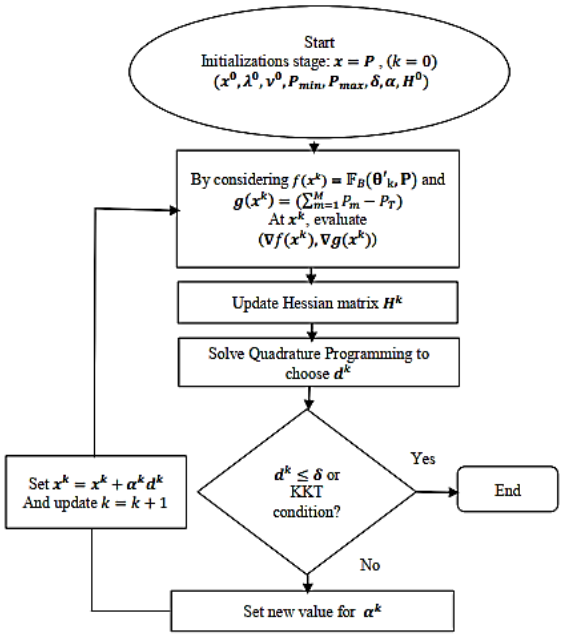


Fig.2. The Flowchart of SQP Algorithm for Solving the Problem of this Paper

5- Simulations

We perform some experiments to evaluate the proposed power allocation scheme. All the simulations are performed by Matlab software. In this paper, two symmetric and one asymmetric geometrical antenna placement scenarios are considered to illustrate the effect of antenna placement on tracking performance. Fig. 3 shows these two symmetric schemes for a MIMO radar with $M = 4$ and $N = 4$. To analyze the effect of the number of antennas on target tracking performance, it is considered another symmetric scenario with $M = 6$ and $N = 6$. Fig.4 shows this antenna placement geometry.

In symmetric cases, all antennas have ten kilometers distance from the origin. P_T equals 10 $Kwatt$. The carrier frequency is considered $9GHz$. σ^2_{mn} is supposed random and unknown parameter and in each transmit-receive path, it is different (in other research, for simplicity, it is usually considered one and known). $l = 0.1$ and $T = 0.2s$. The initial value for target location and velocity in x and y axis is $[500m\ 1000m\ 50\frac{m}{s}\ \ 30\frac{m}{s}]$. $P_{min} = 0.02P_T$ and $P_{max} = 0.8P_T$(watt).

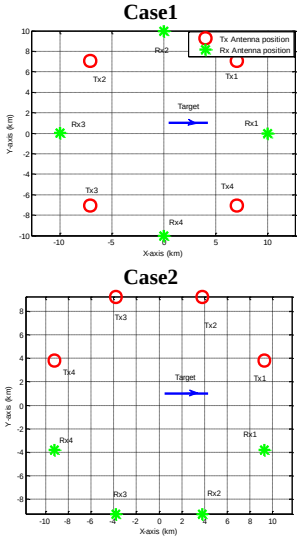


Fig.3 Symmetric Antenna Geometry Placement (M = 4, N = 4)

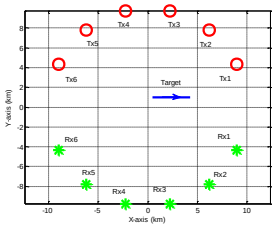


Fig.4 Symmetric Antenna Geometry Placement (M = 6, N = 6)

In Fig.5. We consider asymmetric antenna placement for a MIMO radar with $(M = 6, N = 6)$.

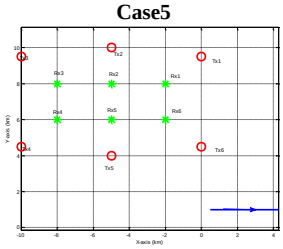
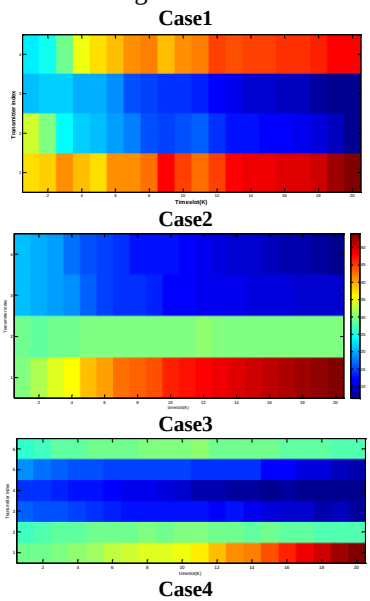


Fig.5 Asymmetric Antenna Geometry Placement (M = 6, N = 6)

To evaluate the proposed power allocation scheme for target tracking in widely separated MIMO radar, first, we apply PSO algorithm and extract the results. In Fig.6, the transmit power percentage of each transmit antenna in five considered cases (1 to 5) is shown. We can realize from Fig.6 that by moving the target toward the transmit antenna, the more power is allocated to that antenna. Therefore, in Case1, the transmit antenna 1and 4, in Case2, the transmit antenna 1and 2, in Case3, the transmit antenna 1and 6, in Case4, the transmit antenna 1and 2, and in Case5, the transmit antenna 1 and 6, take more power to have a better target tracking performance. In symmetric Cases (1 to 4), if the target is placed in (0,0), the power is equally distributed among transmit antennas.



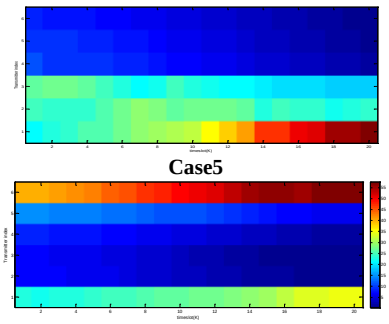
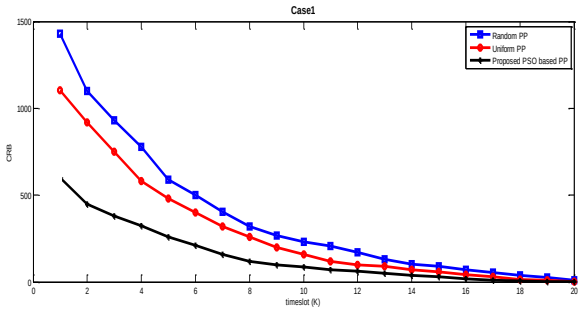
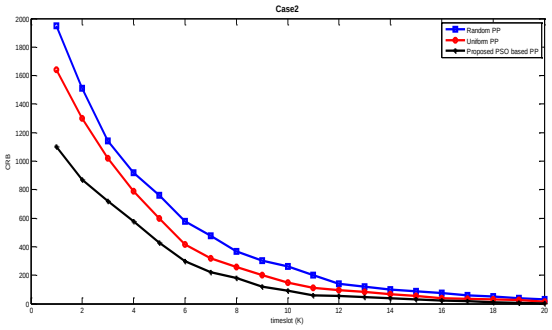


Fig.6. Each Transmit Antenna Power Percentage in Five Cases in Different Timeslots (based on PSO-based PP strategy)

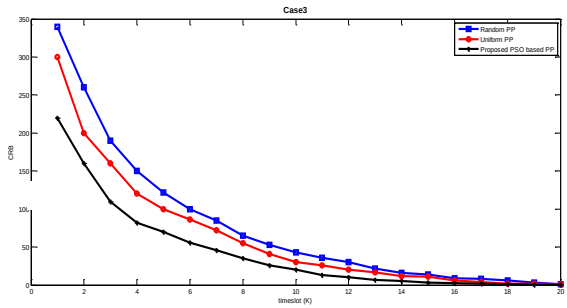
Fig.7 illustrates that the proposed power allocation strategy (based on the PSO algorithm) has better performance and the less CRB of target tracking error than other schemes such as uniform and random power allocation. We can see this priority in all cases (Case1 (a), Case2 (b), Case3 (c), Case4 (d), and Case5 (e) in Fig.7). By attention to this figure, we can realize that by increasing the number of antennas, the performance is growing and the target tracking error is decreasing. In addition, by comparing Case3 and Case4 (symmetric geometry) with Case5 (asymmetric geometry), it is obvious that the symmetric configuration has better performance. (Note that in the all figures and scenarios, the unit of CRB and MSE is m^2).



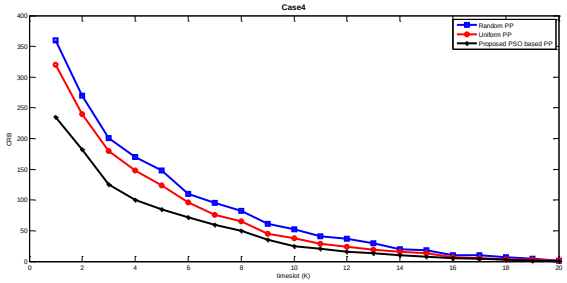
(a)



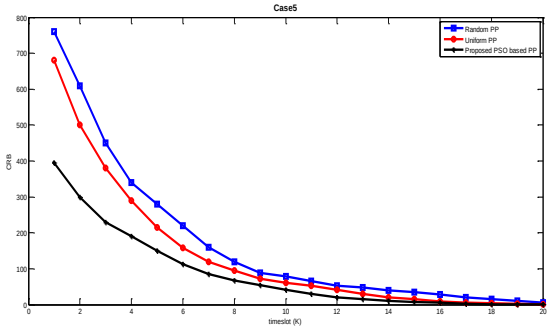
(b)



(c)



(d)



(e)

Fig.7. The Proposed Power Allocation Scheme (based on PSO) Comparison with Random and Uniform Power Allocation Schemes in the all Five Cases

To verify the accuracy of the proposed target tracking schemes, Fig.8 illustrates the tracking MSE and joint CRB in five Cases. The MSE is calculated as:

$$\mathrm{MSE_k} = \frac{1}{N_{MC}} \sum_{i=1}^{N_{MC}} trace(\boldsymbol{\Upsilon}_k(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k^j)(\boldsymbol{\theta}_k - \widehat{\boldsymbol{\theta}}_k^j)^T \boldsymbol{\Upsilon}_k^T) \quad (67)$$

Where N_{MC} is the Monte Carlo number and $\widehat{\boldsymbol{\theta}}_k^j$ is the state estimate in the jth cycle. The joint CRB and MSE results in Fig.8, shows that the proposed tracking scheme results is close to actual conditions. This is true in the all five cases.

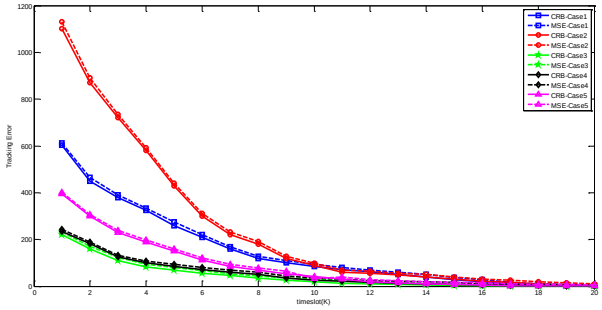


Fig.8. Tracking Errors in Five Cases with Proposed Target Tracking procedure (based on PSO)

In addition, Fig.8 shows that Case3 has the least tracking error and it is the best case. And also, the asymmetric Case5 has the better performance than Case1 and Case2 because the number of antennas in this case is more than those two cases. But in equal number of antennas, the symmetric Cases (Case3 and Case4) has the less target tracking error than asymmetric Case5. In Fig.9. we compare the joint CRB of the tracking error of the proposed power allocation scheme (based on PSO) is compared with the Exhaustive search method [16], which is the best algorithm for finding the result because it considers all possible conditions. This comparison is performed the best Case, Case3.

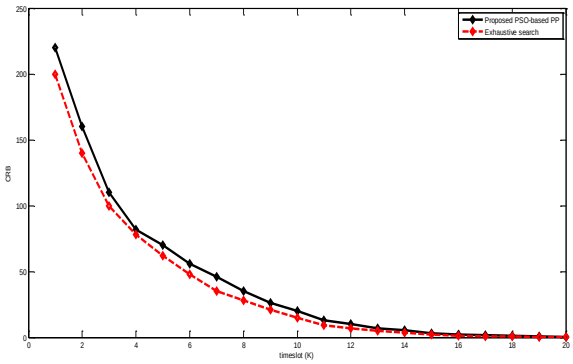


Fig.9. The Proposed Power Allocation scheme in Target Tracking (based on PSO) and Exhaustive Search Performance Comparison in Case3

By attention to Fig.9, it is clear that the proposed scheme result (based on PSO) for Case3 is near to the Exhaustive search method. This can prove the accuracy of the proposed strategy. However, the Exhaustive search has a high computational complexity and it takes about 49583 seconds. However, the proposed scheme (based on PSO) takes 804 seconds. The simulations are run in the system with Intel(R) core(TM) i7-3612QM CPU @2.1GHz and 6 GB RAM. In addition, the number of possible variable P_m for exhaustive search equals 5. To verify the proposed scheme (based on PSO) results, we repeat the experiments with SQP algorithm. We use this algorithm for the best Case in the previous experiment, Case3. Fig.10 shows the joint CRB of tracking error in the proposed power allocation scheme (based on SQP) with uniform and

random power allocation strategies and also the proposed scheme based on PSO algorithm. This figure illustrates that the proposed scheme based on SQP has the best performance. But its results are very close to the proposed schemes based on PSO. This proves that our proposed scheme has high accuracy and performance.

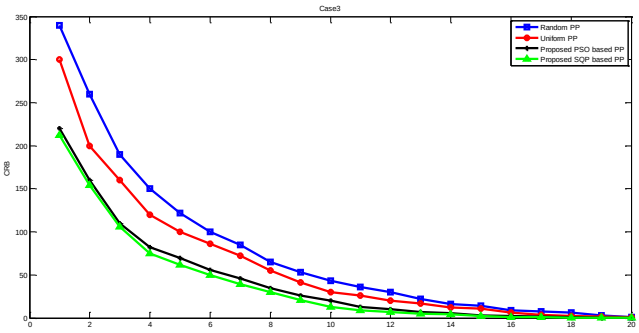


Fig.10. The Comparison of the Proposed power Allocation Scheme in Target Tracking (based on SQP) with Other Strategies in Case3

It should be emphasized that although the two SQP and PSO algorithm has the near performance for our proposed power allocation scheme in target tracking problem, but SQP has the less computational complexity than PSO and it takes 56.8612 seconds. Therefore, it is applicable in real-time scenarios. In Fig.11, the joint CRB of target tracking error of the proposed scheme (based on SQP) is compared with MSE. This shows that two results are near to each other. Therefore, the SQP based scheme is also close to real condition.

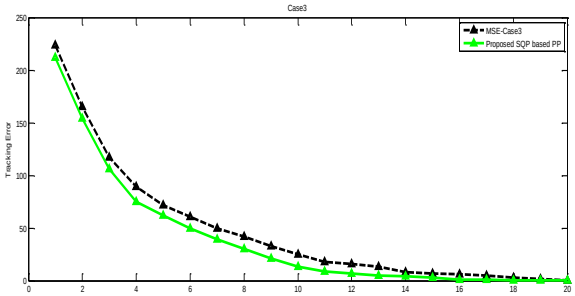


Fig.11 Target Tracking Error Evaluation in Case3 with Proposed Target Tracking Procedure (based on SQP algorithm)

6- Conclusions

The researchers should emphasize more in power allocation strategy on MIMO radar with widely dispersed antennas. Based on the limitations in the total power in the MIMO radar, the power allocation is critical.

In this manuscript, the power allocation scheme performance in widely separated MIMO radar is investigated. A complex Gaussian random RCS with different variance in each transmit-receive path is supposed. And also it is considered an unknown parameter. These are not considered in other papers and this condition is near to real. The simulation results prove that these assumptions enhance the radar performance.

Applying joint estimation of target velocity and position tracking error and also adding RCS estimation to the estimation state vector helps to improve the performance of this kind of MIMO radar. In simulations, five different Cases with symmetric and asymmetric antenna placement are considered to evaluate the proposed power allocation scheme for the target tracking problem in considered MIMO radar. This paper aims to form the power allocation scheme to minimize the tracking errors subject to the total transmit power and transmit power of each transmit antenna limitation. We proved that this problem is convex and used PSO and SQP algorithms to solve it. The simulation experiments are performed in various scenarios and the simulation proves the accuracy of the proposed scheme.

References

[1] M.A. Darzikolaei, A.Ebrahimzade, and E.Gholami, "Classification of radar clutters with artificial neural network." In 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), 2015, pp. 577-581.

[2] M.A. Darzikolaei, A.Ebrahimzade, and E.Gholami,"The Separation of Radar Clutters using Multi-Layer Perceptron.", Information Systems & Telecommunication, Vol.1, No.17,2017,pp 1-10.

[3] E. Fishler, A.Haimovich, R.Blum, D.Chizhik, L.Cimini, and R.Valenzuela,"MIMO radar: An idea whose time has come", In Proceedings of the 2004 IEEE Radar Conference (IEEE Cat. No. 04CH37509), , 2004, pp. 71-78.

[4] A. Pakdaman, and H.Bakhshi, "Separable transmit beampattern design for MIMO radars with planar colocated antennas", AEU-International Journal of Electronics and Communications, Vol.89, No.1, 2018,pp.153-159.

[5] M.J.Jahromi,and H.K.Bizaki, "Target Tracking in MIMO Radar Systems Using Velocity Vector",Journal of Information Systems and Telecommunication (JIST),Vol.3,No.7, 2014,pp. 150-158.

[6] S.H. Mostafavi-Amjad, V. Solouk, and H. Kalbkhani, "Energy-efficient user pairing and power allocation for granted uplink-NOMA in UAV communication systems", Journal of Information Systems and Telecommunication (JIST), Vol. 10, No. 40, 2014, pp.312-323.

[7] M. G. Adian, and H. Aghaeenia, "Joint relay selection and power allocation in MIMO cooperative cognitive radio networks", Journal of Information Systems and Telecommunication (JIST), Vol. 1, No. 9, 2015, pp.1-10.

[8] X.Mingchi, W.Yi, T.Kirubarajan, and L. Kong, "Joint node selection and power allocation strategy for multitarget tracking in decentralized radar networks", IEEE Transactions on Signal Processing, Vol.66, No. 3, 2017, pp.729-743.

[9] H. Godrich, A.P. Petropulu, and H. V.Poor,"Power allocation strategies for target localization in distributed multiple-radar architectures", IEEE Transactions on Signal Processing, Vol.59, No. 7, 2011,pp. 3226-3240.

[10] H.Chen, T.Shiying, and S.Bin,"Cooperative game approach to power allocation for target tracking in distributed MIMO radar sensor networks", IEEE Sensors Journal,Vol.15, No. 10, 2015, pp.5423-5432.
[11] M. Botao, H.Chen, S.Bin, and H,Xiao,"A joint scheme of antenna selection and power allocation for localization in MIMO radar sensor networks", IEEE communications letters, Vol.18, No. 12, 2014,pp.2225-2228.
[12] L.Yanxi, Z.He, X.Zhang, and S.Liu, "Transmit and receive sensors joint selection for MIMO radar tracking based on PCRLB", In 2016 IEEE 13th International Conference on Signal Processing (ICSP),2016, pp. 1551-1555.
[13] S.Xiyu, N.Zheng, and T.Bai,"Resource allocation schemes for multiple targets tracking in distributed MIMO radar systems",International Journal of Antennas and Propagation, Vol.2017 ,No.1,2017, pp.1-12.
[14] Y.Wei, Y.Yuan, R.Hoseinnezhad, and L.Kong,"Resource scheduling for distributed multi-target tracking in netted colocated MIMO radar systems", IEEE Transactions on Signal Processing, Vol.68, No.1, 2020, pp.1602-1617.
[15] Q. Cheng, J.Xie, and H.Zhang, "Joint Antenna Placement and Power Allocation for Target Detection in a Distributed MIMO Radar Network", Remote Sensing, Vol.14, No. 11,2022, pp.2650-2662.
[16] M.A. Darzikolaei, M.R. K.Mollaei, and M.Najimi,"An effective PSO-based power allocation for target tracking in MIMO radar with widely separated antennas", Physical Communication, Vol.51, No.1, 2022,pp.101544-101557.
[17] Y.Wei, Y.Yuan, R.Hoseinnezhad, and L.Kong. "Resource scheduling for distributed multi-target tracking in netted colocated MIMO radar systems",IEEE Transactions on Signal Processing, Vol.68, No.1, 2020,pp.1602-1617.
[18] L.Zhengjie, J.Xie, H.Zhang, H.Xiang, and Z.Zhang, "Adaptive sensor scheduling and resource allocation in netted collocated MIMO radar system for multi-target tracking", IEEE Access, Vol.8 ,No.1, 2020, pp.109976-109988.
[19] H. Qian, R.S. Blum, and A.M. Haimovich,"Noncoherent MIMO radar for location and velocity estimation: More antennas means better performance",IEEE Transactions on Signal Processing, Vol.58, No. 7, 2010,pp.3661-3680.
[20] E. Fishler, A.Haimovich, R.Blum, D.Chizhik, L.Cimini, and R.Valenzuela,"MIMO radar: An idea whose time has come", In Proceedings of the 2004 IEEE Radar Conference (IEEE Cat. No. 04CH37509), , 2004, pp. 71-78.
[21] C., and A.Nehorai,"Scheduling and power allocation in a cognitive radar network for multiple-target tracking",IEEE Transactions on Signal Processing,Vol.60, No. 2, 2012, pp.715-729.
[22] H.Godrich, A.M. Haimovich, and R.S. Blum,"Target localization accuracy gain in MIMO radar-based systems",IEEE Transactions on Information Theory, Vol.56, No. 6,2010, pp. 2783-2803.
[23] V.Trees, and L.Harry, Detection, estimation, and modulation theory, part I: detection, estimation, and linear modulation theory. John Wiley & Sons, 2004.
[24] S.Palin, J.Tang, Q.He, B.Tang, and X.Tang,"Cramer–Rao bound of parameters estimation and coherence performance for next generation radar",IET Radar, Sonar & Navigation, Vol.7, No. 5, 2013, pp.553-567.
[25] E.Russell, and J.Kennedy, "A new optimizer using particle swarm theory",In MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 1995, pp. 39-43.
[26] S.Chenguang, Y.Wang, F.Wang, S.Salous, and J.Zhou,"Joint optimization scheme for subcarrier selection and power allocation in multicarrier dual-function radar-communication system", IEEE Systems Journal, Vol.15, No. 1 ,2020, pp. 947-958.
[27] S.Boyd, and L.Vandenberghe, Convex optimization, Cambridge university press, 2004.
[28] J.Yan, H.Liu, B.Jiu, and Z.Bao,"Power allocation algorithm for target tracking in unmodulated continuous wave radar network", IEEE sensors journal, Vol.15, No. 2 , 2014,pp.1098-1108.

Appendix I.

If we do not consider noise, we will have $r_{mn,k}(t) = \sqrt{P_m}\ \xi_{mn,k}\ s_m\big(t-\tau_{mn,k}\big)\ e^{j2\pi f_{mn,k}t}$, therefore for two independent paths:

$$\mathrm{L}_J\big(\boldsymbol{\theta}_k;\boldsymbol{r}(t)\big) = \sum_{m=1}^{M}\sum_{n=1}^{N}\Big(-\ln\big(\sigma_{mn,k}^2P_m+1\big) + \frac{\sigma_{mn,k}^4P_m^2}{\sigma_{mn,k}^2P_m+1}\Big|\int_{-\infty}^{+\infty} s_m\big(t-\tau_{m'n',k}\big)\ s_m^*\big(t-\tau_{mn,k}\big)e^{+j2\pi(f_{m'n',k}-f_{mn,k})t}\,dt\Big|^2\Big) \tag{A.1}$$

And

$$\mathrm{G}_{13=} -\frac{\partial}{\partial\sigma_{mn,k}^2}\cdot\frac{\partial}{\partial\tau_{mn,k}}\mathrm{L}_J\big(\boldsymbol{\theta}_k;\boldsymbol{r}(t)\big) = -\frac{\partial}{\partial\sigma_{mn,k}^2}\frac{\sigma_{mn,k}^4P_m^2}{\sigma_{mn,k}^2P_m+1}\cdot\frac{\partial}{\partial\tau_{mn,k}}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big|\int_{-\infty}^{+\infty} s_m\big(t-\tau_{m'n',k}\big)\ s_m^*\big(t-\tau_{mn,k}\big)e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\,dt\Big|^2\Big) \tag{A.2}$$

And also by considering

$$\frac{\partial}{\partial\sigma_{mn,k}^2}\frac{\sigma_{mn,k}^4P_m^2}{\sigma_{mn,k}^2P_m+1} \approx \frac{\sigma_{mn,k}^2P_m^2}{\sigma_{mn,k}^2P_m+1} \tag{A.3}$$

And

$$\frac{\partial}{\partial\tau_{mn,k}}\sum_{m=1}^{M}\sum_{n=1}^{N}\Big|\int_{-\infty}^{+\infty} s_m\big(t-\tau_{m'n',k}\big)\ s_m^*\big(t-\tau_{mn,k}\big)e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\,dt\Big|^2\Big) =$$

$$-2\Big(\int_{-\infty}^{+\infty} s_m\big(t-\tau_{m'n',k}\big)\frac{\partial s_m^*\big(t-\tau_{mn,k}\big)}{\partial\tau_{mn,k}}\ e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\,dt\Big)\cdot\Big(\int_{-\infty}^{+\infty} s_m\big(t-\tau_{m'n',k}\big)\ s_m^*\big(t-\tau_{mn,k}\big)e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\,dt \tag{A.4}$$

Based on the paper assumptions and **Appendix II** and by considering $(m=m', n=n')$:

$$\int_{-\infty}^{+\infty} s_m(t - \tau_{m'n',k}) \frac{\partial s_m^*(t-\tau_{mn,k})}{\partial \tau_{mn,k}} e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\, dt = \int_{-\infty}^{+\infty} s_m(t-\tau_{mn,k}) \frac{\partial s_m^*(t-\tau_{mn,k})}{\partial \tau_{mn,k}}\, dt \xrightarrow{D.2} -j2\pi \int f|S_m(f)|^2 df \quad (A.5)$$

And also,

$$\int_{-\infty}^{+\infty} s_m\left(t-\tau_{m'n',k}\right)\ s_m^*\left(t - \tau_{mn,k}\right)e^{+j2\pi\left(f_{m'n',k}-f_{mn,k}\right)t}\, dt = \int_{-\infty}^{+\infty} s_m\left(t-\tau_{mn,k}\right)\ s_m^*\left(t-\tau_{mn,k}\right) dt \xrightarrow{D.1} 1 \quad (A.6)$$

Therefore:

$$G_{13} = G_{31} = -\frac{\sigma_{mn,k}^2 P_m^2}{\sigma_{mn,k}^2 P_m + 1}.(-2).-j2\pi \int f|S_m(f)|^2 df = -j4\pi \frac{\sigma_{mn,k}^2 P_m^2}{\sigma_{mn,k}^2 P_m + 1}\left(\int f|S_m(f)|^2 df\right) \quad (A.7)$$

And also, G_{23}, G_{32} ,and G_{33} are obtained in the same method as:

$$G_{23} = G_{32} = j4\pi\left(\frac{\sigma_{mn,k}^2 P_m^2}{\sigma_{mn,k}^2 P_m + 1}\right).\left(\int_{-\infty}^{+\infty} t \mid s_m^*(t - \tau_{mn,k})\right|^2 dt\Bigg) \quad (A.8)$$

$$\mathrm{G}_{33} == \frac{2P_m^2}{\left(\sigma_{mn,k}^2 P_m + 1\right)^2} \quad (A.9)$$

Since we want to compute FIM and we can choose greater value, and for simplicity we choose $\frac{2P_m^2}{\sigma_{mn,k}^2 P_m+1}$ for G_{33}.

Appendix II.

If the signals are orthogonal [24], they may have the below conditions:

$$\int_T s_k(t)\ s_m^*(t)\, dt \approx \begin{cases} 1 & k = m \\ 0 & k \neq m \end{cases} \quad (B.1)$$

$$\int_T s_k(t-\tau_{lk})\ \dot{s}_k^*(t-\tau_{lk})\, dt = -j2\pi \int f|S_k(f)|^2 df \quad (B.2)$$

$$\int_T |\dot{s_k}(t)|^2\, dt = 4\pi^2 \int f^2|S_k(f)|^2 df \quad (B.3)$$

Appendix III.

we should check the convexity of (46). First, we simplify the problem:

$$\mathbf{J}_{\mathrm{D}}(\boldsymbol{\theta}_{\mathrm{k}}) = \sum_{m=1}^{M-}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2 P_m - 1\right).(\mathrm{G_{mn}})_{5\times5} + \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\ \frac{+1}{\sigma_{mn}^2 P_m + 1}.(\mathrm{G_{mn}})_{5\times5} \quad (C.1)$$

The objective function is as $trace(\mathbf{J}_{\mathrm{D}}^{-1}(\boldsymbol{\theta}_{\mathrm{k}}))$. For proof the convexity, if we suppose objective function as$G(P_m)$, by choosing two value P_m and $P_{m'}$ then we should be have $G(\alpha P_m + (1-\alpha)P_{m'}) \leq \alpha G(P_m) + (1-\alpha)G(P_{m'})$[25].

$$G(P_m) \approx trace\left(\left\{\sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2 P_m - 1\right).(\mathrm{G_{mn}})_{5\times5} + \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\ \frac{+1}{\sigma_{mn}^2 P_m + 1}.(\mathrm{G_{mn}})_{5\times5}\right\}^{-1}\right) \quad (C.2)$$

Since the objective function is as $trace(\boldsymbol{X}^{-1})$, for proof convexity, the $\boldsymbol{X}$ should be affine [28]. First, we investigate the convexity of $\sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2 P_m - 1\right).(\mathrm{G_{mn}})_{5\times5}$:

$$G_1'(P_m) = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2 P_m - 1\right).(\mathrm{G_{mn}})_{5\times5}$$

$$\rightarrow G_1'(\alpha P_m + (1-\alpha)P_{m'}) = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2(\alpha P_m + (1-\alpha)P_{m'}) - 1\right).(\mathrm{G_{mn}})_{5\times5}$$

$$= \alpha \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\left(\sigma_{mn}^2(P_m) - 1\right)\right).(\mathrm{G_{mn}})_{5\times5} + (1-\alpha)\sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\left(\sigma_{mn}^2(P_{m'}) - 1\right)\right).(\mathrm{G_{mn}})_{5\times5}$$

$$= \alpha G_1'(\alpha P_m) + (1-\alpha)G_1'(P_{m'})) \quad (C.3)$$

Therefore, $G_1'(P_m)$ is convex.For the next part:

$$G_2(P_m) = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\ \frac{+1}{\sigma_{mn}^2 P_m + 1}.(\mathrm{G_{mn}})_{5\times5} \quad (C.4)$$

Therefore,

$$G_2(\alpha P_m + (1-\alpha)P_{m'}) = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2(\alpha P_m + (1-\alpha)P_{m'}) + 1\right)^{-1}.(\mathrm{G_{mn}})_{5\times5} = \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\left(\sigma_{mn}^2\alpha(P_m + 1) + \sigma_{mn}^2(1-\alpha)(P_{m'} + 1)\right)^{-1}.(\mathrm{G_{mn}})_{5\times5} \quad (C.5)$$

Therefore, with respect to the value of P_m and parameters of our problem, the condition $(A+B)^{-1} \leq A^{-1} + B^{-1}$ is satisfied for (C.5) and we will have:

$$\sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\,(\sigma_{mn}^2\alpha(P_m+1)+\sigma_{mn}^2(1-\alpha)(P_{m'}+1))^{-1}.(\mathrm{G_{mn}})_{5\times5} \le \alpha * \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\,(\sigma_{mn}^2(P_m+1)^{-1}.(\mathrm{G_{mn}})_{5\times5} + (1-\alpha) * \sum_{m=1}^{M}\sum_{n=1}^{N} 8\pi^2\,(\sigma_{mn}^2(P_{m'}+1)^{-1}.(\mathrm{G_{mn}})_{5\times5} = \alpha * G_2(P_m) + (1-\alpha) * G_2(P_{m'}) \quad \text{(C.6)}$$

Therefore $G_2(P_m)$ is also convex. Then we can conclude that $G(P_m)$ is convex and in result, our problem is convex.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 307-319
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Inferring Diffusion Network from Information Cascades using Transitive Influence

Mehdi Emadi[1*], Maseud Rahgozar[2*], Farhad Oroumchian[3]

[1].Faculty of Electrical & Computer Engineering, Babol Noshirvani University of Technology, Babol, Mazandaran, Iran.
[2].School of Electrical & Computer Engineering, University College of Engineering, University of Tehran, Tehran, Iran.
[3].Faculty of Engineering and Information Sciences, University of Wollongong in Dubai, Dubai, UAE.

Received: 05 Feb 2022/ Revised: 19 Nov 2022/ Accepted: 05 Feb 2023

Abstract

Nowadays, online social networks have a great impact on people's life and how they interact. News, sentiment, rumors, and fashion, like contagious diseases, are propagated through online social networks. When information is transmitted from one person to another in a social network, a diffusion process occurs. Each node of a network that participates in the diffusion process leaves some effects on this process, such as its transmission time. In most cases, despite the visibility of such effects of diffusion process, the structure of the network is unknown. Knowing the structure of a social network is essential for many research studies such as: such as community detection, expert finding, influence maximization, information diffusion, sentiment propagation, immunization against rumors, etc. Hence, inferring diffusion network and studying the behavior of the inferred network are considered to be important issues in social network researches. In recent years, various methods have been proposed for inferring a diffusion network. A wide range of proposed models, named parametric models, assume that the pattern of the propagation process follows a particular distribution. What's happening in the real world is very complicated and cannot easily be modeled with parametric models. Also, the models provided for large volumes of data do not have the required performance due to their high execution time. However, in this article, a nonparametric model is proposed that infers the underlying diffusion network. In the proposed model, all potential edges between the network nodes are identified using a similarity-based link prediction method. Then, a fast algorithm for graph pruning is used to reduce the number of edges. The proposed algorithm uses the transitive influence principle in social networks. The time complexity order of the proposed method is $O(n^3)$. This method was evaluated for both synthesized and real datasets. Comparison of the proposed method with state-of-the-art on different network types and various models of information cascades show that the model performs better precision and decreases the execution time too.

Keywords: Transitive Influence; Network Inferring; Diffusion Network; link Prediction, Random Network.

1- Introduction

Nowadays, online social networks play an undeniable role in propagating information. People are capable of creating contents on a medium to influence other people's opinions. With the increasing importance of online social networks, researchers have been interested in social network analysis. Several methods have been introduced for studying social network behavior on different topics such as information diffusion [1], community detection [2] - [4] link prediction [5] - [8] influence maximization [9], sentiment analysis [10], and expert finding [11]. News, sentiment, rumors, ideas, innovations, and knowledge diffuse over social networks as different types of information. Hence, modeling information diffusion network for any type of information lets researchers apply various social network analysis methods to understand the behavior of people for further social studies. Knowing that information spreads over an underlying network, information diffusion is modeled as a graph in which people are the nodes and the relations between them are the edges.

Like contagious diseases, a diffusion, also called a contagion, occurs when a piece of information is transmitted from one node to another through the edges between them over the underlying network [12]. In this field, any epidemic event disseminated over a social network can be considered as a piece of information. When a member (node) mentions or copies any piece of information from another member (its neighbors), then, it is called to be infected by a contagion. During the process of information diffusion, nodes get infected by a contagion, and an observable footprint is the time of infection. Like

✉ **Maseud Rahgozar**
m.emadi@nit.ac.ir, rahgozar@ut.ac.ir

epidemic diseases, in the outbreak of a disease in a society, the viruses spread from one person to another, while it is unclear by whom each person is really infected. However, the infection time for each person is observable. Similarly, in viral marketing, no one knows who influenced a client, but we know when a client bought the new product.

The main challenge in the field of information diffusion analysis is the lack of knowledge on the structure of underlying network. To study the behavior of people in a social network, the initial requirement is to infer the network structure from the observed data. Inferring the network structure of neurons in neuroscience [13], sentiment in online social networks [14], [15], community detection [16], or the genes in biology [17],[18] are similar points of interest in current researches. The aim of this article is investigating an epidemiology approach to infer the structure of an influence network from a set of information cascades, i.e., the time history of various events occurred in a network.

In recent years some models have been proposed to infer a network from the observed information cascades. In most cases these models try to solve an optimization problem [17]–[20]. This causes a long runtime which is not applicable for real-world networks with a large size. In recent years, with the development of content along with the graph structure, works have placed more emphasis on the use of content. For this reason, less effort has been made to extend pure structure-based algorithms. For example, in the article [21] work is done on the three features: "information, user decision, and social vectors". In the [22] work is done on the 5 different information source and data mining technique to find hidden influence. In this study[23], yang and friends used the community structure in addition to the information cascade. But in this research, we have worked on pure structure and tried to provide an algorithm for this purpose. For this reason, in order to make a fair comparison, we have compared our work with solutions based on pure structure. Of course, this method can be used to continue the work in any of the combined works of structure and content.

In this paper, we propose a method for modeling the diffusion which results in inferring the diffusion network. The approach of this method is algorithmic and non-optimization. With the help of link prediction concept and proposing an algorithm for pruning the transitive edges in a graph, a time-efficient method is proposed. First, we look for a formula for modeling the influence of a user on another user. Various formulas are presented based on social rules to find the appropriate one. Experimental results show that one of these formulas is more suitable for modeling. The selected model has a better result based on the f1 measure. These experiments are based on synthesized data. Second, with the use of the appropriate model, the propagation network will be inferred. Approximately, we have an influence rate between each of the two nodes, and a generated graph seems like a complete graph. In the real network, we have a direct edge among the smaller number of users. These additional edges are due to the indirect influence (transitive influence) [24], [25] of individuals on one another. We present a heuristic algorithm that identifies and eliminates indirect influence. This identification is based on a social rule called transitive influence. The time complexity of this algorithm is O(n3) which, in comparison with similar algorithms, has an efficient execution time. The experiments show that the proposed method outperforms several state-of-the-art models in both synthetic and real dataset.

The remainder of the article is organized as follows: In the second section, the problem of the inference of diffusion network is defined, and in the third section, the related and previous works have been reviewed. In the fourth part, the explanation of the proposed method is discussed. Section five shows the results of the experiments and evaluations of the proposed algorithm are investigated. For this reason, we use synthesized and real data sets. And in the last section, we conclude our work.

2- Problem Definition

For modeling a diffusion process, information cascades can be employed. Assume user A has communication with user B in a social network. If user A joins a social campaign, then the effect of this event on user B is joining the campaign as a similar action. The process when a piece of information or an action spread from one node to another over a network generates information cascade. A cascade can be specified as two vectors of T and Q. The vector T = [t1,…,tn] represent a time series of infection times of nodes and the features of the contagion (such as user identification) represented in the vector Q = [q1,…,qn]. For the cascade C(T,Q), there are two assumptions [26]: it's not obvious which of the nodes is affecting each node, and each node can be affected by many nodes.

Consider a hidden network with graph G' wherein multiple cascades have been spread over that. The main effort is finding graph G which is an estimation of G', from the observed cascades. Assume that we have a set of cascades ({C1(T1,Q1), … CN(TN,QN)}); the main problem in "inferring diffusion network" is finding the underlying network which caused these cascades.

3- Related Works

The problem of inferring influence network or modeling information diffusion network may be divided into multiple sub-groups considering several aspects of this problem. For example, in the assumptions of one model, the set of information cascades are fully observed [26], [27], but in some, there are missing data in the cascades [28], or the

dynamics of network may change over time in some models [28], [29] whereas the other models assume a time-invariant network [26]. Considering a set of information cascades which are infection time series of a network's nodes, some models proposed to infer the underlying network over which the information diffuses. As far as finding the best possible graph is NP-Hard, in most cases, these models apply methods like Maximum Likelihood Estimation (MLE) to solve the optimization problem which causes a long runtime. As the output of these models, two main aspects of diffusion network would be characterized: structure of the network and its temporal dynamics [27].

The CONNIE [30] method uses convex programming to learn a network under a randomly uniform distribution of transmission time and recovery time. NETINF [26] considers a static network, and the proposed model uses a tree-shaped graph to infer the relationships between nodes from information cascades. In contrast to NETINF, new models have been proposed which assume the network is not static, and the pathways would change over time, and they are dynamic. The NETRATE [27] method, having a set of cascades, maps the parameters of the transmission rate models for each edge. Based on NETRATE, a new method called INFOPATH [29] was developed. The INFOPATH method calculates a pairwise transmission probability for edges between nodes based on information cascades. Then an optimization problem is formed to select the best edges. The best edges that model information cascades with the least error. With the use of stochastic convex optimization, INFOPATH solved the problem of inferential network inference in less time. In previous methods, the assumption of the homogeneity of the relationship between people in an area was significant. But, the hypothesis of the researchers in MMRATE [31] is that of individuals who are different in different topics. This approach focuses on the multifaceted relationship between the network members. The main focus of TOPIC CASCADE [32] is the prediction of the transmission time of a publication on the network. This method solves, as in previous methods, an optimization algorithm for estimating the parameters of the transmission model. TOPIC CASCADE uses an efficient proximal gradient algorithm based on a block coordinate descent for estimation. Other methods also take into account information from contexts such as text content and individuals. In NIMFC [33], different dimensions of information cascades, including: "time, and topographic characteristics of cascades," "user attributes," and "information content" are used to infer diffusion network.

4- The Proposed Method

In the proposed method, the goal was to find the influence network of the nodes in the input data with some information cascades as input data. Information cascades have the time for the activation of each person.

Accordingly, in this method, formulas for "modeling the impact of individuals on each other" have been presented. Various parameters extracted from the information cascades have been used to express formulas. The parameters that have been extracted from the cascades are the time interval of activation of two people in a cascade ($\Delta t_{ab,c}$), the number of cascades where the person is activated (c_a), and the number of times a person ***"b"*** has been activated after the person ***"a"*** (h_{ab}). In this research, we have tried to provide a model that is general and capable of responding to different types of networks. For this purpose, various models have been presented with different combinations of extracted parameters. Different models were tested in a variety of ways to achieve an acceptable general model. Of course, the models presented are based on the rules governing human relationships in social networks. Information cascades display the time of activation or the participation of a node in a particular publication in the order of the event time.

