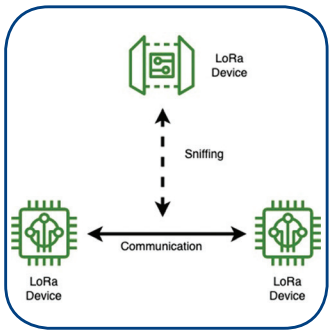
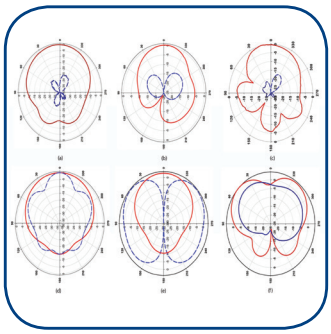
Volume 13, Number 7, 2022 ISSN 1847-6996

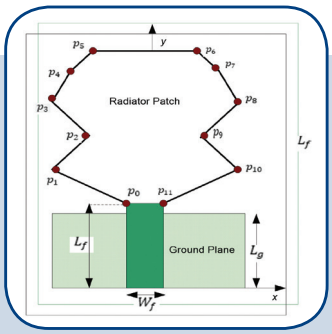
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

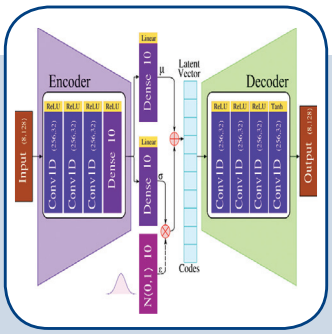
IJECES International Journal of Electrical and Computer Engineering Systems

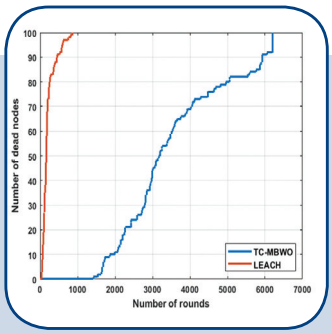
International Journal of Electrical and Computer Engineering Systems











ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

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

Osijek, Croatia | Volume 13, Number 7, 2022 | Pages 493 - 610

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

CONTACT

International Journal of Electrical and Computer Engineering Systems (IJECES)

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

Subscription Information

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

Journal is referred in:

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

Bibliographic Information

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

IJECES online

https://ijeces.ferit.hr

TABLE OF CONTENTS

Analysis of Tripleband Single Layer Proximity Fed 2x2 Microstrip Patch Array Antenna

Original Scientific Paper

Jacob Abraham
Department of Electronics
B P C College, Piravom, Kerala, India
tjacobabra@gmail.com

Kannadhasan Suriyan
Department of Electronics and Communication Engineering
Study World College of Engineering, Coimbatore, Tamilnadu, India
kannadhasan.ece@gmail.com

Abstract *– Microstrip patch antennas that are multiband and downsized are required to suit the high demand of modern wireless applications. To meet this need, a one-of-a-kind triple band array antenna has been proposed. The proposed 2x2 microstrip patch array, which comprises of four hexagon-shaped radiating patches are electromagnetically excited by a centrally positioned microstrip feed line in the same plane along with a slotted ground plane, is investigated. CST Microwave Studio, a powerful 3D electromagnetic analysis programme, was used to design and optimize the array antennas. The 2x2 array antenna was constructed on a FR-4 substrate with a dielectric constant of 4.3, a loss tangent of 0.001, and a height of 1.6mm. To optimize energy coupling from the feed line to the radiating patches, the ground plane has an H-shaped groove cut into it. The suggested 2x2 array antenna's multi-frequency behaviour is shown. Three resonant peaks were detected at 1.891GHz, 2.755GHz, and 3.052GHz. The observed bandwidths for these resonances are 234MHz, 69MHz, and 75MHz, respectively, with measured gains of 7.57dBi, 6.73dBi, and 5.76dBi. The goal of this work is to design, build, and test a single layer proximity fed array antenna. Standard proximity fed array antennas contain two substrate layers; however this array antenna has only one. As a consequence, the impedance matching and alignment are better. Simulated and experimental results showed that the this 2x2 array antenna operates in various important commercial bands, such as L and S bands and the array antenna might be beneficial for a wide range of wireless applications. The proposed antenna has good Impedance, S11, and radiation qualities at resonant frequencies. In this work, the 2x2 array antenna with hexagon-shaped radiating patches was successfully created utilizing the single layer proximity fed antenna concept and gap coupled parasitic patches.*

Keywords: *Triple Band, Single Layer Proximity Fed, 2x2 array, Slotted Ground Plane, Hexagon Shaped Patches*

1. INTRODUCTION

Due to the rapid growth of wireless communication networks and new wireless applications, there is a rising need for miniaturized portable handheld devices. For a portable communication device, the antenna is one of the most critical components. Nowadays, good number of wireless services and devices uses different frequency bands and protocols. To simultaneously cover all of these services, the antennas on portable communication devices should be able to cover many frequency bands at once. Building a multi-band antenna with a small footprint is not only necessary, but also challenging due to the limited space available for antennas in portable devices. Microstrip antennas are appropriate for portable wireless communication devices [1-2] due to benefits such as small weight, cheap cost, and simplicity of production. In contrast, a conventional microstrip antenna has a single resonance frequency and a limited band width [3]. Several ways have been utilized to extend the microstrip antenna's bandwidth [4-6] and enable it to function as a multiband antenna [7-9].