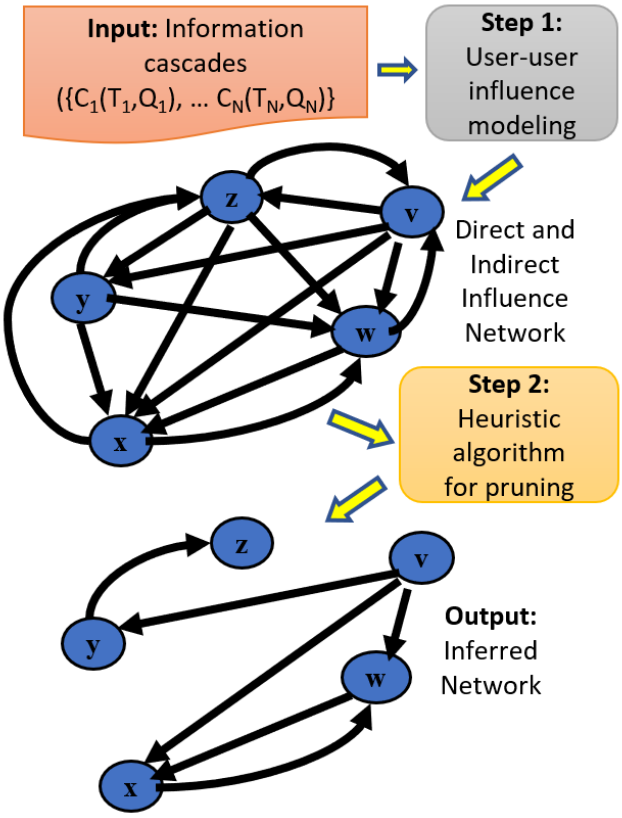


Fig. 1. The schematic image of the suggested method. With regard to the collection of information cascades, at step 1, the impact of individuals on each other is firstly modeled. This action causes all potential edges between the network nodes to be identified and aggregated in a graph. Then, in step 2, by applying the pruning algorithm on the potential graph, the target graph is deduced.

In this research, by evaluating each cascade, the rate of the influence of the nodes on each other, was calculated. We defined *w*(a,b) as the influence rate of *a* on *b*. In each cascade, by observing the activation of b after a, the amount of *w*(a,b) increases. We needed the formula to express how each observation affects the *w*(a,b) calculation. In each information cascade, only the activation time of the node is visible, so the only useful parameter is the activation time. Other parameters can also be used to calculate *w*(a,b), including "the number of caches in which a comes after b," and "the number of cascades where a or b exists."

In Table 1, the parameters extracted from the information cascades have been introduced. "The activation times of a node after another node is activated," "the time interval between the activation of a node with the activation of another node," or "the frequency of activating a node individually" are some of the extracted parameters. In this research, using the same parameters, various models have been provided for calculating *w*(a,b); their list is given in Table 1.

In an output graph between two vertices a and b, where w(a,b) is not zero, we considered an edge (step 1 of Fig. 1). In this way, the output graph had many edges. Most of these edges were derived from our formula of computing the rate of the influence of the nodes on each other(*w*(a,b)), and in fact, we do not have such a direct relationship between the two users.

Table 1. Different functions for different Models of scoring the impact of nodes on each other.

Model	Formula
Model 1 (F. 1) [26]	$\sum \frac{1}{\Delta t}$
Model 2 (F. 2)	$\frac{h_{ab}}{C_a - h_{ba}}$
Model 3 (F. 3)	$\frac{h_{ab}}{C_a - h_{ba}} * \frac{h_{ab}}{C_b}$
Model 4 (F. 4)	$\sum \frac{1}{\Delta t} * \frac{h_{ab}}{C_a - h_{ba}}$
Model 5 (F. 5)	$\frac{h_{ab}}{C_a - h_{ba}} * \frac{h_{ab}}{C_b} * \sum \frac{1}{\Delta t}$
Model 6 (F. 6)	$\sum e^{-\Delta t}$
Model 7 (F. 7)	$\frac{h_{ab}}{C_a - h_{ba}} * \sum e^{-\Delta t}$
Model 8 (F. 8)	$\frac{h_{ab}}{C_a - h_{ba}} * \frac{h_{ab}}{C_b} * \sum e^{-\Delta t}$

With a technique, we must recognize the “real edges” of the “non-real edges”. With the help of the above functions, the influence rate between two nodes is obtained. For some nodes, this number represents the direct effect of these two on each other, which we call “real edge”. But for some nodes, this effect, which has been seen many times in cascades, is due to the indirect effect of two nodes. These edges are called “non-real edge”, which is the result of “Transitive Influence” (Step 2 of **Fig. 1**). For this purpose, a heuristic derived from the social networking space was used. To this end, we tried to find the edges of the transitive influence. The algorithm presented on this heuristic is explained in the following section.

4-1- Proposed user-user Influence Models

Different Models have been presented to calculate the influence rate of one person on another. In all the previous studies, this rate was calculated during an optimization process. However, in this method, the rate has been calculated by reading the cascade information once. In some ways, the method was taken from an optimization problem toward a simple modeling problem and solves the problem in that space.

In most of the conducted researches, they consider Model 1 [26], which is a simple time-based model. And because they raise an optimization problem based on this, they don't need another model. But in this research, we want to determine transitive influence. For this issue, it was necessary to develop and examine different models.

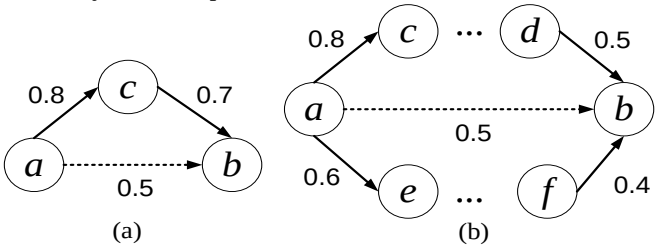


Fig. 2. Transitive influence: (a) node c is affected by node a by 0.8, and node b from node c is affected by 0.7. influence of node c on node b and node a on c causes node a to have an indirect influence on node b of 0.5. (b) As in part a, node b is indirectly affected by node a.

In Table 1., Model 1 and 6 consider the effect of the parameter of the time interval of activation of node b after node a in cascade C in two ways. This formula has been defined exponentially in model 6 for exponential waiting time models and power-law in function 1 for power-law waiting time models. The function 2 defined as division of h_{ab} by $(C_a - h_{ba})$; in other words, "the number of times b is after a" divided by "the number of times b could have come after a." Functions 4 and 7 have been constructed from the combination of functions 1 and 6 with function 2. The simultaneous effect of these two types of functions has been considered in functions 4 and 7. The function 3 multiplies the net effect of b on a in the overall coefficient of influence b to calculate a more normal value than that of function 2. Functions 5 and 8 have been constructed from the combination of functions 1 and 6 with function 3.

4-2- Algorithm for Network Inference

With the aid of the user-user influence model, for all possible edges, the edge weight was calculated. In fact, we obtained a weighted graph close to the complete graph. The weighted graph obtained by the scoring function could not be considered as the actual graph of diffusion network because many of the edges of this graph were derived from indirect influence. But, our goal was to find the direct impact of nodes from each other.

For this purpose, an algorithm was proposed for pruning the indirect edges of the graph and reaching the direct influence graph. For example, in Fig. 2(a), the weight of the edges between the three nodes a, b, and c is calculated using the scoring function. The weight of the edge (a, c) is 0.8, that of the edge (c, b) is 0.7, and that of the edge (a, b) is 0.5. Node b is influenced more by node c and less influenced by node a. Also, node c itself is influenced by node a. The weight of the edge (a,b) is less than these two other edges. According to the indirect influence principle, this edge is due to the indirect influence node a on node b, and it is likely to be said that there is no direct influence between node a and node b. In fact, b through c is influenced by a. So, we can remove this edge.

The indirect influence does not always occur at a distance as large as a node. The distance between two individuals who accept the indirect influence can be more than one node (Fig. **2**(b)). In this case, we define a f(a,b) parameter to calculate the rate of indirect influence. In line 7 of algorithm 1 (Fig. **3.**), parameter f(a,b) is the maximum flow between the edge a and edge b in the graph G(V, E-(a,b)). If f(a,b) is larger than w(a,b), straight edge (a,b) has been achieved on the basis of indirect influence and should be eliminated (line 8-10 of algorithm 1).

Table 2. Table of Notations for Proposed Algorithm.

G	Graph of user of social network and their possible interaction in all information casscades
E	List of possible interaction between users of Social Network (Edge List)
V	List of users of Social Network (Vertex List)
W	Weight list of extracteg graph from Step 1
w(a,b)	Capacity of edge (a,b) derived from Step 1 of proposed method based on formulas of Table 1.
E’	List of interaction between users of Social Network after graph prunning
W’	Weight list of extracteg graph after graph prunning

According to the selected function, indirect influence is smaller than direct influence. For pruning the edges, we start from the edges with high-weight. If we start with light-weight edges, the algorithm does not work properly.

1 **input:** G(V,E,W)
2 **output:** G`(V,E`,W`)
3 **for all** (a,b) ∈ E *w’(a,b)* ←0 //use w’ as capacity of edges
4 E`←{}
5 **sort** the edges of E into decreasing order by weight w
6 **for each** (a,b) ∈ E, taken in decreasing order by weight
7 *f(a,b)* = FindMaxFlow(a,b, G`(V`,E`,W`))
8 **if** *w(a,b)* > *f(a,b)* **then:**
9 E` ← E` ∪ (a,b)
10 *w’(a,b) = w(a,b)*
11 **return** G`(V`,E`,W`)

Fig. 3. Algorithm 1: Pruning the Indirect Edges

4-3- Efficient Algorithm for Pruning Phase

Runtime of the algorithm 1 was not good. We needed a faster algorithm (algorithm 2 in Fig. **4.** In this algorithm, the edges of E were sorted in graph G into decreasing order by weight w (Line 5). Then, in line 6, we started with the maximum weight edge. If a path from node a to node b was not available in the graph G', we added the edge (a,b) to the graph G' (Lines 7-8). The findPath(G,a,b) function in this algorithm is a Boolean function that returns true if there is a path from node a to node b in graph G. The algorithm 2 had a better order in terms of time complexity than algorithm 1. Of course, with the help of various experiments, it was shown that they return the same results.

5- Analysis of Algorithms

In algorithm 1, the order of execution for the sorting (line 5 of algorithm 1 in Fig. **3**) is equal to O(|E| log |E|). For the second part (lines 6-10 of Fig. **3.**) of this algorithm, we can use the Ford–Fulkerson algorithm [34] to calculate the maximum flow. The running time of the Ford–Fulkerson algorithm is equal to O (|E’| max |f|). As a result, the total running time of the second part is equal to O(|E| (|E’| max |f|)).

If we use the Edmonds–Karp algorithm [34] to calculate the maximum flow, the total running time of the second part is equal to $O(|E|(|V'|+|E'|^2))$. Because the running time of the Edmonds–Karp algorithm is equal to $O(|V|+|E|^2)$, the total running time of proposed algorithm is $O((|E| \log|E|+|E|(|V'|+|E'|^2))$. |V|=|V’|. The graph G' is very close to the tree, and consequently, the size of |E’| is equal to c|V|. The size of |E| is equal to $|V|^2$ because G is very close to the full graph. After replacing the new value in the formula, we have $O(|V|^2\log|V|+|V|^4)$. Finally, we have a total time complexity of $O(n^4)$ for algorithm 1. For algorithm 2, we have sorting section (line 5 of algorithm 2

in Fig. **4**) too. For the second section of algorithm 2, we have O(|V'|+|E'|) for finding the path method, and the running time of the second part is O(|E|(|V'|+|E'|)). The total running time of the algorithm 2 is O((|E| log|E|+|E|(|V'|+|E'|)). After replacing a new value to the formula, we have $O(|V|^2 \log|V|+|V|^3)$. Finally, we have the total time complexity of $O(n^3)$ for algorithm 2. In the article on INFOPATH [29], there are no references to running time of its algorithm. But our experiments show that the running time of INFOPATH was longer than that of the proposed algorithm in this research.

```
1  input: G(V,E,W)
2  output: G'(V,E',W')
3  W'←W
4  E'←{}
5  sort the edges of E into decreasing order by weight w
6  for all (a,b) ∈ E do:
7        if NOT findPath(G,a,b) Then:
8                E' ← E' ∪ {a,b}
9  return G'(V,E',W');
```

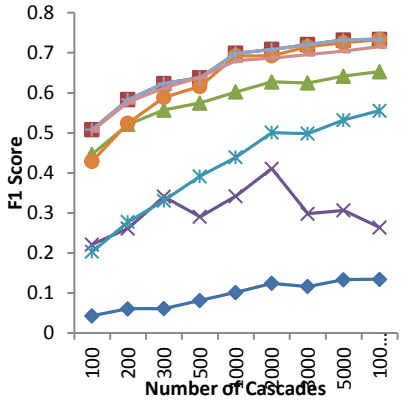
Fig. 4. Algorithm 2: Optimization of the Execution of the Algorithm 1

6- Results and Experiments

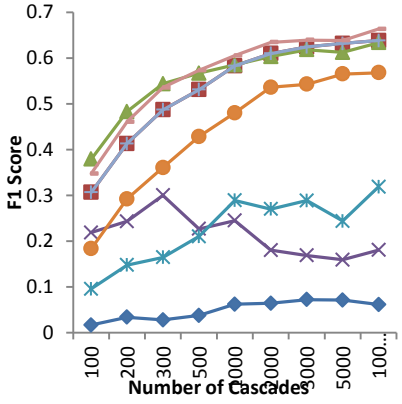
For the evaluation of our work, we needed data sets to evaluate the proposed method. Due to the inaccessibility of the main graph in this type of problem, synthesized data and real dataset were used to evaluate the proposed methods. Three types of synthesized networks were generated using Kronecker [35] graph models: hierarchical, random, and core-periphery. Also, were generated the information cascades with two types of cascade models: Rayleigh and exponential. For assessment of proposed method with real data, a real dataset from BrightKite social network [36] was used. In this section, we retrieve the graph or network structure by examining the information cascades. In the following, we evaluate the correctness of the algorithm by comparing the resulting graph with the ground truth graph. We used three measures, precision, recall, and f1-score, to evaluate the matching of the network with the main network and to evaluate and compare the methods. The precision is the fraction of inferred edges that are inferred correctly. The recall is the fraction of edges in the ground through network that is inferred correctly. F1-score is computed as the combination of precision and recall (Eq. (1)).

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (1)$$

In the remaining sections, we compare the effectiveness of different models in the first subsection, show the experimental results which compare the proposed method with the state-of-the-art method in the next subsection, and in the last subsection, present the result of the experiment on the runtime of the proposed method.



Random (Exponential)



Random (Rayleigh)

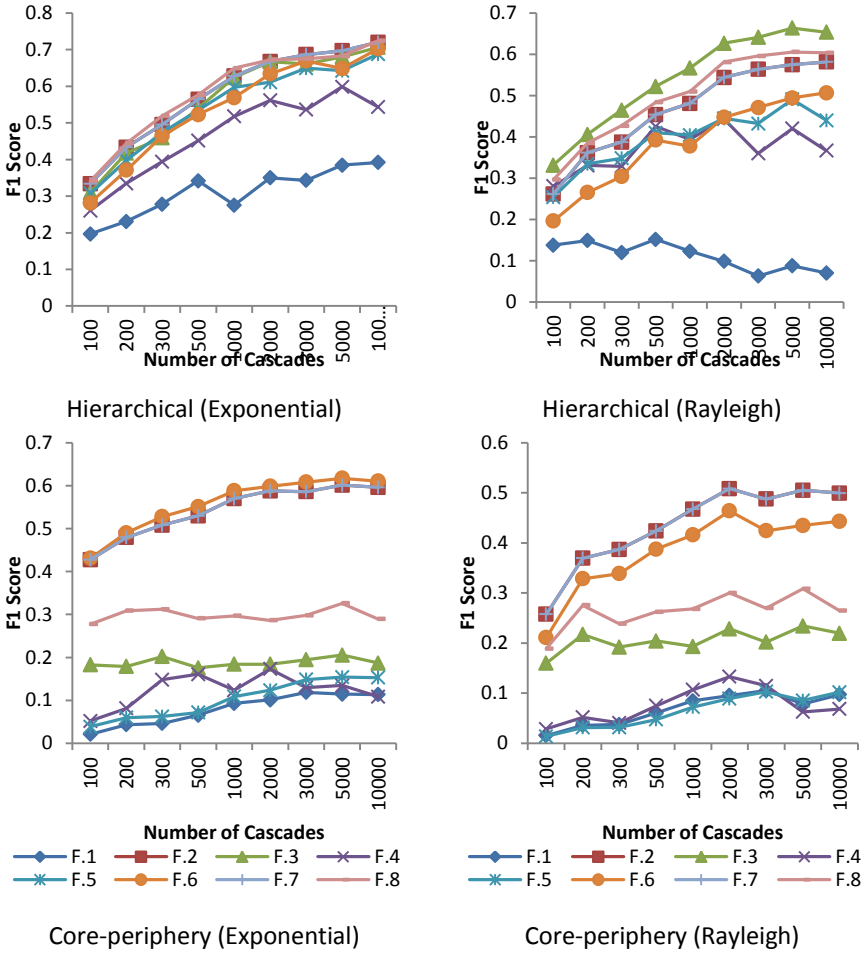
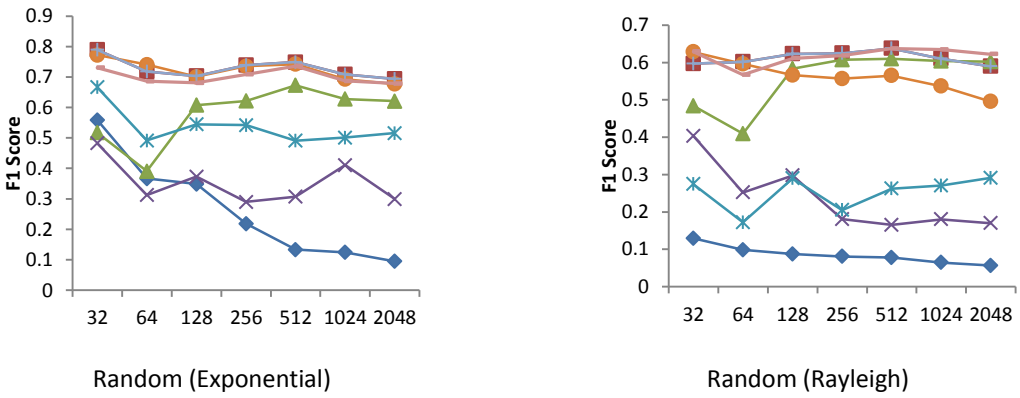


Fig. 5. Comparing the F1-score of the proposed algorithm with the various user-user influence models (influence rate functions) for the synthesized data with constant network user size (1024 user) and various cascade size from 100 to 10000 .

6-1- Effectiveness of Different Models

By determining the impact value of each node on another node, approximately, we will have the weights for all the edges. The proposed algorithm has been used to find the best influence rate model. Here, we compared all the scoring functions introduced in the previous section using the proposed method and algorithm. To this end, we evaluated various models for different modes of cascade size and network size. For comparison, synthetic data was used. The SNAP tool [37] has been used to produce different types of networks and cascades. Three types of networks (random, hierarchical, and core-Periphery) and two types of cascade models (exponential and Rayleigh) were used for this assessment. For the first mode, the network size was constant, and the number of cascades varied.



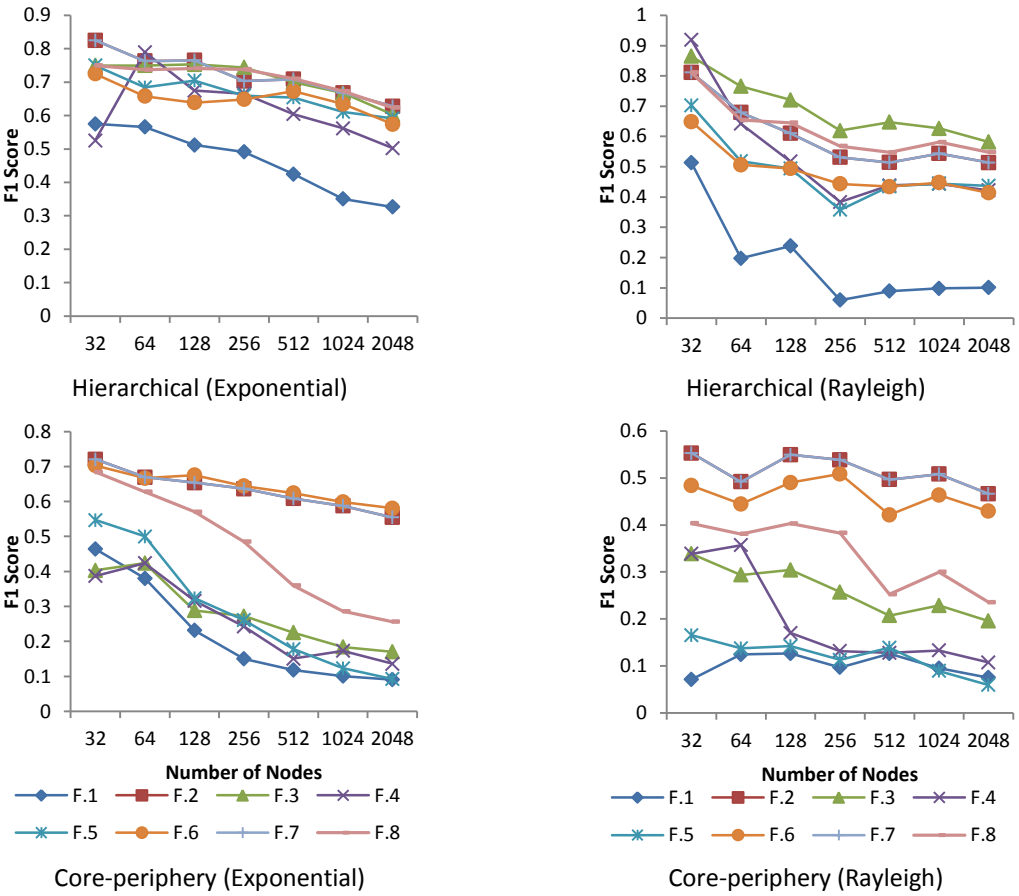
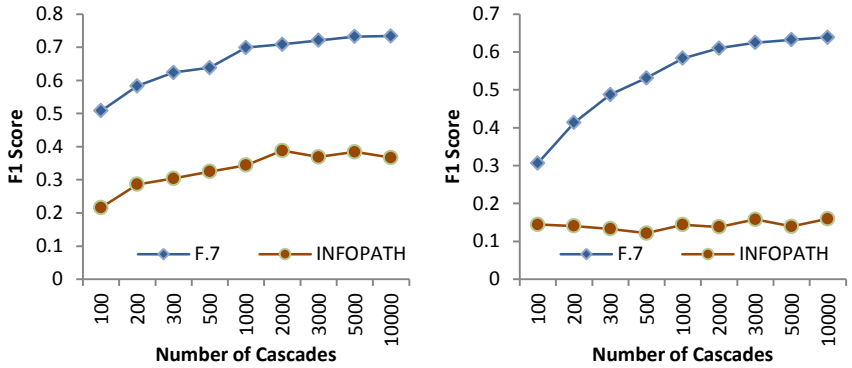


Fig. 6. Comparing f1-score of the proposed algorithm using various user-user influence model (influence rate functions) for synthesized data with constant cascade number (2000 cascades) and various network size from 32 to 2048.

The number of nodes in the network was 1024. The number of cascades varied from 100 to 10000 (100, 200, 300, 500, 100, 2000, 3000, 5000, and 10000). Fig. 5 shows that the increase in the number of information cascades from 100 to 10000 had influenced the performance of the proposed functions. This comparisons show that the proposed functions 3, 7, and 8 were more stable against variation of cascade size and show better performance. By increasing the number of cascades, which is the number of our observations, we will have a better f1measure value. That's why all the diagrams are incremental. Functions 7 and 2 behaved quite similar. This shows that the different parts of the two formulas ($\sum e^{-\Delta t}$) had no effect on performance improvement. Of course, this difference was equivalent to function 6.



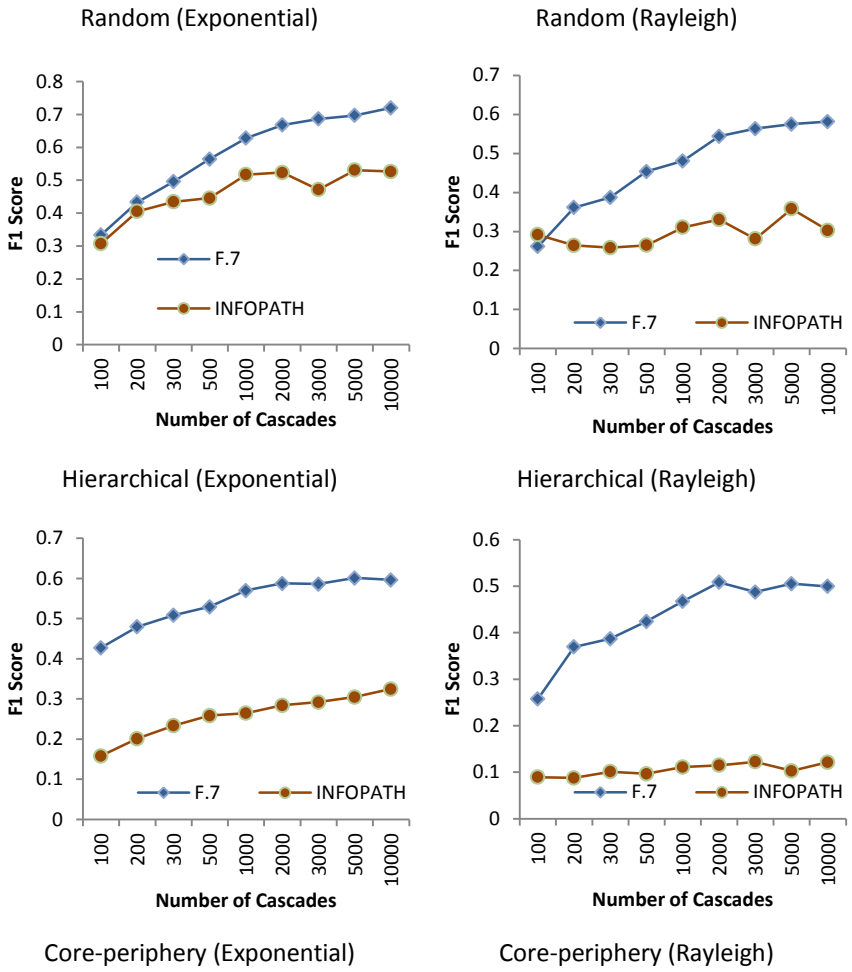


Fig. 7. Comparing f1-score of the proposed algorithm with INFOPATH for the synthesized data with constant network user size (1024 user) and various cascade size from 100 to 10000.

Table 3. Collection of Information Cascades Extracted from the Brightkite Social Network.

Cascade type	How to select users	Number of users	Number of edges	Number of extracted cascades
Type 1	Check in every day	5159	35280	53399
Type 2	Every two days, three times check in	3677	25563	42802
Type 3	Every day twice check in	2805	19871	35826
Type 4	Every two days, five times check in	2241	15690	30386

The value of f1measure for function 6 was also high, and this case shows that function 6 when combined with function 2 does not have much effect on efficiency.

Function 2 has enough information in itself, and adding formula of function 6 will not have much improvement. The function 6 in the eight modes of the nine possible modes of data generated had a lower result, but only in the case of core-periphery exponential, the function 6 had a better answer. Of course, in this case, the function 6 behaved similar to function 7. For other conditions like "core-periphery exponential," function 3 had a lower f1 value. The function 3 had many oscillations in different states of graph size and cascade numbers, and it was not a good option to choose as the selected model. If we want to choose a function that in most cases is close to the best, we can select the function 7 or 2.

For the second mode, the cascade size was constant, and the network size varied. The number of cascade for the network was 2000. The number of nodes varies from 32 to 2048 (32, 64, 128, 256, 512, 1024, and 2048). For this case, the cascade number was constant, and the number of nodes in the network was changed (Fig. 6); the diagrams have a decreasing behavior. The reason for the downside of the charts was that the number of cascades was constant, and the number of nodes increased. As a result, the ratio of the number of nodes to the number of cascades increased, and the accuracy of the detection of the main edges decreased.

As the charts demonstrated, the behavior of the functions in this mode (variable network size) was the same as in the previous state (variable cascade size). Our experiments show that function 7 gives a better result based on the f1-score measure.

6-2- Comparing the Proposed Methods

Different methods have been proposed to infer the network from information cascades. Each of the methods models a few specific models of networks. The goal of all these methods is to find the best network that models the cascades that have happened. For this reason, their optimization method is sometimes applicable for several specific models of the network or some specific models of cascade production, and they do not do well in the rest of the network. Some methods, such as the INFOPATH, have an acceptable behavior for many types of networks. For the same reason, we compared the proposed method with INFOPATH. To compare our method with the INFOPATH method, we used synthetic and real dataset.

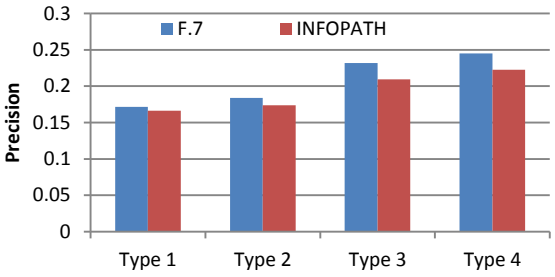


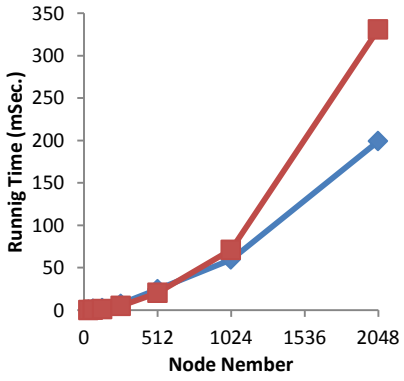
Fig. 8. Precision of proposed method and INFOPATH on BrightKite dataset results for five types of cascades.

6-2-1 Synthesized Data

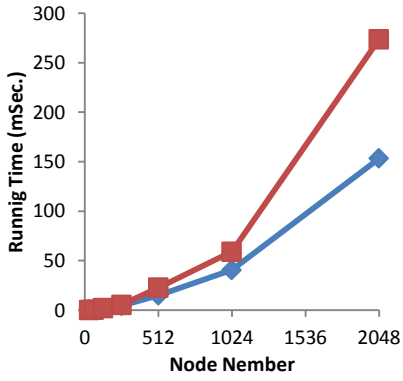
To evaluate the proposed method, we produced synthesized data for all the different modes of the network. The SNAP tool [37] has been used to produce different types of networks and cascades. Three types of network (random, hierarchical, and core-periphery) and two types of cascade models (exponential and Rayleigh) were used for this assessment. The number of nodes in the network was 1024. The number of cascades varied from 100 to 10000 (100, 200, 300, 500, 100, 2000, 3000, 5000, and 10000). Fig. 7 shows that the proposed method was better than INFOPATH in term of f1-score measure. The proposed method is highly accurate compared to the INFOPATH for low cascades count. By calculating the average of all execution modes (which are created from the combination of 6 different random network modes, different number of nodes and different cascade sizes), the presented method has improved by an average of nearly 5% based on f1-score measure.

6-1-1 Real Data

The Bright Kite social networking dataset [36] was used to assess the effectiveness of the proposed method on real social networks. In the dataset which is selected from this social network, there were more than 4 million check-ins from over 58000 users, whose relation network is known. In the data-collection, there were 58228 users. There were 214078 communication links between the users. The number of special places according to their unique geographical coordinates was 772966. This dataset was collected from April 2008 to October 2010. In this social network, each person declared his presence after entering a place. Over time, in the profile of each person, the list of places where he or she checked in would be visible.



Random (Exponential)



Random (Rayleigh)

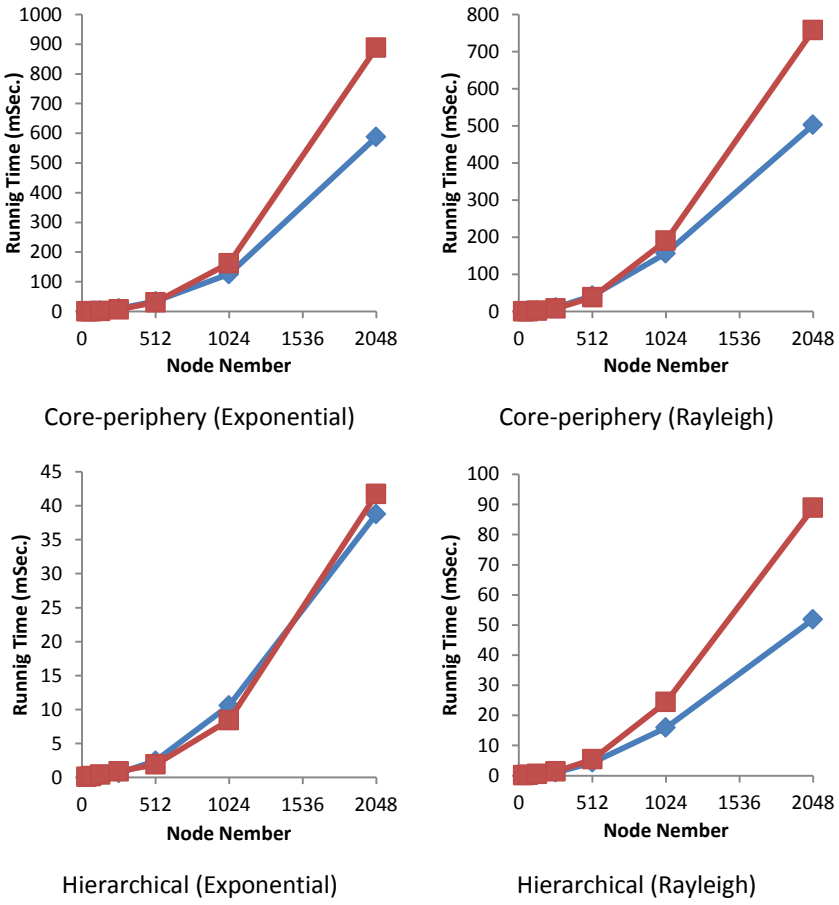


Fig. 9. Comparing running time of the proposed algorithm with INFOPATH for synthesized data with constant cascade number (1000 cascades) and various network size from 32 to 2048 (Redline: INFOPATH;Blueline:F.7).

Also, for each specific location (for example, a restaurant or a cafe), frequent check-ins were recorded from different users. It was necessary to respond to the two challenges, information cascade generation and reducing the size of the datasets, by preprocessing the information to use this dataset to evaluate the proposed method on real social networks.

Checking in the presence of different users for a location is considered as an information cascade. A user will check-in at a specific location (such as a cafe) for the first time, and this will be communicated to his friends on the social network. Then, the friends of that person check-in at that place. So, each information cascade in this dataset is a sequence of checking in for people in a specified location. There are over 700000 unique places in this dataset. But cascade is not made for all of them.

If more than two people are checked in at one place, then a cascade will be created for that place. Due to the large size of this social network, a number of more active users were selected, and their information cascades were extracted (Fig. 8). Four types of cascades were generated. Experiments show that the proposed method had a better precision for all types of the cascade definition (Fig. 8). Averaged over all types of use case, this method provides an improvement of about 2% based on f1-score measure.

6-3- Running Time Analysis

In the previous section, the time complexity analysis of the proposed algorithm was presented. We have shown that the runtime is $O(n^3)$. We have executed proposed algorithm and INFOPATH in the same conditions. The results are showing in Fig. 9. The runtime of INFOPATH was acceptable for smaller values. But by increasing the network volume, INFOPATH runtime increased at high rates. The proposed algorithm for bigger networks has lesser runtime than that for INFOPATH.

Due to the use of stochastic convex optimization to learn the parameters of the information cascade transmission model, the INFOPATH model is relatively faster than the other methods. The time complexity of INFOPATH algorithm is not explained in its article. The experiments to compare the runtime between the INFOPATH model and the proposed method were performed on a machine with 64GB RAM and four processor cores.

6-4- Experimental Environment

All programs are written in C++ language. SNAP library is used for INFOPATH algorithm. The SNAP library has also been used to generate dummy data. All the programs are run in the environment of Ubuntu operating system. The hardware used was a workstation with 32 processing cores and 64 GB of main memory.

7- Conclusion and Future Works

In this research, various functions are proposed to model the impact of individuals on each other in the network. An algorithm based on the indirect influence principle is also presented to infer the graph of the influence of individuals on each other. We evaluated the proposed algorithm based on the various functions affecting the artificial data. As a result of these experiments, we selected model 7 for our proposed method. Then, the proposed method was compared with the INFOPATH method. The INFOPATH model, with hypotheses similar to the proposed method, attempts to infer an influence network based on information cascades. The INFOPATH model has been developed based on the NETRATE model and has been reported to be much faster in terms of runtime. The proposed method has been compared in terms of the accuracy of network inference and runtime with the INFOPATH model on most networks and possible dissemination modes. The comparison of the proposed method with the INFOPATH based on the f1 measure shows that the proposed method can better infer the network. The runtime of the proposed algorithm is of the order of $O(n^3)$. The INFOPATH article does not refer to the time complexity of the algorithm. Therefore, for comparison of these methods, the actual execution time was calculated. From the results of the experiments, the proposed method was found to run faster than the INFOPATH method.

The performed experiments demonstrate that the combination of the time interval and counting parameter creates a better function to calculate the influence rate of individuals on each other. This research only deduces the desired graph based on the information cascades. Therefore, in order to continue to work, information in the content, such as re-tweet or mention, can also be used to improve accuracy. We can also continue to work on functions that improve the performance of the algorithm. Also, specific functions can be provided for real data based on the specific domain of the data.

References

[1] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information Diffusion in Online Social Networks: A Survey," ACM SIGMOD Record, vol. 42, no. 2, pp. 17–28, 2013.

[2] R. Badie, A. Aleahmad, M. Asadpour, and M. Rahgozar, "An efficient agent-based algorithm for overlapping community detection using nodes' closeness," Physica A: Statistical Mechanics and its Applications, vol. 392, no. 20, pp. 5231–5247, Oct. 2013.

[3] Z. Arefian and M.R. Khayyam Bashi. "Scalable Community Detection through Content and Link Analysis in Social Networks," Journal of Information Systems and Telecommunication (JIST), vol. 4, no.12, pp. 1-10, 2015.

[4] A.H. Hosseinian and V. Baradaran. "A multi-objective multi-agent optimization algorithm for the community detection problem," Journal of Information Systems and Telecommunication (JIST), vol. 6, no. 3, pp. 166-176, 2018.

[5] E. Sherkat, M. Rahgozar, and M. Asadpour, "Structural link prediction based on ant colony approach in social networks," Physica A, vol. 419, pp. 80–94, 2015.

[6] V. Martínez, F. Berzal, and J.-C. Cubero, "A Survey of Link Prediction in Complex Networks," ACM Computing Surveys, vol. 49, no. 4, pp. 1–33, Dec. 2016.

[7] M. P. Salvati, J. Bagherzadeh Mohasefi, and S. Sulaimany. "Overcoming the Link Prediction Limitation in Sparse Networks using Community Detection," Journal of Information Systems and Telecommunication (JIST), vol. 3, no. 35, pp. 183-190, 2021.

[8] R. Safa, S. A. Mirroshandel, S. Javadi, and M. Azizi. "Publication venue recommendation based on paper title and co-authors network," Journal of Information Systems and Telecommunication (JIST), vol. 6, no. 21, pp. 33-40, 2018.

[9] K. Rahimkhani, A. Aleahmad, M. Rahgozar, and A. Moeini, "A Fast Algorithm for Finding Most Influential People Based on the Linear," Expert Systems With Applications, vol. 42, no. 3, pp. 1353–1361, Feb. 2014.

[10] M. Emadi and M. Rahgozar, "Twitter sentiment analysis using fuzzy integral classifier fusion," Journal of Information Science, vol. 46, no. 2, 2020.

[11] A.A. Kardan and B. Bozorgi. "Analysis of Main Expert-Finding Algorithms in Social Network in Order to Rank the Top Algorithms," Journal of Information Systems and Telecommunication (JIST), vol. 5, no. 20, pp. 217-224, 2017.

[12] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information Diffusion in Online Social Networks: A Survey," ACM SIGMOD Record, vol. 42, no. 2, pp. 17–28, 2013.

[13] I. Brugere, B. Gallagher, and T. Y. Berger-Wolf, "Network Structure Inference, A Survey: Motivations, Methods, and Applications," Oct. 2016.

[14] O. Gomes, "Sentiment cycles in discrete-time homogeneous networks," Physica A: Statistical Mechanics and its Applications, vol. 428, pp. 224–238, Jun. 2015.

[15] L. Zhao, J. Wang, R. Huang, H. Cui, X. Qiu, and X. Wang, "Sentiment contagion in complex networks," Physica A: Statistical Mechanics and its Applications, vol. 394, pp. 17–23, Jan. 2014.

[16] L. Prokhorenkova, A. Tikhonov, and Y. Nelly Litvak, "When Less Is More: Systematic Analysis of Cascade-Based Community Detection," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 16, no. 4, Jan. 2022.

[17] I. Brugere, B. Gallagher, and T. Y. Berger-Wolf, "Network Structure Inference, A Survey: Motivations, Methods, and Applications," Oct. 2016.