One of the most extensively used methods for enhancing the bandwidth and gain of microstrip patch antennas is to use gap coupled parasitic components. There have been many papers published in the literature on electromagnetically fed gap coupled patch antennas for wireless applications [10-11]. In [12] a triple band proximity fed 2x1 array antenna with defected ground plane is presented. The reported array antenna has two substrate layers and overall thickness of the antenna is 3.2mm. A group of antennas that operate together to transmit and receive radio waves as a single antenna is referred to as an array antenna. As the number of antennas in an array rises, the array's performance improves.

The patch antenna's gain, bandwidth, and emission pattern may be improved using this array antenna configuration [13-15]. Any defect etched in the ground plane of the microstrip array antenna can give rise to increasing effective capacitance and inductance. Slotted ground plane is accomplished by etching a flaw from the ground plane in a basic form. The ground plane faults will disturb the shielded current distribution, resulting in controlled excitation and electromagnetic wave propagation across the substrate layer [16-18]. Using the approaches described above, the proposed 2x2 microstrip array antenna is being constructed.

A single layer proximity fed triple band 2x2 microstrip patch array antenna is described in this paper. Traditional proximity fed patch antennas include two substrate layers. In the current paper a novel design technique is employed to reduce manufacturing complexity. By electromagnetic coupling, a centrally positioned microstrip feedline in the same surface, activates two radiating patches printed on the top surface of the substrate at the same time, as explained in our previous work [19]. In this study, a similar feeding technique is employed to resonate the radiating patches. The array consists of four identically sized radiating components. The lowest two radiating elements are driven patches, whereas the upper two radiating elements are parasitic patches. Parasitic patches are connected to the driven patch through the driven patch's non-radiating edges. Simulation and optimization were performed using CST Microwave Studio. The proposed single layer proximity fed 2x2 microstrip array antenna provides enough gain and impedance bandwidth, as well as triple band capabilities. The array antenna may be utilized for wireless communication equipment and the operating bands are 1.81GHz, 2.707GHz, and 2.962GHz, respectively.

2. ANTENNA GEOMETRY

2.1. SIMULATION BASED ANTENNA EVOLUTION STAGES

The evolution stages of the suggested single layer proximity fed 2x2 microstrip patch array antenna with slotted ground structure is shown in Fig. 1. Initially as illustrated in Fig.1(a), the single layer antenna consists of a hexagonal shaped radiating patch and a microstrip coupling feed line. Both the microstrip patch and feed line are printed on the top surface of the substrate and an H-shaped slot is etched in the ground plane. The microstrip feed line is placed close to the hexagonal shaped radiating patch with a gap G. The hexagon-shaped radiating patch is separated from the microstrip line by a gap G. In this situation, the radio frequency [RF] energy is coupled electromagnetically to the emitting device. The antenna was designed and simulated using CST microwave studio and the findings reveal that it resonates as a single band antenna. From the simulated radiation pattern results, it is observed that the beam maximum of the radiated beam is 37 degree away from the bore sight. In order to correct this pattern behavior, at the next antenna evolution stage another hexagon shaped radiating element is printed on the other side of the microstrip feed line. The new modified antenna design with slotted ground plane is shown in Fig. 1. (b). In this single layer proximity fed 2x1 array antenna configuration, dual band behaviour is seen. The CST Microwave Stimulation tool is used to fine-tune the size of the hexagonal radiating components such that the first band resonates at 1.8GHz. The 2x1 array antenna provides enough bandwidth and gain across both working bands.

The current research looks at the next phase in array antenna development, which aims to enhance the number of functioning bands and bandwidth. In the last evolution stage, two more hexagon-shaped radiating elements with the same dimensions are gap linked with the basic radiating elements, as illustrated in Fig.1. (c). This design is used to create a single layer proximity fed 2x2 microstrip patch array antenna with a slotted ground plane. An H shaped slot cut in the ground plane enhances the coupling to all four radiating patches. The unique features of the proposed single layer proximity fed 2x2 array antenna when compared with reference antenna [12], are triple band bahaviour, small volume, lightweight and improved impedance matching. The slot dimensions in all three antenna configurations have been modified to produce resonance at the desired frequency. A frequency shift or a decrease in impedance bandwidth will arise from a change in any of these design parameters. The length of the feed line is modified in order to improve impedance matching.

2.2 ANTENNA DESIGN PROCEDURE

The key design characteristic of a hexagon shaped radiating patch is its side length s. The equations for side length s, for a hexagonal shaped microstrip patch antenna can be obtained from the resonant frequency equations of the circular shaped microstrip patch antenna as discussed in reference work [12], by equating the respective areas as shown in Fig. 2. A circular patch antenna's basic resonance frequency is determined by

$$fr = \frac{X_{mn}C}{2\pi a_e \sqrt{\varepsilon_r}} \quad (1)$$

Where,

fr = resonant frequency of the patch
X_{mn} = 1.8411 for the dominant mode TM 11
C = velocity of the light in free space
εr = relative permittivity of the substrate
a_e = effective radius of the circular patch and given by

$$a_e = a\{1-2h/\pi a\varepsilon . (\ln\pi a/ 2h + 1.7726)\}^{0.5} \quad (2)$$

In equation (2), 'a' is the actual radius of the circular patch antenna and h is the height of the substrate. By relating the areas of the circular and hexagonal radiating elements as stated in the equation given below, the

aforementioned equations may be used to compute the side length *s*, of a hexagonal shaped microstrip patch antenna.