[18] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring Networks of Diffusion and Influence," ACM Transactions on Knowledge Discovery from Data, vol. 5, no. 4, pp. 1–37, Feb. 2010.
[19] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, 2013, pp. 23–32.
[20] M. G. RODRIGUEZ, J. LESKOVEC, D. BALDUZZI, and B. SCHÖLKOPF, "Uncovering the structure and temporal dynamics of information propagation," Network Science, vol. 2, no. 01, pp. 26–65, Apr. 2014.
[21] LiHuacheng, XiaChunhe, WangTianbo, WenSheng, ChenChao, and XiangYang, "Capturing Dynamics of Information Diffusion in SNS: A Survey of Methodology and Techniques," ACM Computing Surveys (CSUR), vol. 55, no. 1, pp. 1–51, Nov. 2021.
[22] C. Ravazzi, F. Dabbene, C. Lagoa, and A. v. Proskurnikov, "Learning Hidden Influences in Large-Scale Dynamical Social Networks: A Data-Driven Sparsity-Based Approach, in Memory of Roberto Tempo," IEEE Control Systems, vol. 41, no. 5, pp. 61–103, Oct. 2021.
[23] H. Yang et al., "Towards embedding information diffusion data for understanding big dynamic networks," Neurocomputing, vol. 466, pp. 265–284, Nov. 2021.
[24] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, First Ed. Cambridge, United Kingdom: Cambridge University Press, 1994.
[25] K. Chen, P. Luo, and H. Wang, "An influence framework on product word-of-mouth (WoM) measurement," Information and Management, vol. 54, no. 2, pp. 228–240, Mar. 2017.
[26] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring Networks of Diffusion and Influence," ACM Transactions on Knowledge Discovery from Data, vol. 5, no. 4, pp. 1–37, Feb. 2010.
[27] M. G. RODRIGUEZ, J. LESKOVEC, D. BALDUZZI, and B. SCHÖLKOPF, "Uncovering the structure and temporal dynamics of information propagation," Network Science, vol. 2, no. 01, pp. 26–65, Apr. 2014.
[28] S. Shaghaghian and M. Coates, "Online Bayesian Inference of Diffusion Networks," IEEE Transactions on Signal and Information Processing over Networks, vol. 3, no. 3, pp. 500–512, Sep. 2016.
[29] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, 2013, pp. 23–32.
[30] S. A. S. Myers and J. Leskovec, "On the Convexity of Latent Social Network Inference," in NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2, 2010, pp. 1741–1749.
[31] S. Wang, X. Hu, P. S. Yu, and Z. Li, "MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades," Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14, pp. 1246–1255, 2014.
[32] N. Du, L. Song, H. Woo, and H. Zha, "Uncover Topic-Sensitive Information Diffusion Networks," in Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, 2013, vol. 31, pp. 229–237.
[33] D. H. Zhou, W. B. Han, Y. J. Wang, and B. Di Yuan, "Information diffusion network inferring and pathway tracking," Science China Information Sciences, vol. 58, no. 9, pp. 1–15, Sep. 2015.
[34] C. E. Cormen, Thomas H., Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, 3rd ed. MIT Press and McGraw-Hill, 2009.
[35] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker Graphs: An Approach to Modeling Networks," Journal of Machine Learning Research, vol. 11, no. Feb, pp. 985–1042, 2010.
[36] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '11, 2011, pp. 1082-1090.
[37] J. Leskovec and R. Sosič, "SNAP: A General-Purpose Network Analysis and Graph-Mining Library," ACM Transactions on Intelligent Systems and Technology, vol. 8, no. 1, pp. 1–20, Jul. 2016.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 320-330
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Joint Cooperative Spectrum Sensing and Resource Allocation in Dynamic Wireless Energy Harvesting Enabled Cognitive Sensor Networks

Maryam Najimi[1*]

[1].Faculty of Electrical & Computer Engineering, University of Science and Technology of Mazandaran, Behshahr, Iran

Received: 03 Feb 2022/ Revised: 04 Jan 2023/ Accepted: 26 Feb 2023

Abstract

Due to the limitations of the natural frequency spectrum, dynamic frequency allocation is required for wireless networks. Spectrum sensing of a radio channel is a technique to identify the spectrum holes. In this paper, we investigate a dynamic cognitive sensor network, in which the cognitive sensor transmitter has the capability of the energy harvesting. In the first slot, the cognitive sensor transmitter participates in spectrum sensing and in the existence of the primary user, it harvests the energy from the primary signal, otherwise the sensor transmitter sends its signal to the corresponding receiver while in the second slot, using the decode-and-forward (DF) protocol, a part of the bandwidth is used to forward the signal of the primary user and the remained bandwidth is used for transmission of the cognitive sensor. Therefore, our purposed algorithm is to maximize the cognitive network transmission rate by selection of the suitable cognitive sensor transmitters subject to the rate of the primary transmission and energy consumption of the cognitive sensors according to the mobility model of the cognitive sensors in the dynamic network. Simulation results illustrate the effectiveness of the proposed algorithm in performance improvement of the network as well as reducing the energy consumption.

Keywords: Cognitive Sensor Network; Transmission Rate; Mobility Model; Decode-and-Forward (DF) Protocol; Energy Consumption.

1- Introduction

Cognitive radio is a growing technology to overcome the spectrum scarcity in wireless communications. In fact, spectrum sensing is done in cognitive radio networks to determine the spectrum holes for data transmission. In fact, secondary users (SUs) sense the frequency channel to detect the primary user (PU) activity which has the legacy right for the frequency band usage [1]. A new network is the cognitive radio sensor network (CRSN), which has a lot of applications and can manage the spectrum resources. On the other hand, the sensor nodes sense the frequency band and a fusion center (FC) makes a final decision about the status of the frequency band according to the sensors' information. However, sensors are powered with batteries which should be recharged or replaced. This problem leads to decrease the network lifetime. Therefore, improving the network lifetime is very important in cognitive radio sensor networks. In this case, reducing the energy consumption leads to reserve the battery power of the sensors. For this purpose, in [2] and [3], the proper sensor nodes are participated in spectrum sensing to minimize the energy consumption while satisfy the detection performance constraints. In [4], in addition to the energy consumption, the remaining energy of each sensor is an important parameter for selection of the sensors for spectrum sensing. Therefore, extending the network lifetime is the main issue in this paper.

Energy harvesting is another issue to improve the network lifetime. On the other words, cognitive sensors can harvest energy from the environment to charge their batteries. In fact, wireless information and power transfer (SWIPT) is a technology to transfer the energy and information simultaneously as the RF signals to the sensor nodes [5], [6]. In [7], SWIPT protocol is used such that the secondary users harvest their energies from the primary signals to send the primary users' signal and also their signals. In [8], for increasing the energy efficiency, a secondary transmitter is considered as a relay to transmit the signal of the primary user while the secondary receiver does energy

✉ **Maryam Najimi**
Email Address: M.najimi@mazust.ac.ir

harvesting. However, in these papers, the same bandwidth is applied for transmission of the primary and sensor signals. It leads to have the interference in the network.

However, another issue is the mobility of the sensor nodes. In this case, the dynamic topology of the network cannot be applied for static networks. By mobility of the sensor nodes, the distance between each sensor and primary user and also between each sensor and FC is not fixed while it is a random variable. In this case, two approaches including: "wait-and-see" [9] and "here-and-know" [10] are considered for the optimization problem with the random variable. In this paper, we use "here-and-know" method in which Chebyshev's inequality is applied for energy consumption value estimation in a cognitive sensor network. On the other words, the suitable sensor nodes are participated in cooperative spectrum sensing to save more energy. We also use a spectrum sharing protocol in which a cognitive sensor transmitter acts as a relay to transmit the primary signal and harvest energy in the determined accessed bandwidth while in the remaining bandwidth, its signal is transmitted to the corresponding cognitive sensor receiver. In [11], an efficient cooperative spectrum sensing based on Kataoka criterion is stated by node mobility patterns consideration in a dynamic cognitive radio sensor network. In [12], two time slots are considered for spectrum sharing: the first slot is considered for energy harvesting of the cognitive sensor transmitter while in the second slot, amplify-and-forward (AF) or decode-and-forward (DF) relaying protocols are used by the cognitive sensor transmitter to send the signals of the primary transmitter to its primary receiver while in the remaining bandwidth, the signal of the cognitive sensor transmitter is sent. However, the sensor node is located in a fixed position. In [13], optimal disjoint and joint spectrum sensing and power allocation method is considered in a cognitive radio (CR) network with two aims: minimizing the false alarm probability while probability of detection is constant and maximizing the average opportunistic CR data rate under detection probability and CR power budget limitations. In [14], an approach is proposed for cluster head selection and cluster forming such that the coverage and lifetime of the network are improved. In [15], three phases are considered in wireless sensor networks. In the first phase, the position of the sensors are determined. In the second phase, the optimal location of the base station is obtained and in the third phase, the cluster head is selected based on the energy remaining, distance and the number of neighbors .In [16], an energy efficient clustering algorithm is proposed for clustering method and improving the coverage of the network. In [17], a cognitive sensor network is considered in which secondary users relay the information of primary user while primary user leases partial spectrum usage time to secondary users. In this paper, a joint sub channel, power and leasing time allocation algorithm is proposed to maximize the network throughput with constraints on the energy harvesting and transmission outage probability. In [18], the optimization problem of power allocation and spectrum access for maximizing the achievable data rate and minimizing the energy consumption at the secondary network, is formulated and solved using Dinkelbach algorithm. In [19], an energy efficient algorithm applies the gravitational search method to determine the optimal number of clusters and cluster heads. In [20], the problem of the network lifetime is proposed to schedule the network coverage using sleep-awake method for sensors. In [21], a resource allocation scheme is proposed based on the Lyapunov optimization theory while the constraints on the network quality of service (QoS) are satisfied. In [22], a game theory approach is applied in energy efficient wireless sensor networks such that the sensor nodes act as players and decide to sleep or not according to the idle listening time.

Therefore, the main contributions of our work are stated as follows

- A dynamic and energy harvesting cognitive sensor network is considered with the random waypoint model. On the other hand, the cognitive sensor node has the capability of the spectrum sensing and energy harvesting form the primary signal while it transmits the primary signal with a determined accessed bandwidth. In the remaining bandwidth, the cognitive sensor node transmits its signal to the corresponding receiver.
- We propose the problem of maximizing the cognitive system transmission rate by selection of the proper sensors for frequency band sensing and sharing such that the total energy consumption and the primary transmission rate constraints are satisfied.
- We formulate the problem and solve it by applying the convex optimization method and Karush–Kuhn–Tucker (KKT) conditions.
- Simulation results validate the effectiveness of our proposed method for improving the transmission rate of the network and decreasing the energy consumption over the benchmark algorithm.

The rest of the paper is organized as follows. The system model of a cognitive sensor network is stated in section 2 while the formulation of the problem and the problem solution is stated in Section 3. In Section 4, an iterative algorithm is proposed based on the bisection method to solve the problem. Simulation results and conclusions are shown in Section 5 and 6, respectively.

2- System Model

We assume a wireless dynamic cognitive sensor network which consists of a primary system and a cognitive sensor system with the energy harvesting capability. In the primary system, one primary transmitter (PT) and one primary receiver (PR) exist in the network while the cognitive sensor system has N transmitter (ST) and N receiver (SR). The network also has a fusion center (FC) such that the cognitive sensors send their spectrum sensing results to it. It should be noted that each cognitive sensor transmitter has the capability of the energy harvesting in the licensed spectrum of the primary user while it relays the signal of the primary transmitter to the primary receiver. It also transmits its own signal to its corresponding receiver in the determined bandwidth of the channel. We also note that the sensor nodes are moving randomly in the square field of the environment. In fact, we consider two transmission slots. In the first slot, the selected cognitive sensors sense the frequency band to detect the primary user activity, if the primary user is absent, sensors transmit their signal to their receivers while in the presence of the primary signal, the cognitive sensors receive the primary user signal, harvest their energy from it and decode the information from the remaining signal power. In this case, in the second transmission slot, the best selected cognitive sensor acts as a relay and considers a part of bandwidth to forward the primary transmitter signal to the primary receiver while the cognitive sensor uses from the remaining bandwidth for transmission of its signal to the corresponding receiver. For cooperative spectrum sensing, the sensing sensors send their results about the activity of the primary user to the fusion center (FC) to consider a decision about the availability of the spectrum. In this case, two hypothesis are considered. In the first case, $H_1: y_i[n] = h_i[n]x_i[n] + n_i[n]$ shows the presence of the primary user. In the second case, $H_0: y_i[n] = n_i[n]$ states the absence of the primary user. $n\epsilon\{1,2,\dots,\delta f_s\}$ is the time index while δ is the duration of spectrum sensing, f_s is the sampling frequency and T= δf_s states the total number of samples. $h_i[n]$ is the channel gain between the ith sensor transmitter and the primary user. $x_i[n]$ is the transmitted signal of the primary user while $n_i[n]$ is an $i.i.d.$ Gaussian noise with zero mean and variance σ_u^2. The main notations used in this work is also presented in Table 1.

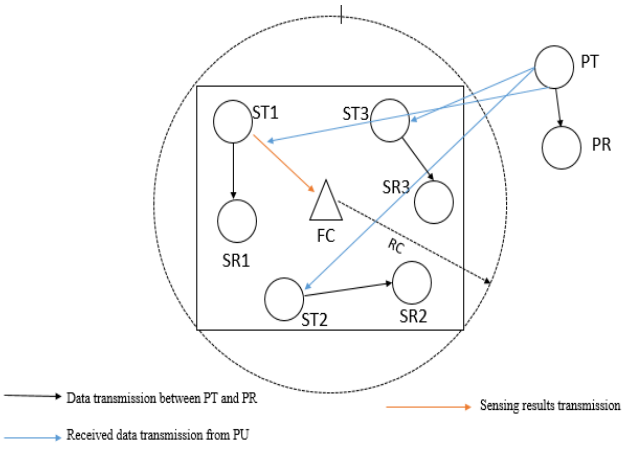


Fig.1 System structure

Table.1. Notations

Notations	Description
N	Number of cognitive sensor transmitter and receiver
δ	Duration of spectrum sensing
f_s	Sampling frequency
σ_u^2	Variance of the Gaussian Noise
P_{d_i}	Probability of detection of the ith sensor
P_{f_i}	Probability of false alarm of the ith sensor
P_D	Total Probability of detection
P_F	Total Probability of false alarm
ρ_i	Assignment index of for sensing
E_s	Energy consumed of the sensor transmitter for spectrum sensing
E_{t-elec}	Energy consumption for the radio electronics
e_{amp}	Power Amplification
d_i	Distance between the ith sensor transmitter and FC
θ	Probability of energy consumption of each sensor transmitter
W	Bandwidth of the licensed spectrum
bW	Part of bandwidth for transmission of the primary signal
P_p	Transmission power of the primary transmitter
P_C	Transmission power of the sensor transmitter
π_i	Assignment index for data transmission
α	Fraction of the received power for harvesting energy
$h_{pt,pr}$	Channel gain between the primary transmitter and primary receiver
$h_{pt,sr}$	Channel gain between Primary transmitter and sensor transmitter
$h_{st,sr}$	Channel gain between sensor transmitter and sensor receiver
${R_p}^1$	Data rate of the primary transmitter to its receiver
${R_s}^{11}$	Data rate of the primary user and sensor transmitter
${R_s}^{12}$	Data rate of the sensor transmitter and sensor receiver

R_p	Achievable rate of the primary user
${R_s}^2$	Achievable rate of the ith cognitive sensor node
R_{pu}	Distance between the primary user and FC
R_c	Cluster radius
$\theta 1$	Exponent of the path loss in Hata model

For simple implementation of the energy detection method and using the received signal energy, the probabilities of the detection and false alarm of the ith sensor are obtained as

$$P_{d_i} = P(E_i > \epsilon | H_1) = Q_T\left(\sqrt{2\gamma_i}, \sqrt{\epsilon}\right) \quad (1)$$

And

$$P_{f_i} = P(E_i > \epsilon | H_0) = \frac{\Gamma(T, \frac{\epsilon}{2})}{\Gamma(T)} \quad (2)$$

Where E_i is the energy obtained from the primary signal to the ith sensor transmitter while ϵ represents the detection threshold. $\Gamma(a,x)$ and $Q_m(a,b)$ denote the incomplete gamma function and the generalized Marcum Q-function, respectively. In fact, higher value of P_{d_i} decreases the probability of interference with the primary signal while lower value of P_{f_i} increases the opportunity of the spectrum usage. In cooperative spectrum sensing, FC can use OR rule as the combination rule to consider a decision about the spectrum status. According to this rule, the channel is considered busy if at least one sensor transmitter detects the existence of the primary signal. Hence, the global probabilities of detection and false alarm are defined as [2]

$$P_D = 1 - \prod_{i=1}^{N}(1 - \rho_i P_{d_i}) \quad (3)$$

And

$$P_F = 1 - \prod_{i=1}^{N}(1 - \rho_i P_{f_i}) \quad (4)$$

Where $\rho_i \epsilon \{0,1\}$. $\rho_i = 0$ denotes that the sensor transmitter is not participated in sensing the frequency channel, otherwise $\rho_i = 1$. However, one of the important issues in cognitive sensors nodes is the constraints of the energy consumption. For cooperative spectrum sensing in static position of the sensors, the energy consumed of the sensor nodes is denoted by [23], [24]

$$E_T = \sum_{i=1}^{N} \rho_i (E_s + E_{t-elec} + e_{amp} d_i^2) \quad (5)$$

Where E_s is the energy consumed of the sensor transmitter for spectrum sensing while E_{t-elec} indicates the energy consumption for the radio electronics. e_{amp} is considered for power amplification. d_i indicates the distance between the ith sensor transmitter and FC. By considering the dynamic position of the sensor nodes according to the random waypoint model, d_i^2 and E_T will be the random values. In this model, the sensor nodes (transmitters and receivers) move from one location to another position. In random waypoint model, the sensors stay in their positions in a duration time and after the time expiration, they can move to their new locations. By definition of E_φ as the upper bound of the energy consumption, we have [11]

$$E_\varphi = \rho_i (E_s + E_{t-elec} + e_{amp} F_{d^2}^{-1}(\theta)) \quad (6)$$

Where $F_{d^2}^{-1}(.)$ states the inverse of the cumulative distribution function (CDF) of the squared distance between the sensors and FC. This function depends on the mobility pattern of the sensors nodes. $\theta \in [0,1]$ represents the probability that the energy consumption of each sensor transmitter less than or equal to E_φ.

For calculation of $F_{d^2}^{-1}(.)$, we consider the probability density function(PDF) of $x = \frac{1}{d^2}$ as the random variable in Fig.2. Then, using the probability density function (PDF) of the random variable of $y = d^2$, it is possible to obtain

$F_{d^2}^{-1}(.)$. Therefore we have,

$$f_Y(y) = \frac{1}{y^2} f_X(\frac{1}{y}) \quad (7)$$

Where $f_X(x)$ and $f_Y(y)$ are the probability density function of x and y, respectively. For the ease of mathematical solution, it is considered the other parameters except the energy consumption are independent from the differences in nodes' locations.

In this paper, we use decode- and- forward strategy in which at the first slot, the selected sensing sensors sense the spectrum to detect the activity of the primary user. When the primary user does not exist, the sensors can forward their signal while in the presence of the primary user, the sensors have the energy harvesting capability and also information decoding. Therefore, the data rates of the primary transmitter to its receiver, primary user and sensor transmitter and also the sensor transmitter and sensor receiver can be stated as follows[12]

$${R_p}^1 = \frac{1}{2} W log_2\left(1 + \frac{P_p |h_{pt,pr}|^2}{\sigma_u^2}\right) \quad (8)$$

And

$${R_s}^{11} = \frac{1}{2} W log_2\left(1 + \frac{\alpha P_p \pi_i |h_{pt,st}|^2}{\sigma_u^2}\right) \quad (9)$$

And

$${R_s}^{12} = \frac{1}{2} W log_2\left(1 + \frac{P_C \pi_i |h_{st,sr}|^2}{\sigma_u^2}\right) \quad (10)$$

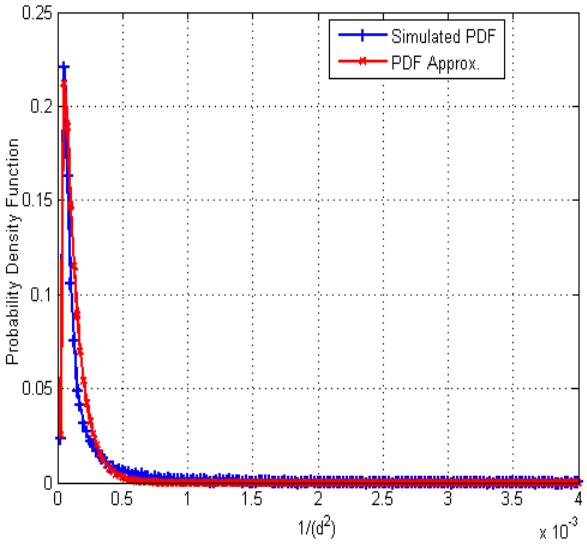


Fig.2. Theoretical and simulated probability density function of $\frac{1}{d^2}$

where $\pi_i \in \{0,1\}$ is similar to ρ_i . It indicates that whether the sensor transmitter is participated in data transmission or not. P_C is the transmission power of the sensor transmitter. $h_{pt,pr}$, $h_{pt,st}$ and $h_{st,sr}$ are the channel gain between the primary transmitter and the primary receiver, the primary transmitter and the sensor transmitter and also between sensor transmitter and the sensor receiver, respectively. α is the fraction of the received power for harvesting energy. W is the bandwidth of the licensed spectrum while bW is the part of bandwidth for transmission of the primary signal. In the second time slot, the selected sensor transmitter acts as a relay to forward the primary transmitter's signal while in the remaining bandwidth, it transmits its own signal to the corresponding sensor receiver. Therefore, by applying maximal ratio combination (MRC) through two slots, we have [12]

$$R_p{}^2 = \frac{1}{2} bW log_2 \left(1 + \frac{\varepsilon(1-\alpha)P_p \pi_i |h_{pt,st}|^2 |h_{st,Pr}|^2}{\sigma_u^2} + \frac{P_p |h_{pt,pr}|^2}{\sigma_u^2}\right) + \frac{1}{2}(1-b)W log_2\left(1 + \frac{P_p |h_{pt,pr}|^2}{\sigma_u^2}\right) \quad (11)$$

Where the term $\varepsilon(1-\alpha)P_p|h_{pt,st}|^2$ is the harvested energy at the cognitive sensor transmitter while ε is the efficiency of the harvested energy at the cognitive sensor transmitter. P_p is the transmission power of the primary transmitter. Hence, the achievable rate of the primary user is obtained as

$$R_p = \min(R_p{}^1, R_p{}^2) \quad (12)$$

We also note the achievable rate of the ith cognitive sensor node is stated as follows

$$R_s{}^2 = \frac{1}{2}(1-b)W log_2\left(1 + \frac{P_C \pi_i |h_{st,sr}|^2}{\sigma_u^2}\right) \quad (13)$$

3- Problem Formulation

The optimization problem is formulated to maximize achievable rate of the cognitive network with constraints on the achievable rate of the primary transmitter, random energy consumption of the sensor nodes and detection performance by selection of the proper sensing nodes and data transmission sensor. Hence, the problem is formulated as follows

$$max_{\pi_i,\rho_i} R_s = (R_s{}^{11} + R_s{}^{12} + R_s{}^2) \quad (14)$$

$$S.t. \quad E_\varphi \le \alpha_1 \quad (14\text{-}1)$$
$$R_p \ge \alpha_2 \quad (14\text{-}2)$$
$$P_F \le \alpha_3 \quad (14\text{-}3)$$
$$P_D \ge \alpha_4 \quad (14\text{-}4)$$
$$\rho_i \in [0,1] \quad (14\text{-}5)$$
$$\pi_i \in [0,1] \quad (14\text{-}6)$$

Where α_1 , α_2 , α_3 and α_4 are the thresholds for the constraints of the problem. We consider $\rho_i \in [0,1]$ and $\pi_i \in [0,1]$ as continuous parameters to reduce the problem complexity in contrast to the NP-complete [2]. We also note that due to the independence of P_{f_i} from γ_i, this metric is the same for all sensor nodes. Although, the problem is not convex; we can use the convex optimization method to solve the problem and obtain the suboptimal solution. In this method, we use the Lagrangian function to express the problem as an unconstrained problem [2]. Therefore, we have

$$L(\pi_i, \rho_i, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = R_s + \lambda_1(E_\varphi - \alpha_1) - \lambda_2(R_p - \alpha_2) + \lambda_3(P_F - \alpha_3) - \lambda_4(P_D - \alpha_4) \quad (15)$$

Where $\lambda_1, \lambda_2, \lambda_3$ and λ_4 are the Lagrangian multipliers which have the nonnegative values. Note that the optimal solution for the problem is the exhaustive search method with high complexity. However, we search a sub optimal method with low complexity by calculation of the cost function for each sensor node and consider the suitable sensors for spectrum sensing and data transmission. Therefore, the cost function of the ith sensor for spectrum sensing is evaluated as follows

$$costf(i) = \lambda_1 E_\varphi + \lambda_3 P_{f_i} - \lambda_4 P_{d_i}$$

(16)

As we know, P_{f_i} is the same for all sensors. However, this metric determines the maximum number of the sensors for spectrum sensors. Hence, we obtain

$$costf(i) = \lambda_1 E_\varphi - \lambda_4 P_{d_i} \quad (17)$$

On the other hand, the sensors which consume the lowest energy and have the higher value of probability of detection, can be considered as the candidates for spectrum sensing. For selection of the sensor node for data transmission, the sensor transmitter with higher R_s and R_p can forward data to its corresponding sensor receiver as the following cost function

$$costf_{DT(i)} = -R_s - \lambda_2 R_p \quad (18)$$

Using Karush Kuhn Tucker conditions (KKT) and the complementary slackness conditions, the optimal conditions for the proposed approach is investigated such that the problem constraints are maintained. It should be noted that sometimes by selection of all sensors for the frequency channel sensing, the detection performance constraints are not satisfied. In this case, the problem has no answer.

We also consider that the cost function should be calculated according to (16) for each sensor transmitter to evaluate the priority of the sensors for sensing. However, (16) is dependent on the inverse probability density function $F_{d^2}^{-1}(\theta)$ which is related to the mobility model. It should be noted that finding a mathematical function for this metric is difficult; Hence, we should consider the numerical methods to find the best function to this metric. On the other hand, the simulations should be run to compute the distance between each sensor and FC. By this method, the probability density function $F_{d^2}^{-1}(\theta)$ is approximated by a polynomial using numerical methods. (using Fig.2). In the next section, we propose an iterative algorithm to find the best sensing sensors and data transmission node as a relay for data transmission of the primary user and itself to the corresponding receiver.

4- Proposed Algorithm

For solving the problem, we propose an iterative algorithm. In the first step, in each iteration, the detection probability is computed for all sensors. By fitting the cumulative density function $F_{d^2}^{-1}(\theta)$, the cost function in (16) is computed for all sensor nodes. Then, this metric is sorted in ascending order and sensors which have the lowest cost function values, are candidates for spectrum sensing. In this algorithm, the suitable sensors are selected one by one for spectrum sensing. If the detection performance constraints are maintained, the selection is stopped otherwise, another candidate sensor with the higher priority is selected. In the second step, the cost function of sensors are calculated according to (18). On the other hand, the node with the highest priority is selected for data transmission. Then, the Lagrangian multipliers are updated using iterative bisection algorithm such that according to their corresponding constraints, the searching space is halved and algorithm is repeated again. The algorithm ends when the desired accuracy of the Lagrangian multipliers is met or the number of iterations reaches a specified number.

We note that sometimes there is not any feasible solutions for the problem. It means that by selection of the all sensors, the detection performance constraints are not met. The implementation of the proposed algorithm is presented in Fig.3. The flowchart for the proposed algorithm is shown in Fig.4.

Note that the complexity of our proposed algorithm with the order of $O(N^2)$ while in exhaustive search method, the order of the complexity is $O(N!)$.

Proposed Algorithm

Initial Parameters($\lambda_{1_{min}}, \lambda_{1_{max}}, \dots$)
Fit function $F_{d^2}^{-1}(\theta)$ for the movement model
ε is a small number
N1=maximum number of iterations
$n1$=number of iterations
M= Maximum number of sensing nodes
While ($n1$ <N1)
$\lambda_k = \frac{\lambda_{k_{min}} \lambda_{k_{max}}}{2}$ $k = 1,2,3,4$
Compute $costf(i) = \lambda_1 E_\varphi - \lambda_4 P_{d_i}$ for each node
While (select n sensor with higher priority< M)
Compute P_d
If $P_d > \alpha_4$, break, end
$n1 = n1 + 1$
end
compute $costf_{DT(i)} = -R_s - \lambda_2 R_p$ for all nodes to obtain the best node for transmitting its data and acts as a relay for primary user's data
Update the Lagrangian multipliers using bisection method
end

Fig.3. Pseudo code for the proposed algorithm

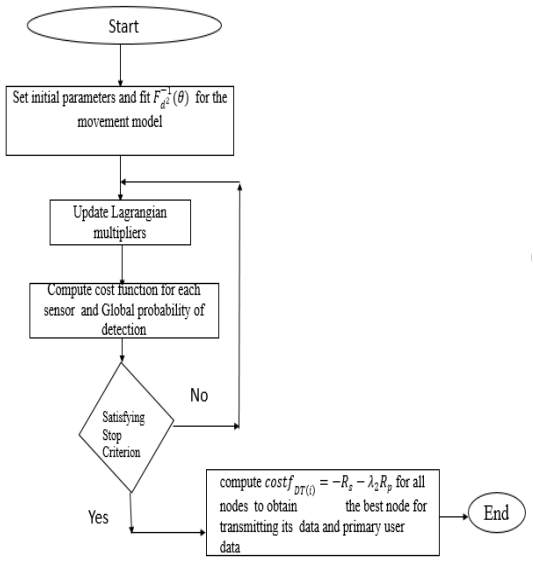


Fig.4. Flowchart for the proposed algorithm

5- Simulation Results

In this section, the performance of our proposed algorithm is evaluated. For this purpose, sensors are distributed randomly over the square field with the length of 100m. The sensors are also moving randomly according to the random waypoint model. FC is located at the center of the square field while the primary user is located outside the cluster which satisfying the following inequality [25]

$$R_{pu} \geq \frac{10^{\frac{0.1}{\theta 1}}+1}{10^{\frac{0.1}{\theta 1}}-1} R_c \qquad (19)$$

Where R_{pu} is the distance between the primary user and FC while R_c is the cluster radius. $\theta 1$ is the exponent of the path loss in Hata model. $\delta = 5\mu s$ is considered as the duration of the sensing time. The inverse cumulative density function $F_{d^2}^{-1}(\theta)$ is calculated in $\theta = 0.9$. The detection performance thresholds are $\alpha_3 = 0.1$ and $\alpha_4 = 0.9$ while $f_c = 2.4$ GHz. The channel gain is modelled according to a free-space path loss model, Raleigh fast fading and large scale log-normal shadowing [26], [27]. According to [28] and [29], the sensing energy (E_s) has two pars: the listening energy which has the value 40 nJ and the signal processing energy. For a data rate of 250kb/s, the signal processing energy is calculated 150 nJ/bit. The remaining energies are defined as E_{t-elec} = 80nJ and e_{amp}=40.4 pJ/m^2. Decision threshold (ϵ) is also selected as a multiple of the noise power [2]. P_p=20mW, P_C= 60mW and ε=1 are considered as the parameters for user's data rates.

The achievable rate of the cognitive network versus different dimensions of the environment is shown in Fig.5. According to the results, our proposed algorithm has the maximum achievable rate due to the proper selection of the sensor node for transmission of its signal and also the primary user signal. It should be noted that increasing the dimension of the environment, decreases the achievable data rate of the cognitive network. Fig.6 illustrates the consumed energy for spectrum sensing versus different dimensions of the environment. The proposed algorithm has the minimum energy consumption in comparison with the random algorithm. In fact, proper selection of the sensing nodes for sensing the frequency channel and transmission of their results to FC has an important role in saving energy.

Fig.7 presents the achievable rate of the primary network versus different dimensions of the environment. According to the results, it is obvious that the proposed algorithm has a better value while by increasing the dimensions of the environment, the value of this metric is decreased due to the decreasing of the receiving power of the receiver.

Fig.8 shows probability of detection for different dimensions of the environment. In fact, this metric states the opportunity of the sensor nodes for spectrum sensing. In fact, the proposed algorithm has the maximum detection probability due to the proper selection of the sensing sensors. We note that by increasing the dimension of the environment, the distances between sensors also increases, therefore, probability of detection of each sensor decreases.

Fig.9 presents the achievable rate of the cognitive network for different values of α. Note that increasing the value of P_C leads to increasing the data rate of the cognitive network. However, by increasing α, this metric also increases.

Fig.10 presents the achievable rate of the cognitive network versus different values of b. As we know, by increasing the value of b, the higher value of the bandwidth is associated to transmit the primary user's signal. Therefore, the achievable rate of the cognitive network is decreased. It is obvious that increasing the value of P_C leads to increasing the data rate of the cognitive network. According to Fig.11, we also note that when the value of b increases, the achievable rate of the primary user is increased due to the increasing value of the associated bandwidth for data transmission of the primary user.

Fig.12 illustrates the convergence analysis for the proposed algorithm in computing the optimal value of λ_4 as the Lagrangian parameter for different iterations. It

should be noted that the convergence is obtained according to the probability of detection for the iterations that reach the optimal value of the Lagrangian parameter. In the 12th iteration, the optimal value of the Lagrangian parameter is obtained. In this experiment, the number of sensor nodes and dimension of environment are set to 20 and 100m, respectively.

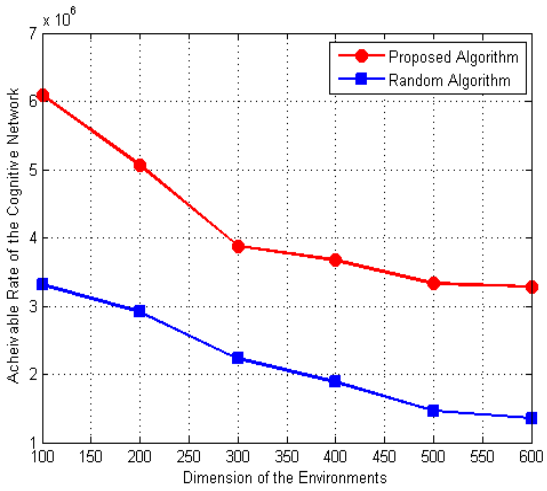


Fig.5. Available rate of the cognitive network versus different dimensions of the environment

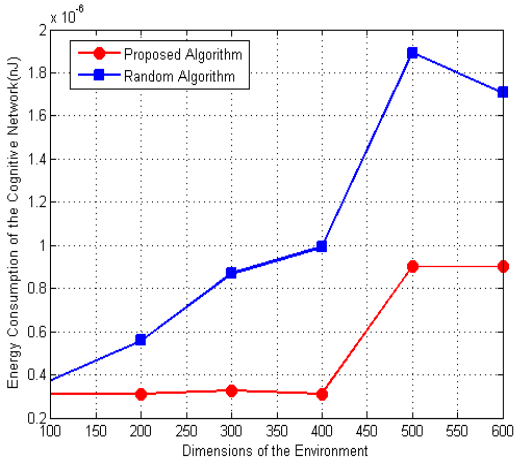


Fig.6. Energy consumption of the cognitive network versus different dimensions of the environment

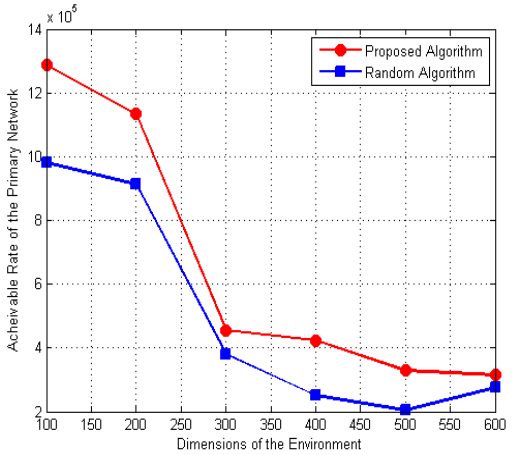


Fig.7. Available rate of the primary network versus different dimensions of the environment

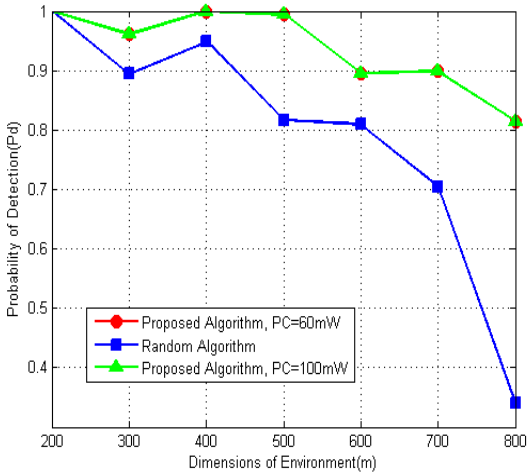


Fig.8. Probability of detection versus different dimensions of the environment

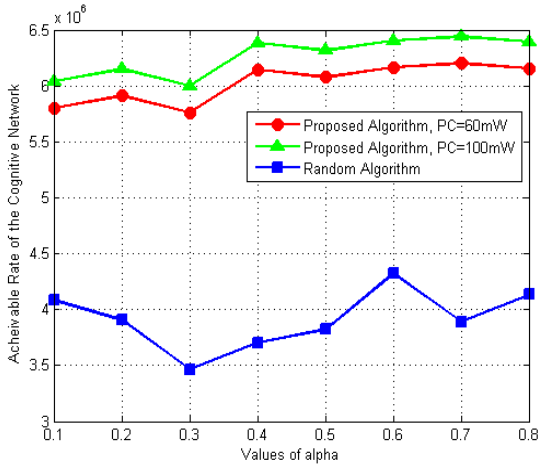


Fig.9. Available rate of the cognitive network versus different values of α

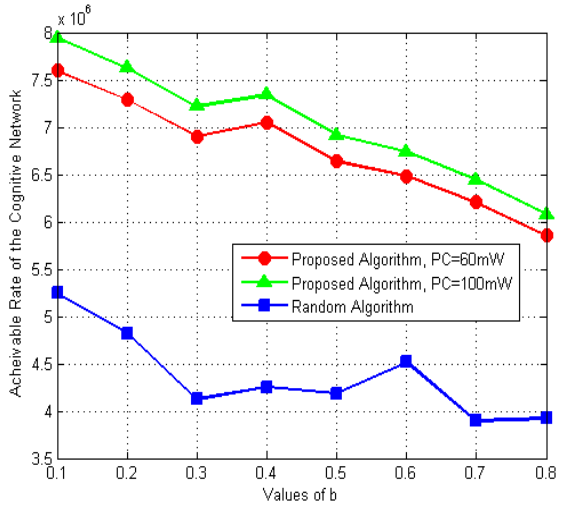


Fig.10. Available rate of the cognitive network versus different values of b

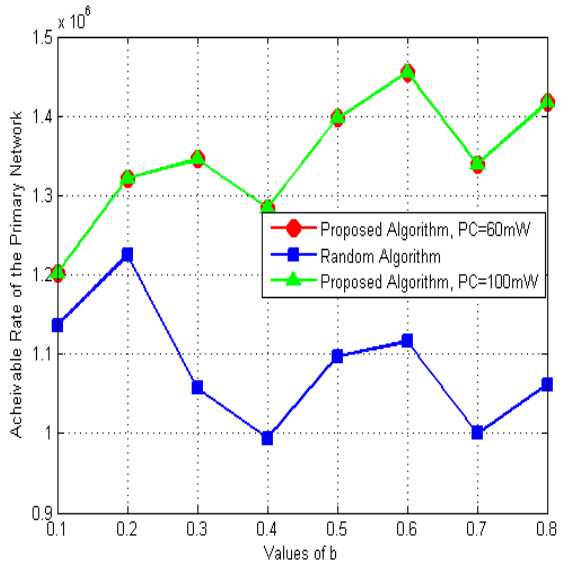


Fig.11. Available rate of the primary network versus different values of b

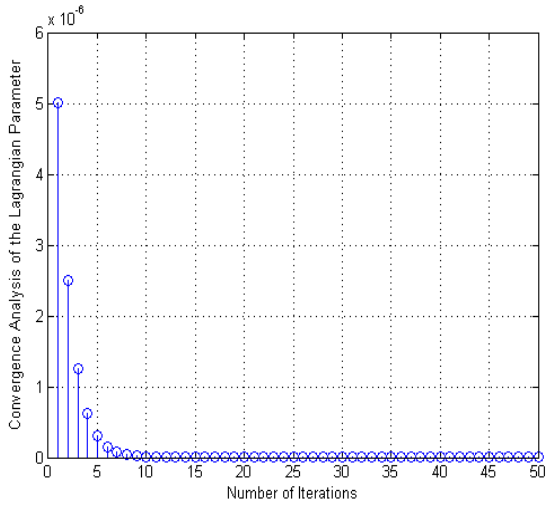


Fig.12. Convergence analysis of the Lagrangian parameter versus different iterations

6- Analysis on Results

We present an iterative algorithm such that the cognitive system transmission rate is maximized by selection of the proper sensors for frequency band sensing and sharing while the total energy consumption and the primary transmission rate constraints are satisfied. We compare our proposed algorithm with the random algorithm in which the sensors are selected randomly for spectrum sensing and data transmission. This algorithm is selected due to its low complexity. In some figures, we compare our proposed algorithm with different transmission power of the sensors ($\boldsymbol{P_C}$). Fig.5, Fig.9 and Fig.10 show the available rate of the cognitive network in different dimensions of the environment, values of $\boldsymbol{\alpha}$ and b. According to figures, our proposed algorithm has the maximum value due to the proper selection of the sensor node for transmission of its signal and also the primary user signal. On the other hand, by increasing the dimension of the environment, the energy consumption of the network is increased (Fig.6) while the available data rate of the network is decreased(Fig.5 and Fig.7). In fact, by increasing $\boldsymbol{\alpha}$, the fraction of the received power for harvesting energy is increased, therefore, the available rate of the cognitive network is increased while by increasing b, more bandwidth is associated to transmit the primary user's signal(Fig.11). Therefore, the achievable rate of the cognitive network is decreased. It should be noted that all algorithms are compared when the constraints of the problem are satisfied.

7- Conclusions

In this paper, a dynamic cognitive sensor network is considered in which mobile sensor nodes have the capability of energy harvesting for spectrum sensing and data transmission. For this purpose, two time slots are considered. in the first slot, the cognitive sensor transmitter participates in spectrum sensing and in the existence of the primary user, it harvests the energy, otherwise the sensor transmitter sends its signal to the corresponding receiver while in the second slot, using the decode-and-forward (DF) protocol, a part of the bandwidth is used to transmit the primary signal and the remained bandwidth is used for transmission of the cognitive sensor. To this end, our optimization problem is proposed to maximize the cognitive network rate subject to the rate of the primary transmission, energy consumption of the cooperative spectrum sensing and the detection performance constraints by proper selection of the cognitive sensor transmitters for spectrum sensing and data transmission. Simulation results show the performance of proposed solution while satisfying the constraints of the problem in comparison with the bench mark algorithms.

References

[1] J. Mitola and G. Q. Maguire, "Cognitive radio: Making software radios more personal, " IEEE Pers. Commun., Vol. 6, No. 4, pp. 13-18, Aug. 1999.

[2] M. Najimi, A. Ebrahimzadeh, S. M. H. Andargoli, and A. Fallahi, "A novel sensing nodes and decision node selection method for energy efficiency of cooperative spectrum sensing in cognitive sensor networks," IEEE Sensors J., Vol. 13, No. 5, pp. 1610-1621, May 2013.

[3] A. Ebrahimzadeh, M. Najimi, S. M. H. Andargoli, and A. Fallahi, "Sensor selection and optimal energy detection threshold for ef_cient cooperative spectrum sensing, " IEEE Trans. Veh. Technol., Vol. 64, No. 4, pp. 1565-1577, Apr. 2015.

[4] A. Bagheri, A. Ebrahimzadeh, and M. Najimi, "Sensor selection for extending lifetime of multi-channel cooperative sensing in cognitive sensor networks " Phys. Commun., Vol. 26, pp. 96_105, Feb. 2018.

[5] S. Kisseleff, X. Chen, I. F. Akyildiz, and W. H. Gerstacker, "Efficient charging of access limited wireless underground sensor networks," IEEE Trans. Commun., Vol. 64, No. 5, pp. 2130-2142, May 2016.

[6] A. Mehrabi, K. Kim, "General framework for network throughput maximization in sink-based energy harvesting wireless sensor networks, " IEEE Trans. Mobile Computing, Vol. 16, No. 7, pp. 1881-1896, July,2017.