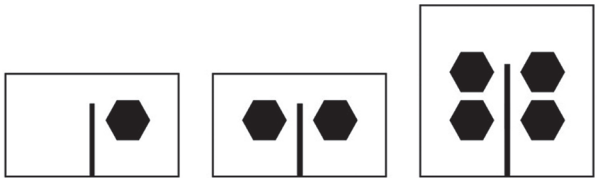


Fig.1. Evolution stages of the proposed single layer proximity fed 2x2 array antenna (a) stage one (b) stage two and (c) stage three

$$\Pi a_e^2 = \frac{3\sqrt{3}\, s^2}{2} \quad (3)$$

where'*s*' is the side length of the hexagonal patch.

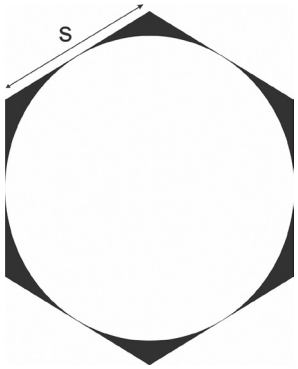


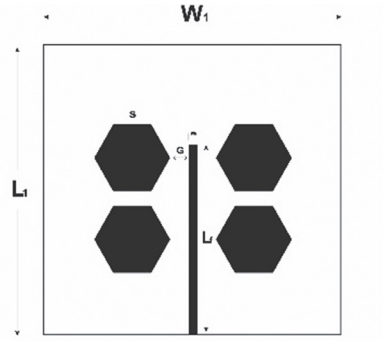
Fig. 2. Formation of a hexagonal shape from a circle

The radiating elements of the proposed array antenna is of hexagonal shape, whose eauation for resonant frequency is obtained using the resonant frequency equation of a circular radiating element by comparing their areas [20]. The resonant frequency of a hexagonal shaped radiating element is,

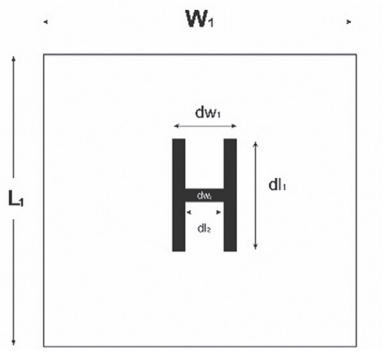
$$fres = \frac{YmnC}{5.714\, ae\sqrt{\varepsilon reff}} \quad (4)$$

Here, Ymn = Y11 (TM11 mode) = 1.841 or Ymn = Y21 = (TM21 mode) = 3.054, ε reff is the effective dielectric constant, C is the veocity of light and ae is the effective radius of the circular patch.

2.3 ANTENNA CONFIGURATION



(a)



(b)

Fig. 3. Geometry of the proposed single layer proximity fed 2x2 array antenna (a) top view and (b) bottom view

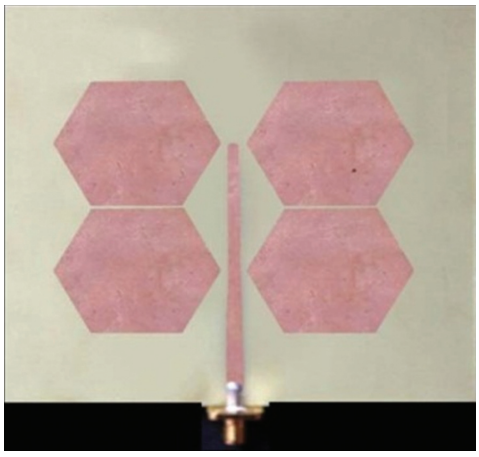
The configuration of the suggested single layer proximity fed 2x2 array antenna with slotted ground plane is shown in Fig.3. A 2x2 array antenna with a dielectric constant of 4.3, with a loss tangent of 0.001, and a height of 1.6mm was designed on a FR-4 substrate. Glass reinforced epoxy laminate sheets, rods, and printed circuit boards are given the grade FR-4. It's a popular and adaptable high-pressurethermoset plastic laminate with excellent strength-to-weight ratios. It's most typically employed as an electrical insulator because of its low water absorption and high mechanical strength. These qualities, together with strong manufacturing capabilities, aided in the selection of thissubstrate for the suggested antenna design. Fig.3 (a) depicts the top view of the suggested single layer proximity fed 2x2 array antenna with slotted ground plane, which consists of four hexgonal shaped radiating patches printed on the top side of the upper substrate with a centrally located microstrip feedline. The distance between the radiating components closest edges and themicrostrip feedline is maintained at λ/12. As illustrated in Fig.3(a), the four hexagon-shaped radiating elements are all activated electromagnatically from the centrally located microstip feed line. Fig. 3(b), shows a metallic ground plane with an H-shaped slot on the rear side of the substrate.

Dimensions of the radiating patches and that of the slot in the proposed array antenna configuration is optimized using CST microwave studio software package, which utilizes the finite integration technique for electromagnetic commutation to operate in three operating bands. Table 1 shows the optimized dimensions of the proposed antenna.

Table 1. Optimized dimensiosns of the proposed 2x2 array antenna (all units in mm).

W1	96	G	3.3	Dl1	23.7
L1	72	Lf	46.4	Dl2	9.6
S	14.6	Wf	4.7	Dw1	30

3. RESULTS AND DISCUSSIONS



(a)

(b)

Fig. 4. Photograph of the fabricated single layer proximityfed 2x2 array antenna (a) top view and (b) bottom view.