[7] G. Zheng, Z. Ho, E. A. Jorswieck, and B. Ottersten, "Information and energy cooperation in cognitive radio networks," IEEE Trans. Signal Process., Vol. 62, No. 9, pp. 2290-2303, May 2014.

[8] J. Yan, Y. Liu, "A dynamic SWIPT approach for cooperative cognitive radio networks," IEEE Trans. Vehicular Technology, Vol. 66, No. 12, pp. 1122-1136, Dec., 2017.

[9] J. R. Birge and F. Louveaux, Introduction to Stochastic Programming 2nd ed. New York, NY, USA: Springer, Jun. 2011.

[10] R. Caballero, E. Cerda, M. M. Muñoz, and L. Rey, "Analysis and comparisons of some solution concepts for stochastic programming problems, " Top, Vol. 10, No. 1, pp. 101_123, Jun. 2002.

[11] H. Kaschel, K. Toledo, J. Torres Gomez and M. Julia Fernandez- Getino Garcia, "Energy-efficient cooperative spectrum sensing base on stochastic programming in dynamic cognitive radio sensor networks, " IEEE Access Journal, Vol.9, pp.720-732, Dec.2020.

[12] W. Lu, T. Nan, Y. Gong, M. Qin, X. Lui, Zh. Xu and Zh. Na, " Joint resource allocation for wireless energy harvesting enabled cognitive sensor networks, " IEEE Access Journal, Vol.6, pp.22480-22488,2018.

[13] M. Karimi, S.M.S. Sadough and M.Torabi, "Improved joint spectrum sensing and power allocation for cognitive radio networks using probabilistic spectrum access, " IEEE Syst. Journal, Vol.13, No.4, pp. 3716-3723, Jan.2019.

[14] A. Pakmehr and A. Ghaffari , "Coverage improving with energy efficient inwireless sensor networks, "Journal of Information Systems and Telecommunication (JIST), Vol.5, No.1, 2017.

[15] M.R. Thaghva, R. Hamlbarani Haghi, A. Hanifi and K. Feizi, "Clustering for reduction of energy consumption in wireless sensor networks by AHP method," Journal of Information Systems and Telecommunication (JIST), Vol.6, No.1, 2018.

[16] M. Bavaghar, A. Mohajer and Sara Taghavi Motlagh, "Energy efficient clkustring algorithm for wireless sensor networks," Journal of Information Systems and Telecommunication (JIST), Vol.7, No. 4 , 2019.

[17] Zh. Liu, M. Zhao, Y. Yuan and X. Guan, "Subchannel and resource allocation in cognitive radio sensor network with wireless energy harvesting , " Computer Networks, Vol.167, Feb. 2020.

[18] M.Sharifi and M. Mohassel Feghhi, "Joint energy and throughput optimization in energy harvesting cognitive sensor networks, " 29th Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, May 2021.

[19] S. Ebrahimi Mood and M.M. Javadi, "Energy-efficient clustering method for wireless sensor networks using modified gravitational search algorithm, "Evolving Systems Journal, Vol.11, pp.575-578, 2020.

[20] J-C Charr , K. Deschinkel, R. Haj Mansour and M. Hakem, " Lifetime optimization for partial coverage in heterogeneous sensor networks, " Ad hoc Networks, Vol. 107, 2020.

[21] X. Deng, P. Guan, C. Hei, F. Li, J. Liu and N. Xiong, "An intelligent resource allocation scheme in energy harvesting cognitive wireless sensor networks, "IEEE Transactions on Network Science and Engineering, Vol.8, No.2, 1900-1912, 2021.

[22] X. Yan, Ch. Huang, J. Gan and X. Wu, "Game theory-based energy efficient clustering algorithm for wireless sensor networks, " Sensors Journal, Vol. 22, No.2, 2022.

[23] A. Bagheri, A. Ebrahimzadeh and M. Najimi, "Lifetime maximization by dynamic threshold and sensor selection in

multi-channel cognitive sensor networks, " Journal of Information Systems and Telecommunication (JIST), Vol.5, No.4, pp.225-235, 2017.

[24] M.Najimi, " Cooperative game approach for mobile primary user localization based on compressive sensing in multi-antenna cognitive sensor networks, " Journal of Information Systems and Telecommunication (JIST), Vol.7, No.2, pp.134-143, 2019.

[25] M. Monemian and M. Mahdavi, "Analysis of a new energy-based sensor selection method for cooperative spectrum sensing in cognitive radio networks, " IEEE Sensors J., Vol. 14, No. 9, pp. 3021_3032, Sep. 2014.

[26] B. Sklar, "Rayleigh fading channels in mobile digital communication systems part1:Characterization, " IEEE Commun. Mag., Jul. 1997.

[27] Y. Ma, D. I. Kim, Zh. Wu, "Optimization of ofdm-based cellular cognitive radio networks, " IEEE Trans. on Communications. Vol. 58, No.8, Aug.2010.

[28] S. Maleki, A. Pandharipande, and G. Leus, "Energy-efficient distributed spectrum sensing for cognitive sensor networks, " in Proc. 35th Annu. Conf. IEEE Ind. Electron. Soc., Nov. 2009, pp. 2642–2646.

[29] S. Maleki, A. Pandharipande, and G. Leus, "Energy efficient distributed spectrum sensing with convex optimization, " in Proc. 3rd Int. Workshop Comput. Advances in Multi-Sensor Adaptive Processing, Nov.2009, pp. 396–399.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 331-346
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Application of Machine Learning in the Telecommunications Industry: Partial Churn Prediction by using a Hybrid Feature Selection Approach

Fatemeh Mozaffari[1], Iman Raeesi Vanani[2], Payam Mahmoudian[1], Babak Sohrabi[1*]

1.Department of Information Technology Management, College of Management, University of Tehran, Tehran, Iran
2.Department of Industrial Management, Faculty of Management and Accounting, Allameh Tabataba'i University, Tehran, Iran

Received: 09 Jul 2022/ Revised: 04 Feb 2023/ Accepted: 07 Mar 2023

Abstract

The telecommunications industry is one of the most competitive industries in the world. Because of the high cost of customer acquisition and the adverse effects of customer churn on the company's performance, customer retention becomes an inseparable part of strategic decision-making and one of the main objectives of customer relationship management. Although customer churn prediction models are widely studied in various domains, several challenges remain in designing and implementing an effective model. This paper addresses the customer churn prediction problem with a practical approach. The experimental analysis was conducted on the customers' data gathered from available sources at a telecom company in Iran. First, partial churn was defined in a new way that exploits the status of customers based on criteria that can be measured easily in the telecommunications industry. This definition is also based on data mining techniques that can find the degree of similarity between assorted customers with active ones or churners. Moreover, a hybrid feature selection approach was proposed in which various feature selection methods, along with the crowd's wisdom, were applied. It was found that the wisdom of the crowd can be used as a useful feature selection method. Finally, a predictive model was developed using advanced machine learning algorithms such as bagging, boosting, stacking, and deep learning. The partial customer churn was predicted with more than 88% accuracy by the Gradient Boosting Machine algorithm by using 5-fold cross-validation. Comparative results indicate that the proposed model performs efficiently compared to the ones applied in the previous studies.

Keywords: Partial Churn; Churn Prediction; Machine Learning; Feature Selection; Telecommunications Industry; The Wisdom of the Crowd.

1- Introduction

Churn prediction that can be defined as identifying which customers would put an end to using a company's services may consider as the most frequent predictive task within the telecommunications industry due to several benefits it could bring in for the company such as the less cost spent on retaining the existing customer in comparison to acquiring a new one [1]. In the customer churn prediction, a churn probability would be assigned to each customer based on their historical data, which leads to identifying the targeted customers for receiving marketing retention campaigns [2].

There are several reasons which make tackling customer churn problem necessary. In today's saturated and competitive market, any company should consider the fact that customers have the option to switch to other service providers [3]. In addition to more costs involved in acquiring new customers than retaining the existing ones [4]–[6], churning might also impact the reputation of a company and could lead to its brand loss [3].

The dynamic relationship between customer satisfaction, service quality, and customer loyalty or switching behavior is the topic of many studies today. For example, satisfied customers will be more accepting of price rises which will, in turn, bring greater profits [7]. In recent years customer management research has followed major categories such as Customer Lifetime Value (CLV) which

✉ **Babak Sohrabi**
bsohrabi@ut.ac.ir

leads to the conclusion that retaining customers is the most crucial work of customer management. Moreover, due to rapid progress in computing science and data mining algorithms, new directions followed by researchers consist of applying various machine learning algorithms, such as Neural Networks, to identify valuable customers and predict customers' churn rate [8].

There are two basic approaches to reduce customer churn, namely untargeted and targeted approaches. In the targeted approach, companies will identify the customers who are likely to churn and offer direct incentives to retain them, while in the untargeted approach, superior products and mass advertisement are used to increase brand loyalty and retain customers. The targeted approach can be divided into two categories, namely, reactive and proactive approaches. In the reactive approach, the company will wait till the customer decides to terminate their subscription, and at that time, the company offers some incentives. However, in the proactive approach, the company tries to identify customers who have high risk of churn in the near future so that they can be targeted with special packages or incentives to keep them from churning. The proactive approach has the advantage of lowering incentive costs. The most important point in taking the proactive approach is to classify the customers accurately since the ineffective classification of customers will only lead to the waste of financial resources for wrongly targeted customers [9].

The customer churn prediction modeling has been studied in various industries such as retailing [10]–[14], banking [11], [15]–[17], e-commerce [18]–[20], media and social networking services [21], [22], financial services [23], [24], and telecommunications [2], [3], [25]–[31]. However, the customer churn predication still is a sophisticated process that contains many decision points for analysts [32].

In recent years, due to high costs related to acquiring new customers, saturated and dynamic market, and continuous new competitive offerings, the concentration of telecom companies has shifted to customer retention strategies [2], [33]. With an increasing interest in customer churn prediction, a wide range of machine learning classifiers have been proposed in the literature to deal with the churn prediction problem [28].

A non-contractual relationship between the customers and the companies suffers from the problem once the customers change their service provider without informing about it [10]. In such a scenario, predicting potential churners by analyzing the pattern in their behavioral features becomes a significant challenge that must be handled. Moreover, in developing a prediction model, there are various feature types with different levels of importance. Another issue that needs to be tackled is gaining an insight into which features have more predictive power and impact on customer churn. Yet another challenge is how to deal with the real-world data. Since the behavioral patterns of different churners are diverse, the company should not treat all customers in the same manner, i.e., the risk or probability of churn is not alike for all churners [3]. Therefore, embedding this difference in churn definition in the form of partial and semi-partial churners, in addition to complete churners, is a crucial point that were not considered in many previous studies in this field.

While numerous research has been conducted to develop an effective classification model and identify the customers with a high risk of churn, defining churn and, more specifically, partial churn in various industries, especially the telecom industry, still needs to be completed. In other words, as mentioned before, inaccurate classification of the customers can lead to wasting financial resources, and achieving an effective model is based on how the churn is defined and the predictability of the features. In this regard, these two points, defining partial churn and feature selection, are focused on in this study.

Encouraged by the aforementioned challenges, the key objectives of this paper are (i) defining partial churn in a new and practical way so that partial and semi-partial churners would be identified and targeted by retention strategies before total defection, (ii) identifying the most effective features in terms of greater power for discriminating between churners and non-churners by using various feature selection methods and the wisdom of the crowd as a novel approach to feature selection, and (iii) developing and evaluating an efficient churn prediction model based on the actual customer data in a telecommunications company.

Hence, this research has been designed to answer the following research questions regarding customer churn prediction:

RQ1: How can customer churn, partial churn, and semi-partial churn, as various categories with different probability of churn, be defined in the telecom company?

RQ2: What are the most predictive features that affect customer churn in the telecom company based on data mining techniques?

RQ3: How can a classification model based on advanced machine learning algorithms be developed to predict the customers with various risks of churn in the telecom company?

RQ4: What are the challenges in dealing with real data in the telecom industry, and how can they be handled to develop a prediction model?

To answer these questions, a novel definition of partial churn is presented in this study. The definition is based on customers' status, determined by usage, plan, and volume features. Four classes of customers are defined by using a data mining approach and classifying customers based on their usage behavior similarity to active or total churners.

Hence, customers can be categorized with more resolution, and retention strategies would be more purposeful. Proposing a hybrid feature selection approach is another novelty of this paper. Comparing various feature selection methods and using the wisdom of the crowd can lead to a significant increase in the accuracy of the predictive model. Finally, a predictive model based on advanced machine learning algorithms is developed, which can efficiently predict partial and semi-partial churners. The model is obtained through handling different challenges such as imbalanced datasets, data cleansing, and feature selection which are tackled in the preprocessing phase of this study. Finally, while defining the four classes of customers, the proposed churn prediction model predicts the class of new customers, i.e., test dataset, with at least 88% accuracy.

The main contribution of this study lies in the partial churn definition based on data mining techniques which can find the degree of similarity between assorted customers with active ones or churners. Also, the hybrid approach in feature selection utilized in this study sheds light on the most predictive features among many features in a real dataset.

The rest of the paper is organized as follows. In Section 2, the background and related works of the customer churn prediction concerning its different aspects are given. Section 3 specifies the methodology used in this work in terms of various steps to develop a churn prediction model. The results are given in Section 4 followed by a discussion in Section 5. Conclusion and directions for further research are in Section 6.

2- Background and Related Work

The literature review indicates about a number of researches already done in the customer churn prediction area. It also highlights techniques from basic as well as advanced algorithms that were used in different industries. This section provides an overview of related works on partial churn, feature selection for the churn prediction modeling, and classification algorithms.

2-1- Partial Churn

Although researchers often define customer churn based on the nature of the organization they examine [34]; but in general, customer churn can be defined as the customers' tendency to stop doing business with one company or organization and switch to products of another company within an assumed period [35]. One of the most fundamental points that should be considered in defining customer churn is to differentiate between contractual and non-contractual businesses since each has its own model. There are relatively few studies that define churn in non-contractual settings [36] probably due to the difficulties in determining the time when customer becomes effectively inactive, when there is no contract. From a business perspective, we can say that partial churn is even more important than the complete churn [37]. The importance of identifying partial churners is that, firstly, these customers are considered to be loyal and the loss of sales related to them, even partially, can be significant. Secondly, the partial defection can lead to complete defection in the long-term. Hence, the sooner the partial churn is detected, the more valuable it is for marketing managers to prevent them from complete churn by adopting appropriate actions [10].

Based on the literature review, most previous studies were conducted in some industries such as telecommunications and retailing, due to their non-contractual settings. Most customers show a partial defection before their complete churn, which may finally lead to a complete switch [10]. Table 1 summarizes some of the most important partial churn definitions in the literature.

Table 1: Partial churn definitions

Study	***Industry***	***Churn Definition***
[10]	Retailing	Change in transaction patterns and stopping the use of a product or service
[25]	Telecommunications	Change in customers' status
[13]	Retailing	Customers who have not made a purchase within a certain period of time or in all subsequent periods have spent less than 40% of the reference period. Also, partial churn is defined as stopping the purchase of certain goods or services
[36]	Retailing	Evaluation of a set of definitions based on economic parameters
[19]	E-commerce	Change in Length, Recency, Frequency, Monetary (LRFM) pattern and purchasing behavior
[38]	Retailing	Considering the Poisson distribution for the customer purchasing pattern, the significant difference between the rate parameter in the two time periods of reference and evaluation indicates churn
Current study	Telecommunications	The amount of similarity to either active or expired

Study	*Industry*	*Churn Definition*
		customers based on the usage, plan, and volume statuses

2-2- Feature Selection

In order to get consistent, unbiased, and a set of explanatory features, the number of features should be reduced by applying feature selection methods [39]. Filtering methods are advantageous due to their high-speed calculations, but there is no guarantee that the subset of the selected features will be optimal. Chi-square test, information gain, Fisher's Score, Anova, Minimum Redundancy Maximum Relevance (MRMR), and Linear Discriminant Analysis (LDA) are among filtering feature selection techniques [5], [33], [40]–[47]. On the other hand, wrapper methods are based on scoring possible subsets of features relying on their predictive potential. Sequential Forward Selection (SFS), Sequential Backward Elimination (SBE), and Boruta algorithm are among the most common wrapper techniques though they are intensive computationally [44], [48]–[50].

Finally, embedded algorithms have their own built-in feature selection methods. These are similar to wrapper methods in terms of dependencies on the classifier. However, embedded methods are computationally faster than wrapper ones. LASSO regression, Random Forest, and Recursive Feature Elimination-Support Vector Machine (RFE-SVM) are some examples of this type of techniques [29], [51]–[54]. Moreover, using the two-phase feature selection method by researchers is reviewed by Jain et al. (2020). The two-phase feature selection method can consist of a combination of experts-advised subset and Markov blanket discovery.

The wisdom of the crowd is an emerging concept that is used for feature selection in this research. The wisdom of the crowd concept refers to the fact that aggregated judgments by individuals are often more accurate than that of the smartest person in the crowd [55]. In this regard, crowd refers to a group of individuals with various perspectives, and wisdom contains purposeful acts, reasonable thought, and effectively dealing with an environment [55]. There are some factors that influence crowd performance, such as diversity, independence, and decentralization which were introduced by Surowiecki (2005) [56].

Possession of varying degrees of knowledge and insight explains the diversity. Independence means no influence on each member's decisions by other members, and decentralization relates to the concept that power does not fully reside in one central location [56]. Moreover, Hong et al. (2020) proposed that crowd size is a moderator between crowd characteristics and crowd performance [55]. In recent years, the wisdom of the crowd has been found to have many applications in some fields, such as the stock prediction domain and financial trade [57]–[59].

The predictive features can be used by the marketing department to implement retention strategies. While data mining techniques are used in this regard, it should be noted that using predictive features to predict and then change the churn intention can find its roots in the Theory of Planned Behaviors (TPB) introduced by Ajzen (1985). This theory, which expands the theory of reasoned action, states that many factors influence the stability of behavioral intentions. Investigating these factors sheds light on how it may be possible to prevent changes in intentions. Hence, a measure of intention is expected to allow precise prediction of intentional behavior unless the intention changes after it is evaluated but before the behavior is observed. This intention is, in turn, a function of two factors which are the attitude toward trying and the subjective norm with regard to trying [60]. The findings of this study are discussed according to this theory in the Discussion section.

2-3- Classification Algorithms

The customer churn prediction can be considered as a management science problem for which we can usually adopt a data mining approach. "Data mining is the process of discovering meaningful new correlations, patterns, and trends sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" [61]. Since the main aim of churn prediction is to classify the type of customers (non-churner, partial churner, churner), a variety of supervised machine learning classifiers have been proposed in the literature. In this subsection, we provide an overview of some of the previous studies on churn prediction. Table 2 provides a review of the classification algorithms used in some of customer churn prediction studies in the telecommunications sector.

Table 2: Literature review on classification algorithm to predict customer churn in the telecom sector

Study	*Classification Algorithm*
[30]	Ensemble
[25]	Logistic Regression (LR)
[62]	Neural Network (NN)
[63]	NN
[64]	Support Vector Machine (SVM), NN, Decision Tree (DT)
[65]	SVM
[26]	Genetic Programming (GP) & Self-Organizing Maps
[31]	SVM, DT, Back Propagation Network (BPN)
[66]	Ordered Fuzzy Rule Induction
[67]	Fast Fuzzy C-Means & GP
[68]	Convolutional Neural Network (CNN)
[69]	Recurrent Neural Network (RNN)
[70]	CNN

Study	*Classification Algorithm*
[71]	Fuzzy classifiers
[2]	LR, DT, NN, Random Forest (RF), SVM
[72]	Long Short-Term Memory (LSTM), Gradient Boosting Tree (GBT), RF, SVM
[73]	Agent-Based Modeling and Simulation (ABMS)
[28]	RNN, LR, RF, LSTM, Probabilistic Neural Network (PNN)
[74]	DT, Naïve Bayes (NB), Rule Induction
[29]	CNN
[3]	LR, Multi-Layer Perceptron (MLP), NB, Bagging and Random Tree, AdaBoostM1, Attribute Selected, Decision Stump, RF, J48, Random Tree, Lazy learning methods (Locally Weighted Learning, lazy k-nearest neighbor)
[27]	NB, Generalized Linear Model (GLM), LR, Deep Learning (DL), DT, RF, GBT
[75]	Ensemble Classifiers (K-Nearest Neighbors, NB, RF, LR) and (Cat Boost, Gradient Boosting, Extreme Gradient Boosting)
[76]	NN, SVM, NB, RF, Adam DL

As Table 2 illustrates, advanced data mining algorithms such as bagging, boosting, and stacking are exploited in recent years due to their better performance compared to the basic ones. Boosting is a technique that improves the accuracy of the classification model by applying the functions in a series iteratively and then combining the result of each function with weighting in order to maximize the total accuracy of the prediction. Friedman (2002), constructed additive regression models by iteratively fitting a base learner or weak classifier to current pseudo-residuals with least squares. Indeed, these pseudo-residuals are the gradient of the loss function, which should be minimized [11].

Implementing Rough Set Theory (RST) which is a technique for dealing with uncertainty and for identifying cause-effect relationships in databases, has also been used for customer churn classification modeling [9], [77]. In [8], knowledge of survival analysis was applied in customer management, while Devriendt et al. (2021) used uplift modeling as prescriptive analytics, which aims a reduction in the likelihood of churn when customers are targeted with the retention campaign based on the net difference in customer behavior [78]. Using unstructured data such as text is another approach that has been used in recent years. For example, [79] utilized a call center dataset to predict customer churn risks and generate meaningful insights using interpretable machine learning with personas and customer segments, and [80] used website data holding customer complaints in this regard.

3- Data and Methodology

This section discusses the main building blocks of the proposed customer churn prediction model, which leads to identify partial churners in a telecommunications company. Since the problem is analyzed through the data mining approach, the methodology in the present study is the data science methodology introduced by IBM [81], which includes ten steps, which is an extension to the CRoss Industry Standard Process for Data Mining (CRISP-DM) method [82]. While using the data mining approach, the findings of the study are discussed through the lens of TPB in the Discussion section.

Figure 1 shows the overall framework of the data analysis. As it can be observed, this framework consists of four main steps. In the first step, a binary classifier is used to model Active versus Expired customers. Expired customers are the ones who were not using the services for more than eight months. In the second step, a partial churn definition can be derived based on the amount of similarity between any customer with either Active or Expired customers. In other words, the "statuses" of the customers are linked to the partial churn. Then, based on the partial churn definition, a multiclass problem is set up to predict whether customers are in the "partial churn" status. Finally, using Partial_Churn and Semi_Partial_Churn labels and prediction model, the most predictive features can be identified.

3-1- Data Definition

Based on the Exploratory Data Analysis (EDA), our variables and their statistical features were obtained through data summarization and visualization. Three datasets were used in this research, which contains customer data on three specific dates, i.e., August and November of 2018 and February of 2019. Table 3 presents the number of records and features of each dataset before and after data cleansing. Three dates were selected so that the customers' status can be tracked at least in a 6-month period.

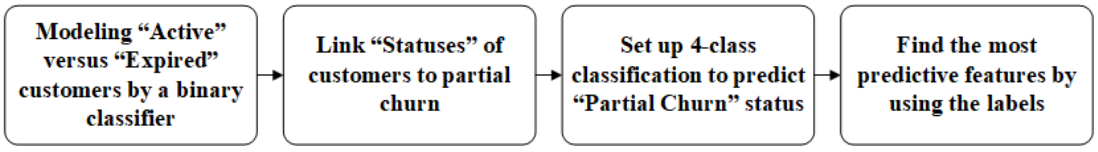


Fig. 1 Overall framework of the data analysis

Table 3: Datasets used in this research

No. of dataset	*Date*	*No. of records before data cleansing*	*No. of features before data cleansing*	*No. of records after data cleansing*	*No. of features after data cleansing*
1	August 2018	213166	110	208585	87
2	November 2018	228941	110	223970	87
3	February 2019	254994	110	249511	87

The resulted independent features consist of 31 nominal variables, nine dates, which then changed to numeric variables, and 47 numeric variables. The dependent variable, which is the STATUS of the customer, i.e., a label indicates the churned or non-churned customers, is a logical variable. In the second phase of data cleansing, some of the nominal features with too many unique values were eliminated by investigating the statistical characteristics of features. Moreover, some dates from which the Status can be derived directly were removed from the dataset, which led to 72 features. The features can be categorized into six main groups: plan-related, customer life-related, usage-related, revenue-related, sale-related, and geographic-related features- the number of each and some examples are provided in Table 4.

Table 4: Categories of features in real-world dataset used to predict customer churn

Category	*Number of features*	*Example*
Plan-related features	10	Volume, Speed, Duration
Customer Life-related features	5	Nominal Life, Real Life
Usage-related features	14	Last_Month_Usage, Last_3_Month_Usage, Burned_Traffic
Revenue-related features	19	Total_Revenue, Monthly_Revenue, Usage_Revenue
Sale-related features	9	Sale_Type, Model_Name
Geographic-related features	15	BTS_City

Training and test datasets are derived from the three datasets mentioned above. In this research, two training and two test datasets were used. The first training and test datasets were used for partial churn definition in a binary classification model, and the second ones were used in 4-class classification. The first training dataset consists of customer records whose status had not been changed during the 6-month period, i.e., from August 2018 to February 2019. In other words, these customers were labeled as Active customers during the 6-months period, which means they were using the services during this period, or they labeled as Expired, which means they were not using the services for more than eight months and maintained this status for at least six months. The second training dataset, which includes the data set of November 2018 with new labels derived from the previous step, was used for training. The second test dataset consists of the customers who were added in the three-month period and were considered new records selected for testing. In both phases, 5-fold cross-validation was applied to training datasets. Table 5 presents training and test datasets for each of the phases.

Table 5: Training and test datasets for each of the phases

phase	*No. of records of the Training dataset*	*No. of records of the Test dataset*
Partial churn definition	96934	152434
	all records from August 2018 which maintained the same status during 6 months	all records from the three dates which are not among training dataset
4-class classification	173706	43280
	all records from November 2018 which were assigned one of the four labels based on Table 6	all new records from February 2019 that were added to the dataset in the last three months

3-2- Partial Churn Definition

In this paper, different customer statuses are grouped into four: active, semi-partial churn, partial churn, and expired. In non-contractual businesses, as in our case, the churn event cannot be determined explicitly since the customer can finish the relationship with the service provider without prior notice. Therefore, we have to clearly define a churn criterion based on which partial churners can be identified before their complete defection. In other words, the churn should be predicted in advance in order to have time for customer retention strategies.

Customers with semi-partial and partial churn status are more likely to churn than customers with active status, hence; these are the customers marketing department can focus on. Optimization of the fraction of customers that should be targeted by the retention campaign can lead to maximize the profit [39]. In our case study, which is an internet service provider company, 14 statuses are defined

for customers based on criteria: (i) usage behavior, (ii) having a plan (i.e., the last main plan purchased by customer), and (iii) having volume. Table 6 summarizes different groups of customers based on their statuses.

In order to define the partial churn, a binary classification model was built based on the training dataset, i.e., a dataset containing the records of active and expired customers who had not changed their status for at least a 6-month period. Therefore, Partial_Churn can be defined as a status that customers have equally shows similarity to either Active customers or Expired customers while Semi_Partial_Churn is assigned to customers with more similarity to Active customers, but they still tend to churn more than the Active ones. As explained, the definition approach is based on the similarity of customers' features with either active customers or total churners.

Table 6: Customers' status based on usage behavior, plan, and volume

Criteria	***Status***
Usage behavior	Active
	without usage 10 to 20 days (10TO20)
	without usage 20 to 30 days (20TO30)
	without usage for more than 30 days (gt30)
Plan	Without plan 0 to 1 month (0TO1)
	without plan 1 to 2 months(1TO2)
	without plan 2 to 3 months (2TO3)
	without plan 3 to 5 months (3TO5)
	without 5 to 7 months (5TO7)
	without plan 7 to 8 months (7TO8)
	without plan more than 8 months (Expired)
Volume	without volume 3 to 7 days (3TO7)
	without volume 7 to 20 days (7TO20)
	without volume more than 20 days (gt20)

Table 7 illustrates the labels which are derived after several moves back and forth between preprocessing and modeling steps in order to obtain a model with more than 99% binary classification accuracy. In Table 7, two new classes were defined: Semi_Partial_Churn and Partial_Churn. Customers labeled as Semi_Partial_Chrun are the ones whose features are more similar to the Active customers, while the customers labeled as Partial_Churn can be either Active or Expired. In other words, the probability of becoming a churner is more with the ones who labeled as Partial_Churn. Moreover, semi_partial churners tend to become churners more than the active ones. Therefore, these labels can categorize the customers with more resolution. The statuses provided in Table 7 are the same as those in Table 6 except for "gt20". This status was removed since it could not be in any of the classes mentioned above.

Table 7: Partial churn definition based on customers' status

Status	***Label***
Active 0TO1 3TO7 7TO20 10TO20	ACTIVE (classified as active with more than 95% probability)
1TO2 gt30	Semi_Partial_Churn (classified as active with 80 to 95 percent probability)
2TO3	Partial_Churn (classified as active (or expired) with nearly 50% probability
3TO5 5TO7 7TO8 Expired	EXPIRED (classified as expired with more than 95% probability)

3-3- Preprocessing

Data preparation methods often refer to the transformation of features into variables that support a particular machine learning algorithm [2]. In the process of knowledge discovery and developing a practical model to predict churn, the data preparation and the feature selection are essential steps. In other words, selecting the most valuable features can reduce over fitting as well as the complexity of the model while improving the interpretability for users [83]. The preprocessing step in this study consists of initial data cleansing, data transformation, normalization, One-Hot encoding, feature selection, and the handling of imbalanced dataset problem. Figure 2 depicts the main tasks of the data preprocessing used in this study.

Data cleansing includes finding the test records, which do not relate to customers, removing the inaccurate data caused by system bugs and data entry mistakes, replacing the inconsistent and missing values, and eliminating irrelevant and zero variance features, all done using experts' knowledge about the data. For example, most of the missing values are replaced by using other databases of the company. As a preprocessing task, the z-score normalization is applied on the dataset to make them appropriate for correlation analysis, i.e., all numerical features are standardized by removing the mean and scaling the values to unit variance. For nominal variables, One-Hot encoding is applied so that each of these variables can be represented with as many binary variables as the number of unique values of that variable. It is worth to note that most of the preparations such as reducing skewness of features distribution, normalization, and a One-Hot encoding are applied to the data for the correlation analysis purpose and not needed when the H2O package is used for modeling since H2O AutoML-function applies all the required preprocessing in accordance with each algorithm before modeling [84].

In many data mining applications such as the churn prediction, rare cases or minority classes, are of main

interest [11]. It may cause classification methods to experience challenges in identifying the churners, which leads to the poor classification power. The best performance of classification techniques can be achieved when the class distribution is approximately even [39]. Considerable works have been done on handling the imbalanced dataset problem [47], [85]–[92] which categorize the methods into oversampling, synthetic data generation which is a type of oversampling, under-sampling, and cost-sensitive learning ones. The Synthetic Minority Oversampling Technique (SMOTE) is one of the synthetic oversampling techniques used in the literature [93], which is also applied in the current study due to its efficiency. Indeed, oversampling methods, on the one hand, focus on improving classifiers' performance on the minority class samples; on the other, boosting methods focus on the hard-to-learn majority class samples. Therefore, as proposed by Barua et al. (2012), boosting and oversampling together can provide an efficient option for learning the imbalanced data [85]. This approach is also useful in this study since we apply a boosting algorithm, i.e., gradient boosting machine, together with oversampling for handling the imbalanced dataset challenge. Therefore, oversampling based on SMOTE technique is adopted in this study to handle imbalanced dataset.

3-3-1- Feature Selection

In this study, feature selection has played a crucial role in achieving the results. So, as a part of the research, the result of some common feature selection methods from various categories, i.e., filtering, wrapper, and embedded methods, are compared to each other. Therefore, the correlation analysis by using Pearson correlation coefficient, Boruta, LASSO, Ridge, and Random Forest methods are used. Moreover, as mentioned before, the concept of the wisdom of the crowd is applied. The individuals who have expert knowledge of the features are from various departments, i.e., marketing, data management, and Customer Relationship Management (CRM). The important features selected by these individuals are then compared to the features selected by algorithms.

As shown in Table 8, two different approaches are used for feature selection: the algorithm-based approach and the wisdom-of-the-crowd-based approach. The features selected by the wisdom of the crowd can help us in two ways. First, in case that there is a difference between the results of various algorithms, it would be observed that using the crowd's wisdom can be helpful. Secondly, most of the features considered as important by the wisdom of the crowd are in the final list of important features with the best performance of the model in terms of accuracy. So, it can be used as a useful method that complements the feature selection algorithms' results.

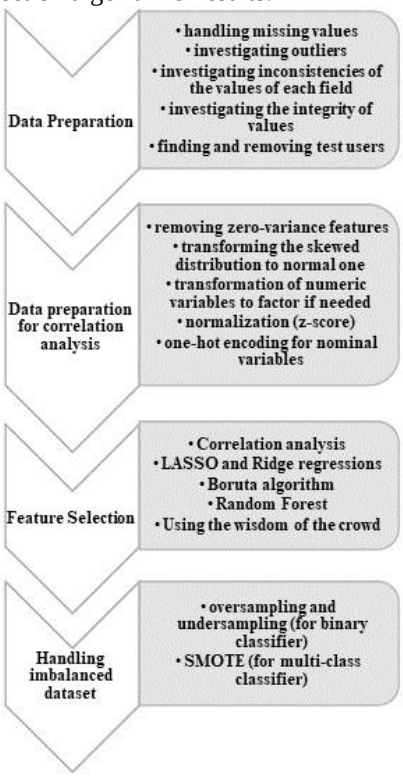


Fig. 2 Preprocessing step for developing a customer churn prediction model

Table 8 presents that some features are considered important by all methods, for example, PLAN_GROUP_NAME, some other features are considered important by some of the methods while their importance weight is low by the others, like REAL_LIFE or BURNED_TRAFFIC. Some features are not considered predictive by the wisdom of the crowd, while feature selection methods considered them important and, therefore, predictive ones such as INSTALLATION_AGENT_NAME. Based on the results by various methods, a hybrid approach was adopted. Based on this hybrid approach, different subsets of features were used to train the prediction model to find the most effective subset, leading to the highest prediction accuracy. The difference between these subsets is based on the selection criterion.

Table 8: Applying various feature selection methods

Feature	*LASSO (absolute coefficients)*	*Ridge (absolute coefficients)*	*RF (weights)*	*Boruta (confirmed/rejected and mean importance)*	*Crowd's wisdom*
CUSTOMER_TYPE	0.8272	0.2385	0.00	Rejected/ -0.2852793	✓
PLAN_GROUP_NAME	0.2057	0.0879	0.40	Confirmed/ 5.9657923	✓
LAST_MONTH_USAGE	0.0002	1.89529E-05	0.47	Confirmed/ 25.9943734	✓
LAST_MONTH_EXTRA_TRAFFIC_REVENUE	0	9.0333E-07	8.87E-05	Rejected/ 1.8492914	✓
BURNED_TRAFFIC	6.72E-06	5.78202E-07	0.00	Confirmed/ 7.5701373	✓
INSTALLATION_AGENT_NAME	0.3188	0.1434	0.07	Confirmed/ 3.9381313	-
REAL_LIFE	0	0.0045	0.72	Confirmed/ 11.5165463	✓

3-4- Gradient Boosting Machine (GBM)

Advanced data mining algorithms are those implemented in the H2O library. The H2O is an open-source advanced machine learning tool that helps us create high-performance models. The algorithms implemented in the H2O library, which have been used for the classification and the churn prediction, include Gradient Boosting Machine (GBM), Stacked Ensemble, Deep Learning (DL) as a fully connected multi-layered artificial neural network, Distributed Random Forest (DRF), Extremely Randomized Trees (XRT), Generalized Linear Model (GLM), and XGBoost [84]. After applying AutoML function on our dataset, GBM outperforms the other algorithms in terms of prediction accuracy, precision vs. recall, Area Under Curve (AUC), and logloss. In the H2O package, the GBM algorithm provided by Ellis et al. (2008) is implemented [94]. This algorithm uses distributed trees in such a way that a tree node is assigned to each row. An in-memory map-reduce task calculates statistical parameters such as the Least Mean Square Error (MSE) and uses it to make an algorithm-based decision [95].

3-5- Evaluation Measures

In this study, in order to evaluate the performance of the classification model, we have used different performance metrics such as accuracy, precision, and recall, which are derived from the confusion matrix and logloss. Area Under Curve (AUC), i.e., the area under the Receiver Operating Characteristic Curve (ROC), is another useful metric for a binary classification since it is not sensitive to imbalanced classes [27]. AUC can be defined as "the estimated probability that a randomly chosen churner has a higher posterior churn probability than a randomly selected non-churner" [2]. Therefore, we have used this metric in the partial churn definition phase.

All the preprocessing, classification, and prediction implementations were done using R language version 3.6.1 mainly with tidyverse [96], recipes [97], and H2O [84] packages. For feature selection, packages such as glmnet [98], Boruta [99], and randomForest [100] were used. The experiment was performed using the classification algorithms of the H2O package in R.

4- Results

This section explores the results obtained through the experimental analysis. Further, the results are used to evaluate the impacts of feature selection, oversampling, and classification algorithms. Figures 3 represent the visualization of comparison between various advanced machine learning algorithms in terms of ROC.

Since GBM outperforms other algorithms with respect to ROC plots, we could select GBM as the best classifier on predicting the correct customer churn for our dataset. We adopted an oversampling technique during the training to handle the imbalanced dataset problem. This led to an increase in the churn rate from 20% to nearly 50%.

Therefore, an approximately uniform distribution of the target variable affects the churn prediction performance with an improvement in the accuracy measured by the ratio of true positive/negative to the total number of samples. Table 9 presents the overall performance of the churn prediction model, which classifies the new customers into four classes with over 88% accuracy for each, and more than 97% overall accuracy rate.

Table 9: Performance of final multi-class GBM classifier

Label (target variable)	*Prediction accuracy*	*Overall accuracy*
ACTIVE	0.9908	97.6%
Semi_Partial_Churn	0.9006	
Partial_Churn	0.8843	
EXPIRED	0.9596	

The features which were selected through the mechanism explained in Section 3.3.1 are provided in Table 10 after

the interpretation step in order to filter them. The results indicate that the 31-numeric features are the best subset of features in terms of accuracy, AUC, and other metrics. 27 out of these 31, which means 87% of the features are amongst the features selected by the wisdom of the crowd. In our case study, the customers classified as Semi_Partial_Churn and Partial_Churn were 10% of the total in test dataset, which equals nearly 4,000 customers who should be targeted for marketing campaigns.

4-1- Interpretation of the Results

To understand the contribution of each feature to the prediction result, we use the Locally Interpretable Model-agnostic Explanations (LIME) package [101]. This is implemented in R language to find the k-most important features which lead to the obtained result for each customer. Figure 4 depicts a sample of the LIME diagram, a local interpreter, and specifies the importance of features in different ranges. In this Section, we want to infer some insights by comparing each feature selection method's results to the results derived by the LIME interpreter. In Figure 4, the blue color indicates the feature positively affecting the category it belongs to, i.e., Active, Partial Churn, or Expired, while the red color shows the contrast. The color intensity illustrates the importance of that feature. Each feature's weight obtained by the LIME interpreter is based on that feature's role in the prediction model. Table 9 shows some of the most predictive features after the interpretation of the results.

The results indicate that among nominal features, the type of the last plan purchased by the customer has a high weight in prediction. Another predictive feature is the usage-related revenue, i.e., dividing the total revenue of the purchased plans to the usage of the customer. The total usage, the number of transactions, and the remaining volume from the last plan are also among the most predictive features. The length of the relationship between customer and company is also found to be important.

Table 10: The most predictive features for customer churn prediction

Category	***Feature***
Revenue	Last month usage-related revenue
Plan	Duration of the last plan
Usage	Remained volume of the last plan
Revenue	Last year average total revenue
Plan	Volume of the last plan
Customer Life	Life of the customer (length of the relationship with the company)
Usage	Burned traffic of the last plan
Usage	Last month usage
Plan	Cluster of the plan

5- Discussion

This section highlights the contributions made to the existing knowledge by comparing the results obtained in Section 4 to some of the previous studies, followed by the current research findings. Finally, we conclude keeping in mind the limitation to the study.

5-1- Contribution to the Existing Knowledge

As described in Section 3.2, we have defined the partial churn in a new way, which was not used by any of the previous studies, to the best of our knowledge. This definition not only exploits the status of customers based on the criteria that can be measured easily in the telecommunications industry, especially internet service provider businesses, but also by defining four classes of customers narrows the targeted customers in accordance with their risk of churning. Hence, retention campaigns would be more purposeful.

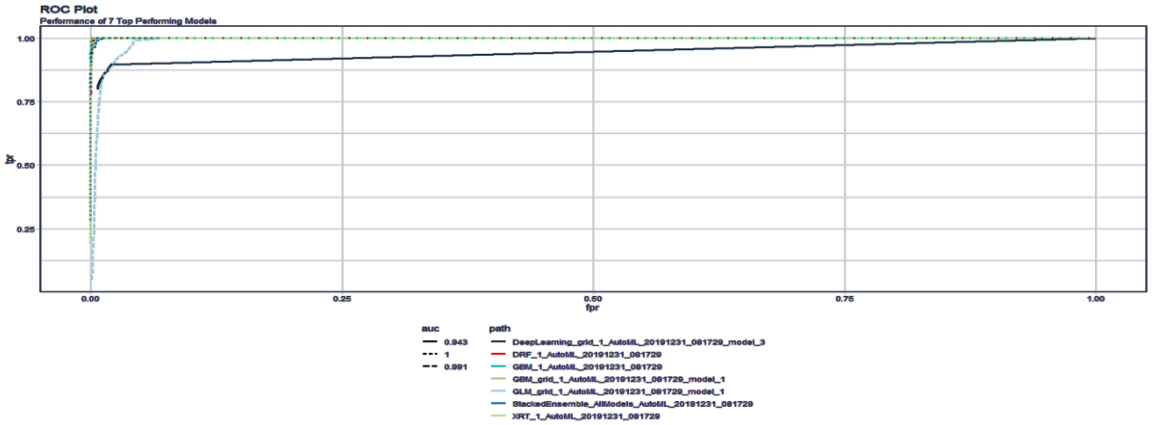


Fig. 3 Comparison between advanced classification algorithms in terms of AUC

As other studies [31] proposed, we too confirmed that boosting can improve the performance of the classifier. Moreover, gradient boosted regression trees can have a better performance compared to other advanced data mining algorithms. This is in line with the results provided by [27], and, therefore, is recommended. Moreover, we also found the features related to recency, frequency, and monetary to be among the best predictors to distinguish between different classes of customers which is in line with other studies [10], [12], [19]. Other features such as customer usage-related ones and the length of relationship are important as well. A complete list of the most important features is presented in Table 10.