The optimal dimensions from the numerically synthesized model as given in Table 1 have been used to fabricate the physical antenna prototype and the performance criteria are measured in a standard anechoic chamber to validate the numerically estimated results. Fig.4 shows the photographs of top and bottom view of the fabricated single layer proximity fed 2x2 array antenna with a slotted ground plane. The antenna is etched on a 1.6mm thick FR 4 substrate with a loss tangent of 0.001 and a dielectric constant of 4.3. Because it is inexpensive and has great mechanical qualities, FR 4 is utilized in this work. Using a commercially available 50Ω SMA coaxial connection, the fabricated microstrip patch array is activated electromagnetically.

The fabricated prototype of the proposed single layer proximity fed 2x2 array antenna is measured for reflection coefficient and gain using vector network analyzer. The measurements are carried out using Agilent E5063A ENA series RF network analyzer, which can be used for testing passive components like antennas up to 18 GHz. The calibration of the network analyzer is performed using short – open – load- through technique. After calibration, the fabricated prototype antenna is connected to analyzer and reflection coefficients are obtained. Broad band standard horn antenna operating between 1GHz and 18 GHz frequencies is used for far field pattern measurements. Both, antenna under test and standard horn antenna are placed inside a fully calibrated anechoic chamber for radiation pattern measurements. The antenna under test or fabricated antenna is placed on a turn table and the attached motor is allowed to rotate the antenna with 100 steps in both principle planes. The chambers are designed to absorb the reflections of electromagnetic radiations and to minimize interfering energy disturbances from external spurious sources.

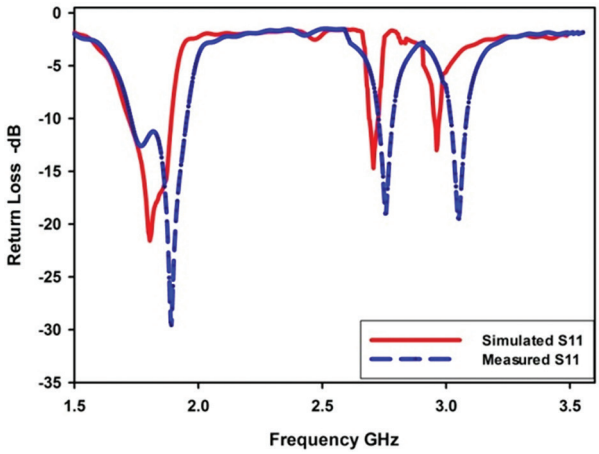


Fig. 5. Simulated and measured return loss characteristics of the proposed single layer proximity fed 2x2 array antenna.

Fig. 5. shows the simulated and measured return loss plot of proposed single layer proximity fed 2x2 array antenna with slotted ground plane for wireless applications. The results can be compared with previous result [12], which shows that the proposed array antenna structure radiates excellent with minimum loss. It is seen from the plot that the proposed array antenna operates for three different frequency bands. The three operating bands achieved during simulation are from 1.717GHz to 1.840GHz, with peak resonance at 1.810GHz; 2.686GHz to 2.731GHz with peak resonance at 2.707GHz and from 2.947GHz to 2.977GHz with peak resonance at 2.962GHz. For the first, second, and third bands, the proposed array antennas simulated bandwidths were 123MHz, 45MHz, and 30MHz, respectively. The measured fundamental resonance of 1.891GHz is observed with an impedance bandwidth of 234MHz (1.723-1.957GHz). The second resonance of 2.755GHz occurs with an impedance bandwidth of 69MHz (2.722 – 2.791GHz). Finally, the third resonance occurs at 3.052GHz with an impedance bandwidth of 75MHz (3.013 – 3.088GHz). Thus, the proposed array antenna is applicable to different wireless communication applications in L and S bands. Changes in inductance are due to manufacturing flaws, connection losses and imperfection in soldering are to be blamed for the minor disparity between the simulated and measured findings.

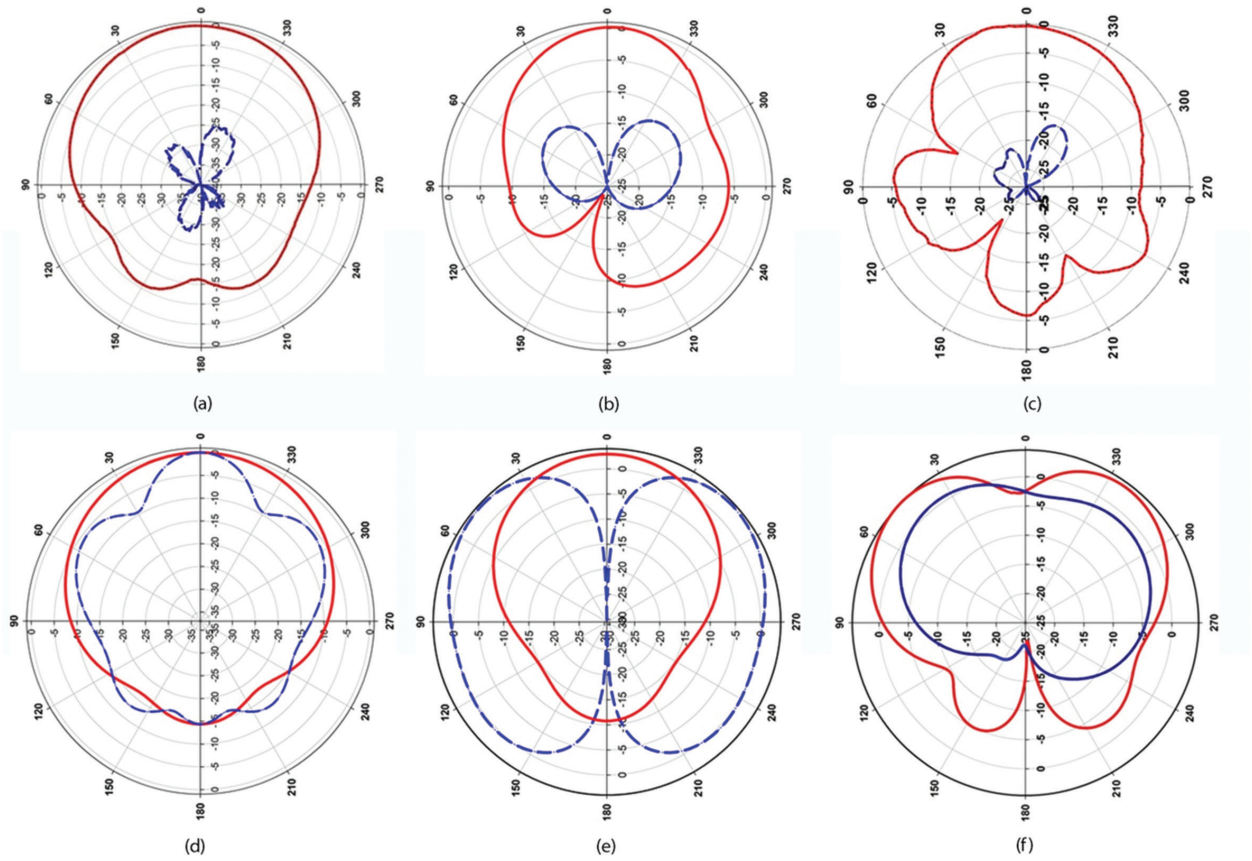


Fig. 6. Measured and normalized radiation patterns of the proposed single layer proximity fed 2x2 array antena at (a) 1.891GHz with Phi=0 (H plane), (b) 1.891GHz with Phi=90 (E plane), (c) 2.755GHz with Phi=0 (H plane), (d) 2.755GHz with Phi=90 (E plane), (e) 3.052GHz with Phi=0 (H plane), (f) 3.052GHz with Phi=90 (E plane)

Fig. 6. shows the observed radiation patterns of the proposed single layer proximity fed 2x2 array antenna in the E (y-z plane) and H (x–z plane) at the resonant frequencies of 1.891GHz, 2.755GHz, and 3.052GHz. Fig.6(a) and 6(b) illustrate the electromagnetic energy distributions of the array antenna in two principal planes at the lower resonant frequency of 1.891GHz. The radiation patterns at this frequency are in the bore sight direction in both planes. In both the E and H planes, the patterns are extremely directed towards the +Z (positive) direction, with modest crosspolarization levels. The antennas HPBW is on the order of 1180 in the E plane and 1050 in the H plane. Fig. 6(c) and 6(d) show the antennas radiation patterns at the second resonant frequency of 2.755GHz . In the H plane, substantial back lobes can be seen. Different established strategies have been employed to diminish the presence of back lobe, and in future study, an appropriate way will be used to reduce back lobe. The antennas half power beam width [HPBW] is in the order of 830 in the E plane and 620 in the H plane. Fig.6(e) and 6(f) illustrate the antennas radiation pattern at the top resonance frequency of 3.052GHz. On both sides, the antenna radiation patterns beam maxima in the H plane are a few degrees distant from bore sight direction. This feature will come handy in non-line of sight applications. The antenna's HPBW is on the order of 810 in the E plane and 690 in the H plane. It should also be noticed that the third operational frequency had a somewhat greater cross polarisation level.

The array antenna realized gain was measured based on the three antenna gain measuremen method [21-22]. The suggested 2x2 array antenna's observed gain at initial resonance frequency is 7.57 dBi, which is close to the simulated gain of 7.31 dBi. At the second and third resonant frequencies, the simulated and observed gains are 6.8dbi and 5.95dbi and 6.73dbi and 5.76dbi, respectively. The observed gain changes over the operational frequency ranges are depicted in Fig.7. The comparison of the proposed work and existing work [12], is discussed in Table 2.

Table 2. Comparison Analysis of the proposed work and existing work.

Frequencies GHz	Reflection Coefficient (dB)		Gain (dBi)	
	Existing Work	Proposed Work	Existing Work	Proposed Work
1.8	-22.56	-28.52	4.8	7.57
2.7	-18.37	-20.87	5.6	6.73
3.0	-19.15	-20.61	5.2	5.76

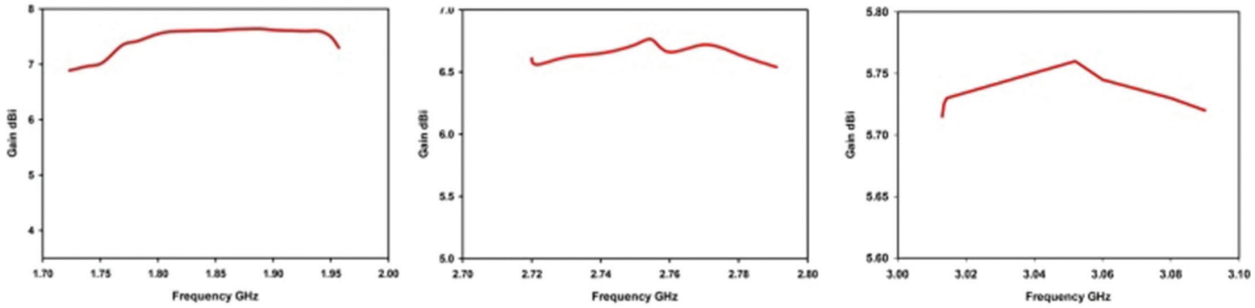


Fig . 7. Measured gain with frequency of the proposed single layer proximity fed 2x2 array antenna at (a) first band (b) second band and (c) third band