The proposed feature selection approach is another important step used to achieve high accuracy and an efficient prediction model. For the feature selection approach, we used at least one method from each category of techniques in order to compare the results and select the features, which are recognized as important by most methods. Additionally, we applied the interpretation of the results and the wisdom of the crowd in the feature selection process. The results showed that the aggregation of this wise crowd's opinions can be used as a complement to the results of feature selection algorithms. Especially in the cases that the results of the algorithms contradict each other, we can rely on the wisdom of the crowd. It can also be observed that 87% of the most important features that form the best subset of features are the same as the features selected by the wisdom of the crowd. So, using the wisdom of the crowd can contribute to the feature selection procedure in two ways. First, if there is any contradiction between the results of various algorithms, it would be observed that using the crowd's wisdom can be helpful. For example, BURNED_TRAFFIC and REAL_LIFE were not selected as important features by LASSO and Ridge algorithms while confirmed as important ones by Boruta and also by the wisdom of the crowd. It can be observed that these features are among the most predictive features in Table 10, which is the final list of features. Secondly, most of the features considered as important by the wisdom of the crowd are in the final list of important features with the best performance of the model in terms of accuracy. Therefore, it can be used as a useful method that complements the feature selection algorithms' results.

Table 11 provides a performance comparison of the present study with some of the other studies in terms of overall accuracy. It can be observed from comparative results that our prediction model performs efficiently as compared to the previously used techniques. The last but not the least, since our experiment was based on the real-world dataset, we faced many challenges to tackle.

As a matter of fact, the results can be considered as applicable to the business. As mentioned, according to the Theory of Planned Behavior, behavioral attitude, subjective norms, and perceived behavioral control influence intention toward the behavior. This cognitive model has become applicable in many research areas, such as predicting loyalty intention. Based on various studies, it was found that behavior attitude is the strongest predictor of intention, while some authors have resembled switching costs as the perceived behavioral control of the TPB because of its ability to predict customer loyalty [102].

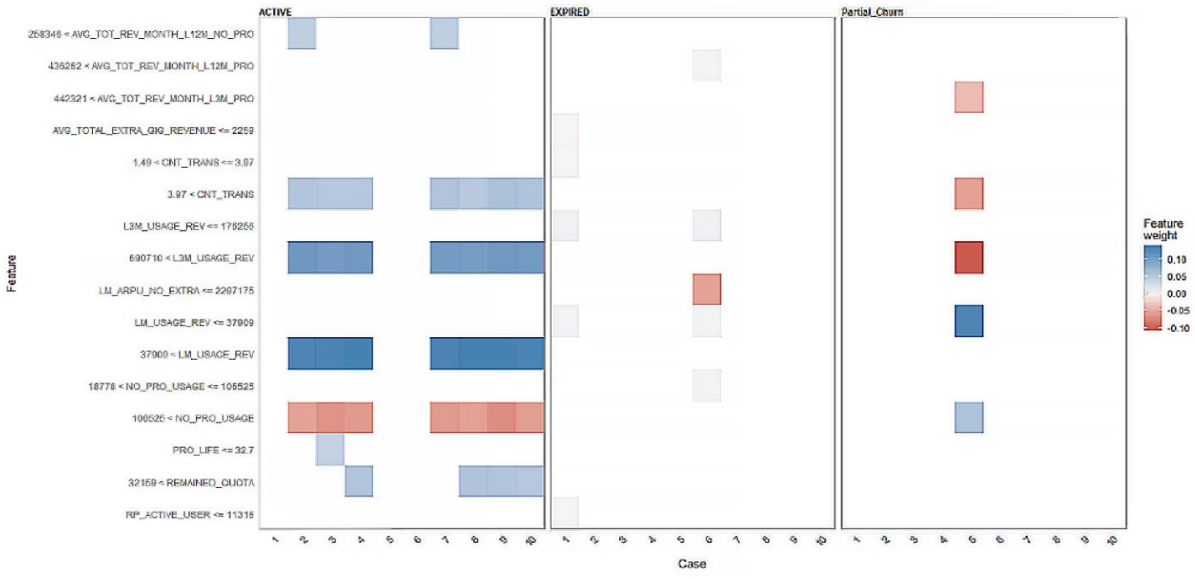


Fig. 4 Interpretation of prediction results with LIME

In recent years, due to the increasing availability of data on customer activities, more complex metrics have been developed to describe customer behavior and how behavioral attributes can be linked to customer retention and company performance [103]. These behavioral attitudes influence the churn intention and can be used as early signs of churn, which can be handled by different retention strategies that affect behavioral attitudes. Based on the findings of this study, Table 10 indicates the most predictive features extracted by the hybrid feature selection approach and interpretation of the prediction model. Among these features, 3 of them, such as last month usage, burned traffic, and remained volume of the last plan, directly relate to customer usage behavior, while the features in plan and revenue categories indirectly indicate this behavior. In other words, by investigating these signs in customer usage behavior, the service provider can offer personalized plans and incentives to customers to influence their attitudes and retain them.

Table 11: Comparison of predictive performance of proposed and previous approaches

Model	***Accuracy***					
	Present study	[31]	[19]	[27]	[3]	[76]
NB	-	-	-	0.7	0.48	0.98
GLM	0.838	-	-	0.8	-	
LR	-	-	-	0.8	0.71	
DL	0.89	-	-	0.7	-	
RF (DRF)	0.941	-	-	0.7	0.89	0.99
Stacked Ensemble/DT Ensemble	0.974	-	0.97	-	-	
GBT (GBM)	0.976	-	-	0.8	-	
Bagging + Random Tree	-	-	-	-	0.89	
BPN/ANN	-	0.95	0.935	-	-	0.974
SVM-Radial Basis Function (RBF)	-	0.96	-	-	-	0.976
SVM-POLY (polynomial kernel)	-	0.968	-	-	-	
DT/DT-C5.0	-	0.95	0.94	0.7	-	
J48	-	-	-	-	0.88	
Adam						0.98

5-2- Implications and Limitations

The results of the present study have some implications for the practice in the churn prediction area of the telecommunications company, such as to improve satisfaction by targeting customers, which are identified as partial or semi-partial churners as well as managing customer's expectations by specifying effective features on customer churn by using proposed feature selection approach through combining various feature selection algorithms and complement it by the wisdom of the crowd. Like others, this study has shortcomings and limitations as well. For that matter, all experimental evaluations are based on a specific real-world dataset and confined to the telecom industry so that not only the results can be verified by other studies but it can also be applied in practice. It must be noted that some of the very important sources of data, such as those on marketing campaigns and call centers, are not used in the process of churn prediction because of their incompleteness.

6- Conclusion and Future Work

The churn prediction in the present telecom market is a compelling issue of the CRM [3] which can be conducted by identifying likely churn customers and providing competitive offers to them. In this article, we first dealt with this problem and its different related aspects by using the real-world data extracted from various sources of a telecommunications company. Then, we proposed that partial churners, i.e., potential ones with medium to high risk of churn, could be identified prior to their decision to a complete churn by analyzing the patterns e.g., customers' usage, revenue-, plan-related features. We used state-of-the-art classification techniques for the customer churn prediction problem for a real-world dataset of an internet service provider company. It is clear from the comparative results that the gradient boosting machine performed better than other classification algorithms in predicting the customer churn. Further, this work sheds some light on the features that should be considered as more important, and it is observed that customer's usage- and plan-related features have more importance and predictive power than other types. While investigating the impact of oversampling technique SMOTE on the performance of the prediction model, the results of the current study suggested that the classifiers could achieve a significantly improved performance applying an oversampling method which also supported the findings of some previous studies [27], [28]. Consequently, this study is unique when it comes to the partial churn definition and its feature selection approach, which uses the feature selection algorithms complemented by the wisdom of the crowd, leading to the high accuracy of the prediction model.

Still some areas need further study. In any future research, we would like to use other feature types such as marketing data and other advanced machine learning algorithms such as CNN and RNN suggested by [29] and [28] respectively.

References

[1] S. Mitrović, B. Baesens, W. Lemahieu, and J. De Weerdt, "On the operational efficiency of different feature types for telco Churn prediction," Eur. J. Oper. Res., vol. 267, no. 3, pp. 1141–1155, 2018.

[2] K. Coussement, S. Lessmann, and G. Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry," Decis. Support Syst., vol. 95, pp. 27–36, 2017.

[3] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," IEEE Access, vol. 7, pp. 60134–60149, 2019.

[4] J. Dyche, The CRM handbook: A business guide to customer relationship management. Addison-Wesley Professional, 2002.

[5] A. Idris and A. Khan, "Customer churn prediction for telecommunication: Employing various various features selection techniques and tree based ensemble classifiers," in 2012 15th International Multitopic Conference (INMIC), 2012, pp. 23–27.

[6] W. Verbeke, D. Martens, C. Mues, and B. Baesens, "Building comprehensible customer churn prediction models with advanced rule induction techniques," Expert Syst. Appl., vol. 38, no. 3, pp. 2354–2364, 2011.

[7] L. Geiler, S. Affeldt, and M. Nadif, "An effective strategy for churn prediction and customer profiling," Data Knowl. Eng., vol. 142, p. 102100, 2022.

[8] Y. Chen, L. Zhang, Y. Zhao, and B. Xu, "Implementation of penalized survival models in churn prediction of vehicle insurance," J. Bus. Res., vol. 153, pp. 162–171, 2022.

[9] M. Makhtar, S. Nafis, M. A. Mohamed, M. K. Awang, M. N. A. Rahman, and M. M. Deris, "Churn classification model for local telecommunication company based on rough set theory," J. Fundam. Appl. Sci., vol. 9, no. 6S, pp. 854–868, 2017.

[10] W. Buckinx and D. Van den Poel, "Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting," Eur. J. Oper. Res., vol. 164, no. 1, pp. 252–268, 2005.

[11] J. Burez and D. Van den Poel, "Handling class imbalance in customer churn prediction," Expert Syst. Appl., vol. 36, no. 3, pp. 4626–4636, 2009.

[12] A. Dingli, V. Marmara, and N. S. Fournier, "Comparison of Deep Learning Algorithms to Predict Customer Churn within a Local Retail Industry," Int. J. Mach. Learn. Comput., vol. 7, no. 5, 2017.

[13] V. L. Miguéis, D. Van den Poel, A. S. Camanho, and J. F. e Cunha, "Modeling partial customer churn: On the value of first product-category purchase sequences," Expert Syst. Appl., vol. 39, no. 12, pp. 11250–11256, 2012.

[14] V. L. Miguéis, A. Camanho, and J. F. e Cunha, "Customer attrition in retailing: an application of multivariate adaptive regression splines," Expert Syst. Appl., vol. 40, no. 16, pp. 6225–6232, 2013.

[15] Y. Chen, Y. R. Gel, V. Lyubchich, and T. Winship, "Deep ensemble classifiers and peer effects analysis for churn forecasting in retail banking," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2018, pp. 373–385.

[16] N. Glady, B. Baesens, and C. Croux, "Modeling churn using customer lifetime value," Eur. J. Oper. Res., vol. 197, no. 1, pp. 402–411, 2009.

[17] Y. Xie, X. Li, E. W. T. Ngai, and W. Ying, "Customer churn prediction using improved balanced random forests," Expert Syst. Appl., vol. 36, no. 3, pp. 5445–5449, 2009.

[18] N. Gordini and V. Veglio, "Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry," Ind. Mark. Manag., vol. 62, pp. 100–107, 2017.

[19] A. D. Rachid, A. Abdellah, B. Belaid, and L. Rachid, "Clustering Prediction Techniques in Defining and Predicting Customers Defection: The Case of E-Commerce Context," Int. J. Electr. Comput. Eng., vol. 8, no. 4, p. 2367, 2018.

[20] A. Tamaddoni, S. Stakhovych, and M. Ewing, "The impact of personalised incentives on the profitability of customer retention campaigns," J. Mark. Manag., vol. 33, no. 5–6, pp. 327–347, 2017.

[21] I. Adaji and J. Vassileva, "Predicting churn of expert respondents in social networks using data mining techniques: a case study of stack overflow," in 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 182–189.

[22] K. Coussement and D. Van den Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," Expert Syst. Appl., vol. 34, no. 1, pp. 313–327, 2008.

[23] D. F. Benoit and D. Van den Poel, "Improving customer retention in financial services using kinship network information," Expert Syst. Appl., vol. 39, no. 13, pp. 11435–11442, 2012.

[24] M. Á. de la Llave, F. A. López, and A. Angulo, "The impact of geographical factors on churn prediction: an application to an insurance company in Madrid's urban area," Scand. Actuar. J., vol. 2019, no. 3, pp. 188–203, 2019.

[25] J.-H. Ahn, S.-P. Han, and Y.-S. Lee, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry," Telecomm. Policy, vol. 30, no. 10–11, pp. 552–568, 2006.

[26] H. Faris, B. Al-Shboul, and N. Ghatasheh, "A genetic programming based framework for churn prediction in telecommunication industry," in International Conference on Computational Collective Intelligence, 2014, pp. 353–362.

[27] A. S. Halibas, A. C. Matthew, I. G. Pillai, J. H. Reazol, E. G. Delvo, and L. B. Reazol, "Determining the Intervening Effects of Exploratory Data Analysis and Feature Engineering in Telecoms Customer Churn Modelling," in 2019 4th MEC International Conference on Big Data and Smart City (ICBDSC), 2019, pp. 1–7.

[28] J. Hu et al., "pRNN: A recurrent neural network based approach for customer churn prediction in telecommunication sector," in 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 4081–4085.

[29] M. Karanovic, M. Popovac, S. Sladojevic, M. Arsenovic, and D. Stefanovic, "Telecommunication Services Churn Prediction-Deep Learning Approach," in 2018 26th Telecommunications Forum (TELFOR), 2018, pp. 420–425.

[30] A. Lemmens and C. Croux, "Bagging and boosting classification trees to predict churn," *J. Mark. Res.*, vol. 43, no. 2, pp. 276–286, 2006.
[31] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," Simul. Model. Pract. Theory, vol. 55, pp. 1–9, 2015.
[32] E. Lima, C. Mues, and B. Baesens, "Monitoring and backtesting churn models," Expert Syst. Appl., vol. 38, no. 1, pp. 975–982, 2011.
[33] A. Amin et al., "Customer churn prediction in the telecommunication sector using a rough set approach," Neurocomputing, vol. 237, pp. 242–254, 2017.
[34] A. Hiziroglu and O. F. Seymen, "Modelling Customer Churn Using Segmentation and Data Mining.," in DB&IS, 2014, pp. 259–271.
[35] V. Bhambri, "Data mining as a tool to predict churn behavior of customers," Int. J. Manag. Res., pp. 59–69, 2013.
[36] M. Clemente-Císcar, S. San Matías, and V. Giner-Bosch, "A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings," Eur. J. Oper. Res., vol. 239, no. 1, pp. 276–285, 2014.
[37] T. Mutanen, V. Österlund, and R. Kinnunen, "Monitoring service adaptation and customer churn in the beginning phase of a new service," in Fourth International Conference on Data Analytics, DATA ANALYTICS 2015, 2015, pp. 69–73.
[38] D. Ringbeck, D. Smirnov, and A. Huchzermeier, "Proactive Retention Management in Retail: Field Experiment Evidence for Lasting Effects," Available SSRN 3378498, 2019.
[39] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, and B. Baesens, "New insights into churn prediction in the telecommunication sector: A profit driven data mining approach," Eur. J. Oper. Res., vol. 218, no. 1, pp. 211–229, 2012.
[40] A. K. Ahmad, A. Jafar, and K. Aljoumaa, "Customer churn prediction in telecom using machine learning in big data platform," J. Big Data, vol. 6, no. 1, p. 28, 2019.
[41] B. Bonev, F. Escolano, and M. Cazorla, "Feature selection, mutual information, and the classification of high-dimensional patterns," Pattern Anal. Appl., vol. 11, no. 3–4, pp. 309–319, 2008.
[42] A. De Caigny, K. Coussement, and K. W. De Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," Eur. J. Oper. Res., vol. 269, no. 2, pp. 760–772, 2018.
[43] T.-H. Hsu, C.-C. Chen, M.-F. Chiang, K.-W. Hsu, and W.-C. Peng, "Inferring potential users in mobile social networks," in 2014 International Conference on Data Science and Advanced Analytics (DSAA), 2014, pp. 347–353.
[44] S. Maldonado, Á. Flores, T. Verbraken, B. Baesens, and R. Weber, "Profit-based feature selection using support vector machines–General framework and an application for customer retention," Appl. Soft Comput., vol. 35, pp. 740–748, 2015.
[45] A. K. Meher, J. Wilson, and R. Prashanth, "Towards a large scale practical churn model for prepaid mobile markets," in Industrial Conference on Data Mining, 2017, pp. 93–106.
[46] K. B. Subramanya and A. Somani, "Enhanced feature mining and classifier models to predict customer churn for an E-retailer," in 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, 2017, pp. 531–536.
[47] J. Van Hulse, T. M. Khoshgoftaar, A. Napolitano, and R. Wald, "Feature selection with high-dimensional imbalanced data," in 2009 IEEE International Conference on Data Mining Workshops, 2009, pp. 507–514.
[48] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," J Stat Softw, vol. 36, no. 11, pp. 1–13, 2010.
[49] H. Li, C.-J. Li, X.-J. Wu, and J. Sun, "Statistics-based wrapper for feature selection: An implementation on financial distress identification with support vector machine," Appl. Soft Comput., vol. 19, pp. 57–67, 2014.
[50] H. Xu, Z. Zhang, and Y. Zhang, "Churn prediction in telecom using a hybrid two-phase feature selection method," in 2009 Third International Symposium on Intelligent Information Technology Application, 2009, vol. 3, pp. 576–579.
[51] K. Cao and P. Shao, "Customer churn prediction based on svm-rfe," in 2008 International Seminar on Business and Information Management, 2008, vol. 1, pp. 306–309.
[52] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, Feature extraction: foundations and applications, vol. 207. Springer, 2008.
[53] Y. Li and G. Xia, "The explanation of support vector machine in customer churn prediction," in *2010 International Con*ference on E-Product E-Service and E-Entertainment, 2010, pp. 1–4.
[54] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2008, pp. 313–325.
[55] H. Hong, Q. Ye, Q. Du, G. A. Wang, and W. Fan, "Crowd characteristics and crowd wisdom: Evidence from an online investment community," J. Assoc. Inf. Sci. Technol., vol. 71, no. 4, pp. 423–435, 2020.
[56] J. Surowiecki, The wisdom of crowds. Anchor, 2005.
[57] W. Pan, Y. Altshuler, and A. Pentland, "Decoding social influence and the wisdom of the crowd in financial trading network," in 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 2012, pp. 203–209.
[58] A. Bari, P. Peidaee, A. Khera, J. Zhu, and H. Chen, "Predicting financial markets using the wisdom of crowds," in 2019 IEEE 4th International Conference on Big Data Analytics (ICBDA), 2019, pp. 334–340.
[59] X. Wu, Q. Ye, Y. Jin, and Y. Li, "Wisdom of Experts and Crowds: Different Impacts of Analyst Recommendation and Online Search on the Stock Market.," in PACIS, 2019, p. 129.
[60] I. Ajzen, "From intentions to actions: A theory of planned behavior," in Action control, Springer, 1985, pp. 11–39.
[61] D. T. Larose and C. D. Larose, Discovering knowledge in data: an introduction to data mining, vol. 4. John Wiley & Sons, 2014.
[62] C.-F. Tsai and Y.-H. Lu, "Customer churn prediction by hybrid neural networks," Expert Syst. Appl., vol. 36, no. 10, pp. 12547–12553, 2009.
[63] P. C. Pendharkar, "Genetic algorithm based neural network approaches for predicting churn in cellular wireless network services," Expert Syst. Appl., vol. 36, no. 3, pp. 6714–6720, 2009.

[64] B. Q. Huang, T.-M. Kechadi, B. Buckley, G. Kiernan, E. Keogh, and T. Rashid, "A new feature set with new window techniques for customer churn prediction in land-line telecommunications," Expert Syst. Appl., vol. 37, no. 5, pp. 3657–3665, 2010.

[65] C. Orsenigo and C. Vercellis, "Combining discrete SVM and fixed cardinality warping distances for multivariate time series classification," Pattern Recognit., vol. 43, no. 11, pp. 3787–3794, 2010.

[66] N. Kamalraj and A. Malathi, "An Ordered Fuzzy Rule Induction Based Churn Mining For Telecom Industry," ICIREIE 2015, p. 17, 2015.

[67] B. Al-Shboul, H. Faris, and N. Ghatasheh, "Initializing genetic programming using fuzzy clustering and its application in churn prediction in the telecom industry," Malaysian J. Comput. Sci., vol. 28, no. 3, pp. 213–220, 2015.

[68] J. Zaratiegui, A. Montoro, and F. Castanedo, "Performing highly accurate predictions through convolutional networks for actual telecommunication challenges," arXiv Prepr. arXiv1511.04906, 2015.

[69] A. Rodan and H. Faris, "Echo state network with SVM-readout for customer churn prediction," in 2015 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), 2015, pp. 1–5.

[70] A. Wangperawong, C. Brun, O. Laudy, and R. Pavasuthipaisit, "Churn analysis using deep convolutional neural networks and autoencoders," arXiv Prepr. arXiv1604.05377, 2016.

[71] M. Azeem, M. Usman, and A. C. M. Fong, "A churn prediction model for prepaid customers in telecom using fuzzy classifiers," Telecommun. Syst., vol. 66, no. 4, pp. 603–614, 2017.

[72] F. Khan and S. S. Kozat, "Sequential churn prediction and analysis of cellular network users—A multi-class, multi-label perspective," in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4.

[73] D. Bell and C. Mgbemena, "Data-driven agent-based exploration of customer behavior," Simulation, vol. 94, no. 3, pp. 195–212, 2018.

[74] L. M. Qaisi, A. Rodan, K. Qaddoum, and R. Al-Sayyed, "Customer churn prediction using data mining approach," in 2018 Fifth HCT Information Technology Trends (ITT), 2018, pp. 348–352.

[75] Y. Beeharry and R. Tsokizep Fokone, "Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry," Concurr. Comput. Pract. Exp., p. e6627, 2021.

[76] S. Baghla and G. Gupta, "Performance Evaluation of Various Classification Techniques for Customer Churn Prediction in E-commerce," Microprocess. Microsyst., vol. 94, p. 104680, 2022.

[77] M. A. Khan, M. A. I. Khan, M. Aref, and S. F. Khan, "Cluster & rough set theory based approach to find the reason for customer churn," Int. J. Appl. Bus. Econ. Res, vol. 14, no. 1, pp. 439–455, 2016.

[78] F. Devriendt, J. Berrevoets, and W. Verbeke, "Why you should stop predicting customer churn and start using uplift models," Inf. Sci. (Ny)., vol. 548, pp. 497–515, 2021.

[79] N. N. Y. Vo, S. Liu, X. Li, and G. Xu, "Leveraging unstructured call log data for customer churn prediction," Knowledge-Based Syst., vol. 212, p. 106586, 2021.

[80] B. Erkayman, E. Erdem, T. Aydin, and Z. Mahmat, "New Artificial intelligence approaches for brand switching decisions," Alexandria Eng. J., vol. 63, pp. 625–643, 2023.

[81] J. B. Rollins, "Foundational methodology for data science," Domino Data Lab, Inc., Whitepaper, 2015.

[82] P. Chapman et al., "The CRISP-DM user guide," in 4th CRISP-DM SIG Workshop in Brussels in March, 1999.

[83] I. Guyon and A. Elisseeff, "An introduction to feature extraction," in Feature extraction, Springer, 2006, pp. 1–25.

[84] M. Landry and B. Angela, "Machine Learning with R and H2O," Mt. View, CA, 2018.

[85] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning," IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, 2012.

[86] P. Cao, O. Zaiane, and D. Zhao, "A measure optimized cost-sensitive learning framework for imbalanced data classification," in Biologically-Inspired Techniques for Knowledge Discovery and Data Mining, IGI Global, 2014, pp. 48–75.

[87] V. Effendy and Z. K. A. Baizal, "Handling imbalanced data in customer churn prediction using combined sampling and weighted random forest," in 2014 2nd International Conference on Information and Communication Technology (ICoICT), 2014, pp. 325–330.

[88] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev., vol. 42, no. 4, pp. 463–484, 2011.

[89] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in International conference on intelligent computing, 2005, pp. 878–887.

[90] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), 2008, pp. 1322–1328.

[91] T. M. Khoshgoftaar, M. Golawala, and J. Van Hulse, "An empirical study of learning from imbalanced data using random forest," in 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007), 2007, vol. 2, pp. 310–317.

[92] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," IEEE Trans. Syst. Man, Cybern. Part B, vol. 39, no. 2, pp. 539–550, 2008.

[93] N. V Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in European conference on principles of data mining and knowledge discovery, 2003, pp. 107–119.

[94] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," J. Anim. Ecol., vol. 77, no. 4, pp. 802–813, 2008.

[95] M. Malohlava and A. Candel, "Gradient boosting machine with H2O." H20 Booklet, http://docs. h2o. ai/h2o/latest-stable/h2o-docs/booklets …, 2017.

[96] H. Wickham and M. H. Wickham, "Package tidyverse," Easily Install Load 'Tidyverse, 2017.

[97] M. Kuhn and H. Wickham, "Recipes: preprocessing tools to

create design matrices.” 2018.
[98] J. Friedman, T. Hastie, R. Tibshirani, and B. Narasimhan, “Package ‘glmnet,’” CRAN R Repositary, 2021.
[99] M. B. Kursa, W. R. Rudnicki, and M. M. B. Kursa, “Package ‘Boruta.’” 2020.
[100] S. RColorBrewer and M. A. Liaw, “Package ‘randomForest,’” Univ. California, Berkeley Berkeley, CA, USA, 2018.
[101] T. L. Pedersen and M. Benesty, “Package ‘lime.’” 2018.
[102] N. Hasbullah, A. J. Mahajar, and M. I. Salleh, “The conceptual framework for predicting loyalty intention in the consumer cooperatives using modified theory of planned behavior,” Int. J. Bus. Soc. Sci., vol. 5, no. 11, 2014.
[103] M. R. Khan, J. Manoj, A. Singh, and J. Blumenstock, “Behavioral modeling for churn prediction: Early indicators and accurate predictors of custom defection and loyalty,” in 2015 IEEE International Congress on Big Data, 2015, pp. 677–680.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 347-358
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Convolutional Neural Networks for Medical Image Segmentation and Classification: A Review

Jenifer S[1*], Carmel Mary Belinda M J[1]

[1].School of Computing, VelTech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.

Received: 31 May 2022/ Revised: 08 Mar 2023/ Accepted: 18 Mar 2023

Abstract

Medical imaging refers to the process of obtaining images of internal organs for therapeutic purposes such as discovering or studying diseases. The primary objective of medical image analysis is to improve the efficacy of clinical research and treatment options. Deep learning has revamped medical image analysis, yielding excellent results in image processing tasks such as registration, segmentation, feature extraction, and classification. The prime motivations for this are the availability of computational resources and the resurgence of deep Convolutional Neural Networks. Deep learning techniques are good at observing hidden patterns in images and supporting clinicians in achieving diagnostic perfection. It has proven to be the most effective method for organ segmentation, cancer detection, disease categorization, and computer-assisted diagnosis. Many deep learning approaches have been published to analyze medical images for various diagnostic purposes. In this paper, we review the works exploiting current state-of-the-art deep learning approaches in medical image processing. We begin the survey by providing a synopsis of research works in medical imaging based on convolutional neural networks. Second, we discuss popular pre-trained models and General Adversarial Networks that aid in improving convolutional networks' performance. Finally, to ease direct evaluation, we compile the performance metrics of deep learning models focusing on covid-19 detection and child bone age prediction.

Keywords: Convolutional Neural Networks; Deep learning; Generative Adversarial Network; Medical Image Analysis; Transfer learning.

1- Introduction

Computer-aided diagnosis (CAD) has emerged as one of the most important research fields in medical imaging. In CAD, machine learning algorithms are often utilized to examine the imaging data from the historical samples of patients and construct a model to assess the patient's condition [1]. The developed model assists clinicians in making quick decisions. The most common imaging modalities used in medical applications are X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Ultrasound. The sole aim of medical image processing would be to improve the interpretability of the information illustrated [2]. The following are the main categories of medical image analysis: enhancement, registration, segmentation, classification, localization, and detection [3].

Earlier, medical images were processed using low-level methods, such as thresholding [4][5], region growing, and edge tracing [6]. Meanwhile, the growth in size and scope of medical imaging data has fueled the evolution of machine learning techniques in medical image analysis. However, since such methods rely on handcrafted features, algorithm design requires manual effort. These constraints of conventional machine learning approaches have risen to the notion of Artificial Neural Networks (ANNs). Factors such as data availability and computational processing capabilities facilitate the deepening of ANNs [7]. The emergence of deep learning techniques like convolutional neural networks has widened the possibilities for the automation of medical image processing.

A Convolutional Neural Network (CNN) is a class of neural networks meant to handle pixel values. CNN makes image classification more scalable by employing linear mathematical concepts to detect patterns inside an image. While traditional CNN architectures consisted solely of convolutional layers placed on top of one another, modern

✉ **Jenifer S**
vtd914@veltech.edu.in

architectures such as Inception, ResNet, and DenseNet come up with new and innovative approaches to build convolutional layers in a way that makes learning more efficient [8].

CNN can be employed as a feature extractor as well. Feature extraction aims to convert raw pixel data into numerical features that can be processed while keeping the information in the original data set. Traditional feature extractors can be replaced with CNNs, which can extract complex features that express the image in much more detail. The resulting features are then fed into a classifier network or used by typical machine learning algorithms for classification [9][10].

Despite the fact that deep CNN architectures exhibit cutting-edge performance on computer vision problems, there are some concerns about using CNN in the radiology field. In 2014, Goodfellow et al. discovered that introducing a little bit of noise to the original information can readily deceive neural networks into misclassifying items [11]. Furthermore, since the efficiency of deep learning is often based on the volume of input data, CNN requires large scale well-annotated radiology images. Building such databases in the medical industry, on the other hand, is costly and labor-intensive.

In this study, we summarize the current developments in deep learning approaches for medical image analysis. The paper is organized as follows: First, survey papers related to medical image analysis are discussed in section 2. Then, in section 3, CNN models employed in the radiology field and approaches for improving CNN performance are described. Following that, the finding of models aimed at detecting Covid-19 and predicting child bone age are reviewed in section 4. And finally, the conclusion is set out.

2- Related Works

This section discusses the survey papers on medical image analysis using deep learning-based algorithms. Hu et al. [12] described four deep learning architectures used for image analysis: CNN, Fully Convolutional Networks (FCN), Deep belief networks, and Autoencoders. They also compiled the recent works on cancer identification and diagnosis.

Liu et al. [13] concentrated on deep learning-based medical image segmentation. They began by explaining the deep learning framework deployed to segment medical images. Then, state-of-the-art segmentation architectures such as FCN, U-Net, and Generative Adversarial Network (GAN) were examined. Shin et al. [14] first studied two medical diagnostic problems, namely interstitial lung disease detection and thoracoabdominal lymph node classification, using three CNN models: AlexNet, CifarNet, and GoogLeNet. Then looked at how transfer learning enhanced the performance of each model.

Kazeminia et al. [15] provided a broad insight into the current studies on GANs for medical applications, discussed the limitations and opportunities of the existing techniques, and elaborated on potential future work.Here, the emphasis was primarily on the segmentation approaches that employed GAN concepts, whereas [13] explained all major segmentation architectures used for medical imaging.On the other hand, Fu et al. [16]divided the approaches reviewed into two main groups: pixel-by-pixel classification and end-to-end segmentation, and discussed the performance, limits, and future potential of each group.

Although previous surveys examined all of the CNN architectures used for medical image analysis, they did not assess the impact of different CNN architectures on a specific application. In this survey, we looked at previous works on medical image segmentation and classification to analyze how CNN performance varied across anatomical regions. Also, we discussed CNN's difficulties as well as solutions for them to enhance CNN's performance.

3- Medical Image Analysis using Deep Learning

The primary focus of medical image analysis is to find out which regions of anatomy are affected by the disease to aid physicians in learning lesion progression. The analysis of a medical image is mostly reliant on four steps: 1. image preprocessing, 2. segmentation, 3. feature extraction, and 4. pattern identification or classification [17]. Pre-processing is to remove unwanted distortions from images or improve image information for further processing [18]. Segmentation refers to the process of isolating regions, such as tumors, organs, etc., for further study. The process of extracting precise details from the Regions of Interest (ROIs) that aid in their recognition is known as feature extraction. Based on extracted features, classification assists in categorizing the ROI [19].

We have compiled a list of research papers primarily concerned with segmentation and classification in medical imaging. Following the review of CNN, we have outlined some techniques for improving CNN's performance.

3-1- Convolutional Neural Network

A CNN is a supervised deep learning framework that can accept the images as input, allocate filters to convert image pixels into features, and apply those features to distinguish one data from another. It is generally composed of three layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. The convolutional layer is the initial layer of a convolutional network. After that, more convolutional layers or pooling layers can be added, with the fully-connected layer being the last.

The convolutional block draws the features from the image, from which the network can analyze and obtain hidden correlations. Pooling layers are applied to reduce the size of the convolved features, referred to as downsampling. Fully-Connected layers execute classification tasks depending on the features retrieved by the preceding layers. While convolutional layers generally adopt Rectified Linear unit function to activate neurons, Fully-connected layers employ a softmax activation function or traditional machine learning classifiers (SVM, KNN, etc.) [20] to classify inputs.

Overview of works:

Despite the deep network's ability to extract features with more precision, it requires a lot of computing resources. Therefore, Badza et al. [21] introduced a simple CNN model with two convolutional blocks for classifying brain tumors using MRI images. While evaluating 3064 MRI images, the model attained the best accuracy of 95.56% using 10-fold cross-validation. Rachapudi et al. presented an efficient CNN architecture with a 22.7% error rate to classify the colorectal cancer histopathological images. To prevent overfitting, the model included five convolutional blocks, each containing a dropout layer [22].

The deep learning architecture for image segmentation comprises an encoder and a decoder. The encoder uses filters to extract features from the image, whereas the decoder is in charge of producing the final output, often a segmentation mask containing the object's shape. A Fully Convolutional Network (FCN) is an encoder-decoder model that lacks dense layers in favor of 1x1 convolutions to serve the function of fully connected layers [23]. Sun et al. developed a 3D FCNN-based model for multimodal brain tumor image segmentation. The encoder had four pathways for extracting multi-scale image features [24]. Then, these four feature maps were fused and fed to the decoder. By experimental validation on the Brain Tumor Segmentation challenge dataset 2019 (BraTS2019), the model segmented the dataset with the dataset with Dice Similarity Coefficient metrics (DSC) of 0.89, 0.78, 0.76 for the complete, core, and enhanced tumor, respectively.

In 2015, Ronneberger et al. introduced U-Net to deal with biomedical image segmentation that can learn from a small number of annotated medical images. U-Net [25] is a U-shaped encoder-decoder-based framework consisting of four encoder and four decoder blocks connected by skip connections. Dharwadkar et al. employed U-Net architecture to design a ventricle segmentation model for heart MRI images. There are four layers in the original U-net, but only three layers were employed in this model [26]. For the Right Ventricle Segmentation Challenge (RVSC) dataset, the proposed model obtained a dice score of 0.91.

For segmenting the left ventricle from cardiac CT angiography, Li et al. introduced U-Net with 8-layer. The exhibited U-Net model comprised eight encoder and eight decoder blocks. To further improve the network's efficiency, residual blocks in the form of skip connections were introduced into each encoder and decoder block [27]. The model was trained using 1600 CT images from 100 patients, resulting in a DSC of 0.9270±139. Li et al. [29] introduced an attention mechanism between nested encoder-decoder paths in U-Net++ [28] architecture to improve the understanding of the study area in liver segmentation. The model achieved a DSC of 98.15% through the experimental analysis of the Liver Tumor Segmentation challenge dataset 2017 (LiTS2017).

V-Net extends U-Net by processing 3D MRI images with 3D convolutions [30]. Guan et al. developed a V-Net-based framework for separating brain tumors from 3D MRI brain images. In the developed framework, the Squeeze and Excite (SE) module and Attention Guide Filter (AG) module were integrated into V-Net architecture to suppress irrelevant information and enhance segmentation accuracy [31]. When tested on the BraTS2020 dataset, the model obtained dice metrics of 0.68, 0.85, and 0.70 for the complete, core, and enhanced tumor, respectively.

Mask Regional CNN is another CNN variant used in medical image segmentation. Mask R-CNN is two-phase object identification and segmentation architecture. The first stage, known as the Region Proposal Network (RPN), returns potential bounding boxes, whereas the second stage generates the segmentation mask from each box [32]. Dogan et al. introduced a hybrid model combining U-Net and MaskR-CNN for pancreas segmentation from CT images. The proposed system was composed of two parts: Pancreas detection and Pancreas segmentation. In pancreas localization, the Region proposal network, in conjunction with the mask production network, was used to determine the bounding boxes of the pancreas portion, and the sub-region centered by the rough pancreas region was sliced [33]. Finally, the cropped sub-region was sent to U-Net for precise segmentation. The average DSC for the two-phase approach demonstrated on the 82 abdominal CT scans was 86.15%.

3-2- Improving the Performance of CNN

The CNN model is often used for image classification because it achieves better accuracy with a low error rate. However, it needs large datasets to generalize the hidden correlations found in the learning data. Here, we have discussed two approaches that may optimize the performance of CNN. 1. Transfer learning 2. General Adversarial Network (GAN)

3-2-1-Transfer Learning

Transfer learning is an effective strategy to train a network with a limited dataset. Here, the model is pre-trained using a large-scale dataset, like ImageNet having 1.4 million images divided into 1000 categories, and then applied to the

problem at hand [34]. The major pre-trained CNN architectures for image classification are as follows:

LeNet-5: LeNet-5 [35], a 7-level convolutional network presented by LeCun et al. in 1998, was the first of its kind. The model was designed to classify handwritten digits and tested on the MNIST standard dataset, with a classification accuracy of roughly 99.2%.

AlexNet: The network's design was quite similar to LeNet, but it was deeper, with more filters per layer. It contains five convolution layers and three fully-connected layers. To control overfitting, it employs a dropout mechanism in fully connected layers [36].

Visual Geometry Group at Oxford (VGGNet): VGGNet typically consists of 16 layers with a lot of 3×3 filters of stride one [37]. It is now the most popular method for extracting features from images. VGGNet, on the other hand, has 138 million parameters, which are difficult to manage.

InceptionV1/GoogLeNet: The inception/GoogleNet architecture, presented by Christian Szegedy et al., has 22 layers. The Inception block does 1×1, 3×3, 5×5 convolutions, and 3×3 pooling at the input, and the outputs of these are stacked to send to the next inception module [38]. By using 1x1 convolutions in each module, GoogleNet can reduce the size of parameters to 4 million compared to AlexNet's 60 million**.**

Residual Network or ResNet: A Residual Network, often known as ResNet, is a 152-layer model. This network employs a VGG-19-inspired network design, with grouped convolutional layers followed by no pooling in between and an average pooling before the fully connected output layer [39]. The design is converted into a residual network by adding shortcut connections. This sort of skip connection has the advantage of training deep networks without problems caused by vanishing gradients.

Moreover, DenseNet(2017), XceptionNet(2017), ShuffleNet(2017), MobileNet(2017), EfficientNet(2019), and ConvNeXt(2020) are some of the latest CNN architectures that have been employed for image classification and feature extraction. In image classification, either the pre-trained model can be used as-is or modified for a given problem. The fine-tuning of a model can be accomplished using one of the following strategies: 1. Train some layers while leaving the others frozen 2. Freeze the convolutional base only.

Overview of works:

LeNet is a popular CNN model because of its simple architecture and shorter training time. Deep neural network models use the concept of the max-pooling layer to extract the most relevant features from a region. However, in medical image analysis, where quality is poor, pixels with lower intensities may hold critical information. Hence, Hazarika et al. introduced the minimum pooling layer in LeNet for Alzheimer's disease (AD) classification. In the modified LeNet [40], the min-pooling and max-pooling layers were merged, and the resulting layer replaced all max-pooling layers. According to the experimental study on 2000 brain images, the original LeNet model classified AD with 80% accuracy, while the revised LeNet attained an accuracy of 96.64%.

Hosny et al. introduced a fine-tuned AlexNet model to categorize skin lesions into seven classes using skin images. In the proposed architecture, the last three layers were replaced by new layers to make them suitable for classifying seven types of skin lesions [41]. The parameters of these new layers were initially set at random and then modified during the training. After training on 10,015 images, the model achieved an accuracy of 98.70% and a sensitivity of 95.60%. Dulf et al. trained and assessed five different models, including GoogleNet, AlexNet, VGG16, VGG19, and InceptionV3, to determine the best model for classifying the eight categories of colorectal polyps. The main criteria for adopting the network were sensitivity and F1-score [42]. Hence, InceptionV3 was chosen with an F1-score of 98.14% and a sensitivity of 98.13%. In InceptionV3 [43], the 5×5 convolutional layer is replaced with two 3×3 convolutional layers to lower the computational cost.

Hameed et al. demonstrated an ensemble deep learning strategy to categorize breast cancer into carcinoma and non-carcinoma using histopathology images. In this case, VGG models, namely VGG16 and VGG19, were used to design the framework. VGG19 has the same basic architecture as VGG16 with three additional convolutional layers. Besides the first block, the remaining four blocks were updated during training to fine-tune the models [44]. Finally, the tunedVGG16 and VGG19 models were ensembled, resulting in an overall accuracy of 95.29%.

Togacar et al. used both VGG16 and AlexNet to extract features for brain tumor classification from MRI images, where each model captured 1000 features [45]. Then, using the Recursive Feature Elimination (RFE) feature selection algorithm, the obtained features were evaluated to identify the most efficient features. Finally, the SVM classifier gave 96.77% accuracy with 200 chosen features. Eid et al. presented ResNet-based SVM for pneumonia detection using X-rays. The developed model preferred ResNet to get features from chest X-rays, then used boosting algorithm to choose the relevant features and an SVM classifier to detect pneumonia based on those features [46]. The model had 98.13% accuracy after being trained on 5,863 X-rays.

Xiao et al. used a Res2Net-based 3D-UNet to segment the left ventricle from echocardiography images. To extract 3D features at multiple scales, the basic residual unit in Res2Net was replaced with a set of 3×3×3 filters [47]. Finally, a group of 1×1×1 filters merged feature maps from all groups. According to an experimental analysis of 1186 lung images from the Lung Nodule Analysis dataset 2016 (LUNA16), the model acquired a DSC of 95.30%. Goyal et al. used the Mask R-CNN for segmenting kidneys from the MRI images. In the proposed work,

InceptionResNetV2 was adopted as the CNN network to segment the kidneys. Then, to refine the segmentation result, post-processing procedures such as eliminating any voxel that was not associated with the kidney and fill operation were performed [48]. The proposed model got a mean dice score of 0.904 after being evaluated with 100 scans. Table 1 lists a few more surveyed works on transfer learning. All of the listed models [59]-[74] underwent pre-training on the ImageNet dataset.