Advantages:

1. 2X2 Array Antenna
2. Good Return Loss is Obtained
3. High Gain is Obtained
4. Close aggrement between the measured and simulated results
5. It is used for Wireless Communication

4. CONCLUSION

A single layer proximity fed triple-band 2 x 2 patch array antenna with a slotted ground plane is described in this paper. Radiating array components and a microstrip feed line are printed on the substrate's top surface. Feed line stimulates printed array components electromagnetically. The 2x2 array antenna was designed and simulated with aid of CST studio suite, a high performance 3D electromagnetic analysis software package. The proposed 2x2 array antenna has been built and tested, and the measured and simulated results are very close. The single layer proximity fed array antenna revealed three bands with centre resonant frequencies of 1.891GHz, 2.755GHz, and 3.052GHz, with measured impedance bandwidths of 234MHz, 69MHz, and 75MHz respectively. The far-field radiation patterns of the principal planes (E and H) were measured in a fully equipped anechoic chamber. The gain of the proposed 2x2 array antenna is estimated using a three-antenna technique. The antenna may receive services from a variety of wireless technology networks that operate in both the L and S bands.

5. REFERENCES:

[1] J. Masroor, S. Ansari, S. Aslam, A. K. Saroj, "Sierpinski-Carpet Fractal Frequency Reconfigurable Microstrip Patch Antenna Design for Ku/k /Ka Band Application", Progress In Electromagnetics Research M, Vol. 106, 2021,pp. 59-69.

[2] P. K. Malik, S. Padmanaban, J. B. H. Nielsen, "Microstrip Antenna Design for Wireless Applications", 1st Edition, CRC Press, 2021.

[3] R. Garg, P. Bhartia, I. J. Bahl, A. Ittipiboon, "Microstrip Antenna Design Handbook", 1st Edition, Artech House Antennas, 2001.

[4] S. Alam, I. Surjati, T. Firmansyah, "Bandwidth Enhancement of Square Microstrip Antennas Using Dual Feed Line Techniques", International Journal of Electrical and Electronic Engineering and Telecommunications, Vol. 10, No. 1, 2021, pp. 60-65.

[5] L. Wang, Z. Zhu, "Performance enhancement of cross dipole circularly polarized antenna using parasitic elements", Microwave and Optical Technology Letters, Vol. 63, No. 12, 2021, pp. 124-127.

[6] M. Asaadi, A. Sebak. "Gain and Bandwidth Enhancement of 2 × 2 Square Dense Dielectric Patch Antenna Array Using a Holey Superstrate", IEEE Antennas and Wireless Propagation Letters, Vol. 16, 2017, pp. 1808-1811.

[7] A. S. Elkorany, A. N. Mousa, S. Ahmad, D. A Saleeb, A. Ghaffar, M. Soruri, Dalarsson, M.; Alibakhshikenari, E. Limiti, "Implementation of a Miniaturized Planar Tri-Band Microstrip Patch Antenna for Wireless Sensors in Mobile Applications", Sensors, Vol. 22, No. 2, 2022, pp. 1-13.

[8] A. Abdalrazik, A. Gomaa, A. A. Kishk, "Hexaband Quad-Circular-Polarization Slotted Patch Antenna for 5G, GPS, WLAN, LTE, and Radio Navigation Applications", IEEE Antennas and Wireless Propagation Letters, Vol. 20, No. 8, 2021, pp. 1438-1442.

[9] A. A. Deshmukh, S. B. Deshmukh, "Wideband Designs of SectoralMicrostrip Antennas Using Parasitic Arc Shape Patches", Progress In Electromagnetics Research C, Vol. 98, 2020, pp. 97-107.

[10] J. H. Lee, J. M. Lee, K. C. Hwang, D. W. Seo, D. Shin, C. Lee, "Capacitively Coupled Microstrip Comb-Line Array Antennas for Millimeter-Wave Applications", IEEE Antennas and Wireless Propagation Letters, Vol.19, No. 8, 2020, pp. 1336-1339.

[11] S. D. Jagtap, R. Thakare, R. K. Gupta, "Low Profile, High Gain and Wideband Circularly Polarized Antennas Using Hexagonal Shape Parasitic Patches", Progress In Electromagnetics Research C, Vol. 95, 2019, pp. 15-27.

[12] J. Abraham, "Proximity Fed Triple Band David Fractal 2×1 Microstrip Patch Antenna with DGS", Progress In Electromagnetics Research M, Vol. 107, 2022, pp. 91-103.

[13] N. Nie, X. Yang, Z. N. Chen, B. Wang, "A Low-Profile Wideband Hybrid Metasurface Antenna Array for 5G and WiFiSystems", IEEE Transactions on Antennas and Propagation, Vol. 68, No. 2, 2020, pp. 665-671.

[14] Z. Niu, H. Zhang, Q. Chen, T. Zhong, "Isolation Enhancement for 1×3 Closely Spaced E-Plane Patch Antenna Array Using Defect Ground Structure and Metal-Vias", IEEE Access, Vol. 7, 2019, pp. 119375-119383.

[15] S. Kannadhasan, R. Nagarajan, "Development of an H-Shaped Antenna with FR4 for 1-10GHz Wireless Communications", Textile Research Journal, Vol. 91, No.2, 2021, pp. 15-18.