3-2-2- Generative Adversarial Network

Goodfellow et al. introduced the Generative Adversarial Network (GAN), a type of neural network meant for unsupervised learning. GANs generally are of two competing neural network models: a generator that creates new data samples that mimic training data and a discriminator that differentiates training data from the generator's output [49].

Cirillo et al. introduced a 3D GAN to segment brain tumors using MRI images from the BraTS2020 dataset. The U-Net architecture-based generator resulted in the segmented tumor region. The GAN discriminator was given a 3D MRI image and its segmentation output from the generator as input and generated a precise segmentation mask [50]. The GAN model segmented the whole, the core, and the enhanced tumor with average dice scores of 87.20%, 81.14%, and 78.67%, respectively. Wang et al. developed a U-Net segmentation network and a discriminant network with multi-scale features extraction to enhance prostate segmentation accuracy [51]. The approach obtained a DSC value of 91.66% by demonstrating it on 220 MRI images.

Wei et al. used a combination of GAN and Mask R-CNN to segment the liver from CT images. In the improved Mask R-CNN, the k-means algorithm was utilized to adjust the bounding box parameters using Euclidean distance [52]. The GAN-based approach yielded an average DSC of 95.3% while evaluating 378 CT images. A V-Net and Wasserstein GAN-based model was explored by Ma et al. [53] to improve the efficiency of liver segmentation. The WGAN [54] model includes Wasserstein distance to fix the issue of GAN training instability. On two training on two abdominal CT scan datasets, LiTS and CHAOS, the method achieved DSC of 92% and 90%, respectively. Zhang et al. proposed Dense GAN coupled with the U-Net to separate lung lesions from covid-19 CT images. A Dense Block with five layers [55] was introduced into the discriminator network to make the model more compact. The proposed model got a mean dice score of 0.683 when tested on 100 lung CT images. Besides this, table 2 includes some more GAN-based techniques [75]-[84] applied to medical images.

Table 1: Overview of pre-trained models

Ref.	***Year***	***Model***	***Findings***	***Modality***	***No of Samples***	***Accuracy***
[59]	2021	VGG16	Brain tumor classification	MRI	3704	95.71%
[60]	2020	ResNet50	Brain tumor classification	MRI	253	97.2%
[61]	2020	GoogleNet	Alzheimer's disease classification	MRI	479	97.15%
[62]	2020	Ensemble of AlexNet, DenseNet121, ResNet18, GoogleNet, InceptionV3	Pneumonia detection	X-ray	5232	96.4%
[63]	2020	AlexNet	Lung nodule classification	CT and X-ray	16,471	99.6%
[64]	2020	ResNet50	Breast tumor classification	Mammogram	1167	85.71%
[65]	2020	ResNet50	Breast tumor classification	Histopathological images	7909	99%
[66]	2021	VGG16	Breast tumor classification	Mammogram	322	98.96%
[67]	2020	DenseNet201	Skin lesion classification	Skin images	10,050	96.18%
[68]	2020	GoogleNet	Skin image classification	Skin images	2376	99.29%

[69]	2021	VGG19	Thyroid nodule cell classification	Cytology images	9209	93.05%
[70]	2021	GoogleNet	Thyroid nodule classification	Ultrasound	3123	96.04%
[71]	2021	GoogleNet	Colorectal polyps classification	Gastrointestinal polyp images	47,238	98.44%
[72]	2020	Faster R-CNN + VGG16	Brain tumor segmentation and classification	MRI	2406	77.60%
[73]	2021	U-Net + InceptionV3	Breast tumor segmentation and classification	Mammogram	1216	98.87%
[74]	2020	Mask R-CNN + ResNet-50	white blood cells detection and classification	Cytological images	145	95.3%

GAN can also be used for data augmentation [56] (i.e., creating plausible examples to add to a dataset) to boost classifier accuracy. GAN was also used [57] to generate realistic skin cancer images. The generator generated high-quality training data, and the discriminator tried to distinguish the original data from the generator's data. Ahmad et al. developed an Auxiliary GAN framework to assess the accuracy of skin cancer categorization. First, the variational autoencoder network was trained to obtain the latent noise vector, and the generator produced skin lesion samples from this informative noise vector [58]. The GAN used here not only decided whether the image was original or not but also predicted the image's class label with 92.5% accuracy.

Table 2: Overview of GAN-based methods

Ref.	***Approach***	***Findings***	***Samples***	***Metrics***
[75]	Capsule GAN + LeNet	Prostate image classification	1400 MRI Images	Accuracy 89.20%
[76]	GAN + AlexNet	Parkinson's disease	504 MRI Images	Accuracy 89.23%
[77]	GAN + DenseNet 121	Skin lesion classification	525 Skin Images	Accuracy 94.25%
[78]	GAN + Inception V3	Breast mass classification	1447 Ultra sound Images	Accuracy 90.41%
[79]	GAN + ResNet50	Brain tumor classification	3064 MRI Images	Accuracy 96.25%
[80]	3D U-Net , VGG16	Brain tumor segmentation	285 MRI Images	DSC 90.1%
[81]	U-Net , Fully connected	Breast tumor segmentation	1062 Ultra sound Images	DSC 88.41%
[82]	DeepLap V2 [104], FCN	Left ventricle segmentation	10,022 MRI Images	DSC 88.0%
[83]	U-Net , FCN	Whole heart segmentation	500 CT Images	DSC 86.32%
[84]	Auto encoder, CNN	Lung lesion segmentation	1936 PET Images	DSC 62.0%

4- Discussion

To make straightforward comparisons, we have summarized the outcomes of the papers based on covid-19 identification and child bone age prediction.

4-1- Covid -19 Detection

COVID-Net, an open-access initiative, was launched in March 2020 to assist healthcare professionals in combating Corona Virus Disease 2019 (COVID-19) by leveraging the advancement of machine learning. Furthermore, it regularly releases deep learning models and benchmark datasets to keep up with the pandemic [105]. In response to this initiative, Wang et al. introduced COVID-Net, a deep CNN for covid-19 identification from chest X-rays.

In the COVID-Net model, residual Projection-Expansion-Projection-Extension (PEPX) blocks which comprise four 1×1 convolutions, were introduced to enhance the efficiency of features while ensuring computational efficiency [85]. The model efficacy was verified using 13,975 X-ray images from the COVIDx dataset. According to the experimental findings, the model attained a precision of 98.9% and a

recall of 91.0% to detect covid-19. The COVIDx is a large-scale dataset of chest X-ray images compiled from publicly available data sources. As of now, COVIDx consists of 30,882 X-ray images from 17,026 patients.

Table 3 shows the deep learning models that were employed for covid-19 detection. We can observe from the methods studied that the classification accuracy is affected not only by the CNN model chosen but also by the size of the dataset, the type of modality, the data augmentation techniques, and the features opted for processing.

4-2- RSNA Pediatric Bone Age Challenge 2017

In 2017, the Radiological Society of North America (RSNA) held a contest to predict the children's bone age from the hand X-rays. The main goal of this challenge was to encourage people to develop machine learning models that could accurately estimate bone age from pediatric hand X-rays.The performance measure was the Mean Absolute Error in months, the average absolute difference between predicted results and ground truth bone age [103]. The bone age dataset [106], consisting of 14,236 left hand X-ray images, was divided into a training set, a validation set, and a test set of 12,611, 1425, and 200, respectively.

Table 3: Deep learning networks for covid-19 identification

Ref.	***Deep learning model***	***Modality***	***Total samples***			***Evaluation Metrics***
			Normal	***pneumonia***	***Covid-19***	
[85]	COVID-Net	X-ray	8,066	5,521	183	Accuracy 94.3%, Precision 90.9%, Recall 96.8%
[86]	EfficientNet	X-ray	8,066	5,521	183	Accuracy 93.9%, Precision 100%, Recall 96.8%
[87]	NASNet	X-ray	533	515	108	Accuracy 95%, Precision 95%, Recall 90%
[88]	GAN and VGG16	X-ray	721	0	403	Accuracy 95%, Recall 90%
[89]	DenseNet103, ResNet18	X-ray	191	20	180	Accuracy 88.9%, Precision 83.4%, Recall 85.9%
[90]	ResNet101, ResNet152	X-ray	8851	9576	140	Accuracy 96.1%
[91]	DenseNet and Graph Attention Network	X-ray	10192	7399	399	Accuracy 94.1%, Precision 94.47%, Recall 91.9%
[92]	VGG19	X-ray	3181	0	2049	Accuracy 98.36%
[93]	ResNet34, HRNet	X-ray	400	0	400	Accuracy 99.99%, Precision 100%, Recall 99.9%
[94]	VGG16	CT	49800	23652	80800	Accuracy93.57%, Precision 89.40%, Recall 94%
[95]	Deep long short-term memory network	CT	547	631	612	Accuracy 97.93%, Recall 98.18%
[96]	VGG19	Ultrasound	235	277	399	Accuracy 100%

Table 4 summarizes some of the recent CNN-based methods that used the RSNA bone age benchmark dataset. The approaches stated were divided into two phases. CNN model was used in the initial stage to carve up the hand region from the X-ray images. The second phase included a pre-trained model for extracting inherent features from the hand region and a regression layer to estimate bone age.

Table 4: CNN frameworks using RSNA bone age dataset

Ref.	***Segmentation***	***Regression***	***Mean Absolute Error (in months)***
[97]	U-Net	VGG16	8.08
[98]	U-Net	Inception-ResNetV2	8.59
[99]	DeepLabV3	MobileNetV1	8.200
[100]	U-Net	VGG16	9.997
[101]	CNN	MobileNetV3	6.2
[102]	Mask R-CNN	VGG19	6.38

Table 4 shows that MobileNetV3 [101] and VGG19 [102] both performed better on bone age prediction, with MAE around six months.VGG16 provided a better MAE [97] when utilising specific bones on the hand region to estimate bone age.Inception-ResNetV2 was employed as a feature extractor [98]. After extracting features, Support Vector Regression (SVR) and Kernel Ridge Regression (KRR) were ensembled to forecast skeletal age.

5- Conclusion

We have presented a detailed overview of newly published deep learning-based methods from 2019 to 2022 in medical imaging. Recent advances in deep learning architectures have the ability to boost diagnostic precision in medical imaging. On the other hand, deep learning necessitates a large volume of data to outperform traditional machine learning models. In practice, however, obtaining such datasets containing medical images is difficult. Transfer learning via pre-trained models can help to solve this problem. There is a clear tendency toward modifying pre-trained models to make them more appropriate for a specific task. This popularity is because pre-trained models expedite training while ensuring good classification accuracy. Another trend is toemploy GAN to enhance segmentation accuracy due to its capacity to generate high-quality medical images and imitate input data distribution. The GAN-based approaches have proven to be effective in resolving discrepancies between ground truth and model-generated segmentation masks. Also, GAN's ability to synthesize data can help solve difficulties such as lack of medical images or imbalanced data distribution, resulting in improved classification model performance.

References

[1] H. P. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era of deep learning," Medical Physics, vol. 47, no. 5, pp. e218–e227, May 2020.

[2] F. Ritter, T. Boskamp, A. Homeyer, H. Laue, M. Schwier, F. Link, and H. O. Peitgen, "Medical Image Analysis," IEEE Pulse, vol. 2, no. 6, pp. 60–70, Nov. 2011.

[3] J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," IEEE Access, vol. 6, pp. 9375–9389, Dec. 2017.

[4] T. Kiyatmoko, "Retinal Vessel Extraction using Dynamic Threshold and Enhancement Image Filter from Retina Fundus," Journal of InformationSystems & Telecommunication, vol. 6, no. 24, pp. 189-196, Jun. 2019.

[5] K. A. Kumar, and R. Boda, "A Threshold-based Brain Tumour Segmentation from MR Images using Multi-Objective Particle Swarm Optimization," Journalof Information Systems and Telecommunication, vol. 9, no. 36, pp. 218-225, Oct. 2021.

[6] M. Jena, S. P. Mishra, and D. Mishra, "A survey on applications of machine learning techniques for medical image segmentation," International Journal of Engineering & Technology, vol. 7, no. 4, pp. 4489–4495, Nov. 2018.

[7] S. Niyas, S. J. Pawan, M. Anand Kumar, and J. Rajan, "Medical image segmentation with 3D convolutional neural networks: A survey," Neurocomputing, vol. 493, pp. 397–413, Jul. 2022.

[8] P. Dutta, P. Upadhyay, M. De, and R. G. Khalkar, "Medical image analysis using deep convolutional neural networks: CNN architectures and transfer learning," in 2020 International Conference on Inventive Computation Technologies (ICICT), Feb. 2020, pp. 175-180.

[9] M. Jogin, Mohana, M. S. Madhulika, G. D. Divya, R. K. Meghana, and S. Apoorva, "Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning," in 2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), May 2018, pp. 2319–2323.

[10] E. Gholam, and S. R. KamelTabbakh, "Diagnosis of Gastric Cancer via Classification of the Tongue Images using Deep Convolutional Networks," Journal of Information Systems and Telecommunication, vol. 9, no. 35, pp.191-196, Jul. 2021.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv [stat.ML], 2014.

[12] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis − A survey,"Pattern Recognition, vol. 83, pp. 134–149, Nov. 2018.

[13] X. Liu, L. Song, S. Liu, and Y. Zhang, "A Review of Deep-Learning-Based Medical Image Segmentation Methods," Sustainability, vol. 13, no. 3, p. 1224, Jan. 2021.

[14] H. C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R.M. Summers, "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1285–1298, May 2016.

[15] S. Kazeminia, C. Baur, A. Kuijper, B. Van Ginneken, N. Navab, S. Albarqouni, and A. Mukhopadhyay, "GANs for medical image analysis," Artificial Intelligence in Medicine, vol. 109, p. 101938, Sep. 2020.

[16] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," PhysicaMedica, vol. 85, pp. 107–122, May 2021.

[17] B. Halalli, and A. Makandar, "Computer Aided Diagnosis - Medical Image Analysis Techniques," Breast Imaging, Dec. 2017.

[18] L. Chandrashekar, and A. Sreedevi, "A two-stage multi-objective enhancement for fused magnetic resonance image and computed tomography brain images, "Journal of InformationSystems & Telecommunication, vol. 8, no. 30, pp. 93-104, Aug. 2020.

[19] S. Zakariapour, H. Jazayeriy, and M. Ezoji, "Mitosis detection in breast cancer histological images based on texture features using adaboost, "Journal of InformationSystems & Telecommunication, vol. 5, no. 8, pp. 1-10, Jul. 2017.

[20] M. Kumar, S. K. Khatri, and M. Mohammadian, "Breast Cancer Classification Approaches-A Comparative Analysis," Journal of InformationSystems & Telecommunication, vol. 11, no. 41, pp. 1-11, Jan. 2023.

[21] M. M. Badža, and M. Č. Barjaktarović, "Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network," Applied Sciences, vol. 10, no. 6, p. 1999, Mar. 2020.

[22] V. Rachapudi, and G. Lavanya Devi, "Improved convolutional neural network based histopathological image classification," Evolutionary Intelligence, vol. 14, no. 3, pp. 1337-1343, Feb. 2020.

[23] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 4, pp. 640–651, Jan. 2017.

[24] J. Sun, Y. Peng, Y. Guo, and D. Li, "Segmentation of the multimodal brain tumor image used the multi-pathway architecture method based on 3D FCN," Neurocomputing, vol. 423, pp. 34- 45, Jan. 2021.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Lecture Notes in Computer Science, pp. 234–241, Oct. 2015.

[26] N. V. Dharwadkar, and A. K. Savvashe, "Right Ventricle Segmentation of Magnetic Resonance Image Using the Modified Convolutional Neural Network," Arabian Journal for Science and Engineering, vol. 46, no. 4, pp. 3713–3722, Jan. 2021.

[27] C. Li, X. Song, H. Zhao, L. Feng, T. Hu, Y. Zhang, J. Jiang, J. Wang, J. Xiang, and Y. Sun, "An 8-layer residual U-Net with deep supervision for segmentation of the left ventricle in cardiac CT angiography," Computer Methods and Programs in Biomedicine,vol. 200, p. 105876, Mar. 2021.

[28] Z. Zhou, M. M. RahmanSiddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation,"inDeep learning in medical image analysis and multimodal learning for clinical decision support, Cham: Springer, Sep. 2018, pp. 3-11.

[29] C. Li, Y. Tan, W. Chen, X. Luo, Y. Gao, X. Jia, and Z. Wang, "Attention unet++: A nested attention-aware U-net for liver CT image segmentation," in 2020 IEEE International Conference on Image Processing (ICIP), Oct. 2020, pp. 345-349.

[30] Milletari, N. Navab, and S. A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in 2016 Fourth International Conference on 3D Vision (3DV),Oct. 2016, pp. 565-571.

[31] X. Guan, G. Yang, J. Yang, X. Xu, W. Jiang, and X. Lai, "3D AGSE-VNet: an automatic brain tumor MRI data segmentation framework," BMC Medical Imaging, vol. 22, no. 1, Jan. 2022.

[32] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," IEEE Transactions on Pattern Analysis andMachine Intelligence, vol. 42, no. 2, pp. 386–397, Mar. 2017.

[33] R. O. Dogan, H. Dogan, C. Bayrak, and T. Kayikcioglu, "A Two-Phase Approach using Mask R-CNN and 3D U-Net for High-Accuracy Automatic Segmentation of Pancreas in CT Imaging," Computer Methods and Programs in Biomedicine, vol. 207, p. 106141, Aug. 2021.

[34] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," Insights into Imaging, vol. 9, no. 4, pp. 611–629, Jun. 2018.

[35] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Communications of the ACM,vol. 60, no. 6, pp. 84–90, May 2017.

[37] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv [cs.CV], 2014.

[38] C. Szegedy, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[40] R. A. Hazarika, A. Abraham, D. Kandar, and A. K. Maji, "An Improved LeNet-Deep Neural Network Model for Alzheimer's Disease Classification Using Brain Magnetic Resonance Images," IEEE Access, vol. 9, pp. 161194–161207, Nov. 2021.

[41] K. M. Hosny, M. A. Kassem, and M. M. Fouad, "Classification of Skin Lesions into Seven Classes Using Transfer Learning with AlexNet," Journal of Digital Imaging, vol. 33, no. 5, pp. 1325–1334, Jun. 2020.

[42] Eva-H. Dulf, M. Bledea, T. Mocan, and L. Mocan, "Automatic Detection of Colorectal Polyps Using Transfer Learning," Sensors, vol. 21, no. 17, p. 5704, Aug. 2021.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2016, pp. 2818-2826.

[44] Z. Hameed, S. Zahia, B. Garcia-Zapirain, J. Javier Aguirre, and A. MaríaVanegas, "Breast Cancer Histopathology Image Classification Using an Ensemble of Deep Learning Models," Sensors, vol. 20, no. 16, p. 4373, Aug. 2020.

[45] M. Toğaçar, Z. Cömert, and B. Ergen, "Classification of brain MRI using hyper column technique with convolutional neural network and feature selection method," Expert Systems with Applications, vol. 149, p. 113274, Jul. 2020.

[46] M. M. Eid, and Y. H. Elawady, "Efficient Pneumonia Detection for Chest Radiography Using ResNet-Based SVM," European Journal of Electrical Engineering and Computer Science, vol. 5, no. 1, pp. 1–8, Jan. 2021.

[47] Xiao, B. Liu, L. Geng, F. Zhang, and Y. Liu, "Segmentation of Lung Nodules Using Improved 3D-UNet Neural Network," Symmetry, vol. 12, no. 11, p. 1787, Oct. 2020.

[48] M. Goyal, J. Guo, L. Hinojosa, K. Hulsey, and I. Pedrosa, "Automated kidney segmentation by mask R-CNN in T2-weighted magnetic resonance imaging," in Medical Imaging2022: Computer-Aided Diagnosis, vol. 12033, pp. 89-94, Apr. 2022.

[49] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio,"Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[50] M. D. Cirillo, D. Abramian, and A. Eklund, "Vox2Vox: 3D-GAN for Brain Tumour Segmentation," Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 274–284, Oct. 2021.

[51] W. Wang, G. Wang, X. Wu, X. Ding, X. cao, L. Wang, J. Zhang, and P. Wang "Automatic segmentation of prostate magnetic resonance imaging using generative adversarial networks," Clinical Imaging, vol. 70, pp. 1–9, Feb. 2021.

[52] X. Wei, X. Chen, C. Lai, Y. Zhu, H. Yang, and Y. Du, "Automatic Liver Segmentation in CT Images with Enhanced GAN and Mask Region-Based CNN Architectures," BioMed Research International, vol. 2021, pp. 1–11, Dec. 2021.

[53] J. Ma, Y. Deng, Z. Ma, K. Mao, and Y. Chen, "A Liver Segmentation Method Based on the Fusion of VNet and WGAN," Computational and Mathematical Methods in Medicine, vol. 2021, pp. 1–12, Oct. 2021.

[54] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein Generative Adversarial Networks," in International Conference on Machine Learning, Jul. 2017, pp. 214-223.

[55] J. Zhang, L. Yu, D. Chen, W. Pan, C. Shi, Y. Niu, X. Yao, X. Xu, and Y. Cheng, "Dense GAN and multi-layer attention based lesion segmentation method for COVID-19 CT images," Biomedical Signal Processing and Control, vol. 69, p. 102901, Aug. 2021.

[56] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation Generative Adversarial Networks," arXiv [stat.ML], Nov. 2017.

[57] B. Beynek, Ş. Bora, V. Evren, and A. Ugur, "Synthetic Skin Cancer Image Data Generation Using Generative Adversarial Neural Network," International Journal of Multidisciplinary Studies and Innovative Technologies, vol. 5, no. 2, pp. 147–150, Nov. 2021.

[58] B. Ahmad, S. Jun, V. Palade, Q. You, L. Mao, and M. Zhongjie, "Improving Skin Cancer Classification Using Heavy-Tailed Student T-Distribution in Generative Adversarial Networks (TED-GAN)," Diagnostics, vol. 11, no. 11, p. 2147, Nov. 2021.

[59] V. K. Waghmare, and M. H. Kolekar, "Brain Tumor Classification Using Deep Learning," in Internet of Things for Healthcare Technologies, Jun. 2020, pp. 155–175.

[60] A. Çinar, and M. Yildirim, "Detection of tumors on brain MRI images using the hybrid convolutional neural network architecture," Medical Hypotheses, vol. 139, p. 109684, Jun. 2020.

[61] S. Chen, J. Zhang, X. Wei, and Q. Zhang, "Alzheimer's Disease Classification Using Structural MRI Based on Convolutional Neural Networks," in 2020 2nd International Conference on Big-data Service and Intelligent Computation, Dec. 2020, pp.7-13.

[62] V. Chouha, S.K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damaševičius, and V.H.C. De Albuquerque, "A Novel Transfer Learning Based Approach for Pneumonia Detection in Chest X-ray Images," Applied Sciences, vol. 10, no. 2, p. 559, Jan. 2020.

[63] C.J. Lin, and Y.C. Li, "Lung Nodule Classification Using Taguchi-Based Convolutional Neural Networks for Computer Tomography Images," Electronics,vol. 9, no. 7, p. 1066, Jun. 2020.

[64] A. S. Abdel Rahman, S. B. Belhaouari, A. Bouzerdoum, H. Baali, T. Alam, and A. M. Eldaraa, "Breast Mass Tumor Classification using Deep Learning," in 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), Feb. 2020, pp. 271-276.

[65] Q. A. Al-Haija, and A. Adebanjo, "Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network," in 2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS),Sep. 2020, pp. 1-7.

[66] A. Saber, M. Sakr, O. M. Abo-Seida, A. Keshk, and H. Chen, "A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique," IEEE Access, vol. 9, pp. 71194–71209, May 2021.

[67] K. Thurnhofer-Hemsi, and E. Domínguez, "A Convolutional Neural Network Framework for Accurate Skin Cancer Detection," Neural Processing Letters, vol. 53, no. 5, pp. 3073-3093, Oct. 2020.

[68] K. M. Hosny, M. A. Kassem, and M. M. Foaud, "Skin melanoma classification using ROI and data augmentation with deep convolutional neural networks," Multimedia Tools and Applications, vol. 9, no. 33, pp. 24029-24055, Jun. 2020.

[69] A. B. Bakht, S. Javed, R. Dina, H. Almarzouqi, A. Khandoker, and N. Werghi, "Thyroid Nodule Cell Classification in Cytology Images Using Transfer Learning Approach," in International Conference on Soft Computing and Pattern Recognition, Dec. 2021, pp. 539–549.

[70] W. Chen, Z. Gu, Z. Liu, Y. Fu, Z. Ye, X. Zhang, and L. Xiao, "A New Classification Method in Ultrasound Images of Benign and Malignant Thyroid Nodules Based on Transfer Learning and Deep Convolutional Neural Network," Complexity, vol. 2021, pp. 1–9, Sep. 2021.

[71] Eva-H. Dulf, M. Bledea, T. Mocan, and L. Mocan, "Automatic Detection of Colorectal Polyps Using Transfer Learning," Sensors, vol. 21, no. 17, p. 5704, Aug. 2021.

[72] Y. Bhanothu, A. Kamalakannan, and G. Rajamanickam, "Detection and Classification of Brain Tumor in MRI Images using Deep Convolutional Network," in 2020 6th International Conference on Advanced Computing and Communication Systems, Mar. 2020, pp. 248-252.

[73] W. M. Salama, and M. H. Aly, "Deep learning in mammography images segmentation and classification:

Automated CNN approach," Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701–4709, Oct. 2021.

[74] A. Khouani, M. El HabibDaho, S. A. Mahmoudi, M. A. Chikh, and B. Benzineb, "Automated recognition of white blood cells using deep learning," Biomedical Engineering Letters, vol. 10, no. 3, pp. 359–367, Jul. 2020.

[75] H. Yu, and X. Zhang, "Synthesis of Prostate MR Images for Classification Using Capsule Network-Based GAN Model," Sensors, vol. 20, no. 20, p. 5736, Oct. 2020.

[76] S. Kaur, H. Aggarwal, and R. Rani, "Diagnosis of Parkinson's disease using deep CNN with transfer learning and data augmentation," Multimedia Tools and Applications, vol. 80, no. 7, pp.10113-10139,Nov. 2020.

[77] B. Mondal, N. Das, K. C. Santosh, and M. Nasipuri, "Improved Skin Disease Classification Using Generative Adversarial Network," in 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS),Jul. 2020, pp. 520-525.

[78] T. Pang, J. H. D. Wong, W. L. Ng, and C. S. Chan, "Semi-supervised GAN-based Radiomics Model for Data Augmentation in Breast Ultrasound Mass Classification," Computer Methods and Programs in Biomedicine, vol. 203, p. 106018, May 2021.

[79] B. Ahmad, J. Sun, Q. You, V. Palade, and Z. Mao, "Brain Tumor Classification Using a Combination of VariationalAutoencoders and Generative Adversarial Networks," Biomedicines, vol. 10, no. 2, p. 223, Jan. 2022.

[80] Y. Li, Y. Chen, and Y. Shi, "Brain Tumor Segmentation Using 3D Generative Adversarial Networks," International Journal of Pattern Recognition and Artificial Intelligence, vol. 35, no. 4, p.2157002,Aug. 2020.

[81] A. Negi, A. N. J. Raj, R. Nersisson, Z. Zhuang, andM. Murugappan, "RDA-UNET-WGAN: An Accurate Breast Ultrasound Lesion Segmentation Using Wasserstein Generative Adversarial Networks," Arabian Journal for Science and Engineering, vol. 45, no. 8, pp. 6399–6410, Apr. 2020.

[82] C. Decourt, and L. Duong, "Semi-supervised generative adversarial networks for the segmentation of the left ventricle in pediatric MRI," Computers in Biology and Medicine, vol. 123, p. 103884, Aug. 2020.

[83] Z. Lou, W. Huo, K. Le, and X. Tian, "Whole Heart Auto Segmentation of Cardiac CT Images Using U-Net Based GAN," in 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Oct. 2020, pp. 192-196.

[84] X. Wu, L. Bi, M. Fulham, and J. Kim, "Unsupervised Positron Emission Tomography Tumor Segmentation via GAN based Adversarial Auto-Encoder," in 2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV), Dec. 2020, pp. 448-453.

[85] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," Scientific Reports, vol. 10, no. 1, pp. 1-12, Nov. 2020.

[86] E. Luz, P. Silva, R. Silva, L. Silva, J. Guimarães, G. Miozzo, G. Moreira, and D. Menotti, "Towards an effective and efficient deep learning model for COVID-19 patterns detection in X-ray images," Research on Biomedical Engineering, Apr. 2021, pp. 1-14.

[87] N. S. Punn, and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," Applied Intelligence, vol. 51, no. 5, pp. 2689-202, Oct. 2020.

[88] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "CovidGAN: DataAugmentation using Auxiliary Classifier GAN for Improved Covid-19 Detection," IEEE Access, vol. 8, pp. 91916-91923, May 2020.

[89] Y. Oh, S. Park, and J. C. Ye, "Deep Learning COVID-19 Features on CXR using Limited Training Data Sets," IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2688-2700, May 2020.

[90] N. Wang, H. Liu, and C. Xu, "Deep learning for the detection of COVID-19 using transfer learning and model integration," in 2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC), Jul. 2020, pp. 281-284.

[91] J. Li, D. Zhang, Q. Liu, R. Bu, and Q. Wei, "COVID-GATNet: A deep learning framework for screening of COVID-19 from chest X-ray images," in 2020 IEEE 6th International Conference on Computer and Communications (ICCC), Dec. 2020, pp. 1897-1902.

[92] M. Ahsan, M. Based, J. Haider, and M. Kowalski, "COVID-19 detection from chest X-ray images using feature fusion and deep learning," Sensors, vol. 21, no. 4, p.1480, Jan. 2021.

[93] A. S. Al-Waisy, S. Al-FahdawiS, M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. S. Maashi, M. Arif, and B. Garcia-Zapirain, "COVID-CheXNet: hybrid deep learning framework for identifying COVID-19 virus in chest X-rays images," Soft Computing,Nov. 2020, pp. 1-16.

[94] X. Li, W. Tan, P. Liu, Q. Zhou, and J. Yang, "Classification of COVID-19 Chest CT Images Based on Ensemble Deep Learning," Journal of Healthcare Engineering, vol. 2021, pp. 1–7, Apr. 2021.

[95] Y. Pathak, P. K. Shukla, and K. V. Arya, "Deep Bidirectional Classification Model for COVID-19 Disease Infected Patients," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 18, no. 4, pp. 1234–1241, Jul. 2021.

[96] M. J. Horry, S.Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, and N. Shukla, "COVID-19 Detection Through Transfer Learning Using Multimodal Imaging Data," IEEE Access, vol. 8, pp. 149808–149824, Aug. 2020.

[97] V. I. Iglovikov, A. Rakhlin, A. A. Kalinin, and A. A. Shvets, "Paediatric bone age assessment using deep convolutional neural networks," in Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Cham: Springer International Publishing, Sep. 2018, pp. 300-308.

[98] X. Pan, Y. Zhao, H. Chen, D. Wei, C. Zhao, and Z. Wei, "Fully Automated Bone Age Assessment on Large-Scale Hand X-Ray Dataset," International Journal of Biomedical Imaging,vol. 2020, pp. 1–12, Mar. 2020.

[99] M. A. Zulkifley, S.R. Abdani, and N.H. Zulkifley, "Automated Bone Age Assessment with Image Registration Using Hand X-ray Images," Applied Sciences, vol. 10, no. 20, p. 7233, Oct. 2020.

[100] Y.Gao, T. Zhu, and X. Xu, "Bone age assessment based on deep convolution neural network incorporated with segmentation," International Journal of Computer Assisted Radiology and Surgery, vol. 15, no. 12, pp.1951-1962, Sep. 2020.

[101] S. Li, B. Liu, S. Li, X. Zhu, Y. Yan, and D. Zhang, "A deep learning-based computer-aided diagnosis method of X-ray images for bone age assessment," Complex & Intelligent Systems, pp.1-11, Apr. 2021.
[102] I. Salim, and A. B. Hamza, "Ridge regression neural network for pediatric bone age assessment," Multimedia Tools and Applications, vol. 80, no. 20, pp. 30461–30478, May 2021.
[103] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A.B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, and F.C. Kitamura, "The RSNA Pediatric Bone Age Machine Learning Challenge," Radiology, vol. 290, no. 2, pp.498-503, Feb. 2019.
[104] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[105] AlexSWong, "AlexSWong/COVID-Net," GitHub,Feb. 2022, https://github.com/AlexSWong/COVID-Net.
[106] "RSNA Bone Age", www.kaggle.com. https://www.kaggle.com/datasets/kmader/rsna-bone-age.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 359-367
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Comparing the Semantic Segmentation of High-Resolution Images Using Deep Convolutional Networks: SegNet, HRNet, CSE-HRNet and RCA-FCN

Nafiseh Sadeghi[1,2], Homayoun Mahdavi-Nasab[*1,2], Mansoor Zeinali[1,2], Hossein Pourghassem[1,2]

[1].Department of Electrical Engineering, Najafabad Branch, Islamic Azad University, Najafabad, Iran
[2].Digital Processing and Machine Vision Research Center, Najafabad Branch, Islamic Azad University, Najafabad, Iran

Received: 11 Oct 2022/ Revised: 04 Mar 2023/ Accepted: 25 Apr 2023

Abstract

Semantic segmentation is a branch of computer vision, used extensively in image search engines, automated driving, intelligent agriculture, disaster management, and other machine-human interactions. Semantic segmentation aims to predict a label for each pixel from a given label set, according to semantic information. Among the proposed methods and architectures, researchers have focused on deep learning algorithms due to their good feature learning results. Thus, many studies have explored the structure of deep neural networks, especially convolutional neural networks. Most of the modern semantic segmentation models are based on fully convolutional networks (FCN), which first replace the fully connected layers in common classification networks with convolutional layers, getting pixel-level prediction results. After that, a lot of methods are proposed to improve the basic FCN methods results. With the increasing complexity and variety of existing data structures, more powerful neural networks and the development of existing networks are needed. This study aims to segment a high-resolution (HR) image dataset into six separate classes. Here, an overview of some important deep learning architectures will be presented with a focus on methods producing remarkable scores in segmentation metrics such as accuracy and F1-score. Finally, their segmentation results will be discussed and we would see that the methods, which are superior in the overall accuracy and overall F1-score, are not necessarily the best in all classes. Therefore, the results of this paper lead to the point to choose the segmentation algorithm according to the application of segmentation and the importance degree of each class.

Keywords: Semantic Segmentation; Convolutional Neural Network; Deep Neural Network; High-Resolution Image Processing.

1- Introduction

Segmentation is the task process of assigning a label to every pixel in the image, based on features such as pixel intensity, color, texture, etc [1]. Nowadays, the subject of interest is semantic segmentation, predict the semantic category of each pixel from a given label set.

There are learning and anti-learning methods frequently used for segmentation [2]. Anti-learning methods, typically include graph cuts, level set, region growing, etc. and learning methods include fuzzy, neural, genetic algorithms and derivations [3]. Various learning methods have been created and developed in recent years, due to their considerable success in learning a hierarchy of features from high to low [2,4]. These methods were inspired by human brain's ability to receive, learn, and organize input information, especially visual data [5].

Convolutional neural network (CNN) is a form of learning techniques, in which local neighborhood pooling operations and trainable filters are alternatingly applied on the input images, resulting in a hierarchy of increasingly complex features [6-8]. Convolution layers in CNN try to find patterns in an image by convolving over it. So CNN may detect nonlinear mappings between the inputs and outputs [9].

LeCun et al. (1998) introduced the first structure of convolutional neural networks named LenNet. In the same year, they received an award for simulating their proposed network on the ImageNet dataset. LeNet had six

✉ **Homayoun Mahdavi Nasab**
mahdavinasab@iaun.ac.ir

convolutional layers, a pooling layer, and two FC[1] layers [10].

After introducing basic convolutional neural network architectures, researches in this field were continued in two directions: some studies focused on designing new convolutional network architectures and others focused on implementing techniques and strategies to optimize existing architectures [11,12]. In the following, we will introduce some of the most important architectures that were proposed after the creation of convolutional neural network.

In 2012, Krizhevsky et al. proposed a CNN structure called AlexNet with five convolutional layers, three pooling layers, two normalization layers, and three FC layers [13]. The innovation of AlexNet was its use of ReLU to reduce training time. Some have criticized this structure for implementing very heavy data augmentation.

Then, Simonyan et al. (2014) designed a deeper network named VGG with smaller filter sizes [14]. They design two structures of VGG with 16 and 19 layers. The proposed structure showed promising performance and could also be generalized to other datasets. Although the architectures introduced so far focused on window size and smaller steps in the first convolutional layer, VGG focused on an important aspect of convolution neural networks, called depth [15].

The VGG architecture included 1×1 convolutional filters, acting as linear transformation of the input. All hidden layers of the VGG had a ReLU unit for reducing training time. To have no change in spatial resolution after convolving, this architecture kept the convolution step fixed on one pixel. VGG had three FC layers follow a stack of convolutional layers: the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (means a channel for each class). The final last layer is soft-max.

Since its developers believed that localized response normalization (LRN) increased memory consumption and training time with no significant resolution improvement, VGG did not use LRN.

In the same year, Szegedy et al. designed a deeper and more computationally-efficient network called GoogleNet [16]. It had twenty-two layers, no FC layer, and a new module called Inception to increase efficiency. The developers claimed that it was twelve times faster than AlexNet with increased depth and width at the same computational cost. Although there were benefits to the increased size, it also increased the number of parameters. This made the network more susceptible to overfitting, especially with a limited number of labeled samples in the dataset.

SegNet was an important architecture proposed in 2015 on a set of camera images [17]. The SegNet architecture was based on encoder-decoder network with thirteen convolutional layers in the VGG16. Since the decoder part of SegNet is identical to the VGG, it was possible to achieve the pre-training benefits in this architecture. The decoder block consisted of five sub-blocks, each with convolutional layers and a downsample layer. Likewise, the decoding block had five sub-blocks with deconvolutional layers and an upsample layer. In fact, the innovation of the SegNet architecture was its use of upsampling layers (reconstructing the image in the original dimensions). In terms of memory use, accuracy, and reducing network parameters, the SegNet architecture demonstrated excellent performance compared to other architectures [18].

HRNet was a successful network that used a parallel integration strategy [19]. The first stage in this network was a high-resolution subnetwork, then high-to-low resolution subnetworks one after another adding to form other stages. The multi-resolution subnetworks were connected in parallel. Each high-to-low resolution representation gets information from other parallel representations, again and again, resulting into a rich high-resolution representation. The convolutional layers in this network were placed in parallel from high to low accuracy [20]. The network had a main subnetwork that produced feature maps with the same accuracy and a series of step-by-step convolutions that reduced accuracy. This network has a multiresolution composition.

After HRNet, Wang proposed its enhanced model built on the backbone network of HRNet. It used NDRB as the generic extractor for multi-scale contextual features. So CSE-HRNet could resolve intra-class heterogeneity and inter-class homogeneity.

Another recent architecture is RCA-FCN consisted of two network units, namely the spatial relation module and the channel relation module [21]. These two modules learn and reason about global relationships between any two spatial positions or feature maps. So they produce RA[2] feature representations. In other word, this model convolutions combine spatial information and channel relation information to record both spatial and channel relations.

Here we aimed to compare segmentation performance with four recent successful methods, SegNet, HRNet, CSE-HRNet, and RCA-FCN. We tried to reference important prior contributions that claim to be particularly successful despite being simple and convolutional-based. This paper is useful to choose the best algorithm for semantic segmentation in our desired application. Thus, section 2 will discuss semantic segmentation and introduce the four structures in more detail. Section 3 will present the implementation and the dataset in more details, and the

[1] Fully-Connected

[2] Relation-Augmented

segmentation results will be compared. Finally, section 4 will present a summary of results.

2- Semantic Segmentation of ISPRS Images

The Vaihingen 2-D semantic segmentation dataset of ISPRS includes 33 images with 3-10 million pixels. These are True Ortho Photos (TOP) taken from Vaihingen, Germany. The ground sampling distance is 9 cm and all pixels in these images are labeled in six classes: building, car, road, tree, low veg, and clutter. The images are eight-bit .tiff files with three bands corresponding to green, red, and near-infrared.

In the following, four new and important semantic segmentation methods for high-resolution remote sensing images will be evaluated. All of these methods lead to six class segmentation, as shown in Fig. 1.

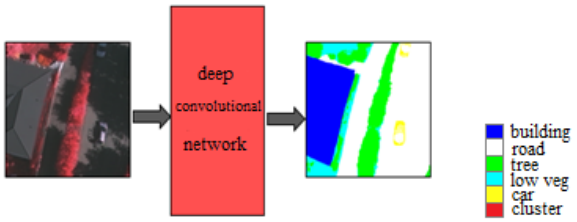


Fig. 1 Six Class Segmentation of Input Dataset

2-1- SegNet

The SegNet architecture was first introduced in 2015 for semantic segmentation on a set of camera images. Its topology was based on a decoder network with 13 convolutional layers in the VGG16 network. As with VGG16, this architecture can also achieve the benefits of pre-learning.

The SegNet topology is comprised of two parts: encoder and decoder. Each encoder performs convolution with a filter bank to produce a set of feature maps, and then they are batch normalized [22,23]. After that, an element-wise rectified linear activation function (ReLU), max(0, x), is applied. Next, a non-overlapping max-pooling, with the window size 2×2 and stride 2, is performed. The output is sub-sampled by a factor of 2. The purpose of using max-pooling is to achieve translation invariance over small spatial shifts in the input image.

Then the decoder upsamples the input feature maps using the memorized max-pooling indices from the corresponding encoder feature map. So sparse feature maps are produced, and then they are convolved with a trainable decoder filter bank. In this way dense feature maps will be produced. So the decoder upsamples and normalizes the stored feature maps. The softmax layer classifies each pixel independently and its output is an image with k channels, where k represents the number of classes.

Fig. 2 shows the schematic representation of the SegNet structure. Determination of boundaries was a success in SegNet architecture. It also showed great performance in terms of the number of network parameters [24], and the most important feature was its memory requirement, which was significantly lower than previous architectures. Therefore, due to its ability to quickly process a large area, SegNet matters when large-scale processing is necessary.