[16] M.C. Derbal, A. Zeghdoud, M. Nedil, "A Dual Band Notched UWB Antenna with Optimized DGS Using Genetic Algorithm", Progress In Electromagnetics Research Letters, Vol. 88, 2020, pp.89-95.

[17] C. Kumar, D. Guha, "Asymmetric and Compact DGS Configuration for Circular Patch With Improved Radiations", IEEE Antennas and Wireless Propagation Letters, Vol. 19, No. 2, 2020, pp. 355-357.

[18] B. S. H. Prasad, M. V. Prasad, "Design and Analysis of Compact Periodic Slot Multiband Antenna with Defected Ground Structure for Wireless Applications", Progress In Electromagnetics Research M, Vol. 93, 2020, pp. 77-87.

[19] J. Abraham, "Investigations on multiband microstrip antennas and arrays for wireless communication applications," Ph.D. Thesis, School of Technology and Applied Science, M G University, Kerala, India, 2018.

[20] A. Azari, "A New Super Wideband Fractal Microstrip Antenna," IEEE Transactions on Antennas and Propagation, Vol. 59, No. 5, 2011, pp. 1724-1727.

[21] "IEEE Standard Test Procedures for Antennas", Technical Report, ANSI/IEEE Std. 149-1979, pp. 1-144, 1979.

[22] K. Harima, M. Sakasai, K. Fujii, "Determination of gain for pyramidal-horn antenna on basis of phase center location", Proceedings of the IEEE International Symposium on Electromagnetic Compatibility, Detroit, USA, 18-22 August 2008, pp. 1-5.

Hybrid Evolutionary Computing Assisted Irregular-Shaped Patch Antenna Design for Wide Band Applications

Original Scientific Paper

Rohini Saxena
Department of Electronics and Communication, University of Allahabad
Prayagraj, U.P., India.
rohini.saxena@gmail.com

J. A. Ansari
Faculty of Department of Electronics and Communication, University of Allahabad
Prayagraj, U.P., India.
jaansari@rediffmail.com

Mukesh Kumar
Faculty of Department of Electronics and Communication, SHUATS
Prayagraj, U.P., India.
mukesh044@gmail.com

Abstract *– A novel optimization concept for modeling irregular-shaped patch antenna with high bandwidth and efficient radiation attributes is proposed in this paper, along with the ability to accomplish the design at a reduced computational and cost burden. A revolutionary computing perception is established with Gravitational Search Algorithm (GSA) and Quantum Based Delta Particle Swarm Optimization (QPSO), now known as GSA-QPSO. The suggested model employed the GSA-QPSO algorithm strategically interfaced with a high-frequency structure simulator (HFSS) software through a Microsoft Visual Basic script to enhance irregular-shaped antenna design while maintaining wide bandwidth with suitable radiation efficiency over the target bandwidth region. The optimally designed microstrip patch antenna is fabricated on an FR-4 substrate with a surface area of 30×30×1.6 mm³. The evaluated outcome shows 96 % supreme radiation efficacy at 2.4 GHz whereas overall effectiveness is above 84% over the entire frequency range, with a nearly omnidirectional radiation pattern. In terms of impedance bandwidth, the suggested antenna offers 126.6 % over the operational frequency range from 2.34 GHz to 10.44 GHz. Fabrication and measurement results are also used to validate the simulated results. It exhibits the proficiency of the offered antenna design to be used for real-world wideband (WB) communication drives.*

Keywords*: Evolutionary Computing, GSA, QPSO, Microstrip Patch Antenna, Wideband, Radiation Efficiency*

1. INTRODUCTION

The demand for portable wireless communication systems has risen exponentially over the past few years due to high-speed demands across socio-industrial, defense, and scientific applications. Consequently, academia and industry have been encouraged to develop more effective solutions for Quality-of-Service (QoS) and Quality-of-Experience (QoE) purposes. To meet aforesaid demands, wireless antenna and allied technology have a decisive role to enable QoS and/or QoE-centric communication under the different operating conditions. The increased use of WB/UWB technologies in recent years has required antennas to be more efficient and robust to provide communications services within expanded bandwidths. Noticeably, a higher bandwidth doesn't guarantee optimal communication, as an antenna requires robustness towards a wider frequency range, and sufficient radiation even at the reduced design cost and size. This as a result has broadened the horizon for academia-industries to achieve an antenna solution with optimal design, and radiation performance even over the larger bandwidth [1]. Considering available antenna technologies, the microstrip patch antenna (MPA) is perceived as one of the most favorable antennas for wireless transmission. Being cost-efficient, lightweight, and easier to implement MPA is used in major contemporary UWB applications [1-4]. Moreover, the MPAs can be found on both planar as well as non-planar surfaces, making them suitable for the applications serving satellite communication, remote sensing, defense communication, and radio-frequency iden-