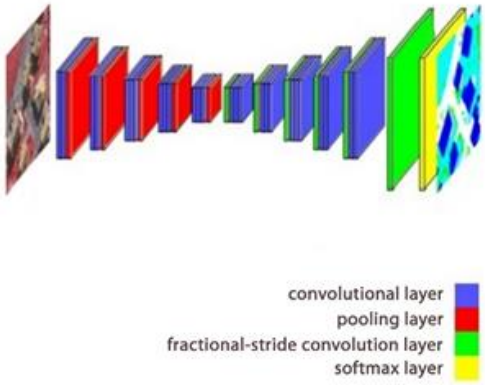


Fig. 2 SegNet Architecture [17]

2-2- HRNet

The main part of HRNet contains four stages with four parallel subnetworks. Each subnetwork is composed of a sequence of convolutions, also there is a down-sample layer across adjacent subnetworks to decrease the resolution to half. So the resolution is step by step decreased and the width (or the number of channels) is proportionally increased. The first stage contains four residual units. Each unit, the same as ResNet-50, is formed by a bottleneck with a width of 64, followed by one 3×3 convolution reducing the width of feature maps.

$$N_{11} \rightarrow N_{22} \rightarrow N_{33} \rightarrow N_{44}$$

In fact, Nsr represents the subnetwork in stage s, and r is the resolution index (Its resolution is $\frac{1}{2^{r-1}}$ of the resolution of the first subnetwork). It is obvious that the precision of each stage is $\frac{1}{2^{r-1}}$ of the first subnetwork's precision.

$$\begin{array}{ccccccc} \mathcal{N}_{11} & \rightarrow & \mathcal{N}_{21} & \rightarrow & \mathcal{N}_{31} & \rightarrow & \mathcal{N}_{41} \\ & \searrow & \mathcal{N}_{22} & \rightarrow & \mathcal{N}_{32} & \rightarrow & \mathcal{N}_{42} \\ & & & \searrow & \mathcal{N}_{33} & \rightarrow & \mathcal{N}_{43} \\ & & & & & \searrow & \mathcal{N}_{44}. \end{array}$$

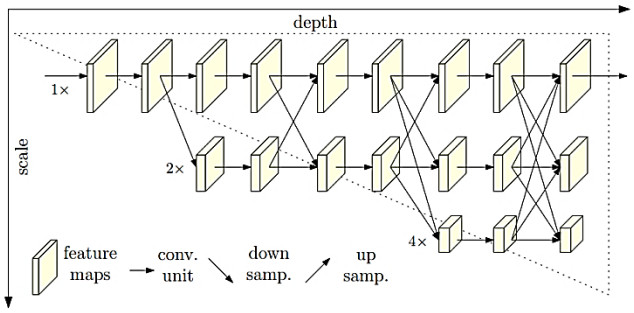


Fig. 3 HRNet architecture [20]

The multi-resolution subnetworks are in parallel. The resolution of parallel subnetworks in each stage will include the resolution of the previous stages and one stage below. An example of a network structure with four subnetworks is shown below:

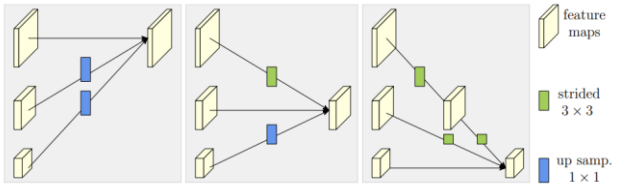


Fig. 4 From left to Right, the Exchange unit Aggregates the Information for High, Medium, and low Resolutions [20]

There are two versions of this network, HRNet-W32 and HRNet-W48. Here, 32 and 48 respectively represent the width (C) of the high-resolution subnetworks in the last three stages. The other three parallel subnetworks have widths of 64, 128, 256 for HRNet-W32 and 96, 192, and 384 for HRNet-W48. HRNet keeps high-resolution representation on the main stem throughout the network and lower-resolution parallel stems are produced via downsampling operations.

2-3- CSE_HRNet

Sometimes objects of the same class in aerial images acquired with high spatial resolutions show various shapes, scales, colors, and structures. Fig. 5 demonstrates some examples of this issue namely intra-class heterogeneity. In Fig. 5(a) cars have different colors, although they all belong to the car class. Similarly, in Fig. 5(b) buildings of the same category vary in texture and shape.

Meanwhile, we may face objects of the different classes having the same colors or interacted with cast shadows that present similar visual characteristics. This would lead to inter-class homogeneity problem.

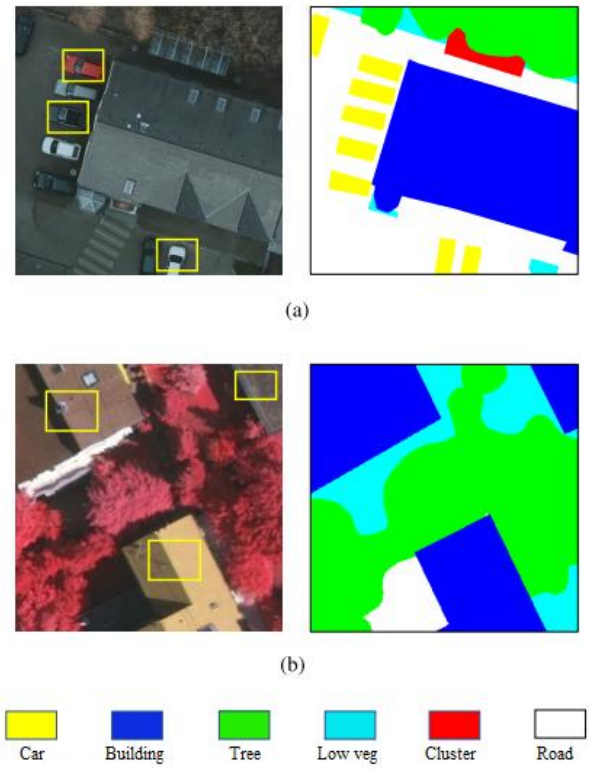


Fig. 5 Intra-class Heterogeneity (a) Cars have Different Colors (b) Buildings are Different in Appearance [25]

Fig. 6 shows objects which are similar in appearance while they should be categorized into separate semantic classes [26,27]. This issue named inter-class homogeneity is shown in Fig. 6. In Fig. 6(a) there are some areas of low veg and trees, which belong to two separate classes, have similar appearances [26,27]. Also, in Fig. 6(b) there are buildings and impervious surfaces are quite similar in appearance. These confusing objects pose extreme challenges for accurate and coherent segmentation [25,28].

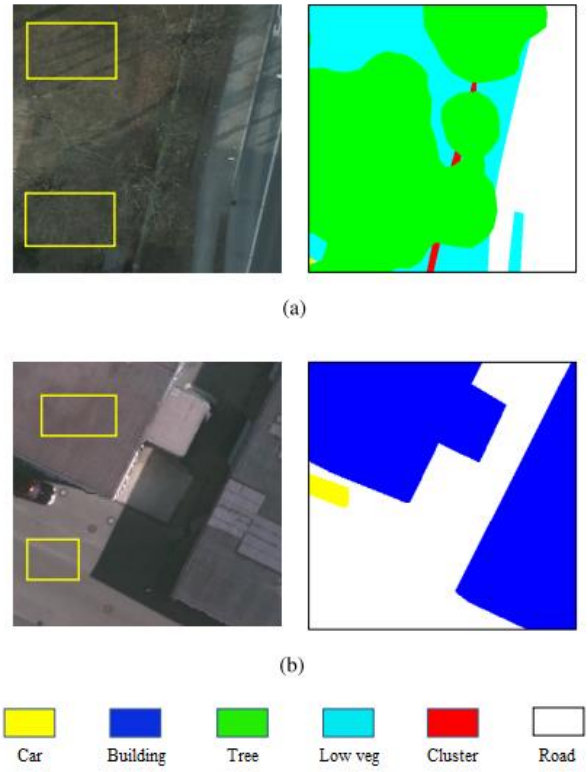


Fig. 6 Inter-Class Homogeneity (a) Trees and low veg are Similar (b) Buildings and Impervious Surface are Analogous [25]

The CSE-HRNet architecture was designed based on the backbone network of HRNet-W32. As with HRNet-W32, "W32" in CSE_HRNet32 represents the feature dimensions of high-resolution or the number of channels representations in the main sub-branch, and the number of other parallel channels will be 64, 128, and 256.

The pyramid structure can exploit the inherent multi-level features, and provide adequate semantic knowledge at all levels. So, the pyramidal feature hierarchy was introduced in this architecture to enhance multi-level semantic representations of the model [29-33]. CSE-HRNet can resolve intra-class heterogeneity and inter-class homogeneity simultaneously by using NDRB combined with the pyramidal multi-level feature hierarchy.

The hierarchy adopts a four-level top-down architecture where the strided convolution as the downsampling method is applied (the stride is set to 2). Widths and heights of feature maps (spatial resolutions) are then reduced by half after each downsampling, whereas the numbers of channels (feature dimensions) are doubled.

The first-level feature map of the pyramid is directly fed into the main high-resolution branch of the network. The second-, third-, and fourth- level feature maps are fused with the counterparts from the multi-resolution branches via the element-wise addition.

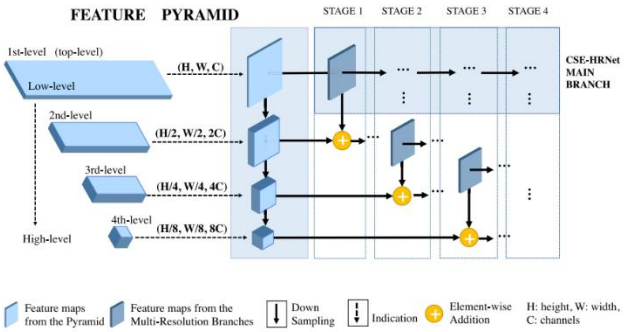


Fig. 7 CSE_HRNet Architecture [25]

2-4- RCA_FCN

Although it has been recognized that contextual relation can offer important cues for semantic segmentation tasks, but using convolution operations in prior convolutional neural networks leads to failure in modeling contextual spatial relations, due to their local valid receptive field [34-39]. However, some convolutional algorithms tried to address this problem using spatial propagation modules or graphical models, but they seek to capture global spatial relations implicitly with a chain propagation way. The effectiveness of these methods depends highly on the learning impact of long-term memorization [40]. Consequently, such models don't work well when long-range spatial relations exist. So, these models most of the time fail to capture long-range spatial relationships between entities, which leads to spatial fragmented prediction [26].

The most important goal of designing the RCA-FCN architecture is to solve spatial relation problems and access channel information. This structure introduces simple effective network units, namely, the spatial relationships module and the channel relationships module. So it can learn and reason about global relationships between every two feature maps or spatial positions, and produce RA feature representations. This network takes VGG16 as a backbone for multilevel feature extraction. Fig. 8 shows a representation of this architecture.

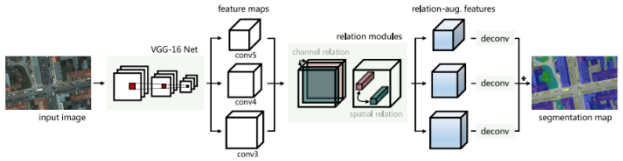


Fig. 8 Overview of the Relation Module [26]

As shown in Fig. 9, outputs of convolve3, convolve4, and convolve5 were fed into the channel relationships module and the spatial relationships module for generating RA features. Then these features were fed into convolutional

layers with 1×1 filters to squash the number of channels to the number of categories.

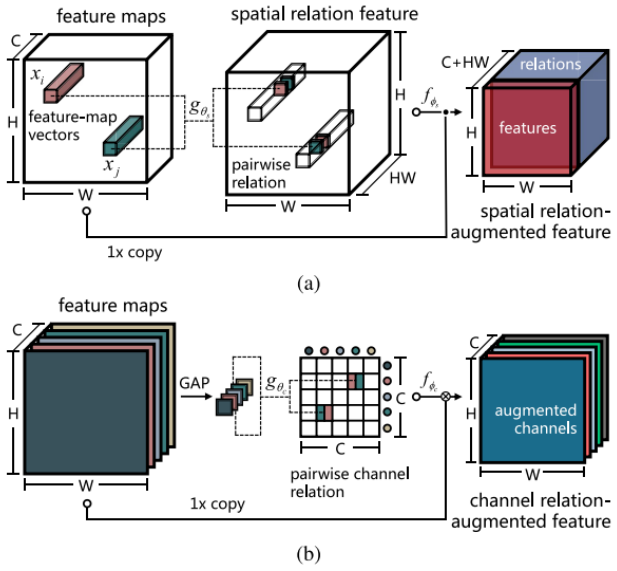


Fig. 9 (a) spatial relation module (b) channel relation module [26]

The convolved feature maps were finally upsampled to desired full resolution and element-wise added to generate final segmentation maps.

3- Implementation

The ISPRS segmentation dataset in Vaihingen was used in this implementation. This dataset is consist of 33 images collected over a 1.38 km^2 area and the average image size of 2494 × 2064 pixels. The spatial resolution is 9 cm. They have green (G), red (R), and near infrared (NIR) bands. Vaihingen dataset was provided by ISPRS-Commission III [26]. Images were captured using digital aerial cameras and mosaicked with Trimble INPHO OrthoVista [41].
Due to the large and diverse dimensions of images in the dataset, five random 240×240 crop were created from each image. Thus, a dataset of images with the same 240×240 resolution was obtained. This dataset was divided into the training dataset and the test dataset. So 60% of images were randomly selected and allocated to the training dataset and remaining 40% allocated to the test dataset. We train the four studied networks architectures in MATLAB 2021.

4- Comparison

The following metrics will be necessary for comparing and evaluating the segmentation performance of these models. They will be explained as follows:

4-1- Accuracy

Accuracy is the percentage of correctly classified instances, or in other words, the ratio of the true results to the total number of cases examined [42,43]. This factor cannot differentiate between FN and FP error and considers them the same.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+FP+TN}} \qquad (1)$$

4-2- Precision

This factor can determine how many of the correctly predicted cases really turned out to be positive [44]. Precision usually uses when the False Positive is a higher concern than the False Negatives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2)$$

4-3- Recall

Recall determines how many True Positive cases can be predicted correctly with our model. This factor is also called sensitivity and it is a good choice for the unbalanced classes.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (3)$$

4-4- F1

F1 can give a combined idea about two metrics, Precision and Recall. It is maximum when the Precision becomes equal to the Recall.

$$F1 = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}} \qquad (4)$$

In the following, the segmentation results of each model is presented according to the aforementioned metrics. Table (1) shows that no model is conclusively superior to others in terms of segmentation accuracy. A model may have high accuracy in some classes and wouldn't be so in some other classes. For example, although SegNet shows high accuracy in the car class, but it's not very good in classes such as tree and low veg. Compared to the other models, the RCA-FCN was also the most accurate in the building, tree, and low veg classes.

Table 1: Segmentation Accuracy for Each Class

Method	***classes accuracy (%)***				
	building	***car***	***road***	***tree***	***low veg***
SegNet	76.74	**95.93**	81.26	64.52	63.59
HRNet	93.34	63.32	87.35	84.97	74.96
CSE-HRNetsi	94.07	63.34	**88.03**	85.47	73.45
RCA-FCN	**94.12**	70.81	87.22	**89.25**	**87.67**

After observing the segmentation accuracy of each model in different classes, Table (2) shows the overall accuracy (OA) of the networks. The table shows that the CSE-

HRNet algorithm's overall accuracy is superior to the others, followed closely by RCA-FCN in second position.

Table 2: Segmentation Overall Accuracy

method	*Overall accuracy (%)*
SegNet	76.41
HRNet	85.06
CSE-HRNet	**89.23**
RCA-FCN	89.03

F1 is another metric for evaluating the performance of segmentation algorithms, and Table (3) shows its values for different algorithms. As what is said about accuracy, Table (3) clearly shows that no model is definitively superior in all classes. In terms of this metric, SegNet has not performed well in any class. HRNet has the highest F1 (88.94% and 83.19%) in the tree and low veg classes, and CSE-HRNet has the highest F1 (95.41% and 91.92%) in the building and road classes. However, the F1 score of these two architecture for other classes has minor differences with the maximum value. RCA-FCN achieved the best performance with the car class (87.16%). Therefore, one of these networks can be selected for segmentation based on the importance classes.

Table 3: Segmentation F1 for Each Class

Method	*Classes F1-score (%)*				
	building	*car*	*road*	*Tree*	*low veg*
SegNet	71.12	57.89	68.20	34.70	34.98
HRNet	92.91	84.28	91.68	**88.94**	**83.19**
CSE-HRNet	**95.41**	86.79	**91.92**	88.53	80.18
RCA-FCN	94.86	**87.16**	91.01	88.74	80.01

Table (4) shows an overall comparison of these models in terms of F1 and suggests that with an overall F1 score of 89.36%, HRNet can be considered the best network architecture.

Table 4: Segmentation Overall F1

Method	*Overall F1-score (%)*
SegNet	65.79
HRNet	**89.36**
CSE-HRNet	88.57
RCA-FCN	88.36

5- Conclusion

A network can be excellent for distinguishing a specific class of an image dataset and perform poorly in detecting the other classes from the dataset.

Unlike SegNet, the HRNet and CSE-HRNet architectures had generally acceptable results in the F1 and accuracy factors. The RCA-FCN structure can also be considered important not only for its near-ideal evaluation with the general factors, but for properly distinguishing some small classes, such as car and tree.

Finally, selecting the best structure for segmentation is fully dependent on image type and the class' importance for different applications. For studying the state of regional roads and traffic, a model with good accuracy for distinguishing the car and road classes is preferable. However, if the goal is to study the regional vegetation, the segmentation performance of the tree and low veg classes becomes more important.

References

[1] K. Farajzadeh, E. Zarezadeh, J. Mansouri, "Concept detection in images using SVD features and multi-granularity partitioning and classification", Journal of Information Systems & Telecommunication (JIST), 2017, pp. 172.

[2] M.J. Hasan, M. Sohaib, J.M. Kim, "An explainable ai-based fault diagnosis model for bearings", Sensors, 2021, Vol. 21, No. 12, pp. 4070.

[3] M. Ahmad, S. F. Qadri, S. Qadri, I. A. Saeed, S. S. Zareen, Z. Iqbal, A. Alabrah, H. M. Alaghbari, M. Rahman, S. A. Md, "A lightweight convolutional neural network model for liver segmentation in medical diagnosis", Computational Intelligence and Neuroscience, 2022.

[4] M. S. Al-Rakhami, M. M. Islam, M. Z. Islam, A. Asraf, A. H. Sodhro, and W. Ding, "Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning", MedRxiv, 2020, pp. 20181339.

[5] M. Islam, "An efficient human computer interaction through hand gesture using deep convolutional neural network", SN Computer Science, 2020, Vol. 1, No. 4, pp. 1-9.

[6] W. Li. R. Zhang, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation", NeuroImage, 2015, Vol. 108, pp. 214-224.

[7] A. Sandooghdar, F. Yaghmaee, "Deep Learning Approach for Cardiac MRI Images", Journal of Information Systems and Telecommunication (JIST), 2022, Vol. 1, No. 37, pp. 61.

[8] E. Gholam, S.R. Kamel Tabbakh, "Diagnosis of Gastric Cancer via Classification of the Tongue Images using Deep Convolutional Networks", Journal of Information Systems and Telecommunication (JIST), 2021, Vol. 3, No. 35, pp. 191.

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradientbased learning applied to document recognition", Proceedings of the IEEE, 1998, Vol. 86, No. 11, pp. 2278-2324.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition", Proceedings of the IEEE, 1998, VOL. 86, No. 11, pp. 2278-2324.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection

and segmentation", IEEE transactions on pattern analysis and machine intelligence, 2015, Vol. 38, No. 1, pp. 142-158.

[12] N. Audebert, B. Le Saux, and S. Lef`evre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks", ISPRS Journal of Photogrammetry and Remote Sensing, 2018, Vol. 140, pp. 20-32.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Advances in neural information processing systems, 2012, Vol. 25.

[14] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[15] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning", Neurocomputing, 2022, Vol. 493, pp. 626-646.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", in Thirty-first AAAI conference on artificial intelligence, 2017.

[17] V. Badrinarayanan, A. Handa, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling", arXiv preprint arXiv: 1505.07293, 2015.

[18] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", IEEE transactions on pattern analysis and machine intelligence, 2017, Vol. 39, No.12, pp. 2481-2495.

[19] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, "High-resolution representations for labeling pixels and regions", arXiv preprint arXiv:1904.04514, 2019.

[20] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation", in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5693-5703.

[21] D. Marmanis, J. D. Wegner, S. Galliani, K. Schindler, M. Datcu, and U. Stilla, "Semantic segmentation of aerial images with an ensemble of CNSS. ISPRS Annals of the Photogrammetry", Remote Sensing and Spatial Information Sciences, 2016, Vol. 3, pp. 473-480.

[22] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", In International conference on machine learning, 2015, pp. 448-456.

[23] V. Badrinarayanan, B. Mishra, and R. Cipolla, "Understanding symmetries in deep networks", arXiv preprint arXiv:1511.01029, 2015.

[24] H. Zamanian, H. Farsi, S. Mohamadzadeh, "Improvement in accuracy and speed of image semantic segmentation via convolution neural network encoder-decoder", Information Systems & Telecommunication (JIST), 2018, Vol. 6, No. 3, pp. 128-135.

[25] F. Wang, S. Piao, and J. Xie, "CSE-HRNet: A context and semantic enhanced high-resolution network for semantic segmentation of aerial imagery", IEEE Access, 2020, Vol. 8, No. 2, pp. 182475-182489.

[26] L. Mou, Y. Hua, and X. X. Zhu, "Relation matters: Relational context-aware fully convolutional network for semantic segmentation of high-resolution aerial images", IEEE Transactions on Geoscience and Remote Sensing, 2020, Vol. 58, No. 11, pp. 7557-7569.

[27] H. Luo, C. Chen, L. Fang, X. Zhu, and L. Lu, "High-resolution aerial images semantic segmentation using deep fully convolutional network with channel attention mechanism", IEEE journal of selected topics in applied earth observations and remote sensing, 2019, Vol. 12, No. 9, pp. 3492-3507.

[28] N. Mboga, S. Georganos, T. Grippa, M. Lennert, S. Vanhuysse, and E. Wolff, "Fully convolutional networks and geographic object-based image analysis for the classification of VHR imagery", Remote Sensing, 2019, Vol. 11, No. 5, pp. 597.

[29] G. Zhang, T. Lei, Y. Cui, and P. Jiang, "A dual-path and lightweight convolutional neural network for high-resolution aerial image segmentation", ISPRS International Journal of Geo-Information, 2019, Vol. 8, No. 12, pp. 582.

[30] Z. Tu, X. Chen, A. L. Yuille, and S. C. Zhu, "Image parsing: Unifying segmentation, detection, and recognition", International Journal of computer vision, 2005, Vol. 63, No. 2, pp. 113-140.

[31] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman, "Using multiple segmentations to discover objects and their extent in image collections", In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006, Vol. 2, pp. 1605-1614.

[32] E. Borenstein, and S. Ullman, "Combined top-down/bottom-up segmentation", IEEE Transactions on pattern analysis and machine intelligence, 2008, Vol. 30, No. 12, pp. 2109-2125.

[33] J. Wu, J. Zhu, and Z. Tu, "Reverse Image Segmentation: A High-Level Solution to a Low-Level Task", In BMVC, 2014.

[34] Q. Zhao, and L. D. Griffin, "Better image segmentation by exploiting dense semantic predictions", arXiv preprint arXiv:1606.01481, 2016.

[35] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks", In Proc. IEEE Int. Conf. Mach. Learn. (ICML), 2011, pp. 129-136.

[36] J. Yao, S. Fidler, and R. Urtasun, "Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation", In IEEE conference on computer vision and pattern recognition, 2012, pp. 702-709.

[37] A. Kae, K. Sohn, H. Lee, and E. Learned-Miller, "Augmenting CRFs with Boltzmann machine shape priors for image labeling", In Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2019-2026.

[38] H. Myeong, and K. M. Lee, "Tensor-based high-order semantic relation transfer for semantic scene segmentation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3073-3080.

[39] J. J. Corso, "Toward parts-based scene understanding with pixel-support parts-sparse pictorial structures", Pattern Recognition Letters, 2013, Vol. 34, No. 7, pp. 762-769.

[40] Q. Li, Y. Shi, and X. Huang, "Building footprint generation by integrating convolution neural network with feature pairwise conditional random field (FPCRF)", IEEE Transactions on Geoscience and Remote Sensing, 2020, Vol. 58, No. 11, pp. 7502-7519.

[41] M. Cramer, "The DGPF-test on digital airborne camera evaluation overview and test design", Photogrammetrie-Fernerkundung-Geoinformation, 2010, pp. 73-82.
[42] M.J. Hasan, J.M. Kim, "Bearing fault diagnosis under variable rotational speeds using stockwell transform-based vibration imaging and transfer learning", Applied Sciences, Vol. 8, No. 12, pp. 2357.
[43] M.J. Hasan, J. Uddin, S.N. Pinku, "A novel modified SFTA approach for feature extraction", In 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 2016, pp. 1-5.
[44] M. Ghasemi, M. Kelarestaghi, F. Eshghi, A. Sharifi, "D 3 FC: deep feature-extractor discriminative dictionary-learning fuzzy classifier for medical imaging", Applied Intelligence, 2022, pp. 1-17.

Journal of Information Systems and Telecommunication
Vol.11, No.4, October-December 2023, 368-382
http://jist.acecr.org
ISSN 2322-1437 / EISSN:2345-2773

Software-Defined Networking Adoption Model: Dimensions and Determinants

Elham Ziaeipour[1], Ali Rajabzadeh Gatari[2*], Alireza Taghizadeh[3]

[1].Department of Management of Information Technology, Faculty of Management and Economics, Science and Research Branch, Islamic Azad University, Tehran, Iran
[2].Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran. (Visiting Professor at SRb)
[3].Faculty of Computer Science, Parand Branch, Islamic Azad University, Tehran, Iran

Received: 21 Nov 2022/ Revised: 4 Feb 2023/ Accepted: 20 Feb 2023

Abstract

The recent technical trend in the field of communication networks shows a paradigm change from hardware to software. Software Defined Networking as one of the enablers of digital transformation could have prominent role in this paradigm shift and migration to Knowledge-based network. In this regard, telecom operators are interested in deploying SDN to migrate their infrastructure from a static architecture to a dynamic and programmable platform. However, it seems that they do not consider SDN as one of their priorities and still depend on traditional methods to manage their network (especially in some developing countries such as Iran). Since the first step in applying new technologies is to accept them, we have proposed a comprehensive SDN adoption model with the mixed-method research methodology. At first, the theoretical foundations related to the research problem were examined. Then, based on Grounded theory, in-depth interviews were conducted with 12 experts (including university professors and managers of the major telecom operators).
In result, more than a thousand initial codes were determined, which in the review stages and based on semantic commonalities, a total of 112 final codes, 14 categories and 6 themes have been extracted using open, axial and selective coding. Next, in order to confirm the indicators extracted from the qualitative part, the fuzzy Delphi method has been used. In the end, SPSS and Smart-PLSv.3 software were used to analyze the data collected from the questionnaire and to evaluate the fit of the model as well as confirm and reject the hypotheses.

Keywords: Adoption; Fuzzy Delphi; Grounded Theory; Service provider; Software Defined Network; Technology.

1- Introduction

The trend of technology in communication networks clearly shows that the different sectors of this industry are transforming from hardware to software [1]. In this regard, in addition to witnessing the development of software in various sectors, the architecture and overall image of the communication network space will also undergo transformation and change [1].
The importance of software development in telecommunications industry was highlighted during the international conference which was held at St. Louis University by the IEEE Organization in 2018. In this conference, SDN technology was identified as the most vital part among 11 key aspects of this trend [1].
By separating the control plane, SDN redefines network architecture and provides a flexible way to manage and control complex networks through efficient management of resources [3].
SDN will provide some important benefits such as new model of service creation and delivery, efficient and software-based management of resource and energy, agility and quick response to changes, operation automation and customization. In addition, SDN manages all domains, layers and vendors in an integrated fashion, which is able to analyze traffic, detect failures, respond promptly to user needs and reduce downtime. It also provides the possibility of using cognitive techniques, optimization and virtualization, reducing implementation costs, independence of services from the underlying hardware layer, faster procurement and equipment configuration time, increasing network intelligence [4]. Software Defined Networks will not only facilitate the fulfillment of the promises of other technologies such as 5G and Cloud computing, but also as an enabler [4][5], it will play a key role in providing concepts such as digital transformation, knowledge-based networks and business intelligence [6].
The importance of this research is due to the benefits of SDN and the global trend towards this technology, although telecom operators in some developing countries

✉ **Ali Rajabzadeh Gatari**
alirajabzadeh@modares.ac.ir

such as Iran are still hesitant to use this technology. It seems that the administrators prefer to manually configure and set up their networks instead of getting rid of their legacy networks and using new technology [4]. However, according to Martec's law, the main challenge of managing organizations is that technological changes are exponential and very fast, while the internal changes of organizations are slow. In this way, this distance becomes more and more over time, so that finally organizations must adjust themselves and align with new technologies [7].

The first step in applying new technologies is to accept them. Identifying the key factors of the information technology adoption is an important research area. In other words, the question why people accept and use a technology or conversely do not, is one of the most important issues in information systems [8].

Technology adoption is a multidimensional phenomenon that includes a wide range of key variables such as values, perceptions, personal perspectives, intellectual preferences, beliefs, attitudes, desires and the degree of their involvement with technology as well as the ability to change on adopting new technology [9][10]. In recent decades, in accordance with the development of information technology, several models have been introduced in the area of technology adoption [10]. According to [12], each of these models has its shortcoming and boundaries and does not complementary to the rest of the models [13]. There are two important concerns with these theories. First, each theory uses distinct term in its constructs, although they are basically under the same concepts. Secondly, there is not even a single theory that covers all behavioral variables [12][13] [13]. In General, the research shows that these models can be developed according to different technologies, context and situation of each country.

To the best of our knowledge, In Iran there has not been any previous proper activity to SDN adoption to date. However, a few studies have been done on SDN from "technical point of view".

One of the reasons for the lack of appropriate research in this field is mainly due to its multidisciplinary nature.

This study intends to fill the existing theoretical gap by identifying the aspects of SDN adoption and examining the different dimensions of this process. In this regard, in literature review, three main topics have been studied:1) the theoretical foundations of SDN technology 2) the theoretical foundations of technology acceptance models and 3) researches on SDN adoption models (Fig.1).

In the first stage, all parameters extracted from the literature review were considered as a basis for preparing the initial questionnaire. Next, based on Grounded theory, an in-depth interview was conducted with 12 experts, which with using open, axial and selective coding the factors affecting SDN adoption have been recognized. Therefore, a proposed model for SDN adoption was presented based on the extracted codes and the Strauss and Corbin's paradigm model [14].

The methodology of research, data collection and analysis in addition to the identified dimensions and categories are presented in the relevant sections.

Then, in order to confirm the indicators extracted from the qualitative part, the fuzzy Delphi method has been used.

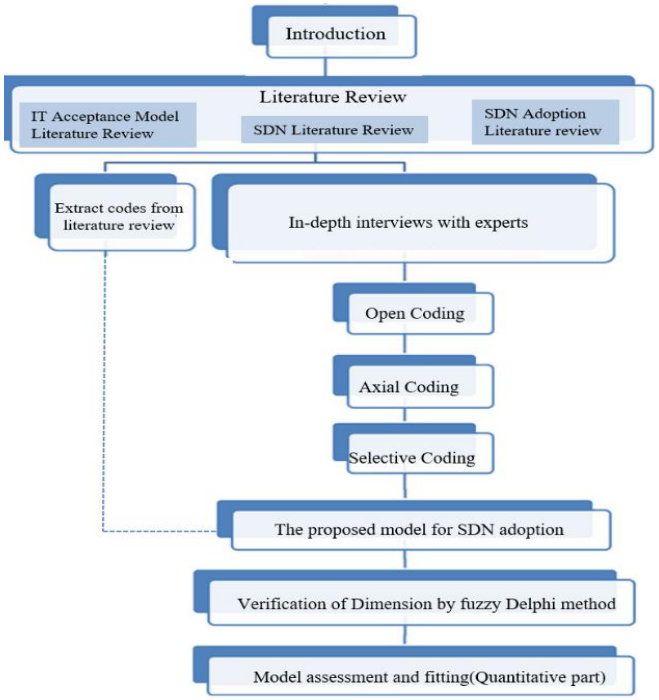


Fig.1: Steps of conducting research

In the quantitative part, by analyzing the data extracted from the questionnaire (with a Likert scale), the structure and fit of the model has been checked. Also, CFA[1], confirmation and rejection of hypotheses has also been done using PLS[2] software. Finally, after the validation, conclusions and suggestions for future research are given.

The expected outcome of current research can be also useful for countries that have similar telecom and geopolitical context. Also, same as the adoption of other technologies such as cloud, 5G, IOT, Big data, etc. it will be useful for telecommunication operators and all entities in the SDN ecosystem (Fig.1).

2- Literature Review

In literature review, three main topics have been studied:1) the theoretical foundations of SDN technology 2) The theoretical foundations of technology acceptance models and 3) researches on SDN adoption models.

[1] Confirmatory Factor Analysis
[2] partial least squares

2-1- Theoretical Foundations of SDN Technology

Software Defined networks are an alternative architecture to common inflexible networks. Generally, SDN can be simply defined as "a new generation of networks that uses software-based switches and controllers alongside high-level APIs to control and manage network infrastructure" [15].
The SDN architecture separates the control plane from the transmission plane and provides an abstraction of the network infrastructure for services and applications. It also allows network programming using the open APIs [16].
SDN networks consist of three main layers with north and south interfaces for communication between layers, as well as western and eastern interfaces for communication with other SDN networks or non-SDN networks [15][17].
Standard architectures for SDN have been provided by standardization organizations such as ITU[۱], OIF[۲], IETF[۳], ETSI[۴], ONF[۵], etc., the most common of which is the architecture provided by ONF.

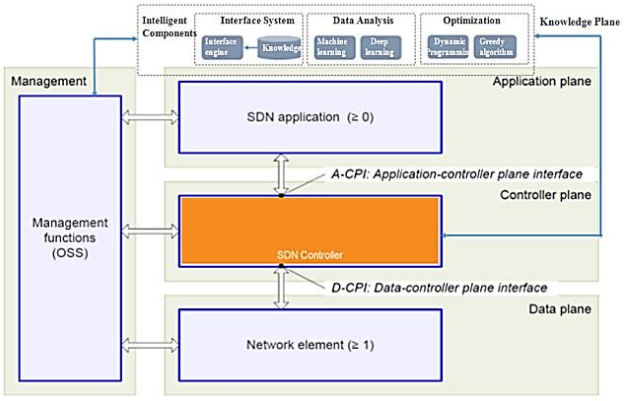


Fig.2: ETSI/ONF SDN architecture (ETSI, 2017) [17]

ETSI has also used ONF reference model and added the knowledge layer to it [17]. In this way, machine learning and cognitive techniques will be used along with neural networks, and operations such as learning, normalization, analysis, decision making, forecasting, judgment, and knowledge extraction will be performed in this intelligent layer (Fig.2) [17].
The new SDN architecture was introduced in 2019 to bring open-source architecture to SDN [18], and work is ongoing to operationalize it by standards-setting organizations.
In 2019, the SDN share was $9.995 million from the global market. Its value could be up to $72.630 million by 2027 (growing at a CAGR of 28.2% from 2020 to 2027) [3] and at $112.95 million by 2028 [19].
The software defined networking market has different sectors based on components, organization size, end users, industry vertical and regions. Telecom service providers, cloud service providers and companies are considered as the end users of this technology [3]. Also, from regional point of view, North America is a major SDN market due to ever increasing network traffic, mobility solutions and cloud applications [19].
Meanwhile, shift toward cloud by various organizations would increase the adoption of SDN among cloud providers and brings new opportunity for the market [3].
Thus, many students, research centers and leading vendors are working on various aspects of this technology. The main concern of the current research is lack of desire to implement SDN in practice.

2-2- Theoretical Foundations of IT Acceptance

Since the 80's, in parallel with the development of information technology in organizations, numerous researches in the field of technology adoption have been done. The foundation of all technology adoption models is shown in (Fig.3) [20].

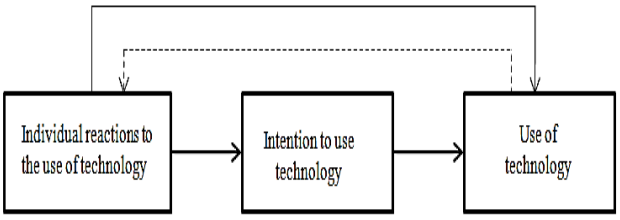


Fig.3: fundamentals of technology adoption models [20]

Technology adoption is a process that begins with the user's awareness of the technology and ends with the user embracing the technology and its full use [21]. According to Rogers (1995), the adoption of a technology involves a rational process in which investment decisions are made about the use of that new technology. Technology acceptance is an Individual's attitude towards a technology [21]. It looks more like acceptance would refer to the intentions towards use and adoption refers to the degree of actual use. Therefore, Technology acceptance is the first step of technology adoption (Fig.4) [21] [22].

Fig.4.Innovation Diffusion Process [22]

Table 1 lists some of the common information technology acceptance models and their determinants of IT adoption [23][11]. The last column of the table shows peer categories of our suggested model which are based on the Strauss-Corbin model.

Table1: Evolution of Theories and Models of Technology Adoption

[1] International Telecommunication Union-Telecommunications Standardization
[2] Organization international de la Francophone
[3] Internet Engineering Task Force
[4] European Telecommunications Standards Institute
[5] Open Networking Foundation

Year	*Theory/Model*	*Developed By*	*Determinants of IT adoption*	*categories*
1975	Theory of Reasoned Action (TRA)	Ajzen & Fishbein	Attitude, Subjective Norm, Behavioral Intention, Behavior	Causal, Phenomenon, Factors
1991	Theory of Planned Behavior (TPB)	Ajzen	Attitude, Subjective Norm, Perceived Behavioral Control, Intention	Causal, Phenomenon, Factors
1989	Social Cognitive Theory (SCT)	Bandura	Personal Factors (Cognitive, affective and biological events), Environmental Factors, Behavior	Causal, Phenomenon, Contextual Factors
1992	Motivation Model (MM)	Davis et al.	Extrinsic Motivation, Intrinsic Motivation, Emotional style	Casual Factors
1983	DOI	Rogers & Shoemaker	Knowledge, Persuasion, Decision, Implementation, Confirmation	Causal, Phenomenon, Strategies Factors
1990	TOE	Tornatzky& Fleischer	Technological Context (Availability, Characteristics, Internal and External) Environment Context (Industry Characteristics, Government Role, Completion, Structure) Organizational Context (Size, Process and Practices, Linking structures (formal and informal)	Causal, Intervening, Contextual Factors
1989	TAM	Davis	External Variables, Perceived Usefulness, Perceived Ease of Use, Attitude, Intension to use, Actual Use	Causal, Phenomenon, Factors
2000	TAM2	Venkatesh & Davis	Subjective Norm, Experience, Image, Job Relevance, Output Quality, Perceived Usefulness, Perceived Ease of Use, Intension to use, Usage Behavior	Causal, Phenomenon, Factors
2008	TAM3	Venkatesh etal.	Subjective Norm, Experience, Image, Job Relevance, Output Quality, Perceived Usefulness, Perceived Ease of Use, Intension to use, Usage Behavior Computer Anxiety, Computer playfulness, Perception of external control, Computer self-efficiency	Causal, Phenomenon, Factors
2003	UTAUT	Venkatesh, Morris, Davis	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Gender, Age, Experience, Voluntariness of Use, Behavioral Intention, Use Behavior	Causal, Phenomenon, Factors
2012	UTAUT2	Venkatesh,Thong & Xu	Performance Expectancy, Effort Expectancy, Social Influence, Facilitating Conditions, Price Value, Habit, Hedonic Motivation, Gender, Age, Experience, Behavioral Intention, Use Behavior	Causal, Phenomenon, Factors

2-3- Literature Review of SDN Adoption Models

Numerous articles have addressed SDN technology from different dimensions such as optimization, simulation, implementation, development, etc. This amount of study on SDN clearly shows its important from the perspective of academic societies. Table 2 shows a brief review of articles related to the research topic and the parameters extracted from them. The last column of the table shows peer categories of our suggested model which are based on the Strauss-Corbin model.

All the options considered in this section have examined the acceptance parameters, development process, advantages, opportunities and challenges of SDN technology. Some of them are the adoption parameters of similar technologies such as virtualization (NFV), cloud computing, big data, Internet of Things and blockchain. This group of research is intended for application in the field of SDN.

Table2: Summary of previous studies regarding SDN adoption

Reference	***Purpose/ Domain***	***Method***	***Findings***	***categories***
[24]	SDN Adoption /IT cloud integrators	Quantitative method (SPSS, PLS) + Questionary (167 cloud integrators)	-Examine the relationship between IT cloud integrators' perceptions of performance expectancy, effort expectancy, social influence, facilitating conditions, and their intention to use SDN by using "UTAUT". -social influence and facilitating conditions were statistically significant; performance expectancy and effort expectancy were not.	Phenomenon, Causal Factors
[25]	Factors affecting Cloud Computing adoption / companies in Turkey	Combination of DOI and TOE + Quantitative method (Smart PLS V.3)	-parameters are: Comparative advantage (including security, cost savings), complexity, compatibility from DOI model and technical parameters (including technical readiness), organizational context (support of senior managers, company size), environmental context (competitive pressure, rule support) from TEO model.	Causal, Intervening, Contextual Factors
[26]	Factors influencing SDN adoption/ Research and Educational Networks (REN)	Qualitative method (Combination DOI and TOE models)	1-Human resources (Leaders' opinion, Team skills) 2-SDN technology (advantages, compatibility, complexity, testability, security) 3-Organization (REN size and Resources, REN Global scope, Network users) 4-Environment (Regulation Policies, technology support)	Causal, Intervening, Contextual Factors
[27]	SDN Adoption Motivations/Telecom Networks	N/A	-Motivations for SDN market: complexity in Data Centers and Enterprises, inability of traditional network architecture to support migration to the cloud and the need for improving application performance. -Some technology trends propelling the adoption of SDN are 5G, IoT, Cloud native, edge computing and Big Data. -IDC enumerates top 6 motivations for SDN adoption: policy-based control and WAN optimization, network agility and flexibility, Optimized cost, Consistent security, Enhanced operational efficiency and Faster deployment.	Causal, Phenomenon, Intervening, Consequences Factors
[28]	Reasons for not accepting SDNs and reviewing studies/ Telecom Networks	Literature review	-SDN adoption limitations: Budget constraints, Performance and reliability, Scalability, maturity, Lack of experts and certification programs. -network operators can incrementally deploy SDN and avoid replacing legacy equipment at once.	Causal, Intervening Factors
[29]	SDN Adoption/Telecom Service Provider	Mixed Method + Interview (14 Service Provider)	-The most important motivate for using SDN is agility, unlike previous research that has shown cost reductions. -Proposed SDN adoption solution: technology evolution with 29%, innovative services with 21%, organizational transformation with 50% impact on SDN adoption	Causal, Consequences Factors

Reference	*Purpose/ Domain*	*Method*	*Findings*	*categories*
[30]	Cloud Computing Adoption / SMEs in Malaysia	Combination of DOI and TAM + Quantitative method (SPSS, SmartPLS) + Questionary (114 responses)	-The parameters are: Comparative advantage parameters (including security issues, cost savings), complexity, compatibility of the DOI model and perceived usefulness and perceived ease of use from TAM model that has strong influence on the user intention to adapt new technology. -Three predictors of relative advantage, compatibility and complexity yielded a significant influence on perceived usefulness. -Relative advantage revealed no significant relationship in cloud computing adoption. -Perceived usefulness was found to have a mediating effect between compatibility and adoption attention, while perceived ease of use yielded a mediating effect between complexity and adoption attention.	Causal Factors
[31]	The impact of TOE Parameters on BDA adoption and its effect on the financial performance and marketing performance/SMEs	Combination of DOI, TOE and RBV[1] models + Quantitative method (PLS) + Questionary (171responses)	-Technological Factors (Relative Advantage, Compatibility, Complexity, Uncertainty and Insecurity, Trial ability, Observability), Organizational Factors (Top management Support, Organizational Readiness), Environmental Factors (Competitive Pressure, External Support from Vendors, High degree of regulatory) -Complexity and uncertainty and insecurity have negative effect, while trial ability and observability affect oppositely. Relative advantage and compatibility show no effect. Top management support and Organizational resources, positively influence on BDA adoption.	Causal, Intervening, Contextual Factors
[32]	IoT Adoption (IoTAM)	Quantitative method (IoTAM) + Questionary (812 survey participant) + Quantitative method (SPSS, SEM [2], PLS)	-Model parameters: User character, Cyber resilience, social influence, Cognitive instrumentals, Trust, Long-term orientation, Flexibility, Perceived Usefulness, Perceived Ease of Use, Attitude, Behavioral intention. -It was demonstrated that facilitated appropriation, Perceived Usefulness and Perceived Ease of Use significantly influence consumers' attitude and BI. User character, cyber resilience, cognitive instrumentals, social influence and trust exhibited a significantly indirect effect on attitude and Behavioral intention, through the three main mediators.	Causal, Phenomenon Factors
[33]	SDN adoption factors / Telecom Networks	Qualitative and Quantitative method + Semi-structured interviews	-Barriers to SDN adoption: a) challenges to integrating SDNs with legacy networks b) the immaturity of vendor solutions c) Technology shortcomings (proper Interfaces) -Top SDN adoption Drivers: a) the simplification of network provisioning b) the better utilization of network resources	Intervening Factors
[34]	SDN Adoption Factors	Questionnaire (Intel experts)	-benefits of SDN: speeding up service provision, creating services automatically, reducing costs -Factors affecting SDN adoption: a proper understanding of SDN performance, adaptability, remove obstacles	Causal, Intervening, Consequences Factors
[35]	Application of SDN on network security/ Telecom Networks	Literature Review	-Benefits such as network security, attack detection and troubleshooting, traffic control, configuration and policy management, and service change	Causal, Intervening, Consequences
[36]	Block chain Adoption	qualitative Method + semi-structured interviews	Model: Technical parameters (comparative advantage, uncertainty), organizational parameters (managerial support, organizational readiness), environmental parameters (competitive pressure, regulatory environment, industry) and trust are added to the TOE framework.	Causal, Intervening, Contextual Factors

[1] Resource Based View

[2] Structural Equation Modeling

The motivating factors for conducting this research are:

- Need for a comprehensive and integrated framework that focuses on all levels (national, organizational, individual, etc.)
- Dispersion of activities and the need to complete previous activities.
- Combining the strengths of the models presented for SDN adoption and the adoption model of similar technologies.
- Update the existing knowledge about SDN, especially in the field of its adoption and express management problems.
- Considering management processes, strategies, contextual conditions and Consequences that previous models had not addressed nor had very little reference to.
- Few researches have used the qualitative-quantitative mixed method. Most of them have extracted factors affecting the acceptance of this technology through content analysis, literature review, questionnaires and interviews.

3- Method

The review of previous studies indicates the weakness of existing theories and models regarding the adoption of SDN technology. Thus, at the beginning, the theoretical foundations related to the research problem were examined. Then, based on the Grounded theory, in-depth interviews have been conducted with the managers and experts of the country's major telecom operators.
In result, dimensions and categories have been extracted using open, axial and selective coding. In order to confirm the indicators extracted from the qualitative part, the fuzzy Delphi method has been used. Then by analyzing the data extracted from the questionnaire (with a Likert scale), the structure and fit of the model has been checked. Also, CFA, confirmation and rejection of hypotheses has also been done using Smart PLS.3 software.

3-1- Data Collection Process

In order to collect the initial data in the qualitative section, three different sources have been considered: 1) Interviews 2) Articles and documents 3) Technical reports, meetings.
A community of experts have been purposefully selected who have rich information about the research topic and are working individually or organizationally in this field.
According to the research methodology and considering the available time and resources, between 10 to 15 interviews would be enough, as the purpose of the interview is to explore the ideas and attitudes of the interviewee. Thus, in order to extract the initial data, current study has been done by 12 interviews (6 university professors and 6 senior managers and experts with PhD in computer, telecommunications, industry and management). Data collection has continued until the research reaches the saturation limit in the data and the concepts related to SDN adoption and there was no new one to be added to the model. Finally, at each stage, the codes were finalized with 5 of interviewees who had the following expertise (Table 3):

Table3: Characteristics of the Interviewed Experts

Expert	*Degree*	*Skills*
Expert1	PHD-Computer	Communication Net-SDN
Expert2	PHD-Telecomm	Telecomm -5G- SDN
Expert3	PHD-Electricity	Telecomm-Cloud/NFV/SDN
Expert4	PHD-Industrial	MCDM-E. Readiness-Technology Adoption-Management
Expert5	PHD-Management	Organizational-Public Administration

In the quantitative section**,** the statistical population includes experts and managers of various Iranian telecommunication Sectors, including: ITRC[1], TIC[2], TCI[3], MCI[4], etc. In this section, the sample size was calculated using Cochran's formula. The statistical population was about 140 individuals, and considering Cochran's formula, a total of 101 experts were selected as a statistical sample. Demographic information of respondents illustrated in table 4.

Table4: Demographic Information of Respondents

Characteristic	*Frequency*	*Percentage*
Age		
30-40 years old	34	33.7%
40-50 years old	43	42.6%
50 years old or older	24	23.8%
Gender		
Male	75	74.3%
Female	26	25.7%
Degree of education		
Bachelor's degree	9	8.9%
Master degree	55	54.5%
PHD	37	36.6%
Field of education		
Telecommunications	19	18.8%
Management	24	23.8%
Computer	23	22.8%
IT	21	20.8%
Industrial Engineering	9	8.9%
other	5	5.0%
Position		
Executive/Senior manager	22	21.8%
Junior manager	70	69.3%
University professor	9	8.9%

[1] Iran Telecommunication Research Center
[2] Telecommunication Infrastructure Company
[3] Telecommunication Company of Iran
[4] Mobile Telecommunication Company of Iran

3-2- Data Analysis Process in Qualitative Section

In this section the major steps of qualitative part of the research are explained.

- Review of research background:

Before conducting the interview, the background of the research was investigated (table 2) so that the findings of the initial stages were the basis for preparing the questionnaire and the interview.
In this regard, databases such as IEEE, Science Direct, Emerald Insight and etc. have been examined and it is important that the contents were not repeated and have been done in the last 15 years.

- Conducting research according to Grounded Theory:

By conducting the interviews and analyzing the data, a deeper understanding of the interviewees' experience and knowledge was gained. In this way, qualitative data analysis was performed in three stages: open, axial and selective coding [41].
Open Coding: Open coding was done in two stages. Initial coding was done by line-by-line coding of the data, and a concept or code was attached to each of them. Thus, more than a thousand initial codes were extracted. In secondary coding by comparing all extracted open codes, the items that had semantic similarity were placed in one category [40]. Then, in order to make the results validate, it has been reviewed by 5 experts who were highly specialized in this field. At the end, a set of 112 final codes is extracted.
Axial Coding: The data decomposed into concepts and categories are examined in a new way so that a link can be made between a category and the concepts in it and even other categories [40].
Selective Coding: Selective coding uses the results of the previous coding steps, selects the main category and links it systematically to other categories, validates the connections, and develops the categories that need further refinement and development [40] [41]. Fig.5 shows the final coding done by MAXQDA software.

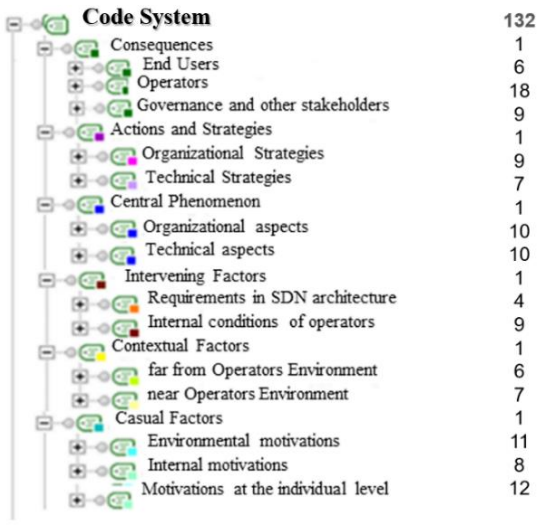


Fig.5: Final coding done by MAXQDA software

- Writing theory and theorizing:

Finally, with the help of the developed theory, 6 hypotheses have been extracted for testing in the research. In order to confirm the indicators extracted from the qualitative part, the fuzzy Delphi method has been used.

3-3- Data Analysis Process in Quantitative Section

After extracting the conceptual model, the model itself and the research hypotheses are tested in order to obtain deeper information about the research model in the statistical population. After collecting information with a questionnaire, the data was analyzed with SPSS and PLS software. Frequency distribution, mean and variance were used to describe the opinions of the statistical sample regarding the questions. This process was carried out at a significance level of 0.05. For data analysis, using Smart PLS.V3, the fit of the model and then the research hypotheses were evaluated. Before using the statistical tests, the normality of the data should be evaluated, and for this purpose, the Kolmogorov-Smirnov test was used.

4- Findings and Results

In order to present the SDN adoption model, the Strauss-Corbin paradigm has been used, which is based on a systemic approach and includes [14] [41]:
Causal Condition: Causal conditions are the motivating factors to encourage organizations to adopt SDN technology. Three categories of motivation identified includes: Environmental motivations of operators, internal motivations of operators [42] and motivations at the individual level.
Contextual Conditions: special conditions that affect strategies and are not under the control of organizations, but awareness of them can lead to appropriate response and understanding why some events related to the process

of accepting SDN technology. Far and close environment of the operators is considered in this category [42]. The far environment is a situation that the operator has no control over that. The close environment of operators includes the competitors and other stakeholders [42].

Phenomenon: the main phenomenon refers to the adoption of SDN technology and its dimensions, which has been the main topic of this research. Use of SDN has been identified in two categories: organizational and technical.

Intervening Conditions: unlike Contextual conditions, intervening conditions are under the control of organizations and network administrators that affect the strategies. The internal conditions of the operators and the requirements of the SDN architecture were considered as intervening conditions [42].

Strategies: special actions that result from the main phenomenon and can be helpful in promoting SDN adoption. The difference between this and phenomenon is that the strategies are not process but action type and help execute processes. Technical and organizational strategies have been considered [42].

Consequences: the outcomes that emerge as a result of strategies. There are three levels of consequences for users, operators, government and other stakeholders [42]. These consequences include positive and negative impacts. Thus, 112 final codes, 14 concepts and 6 main categories were obtained, which is a total of 132 options. The SDN adoption model proposed by the research is shown in the (fig.6).

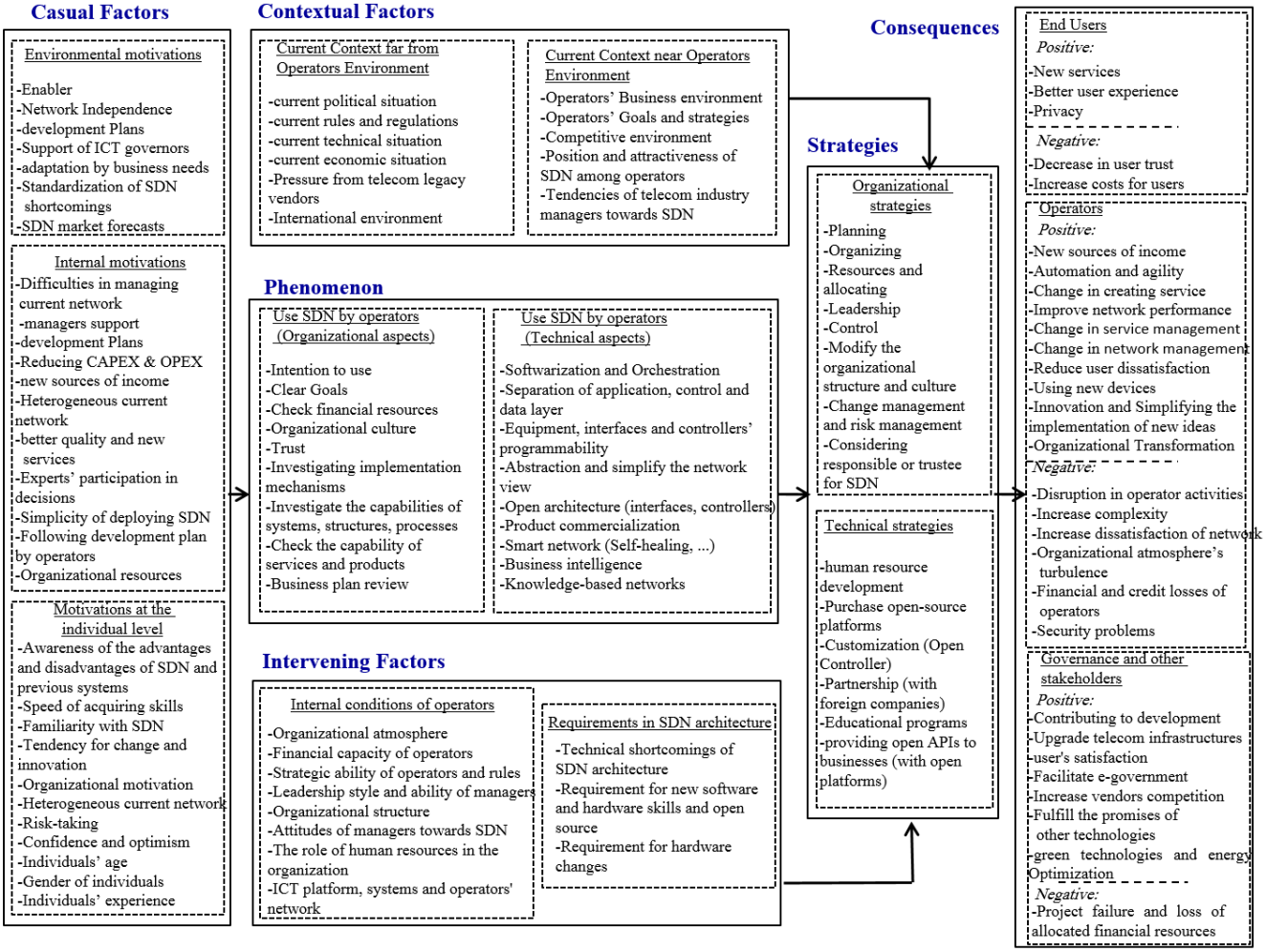


Fig.6: The proposed model for SDN adoption

4-1- Hypothesis

Based on the conceptual model and the main research questions, the following hypotheses are extracted:

- **H1:** Causal Factors have a significant effect on the Phenomenon, which is the use of SDN by operators (organizational and technical).
- **H2:** The Phenomenon, that is, the use of SDN by operators has a significant effect on the strategies.
- **H3:** The Contextual Factors (Current Context far from and near Operators Environment) has a significant effect on the strategies.
- **H4:** The Intervening Factors (internal conditions of operators and requirements in SDN architecture) has effect on the strategies.
- **H5:** The strategies have a significant effect on the consequences (for users, operators, governance and other stakeholders).

4-2- Fuzzy Delphi Method

At this stage, a questionnaire consisting of 112 indicators of the proposed SDN adoption model with the focus on large telecommunications operators, which was extracted from literature review and interviews, was provided to expert members [43]. In order to do fuzzy, the opinions of experts, triangular fuzzy numbers have been used. Experts' views on the importance of each indicator are collected with a 5-point fuzzy Likert spectrum. Then the fuzzy mean is taken from the scores and converted to a definite number according to the relation belong to the fuzzy mean. In the next step, the Delphi first stage questionnaire was given to the experts again. Also, in this round, the definite average of the first round is provided so that the experts are informed about the average of each index in the previous stage. In this round, the fuzzy mean of the scores was calculated in a similar way using the above-mentioned equation [43].

According to Cheng, if the difference between the two stages of the poll is less than the threshold (0.1), the polling process will stop, we have reached a consensus [43]. In the first phase of fuzzy Delphi, considering the threshold number of 0.7 [44], the results show the confirmation of all indicators (Table 5.a).

In the next round, considering the same threshold number of 0.7 [44], the results still show the confirmation of all indicators (Table 5.b).

Considering that the difference between the two stages of the survey is less than the threshold (0.1), the survey process has been stopped, we have reached a consensus.

Table5: Results of the first and second phase of fuzzy Delphi (23 out of 112 codes)

row	Indicator	Fuzzy weight	Non fuzzy weight	status
1	Enabler (For other technologies or other concepts)	(0.5,0.75,0.917)	0.722	√
2	National Network independence	(0.521,0.771,0.958)	0.750	√
3	development Plan and migration towards digital transformation	(0.542,0.792,0.917)	0.750	√
4	Support of ICT governors	(0.5,0.75,0.938)	0.729	√
5	Faster adaptation by changing business needs	(0.5,0.75,0.875)	0.708	√
6	Standardization and elimination of SDN shortcomings	(0.5,0.75,0.875)	0.708	√
7	SDN market forecasts	(0.521,0.771,0.896)	0.729	√
8	Difficulties in managing and controlling the current network	(0.5,0.75,0.917)	0.722	√
9	Positive attitude and managers support from SDN	(0.521,0.771,0.938)	0.743	√
10	Reducing CAPEX & OPEX Costs	(0.479,0.729,0.917)	0.708	√
11	Requirement for new sources of income	(0.542,0.792,0.938)	0.757	√
12	Heterogeneous current network	(0.604,0.854,0.958)	0.806	√
13	Requirement for better quality and new services	(0.604,0.854,0.958)	0.806	√
14	Experts' participation in decisions and giving positive feedback from SDN	(0.521,0.771,0.917)	0.736	√
15	Simplicity of implementing SDN architecture	(0.5,0.75,0.917)	0.722	√
16	Operators follow the development plan and digital transformation	(0.625,0.875,0.958)	0.819	√
17	Ability of organizational resources	(0.625,0.875,1)	0.833	√
18	Awareness and understanding of the advantages and disadvantages of SDN and previous systems	(0.521,0.771,0.917)	0.736	√
19	Familiarity of individuals involved with SDN in English, software skills and networking	(0.542,0.792,0.958)	0.764	√
20	Speed of acquiring skills	(0.5,0.75,0.938)	0.729	√
21	Tendency for change and innovation	(0.5,0.75,0.896)	0.715	√
22	Organizational motivation and commitment	(0.542,0.792,0.917)	0.750	√
23	Risk-taking	(0.583,0.833,0.938)	0.785	√

a) Results of the first phase of fuzzy Delphi (23 out of 112 codes)

row	Indicator	Fuzzy weight	Non-fuzzy weight	status
1	Enabler (For other technologies or other concepts)	(0.563,0.813,0.938)	0.771	√
2	National Network Independence	(0.5,0.75,0.958)	0.736	√
3	development Plan and migration towards digital transformation	(0.5,0.75,0.917)	0.722	√
4	Support of ICT governors	(0.521,0.771,0.958)	0.750	√
5	Faster adaptation by changing business needs	(0.583,0.833,0.958)	0.792	√
6	Standardization and elimination of SDN shortcomings	(0.521,0.771,0.917)	0.736	√
7	SDN market forecasts	(0.521,0.771,0.896)	0.729	√
8	Difficulties in managing and controlling the current network	(0.542,0.792,0.938)	0.757	√
9	Positive attitude and managers support from SDN	(0.521,0.771,0.958)	0.750	√
10	Reducing CAPEX & OPEX Costs	(0.542,0.792,0.938)	0.757	√
11	Requirement for new sources of income	(0.542,0.792,0.938)	0.757	√
12	Heterogeneous current network	(0.563,0.813,0.958)	0.778	√
13	Requirement for better quality and new services	(0.604,0.854,0.958)	0.806	√
14	Experts' participation in decisions and giving positive feedback from SDN	(0.5,0.75,0.917)	0.722	√
15	Simplicity of implementing SDN architecture	(0.521,0.771,0.917)	0.736	√
16	Operators follow the development plan and digital transformation	(0.583,0.833,0.958)	0.792	√
17	Ability of organizational resources	(0.604,0.854,1)	0.819	√
18	Awareness and understanding of the advantages and disadvantages of SDN and previous systems	(0.583,0.833,0.958)	0.792	√
19	Familiarity of individuals involved with SDN in English, software skills and networking	(0.542,0.792,0.958)	0.764	√
20	Speed of acquiring skills	(0.5,0.75,0.938)	0.729	√
21	Tendency for change and innovation	(0.5,0.75,0.896)	0.715	√
22	Organizational motivation and commitment	(0.5,0.75,0.917)	0.722	√
23	Risk-taking	(0.583,0.833,0.938)	0.785	√

b) Results of the second phase of fuzzy Delphi (23 out of 112 codes)

5- Model Analysis

5-1- Data Preprocessing and Data Normality Test

Since the researcher used an electronic questionnaire, there were no missing and no outlier data, and all questionnaires were fully answered. To identify indifferent cases, they were identified by the formula STDEV.P>. /3 in Excel and the answers of 13 experts were removed.

Kolmogorov-Smirnov test was used to check the normality of research variables. If the significance level is greater than 5%, the variables are normal.

Table6: Assessing the Normality of the data

Variables	*Causal Factors*	*Phenomenon*	*Contextual Factors*	*Intervening Factors*	*strategies*	*consequences*
Kolmogorov- Smirnov z	.113	.202	.098	.137	.173	.098
Sig	.003	.000	.017	.000	.000	.017

According to Table 6, the data are not normal, and thus the PLS approach should be used [45][46]. Also, the normality test of each question has been evaluated with mean and variance. The average of the questions was more than 3 and the variance was more than 0.5, and their significance level was less than 0.05.

5-2- Evaluation of the Proposed Model

The factor load test for each question (112 final indicators) was higher than 0.4 and thus none of the questions were removed. The reliability of this measurement model is acceptable. The significance of the factor load was checked with the t-value statistic, none of the 112 questions was smaller than 1.96 and were not removed. Cronbach's alpha, CR[1], SR[2] and rho-a tests were also used to evaluate reliability (Table 7).

[1] Combined Reliability

[2] Shared Reliability

Table7: Reliability of constructs (Cronbach's alpha, CR, SR, rho-a) [46] [47]

Variable	*Cronbach's alpha > 0.7*	*CR > 0.7*	*SR > 0.5*	*rho-a 0.7 or 0.6*
Use SDN by operators (Organizational)	0.890	0.912	0.542	0.910
Telecom Operator	0.971	0.973	0.682	0.971
Current Context far from Operators Environment	0.907	0.931	0.696	0.927
Current Context near Operators Environment	0.890	0.919	0.698	0.914
Use SDN by operators (Technical)	0.889	0.910	0.534	0.900
Governance and other stakeholders	0.888	0.911	0.562	0.898
Strategies	0.955	0.947	0.899	0.960
Organizational strategies	0.914	0.930	0.625	0.919
Technical strategies	0.959	0.967	0.830	0.960
Internal conditions of operators	0.918	0.934	0.640	0.926
Contextual Factors	0.924	0.912	0.838	0.938
Casual Factors	0.960	0.939	0.839	0.964
Intervening Factors	0.938	0.950	0.905	0.944
Motivations at the individual level	0.927	0.938	0.583	0.936
Internal motivations of operators	0.918	0.932	0.577	0.921
Environmental motivations of operators	0.881	0.908	0.586	0.890
Requirements in SDN architecture	0.836	0.902	0.754	0.841
Phenomenon	0.923	0.909	0.834	0.934
Consequences	0.980	0.972	0.921	0.982
End users	0.958	0.967	0.855	0.958

Convergent validity was extracted with AVE [1] and divergent validity was done by Fornell and Larker method [48]. The AVE for all variables is higher than 0.5 which indicates the appropriate convergent validity of the constructs. Also, the divergent validity was at a reasonable level [45].

[1] Average Value of Extracted variance

5-3- Evaluation of the Structural Part of the Model

To evaluate the structure of the model, R Squares, CV Red and CV Com have been used. As the table 8 shows, the endogenous constructs of the model (strategies, consequences, and Phenomenon) with their exogenous constructs (causal, contextual, and intervention factors) with a value of more than 0.76, 0.80 and 0.699, It has a strong structural relationship and this indicates the strength of the structural part of the model. The CV Red index for all variables was higher than the average and in the strong range. Also, the CV Com index for all variables was within the acceptable range [45] (table8).

Table8: Evaluation of the Structural Part of the Model

Variable	*R Squares > 0.19*	*CVRed Q^2> 0.01*	*CV Com Q^2> 0.001*
Use SDN by operators (Organizational)	0.856	0.426	-
Telecom Operator	0.972	0.609	0.422
Current Context far from Operators Environment	0.870	0.547	0.601
Current Context near Operators Environment	0.806	0.516	0.378
Use SDN by operators (Technical)	0.813	0.396	0.457
Governance and other stakeholders	0.895	0.463	0.461
Strategies	0.762	0.431	0.362
Organizational strategies	0.896	0.503	0.532
Technical strategies	0.903	0.692	-
Internal conditions of operators	0.969	0.577	0.485
Contextual Factors	-	-	-
Casual Factors	-	-	0.385
Intervening Factors	-	-	0.653
Motivations at the individuals	0.898	0.478	0.682
Internal motivations of operators	0.827	0.440	0.430
Environmental motivations of operators	0.790	0.403	0.522
Requirements in SDN architecture	0.843	0.597	0.575
Phenomenon	0.807	0.326	0.613
Consequences	0.699	0.403	0.649
End users	0.895	0.710	0.577

6- Conclusions

Today's telecommunication networks are intensively expensive, manual and inflexible. To transform to SDN, which could lead to cost reductions, automation, greater processing capacity, and service orchestration through programmability, system developers and industry leaders

must perceive its necessity and adjust to the intricacies of its adoption.

Although some researches have been done in this regard, most of them have reviewed the advantages, challenges and development process of this technology. In this research, using Grounded theory, the adoption model of SDN technology has been extracted, focusing on the country's telecommunication sectors. The proposed Model include 112 codes, 14 categories and 6 themes which extracted in the qualitative section and confirmed by fuzzy Delphi method.

A comprehensive and integrated framework focusing on all levels (national, organizational, individual, etc.) has been presented, which includes, actions and consequences in addition to environmental factors, input, processing and output.

Among the advantages of the proposed model, we can mention the combination of the strengths of the models presented in adoption of SDN, adoption of similar technologies such as Cloud, and adoption of information technology. Also, attention to management processes, strategies, contextual factors that were not considered in the previous models or were mentioned very little.

The use of grounded theory (Strauss and Corbin model) along with the fuzzy Delphi method and quantitative analysis were very helpful and effective in conducting the research. It is also important to use SPSS and Smart-PLS 3 software to analyze the data collected from the questionnaire and evaluate the fit of the model as well as confirm and reject the hypotheses.

Gall has suggested that there is not much agreement among researchers on determining the criteria for validity and reliability in qualitative research. This disagreement arises from the fact that in the qualitative approach, the findings are based on the researcher's mental reflection and his interpretation of events. Therefore, in the qualitative approach, the definition of validity and reliability a little bit is different [49]. Maxwell has defined different types of validity [52]: Descriptive, Interpretative and Theoretical validity. Goulding believes that the theoretical saturation point indicates the reliability of the grounded theory research method. Theoretical saturation occurs when the data that helps to define the characteristics of a class is no longer included in the research and all the desired comparisons have taken place [51]. in this research these tactics have been used to achieve validity and reliability in qualitative part [50]:

- Data collection from several sources.
- Long-term observation.
- Preventing subjective assumptions in drawing conclusions and ensuring results through feedback. In fact, the researcher presented his interpretations to the participants and identified and corrected the misunderstood areas.
- Using experts with management and human resource experience along with IT.
- Providing various Presentations on SDN technology
- The collection and analysis process have been regularly modified by several experts, and the initial draft of the research findings was provided to research colleagues (supervisors and consultants).
- Quality analysis software (MAXQDA 2012) has been used and thus no data has been ignored in this process.
- In order to confirm the indicators extracted from the qualitative part, the fuzzy Delphi method has been used.
- Issues such as comprehensibility and accuracy of components, adaptation of the result with the phenomenon under study, control of new conditions, inclusion of different dimensions, have been considered in different stages.

In Quantitative part, in order to analyze the collected data, descriptive statistics of demographic variables were presented using SPSS software. Reliability (with factor loading, Cronbach's alpha, composite reliability), convergent and divergent validity (with AVE and Fornell-Larcker method) and measurement of model structure (with R2, CVRed and CV Com (Q2) criteria) were evaluated in inferential statistics. All these criteria were evaluated at a reasonable and acceptable level.

Finally, the fitting of the model was done through the GOF[1] and SRMR[2] criterion. $\sqrt{\overline{\text{communality}}}$ is the average of the shared values of each structure. $\sqrt{\overline{R^2}}$ is the average of the R Squares just for endogenous structures of the model. Wetzles (2009) introduced three values of 0.01, 0.25 and 0.36 as weak, medium and strong values for the overall fit of the model.

$$\sqrt{\overline{\text{communality}}} \times \sqrt{\overline{R^2}} = \sqrt{0.8629} \times \sqrt{0.720} =./7878 \quad (1)$$

Obtaining 0.787 for the overall fit of the model indicates a strong overall fit of the model. Also, the value of SRMR index is equal to 0.979, which is less than the value of 0.1. Thus, the model has a suitable fit.

The output of the model shows that the coefficient of significance regarding research hypotheses is out of the range of ± 1.96, which means that all research variables are confirmed at the 95% confidence level in the statistical sample.

Table9: Summary of the Findings

Hypothesis	***Path coefficient***	***t-value***	***Result***

[1] Goodness Of Fit

[2] Standardized Root Mean Residual

H1	0.898	33.170	Confirm
H2	0.298	2.826	Confirm
H3	0.695	7.994	Confirm
H4	0.302	3.305	Confirm
H5	0.836	24.696	Confirm

In future, the proposed model can develop further by including SDN technology diffusion models. Also, prioritizing the identified categories to understand the importance of each, as well as model simulation to determine the effect of different policies on changing system behavior are among the important items that are proposed to complete the research. In Table 10, some suggestions are presented at three levels: governance, administrators of telecommunication operators and the experts. To avoid this risk, the assessing of company‘s E-Readiness before starting main project is necessary [53].

Table 10: research Suggestions

Level	***suggestions***
Macro level (Governance)	▪ Support and create motivation for operators. ▪ Increasing competitiveness between operators. ▪ Persuading investors to invest in infrastructure technologies. ▪ Establishing cooperation systems between research centers, industry and government. ▪ Creating a trustee in the country for SDN deployment and preparing a detailed roadmap for the operationalization of SDN implementation. ▪ Guiding the industry towards the commercialization and export the products of this technology. ▪ Reducing the impact of the country's macro changes on the implementation of projects.
Telecom Operator (Manager)	▪ Preparing an accurate business plan ▪ Clarify clear goals ▪ Identify the main risks ahead ▪ Increasing awareness and training personnel ▪ providing a roadmap for SDN deployment ▪ Creating alignment between the country's ICT plans with their business activities (operators) ▪ Considering change management ▪ Upgrade hardware and software infrastructure ▪ Plan and resource allocation to deploy SDN ▪ Cultivating expert human resources in the field of SDN technology and related technologies ▪ Increasing the agility of operators ▪ Participation of experts in international scientific forums and standard-setting organizations
Industry Experts	▪ Acquisition of network, software, hardware skills ▪ Trying to disambiguate the shortcomings of this technology ▪ Providing correct feedback based on study and knowledge and trying to create the right attitude of technology to senior managers ▪ Creating partnership and cooperation with other experts
	▪ Enhance flexibility and acceptance of changes ▪ Improving the international language level ▪ Communication with academic centers, conferences, domestic and international associations

References

[1] R.Jain, "Trends and Issues in Softwarization of Networks: What's In, What's Out," IEEE Conf. Netw. Softwarization, Washingt. Univ. Saint Louis, 2018.

[2] D. Kreutz, F. M. V Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," Proc. IEEE, vol. 103, no. 1, pp. 14–76, 2014.

[3] R. Gaikwad, V., Rake, "Software defined networking Market Statistics:2027,"2020. https://www.alliedmarketresearch.com/software-defined-networking-market.

[4] J. H. Cox et al., "Advancing software-defined networks: A survey," IEEE Access, vol. 5, pp. 25487–25526, 2017.

[5] N. Bhalani, M. Chavan, "A Survey on Software Defined Network with 5G", International Journal of scientific & Technology Research, 2020.

[6] A. khamseh, M. Lialestani, and reza radfar, "Digital Transformation Model, Based on Grounded Theory," J. Inf. Syst. Telecomm. , no. 1, pp. 275–284, 2021, doi: 10.52547/jist.9.36.275.

[7] S. Brinker, "Martec's Law: the greatest management challenge of the 21st century," Chiefmartec. com, 2016.

[8] M. Abdulrab, "Factors Affecting Acceptance and the Use of Technology in Yemeni Telecom Companies," Int. Trans. J. Eng. Manag. Appl. Sci. Technol., vol. 11, no. 6, pp. 1–16, 2020

[9] M. K. Chang and W. Cheung, "Determinants of the intention to use Internet/WWW at work: a confirmatory study," Inf. Manag., vol. 39, no. 1, pp. 1–14, 2001.

[10] F. Nouri, R., Hatami, M. and Ebrahimian, "Effective factors on the acceptance of information technology and its impact on human resources," Hum. Resour. Manag. Res. Imam Hossein Univ., vol. 9, no. 4, pp. 27–152, 2018.

[11] P. C. Lai, "The literature review of technology adoption models and theories for the novelty technology," JISTEM-Journal Inf. Syst. Technol. Manag., vol. 14, pp. 21–38, 2017.

[12] F. B. A. Rahman, M. H. M. Hanafiah, M. Salehuddin, M. Zahari, and L. B. Jipiu, "Systematic Literature Review on The Evolution of Technology Acceptance and Usage Model used in Consumer Behavioural Study," Int. J. Acad. Res. Bus. Soc. Sci., vol. 11, no. 13, pp. 272–298, 2021.

[13] A. M. Momani and M. Jamous, "The evolution of technology acceptance theories," Int. J. Contemp. Comput. Res., vol. 1, no. 1, pp. 51–58, 2017.

[14] J. Corbin and A. Strauss, Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications, 2014.

[15] A. shirmarz and A. Ghaffari, "An Autonomic Software Defined Network (SDN) Architecture With Performance Improvement Considering," J. Inf. Syst. Telecomm. , no. 1, pp. 121–129, 2020, doi: 10.29252/jist.8.30.121.

[16] A. S. Thyagaturu, A. Mercian, M. P. McGarry, M. Reisslein, and W. Kellerer, "Software defined optical networks (SDONs): A comprehensive survey," IEEE Commun. Surv. Tutorials, vol. 18, no. 4, pp. 2738–2786, 2016.

[17] ETSI, "Improved Operator experience through Experiential Networked Intelligence (ENI)", 2017, 1st Edition, ISBN No. 979-10-92620-16-0

[18] ONF, "NG-SDN™", Open Networking Foundation, 2022, Available: https://opennetworking.org/reference-designs/ng-sdn/

[19] Data Bridge Market, "Global Software-Defined Networking Market – Industry Trends and Forecast to 2028,"2021. https://www.databridgemarketresearch.com/reports/global-sdn-market.

[20] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," MIS Q., pp. 425–478, 2003.

[21] IGI Global, "What is Technology Adoption Model? https://www.igi-global.com/dictionary/technology-acceptance-model/29484

[22] E. M. Rogers and F. F. Shoemaker, "Communication of Innovations; A Cross-Cultural Approach.," 1971.

[23] F. Masimba and T. Zuva, "Individual Acceptance of Technology: A Critical Review of Technology Adoption Models and Theories," Indiana J. Humanit. Soc. Sci., vol. 2, no. 9, pp. 37–48, 2021.

[24] W. Russ, "The Relationship between Technology Adoption Determinants and the Intention to Use Software-Defined Networking." Walden University, 2021.

[25] C. Sayginer and T. Ercan, "Understanding determinants of cloud computing adoption using an integrated diffusion of innovation (doi)-technological, organizational and environmental (toe) model," Humanit. Soc. Sci. Rev., vol. 8, no. 1, pp. 91–102, 2020.

[26] V. Chergarova, J. Bezerra, J. Ibarra, and H. Morgan, "Factors influencing the adoption of Software Defined Networking by Research and Educational Networks," 2019.

[27] R. Jayaraman, V., Manickam, A., Rajappa, "The Role of SDN in Network Transformation," 2019. https://www.tataelxsi.com/news-and-events/the-role-of-sdn-in-network-transformation.

[28] N. Shah, P. Giaccone, D. B. Rawat, A. Rayes, and N. Zhao, "Solutions for adopting software defined network in practice," International Journal of Communication Systems, vol. 32, no. 17. Wiley Online Library, p. e3990, 2019.

[29] S. Bekele, B., Kriger, "SP NFV/SDN Adoption". STL Partners Research for Cisco," 2017.

[30] S. S. S. Mokhtar, A. S. B. Mahomed, Y. A. Aziz, and S. A. Rahman, "Industry 4.0: the importance of innovation in adopting cloud computing among SMEs in Malaysia," Polish J. Manag. Stud., vol. 22, 2020.

[31] P. Maroufkhani, M.-L. Tseng, M. Iranmanesh, W. K. W. Ismail, and H. Khalid, "Big data analytics adoption: Determinants and performances among small to medium-sized enterprises," Int. J. Inf. Manage., vol. 54, p. 102190, 2020.

[32] M. Tsourela and D.-M. Nerantzaki, "An internet of things (IoT) acceptance model. Assessing consumer's behavior toward IoT products and applications," Futur. Internet, vol. 12, no. 11, p. 191, 2020.

[33] A. M. Aranda, "Software-defined networking: Current state, adoption factors and future impact on network engineers," 2016.

[34] S. Mahankali, I. I. T. Cloud Network Engineer, and S. Rungta, "Adopting Software-Defined Networking in the Enterprise," White Pap. April, 2014.

[35] R. Sahay, W. Meng, and C. D. Jensen, "The application of software defined networking on securing computer networks:

A survey," J. Netw. Comput. Appl., vol. 131, pp. 89–108, 2019.
[36] S. Seshadrinathan and S. Chandra, "Exploring Factors Influencing Adoption of Blockchain in Accounting Applications using Technology–Organization–Environment Framework," J. Int. Technol. Inf. Manag., vol. 30, no. 1, pp. 30–68, 2021.
[37] D. Katsianis, I. Neokosmidis, A. Pastor, L. Jacquin, and G. Gardikis, "Factors influencing market adoption and evolution of NFV/SDN Cybersecurity Solutions. Evidence from SHIELD Project," in 2018 European Conference on Networks and Communications (EuCNC), 2018, pp. 1–5.
[38] M. S. Alhilal, A. M. Aldammas, and A. Y. Alnasheri, "Investigation of Critical Success Factors for Adopting Software-Defined Networking," in 2018 1st International Conference on Computer Applications & Information Security (ICCAIS), 2018, pp. 1–6.
[39] B. Gupta, S. Dasgupta, and A. Gupta, "Adoption of ICT in a government organization in a developing country: An empirical study," J. Strateg. Inf. Syst., vol. 17, no. 2, pp. 140–154, 2008.
[40] A. Bazargan, Introduction to qualitative research methods and a combination of common approaches in behavioral sciences. Tehran: Didar, 2009.
[41] J. W. Creswell, Educational research: Planning, conducting, and evaluating quantitative. Prentice Hall Upper Saddle River, NJ, 2002.
[42] E. Ziaeipour, "Explaining the adoption process of software-oriented networks (SDN) using the Grounded Theory and systems approach," J. Inf. Communication. Technol., vol. 14, no. 52, pp. 172–194, 2022.
[43] C.-H. Cheng and Y. Lin, "Evaluating the best main battle tank using fuzzy decision theory with linguistic criteria evaluation," Eur. J. Oper. Res., vol. 142, no. 1, pp. 174–186, 2002.
[44] C.-H. Wu and W.-C. Fang, "Combining the Fuzzy Analytic Hierarchy Process and the fuzzy Delphi method for developing critical competences of electronic commerce professional managers," Qual. Quant., vol. 45, no. 4, pp. 751–768, 2011.
[45] J. Henseler, C. M. Ringle, and R. R. Sinkovics, "The use of partial least squares path modeling in international marketing," in new challenges to international marketing, Emerald Group Publishing Limited, 2009.
[46] J. F. Hair, D. J. Ortinau, and D. E. Harrison, Essentials of marketing research, vol. 2. McGraw-Hill/Irwin New York, NY, 2010.
[47] C. M. Ringle, M. Sarstedt, R. Schlittgen, and C. R. Taylor, "PLS path modeling and evolutionary segmentation," J. Bus. Res., vol. 66, no. 9, pp. 1318–1324, 2013.
[48] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," J. Mark. Res., vol. 18, no. 1, pp. 39–50, 1981.
[49] M. Gall, W. Borg, and J. Gall, "Quantitative and qualitative research methods in educational sciences and psychology," Vol. I. Transl. by Ahmad Reza Nasr al. Tehran SAMAT Shahid Beheshti Univ. Publ., 2004.
[50] M. Abbaszadeh, "Validity and reliability in qualitative researches," J. Appl. Sociol., vol. 23, no. 1, pp. 19–34, 2012.
[51] K. Charmaz, Constructing grounded theory: A practical guide through qualitative analysis. Sage, 2006.
[52] J. Maxwell, "Understanding and validity in qualitative research," Harv. Educ. Rev., vol. 62, no. 3, pp. 279–301, 1992.
[53] A. Kamanghad, G. Hashemzade, M. A. kazemi, and N. Shadnoosh, "Assessing the Company's E-Readiness for Implementing Mobile-CRM System," J. Inf. Syst. Telecomm., no. 1, pp. 65–73, 2019, doi: 10.7508/jist.2019.01.006.