tification drives [5-8]. Structurally, the aforesaid patches were connected through a specific patch, often called element-step sized. In order to operate it over the target bandwidth, the far-field radiation characteristics are changed by selecting the optimal shape and size value. Undeniably, it expands the horizon for further research. However, in practice, finding an optimal design with different constraints and objectives, such as resonant frequency, BW, miniaturized design, or optimal material is a complex task [2-7]. A design optimization, especially when applied to WB applications, has been recognized to be more complex [7]. This is because inappropriate design parameters and material selection may cause a shift in bandwidth and corresponding resonant frequency that eventual can impact the overall performance of desired communication systems [9]. As above specified fact as motivation, different MPA design approaches have been recommended, which are primarily based on analytical and numerical approaches. These design approaches have their restraints. For example, the analytical approach is easy and suitable for fixed structures of patches. Whereas, numerical approaches are relevant for all shapes of the patch but undergo tedious cycle events. However, both design concepts undergo low controllability over higher operating frequency and BW [10]. Consequently, it limits their application in major at-hand wireless communication purposes. Typically, a patch antenna undergoes such detrimental performance due to random size selection (in the case of smaller size, it forces it to incline towards higher frequency and lower bandwidth) and shift in operating frequency, which is common in contemporary adaptive modulation and multi-rate transmission systems [1] [8-9]. Though, optimal selection of design parameters of MPA such as patch size, substrate thickness, dielectric constant and feeding method, etc., can help achieve to operate over the desired frequency range [10-12]. There have been several investigations conducted with such motives; however, retaining stable and controllable performance with optimal trade-offs between dimension and performance has remained a challenge [9] [13]. During the past few years, artificial neural network (ANN) or data-driven machine learning (ML) based approaches have been proposed to estimate a suitable set of design parameters for MPA design. Some of the key methods such as neuro-computing concepts [14-26] performed design parameter estimation; however, their optimality towards a robust solution remained an unexplored story. This is because most of the existing systems focused on design optimization by learning a set of input patterns, irrespective of the end outcomes and its realistic patterns or performance for WB application. As well, these approaches didn't consider the key limitations of ML methods like local minima and convergence. Though few efforts have been made toward using heuristic-based design parameters estimation; however, the majority of the existing approaches failed in addressing inherent problems such as convergence, inappropriate fitness estimation, and more importantly varying operating condition centric optimization.

Considering it as motivation, in this research paper a state-of-art new and robust hybrid evolutionary computing assisted polyline MPA design is proposed to be used for WB applications. This paper contributed a new hybrid evolutionary computing concept by using GSA and QPSO, here onwards called GSA-QPSO for irregular-shaped MPA design optimization. Noticeably, the key intention behind the use of the hybrid heuristic co-evolutionary concept is to improve the accuracy of the radiator design results, while avoiding any possible local minima and convergence issues. The proposed model used the GSA-QPSO algorithm strategically interfaced with HSFF through Microsoft Visual Basic (VB) script to enhance irregular-shaped MPA design while maintaining reflection coefficient (S_{11}) below -10 dB over the target BW region. The proposed model is simulated for the different test cases, which deliver the varying coordinates of the MPA. In terms of the reflection coefficient, three different cases are observed to get the optimum coordinate parameters of the designed MPA. Finally, for case 3, the exhibited MPA provides 126.6 % impedance BW above the frequency scope of 2.34 to 10.44 GHz. The optimum designed model is fabricated on an FR-4 substrate with an area of $30\times30\times1.6$ mm^3.

The remaining sections of this research manuscript are divided as follows. Section II discusses some of the key literature about MPA design optimization, followed by a proposed system in Section III. Section IV presents the implementation while the simulation results and discussion are given in Section V. The overall research conclusion is discussed in Section VI. References used in this manuscript are given at the end of the manuscript.

2. RELATED WORK

This section reviews existing literature along with a discussion of future research challenges. A genetic algorithm (GA) technique is proposed to optimize the geometry of square-shaped MPA for WB and broadband applications [27-28]. Therefore, Silva et al. [29] suggested a circular-ring MSA design optimization by using the self-organizing GA (SOGA) technique with UWB features. The simulated and experimented antenna provides wide bandwidth of more than 9 GHz and 6 GHz. In [30], the authors introduced a multi-adaptive neuro-fuzzy inference system (MANFIS) to predict the bandwidth of the U-shape slot-loaded RMPAs design. They found the PSO-MANFIS model provided more accurate results than the GA-MANFIS model. Conversely, Mir et al. [31] investigated an automated optimization procedure based on Bayesian-optimization (BO) and bottom-up-optimization (BUO) to design the MPA for broadband with high flat-gain features. The initial structure of the MPA is designed by the BUO approach while the BO process is applied to forecast appropriate dimensional parameters. In [32], the curve fitting-based PSO technique is presented to design a "plus" shape slotted MPA for BW improvement with resonant frequency at 2.4 GHz.

Table 1. Comparison of the proposed antenna with some recently reported optimization techniques.

Ref.	Size (mm²)	Operational frequency (GHz)	Bandwidth (%)	Optimization technique	Applications
23.	33.4×40.6	3.1 to 5.495	55.73	MLPFFBP and RBF	WLAN
24.	15.528×18.40	9.91 to 10.5	5.78	MLPFFBP and RBF	Wideband
27.	38.4×38.4	10.6 to 11.15	5.05	GA	Wideband
28.	30×30	2.3 to 2.6	12.24	GA	Broadband
29.	33×28	3.8 to 9.6	88.23	SOGA	UWB
30.	34.9×31.3	2.0 to 10.75	137.25	MANFIS–PSO	UWB
31.	20×18	8.7 to 10	13.90	BUO and BO-ANN	Broadband
32.	38×47.6	1.795 to 2.95	46.99	PSO	Wi-MAX, WLAN
33.	34×33.35	2.94 to 10.98	115.35	PSO	UWB
36.	24×24.4	3.1 to 10.6	109.48	PSO	UWB
39.	30×30	3.1 to 10.6	109.48	GSA	UWB
Proposed	**30×30**	**2.34 to 10.44**	**126.6**	**GSA-QPSO**	**Bluetooth, WLAN, Wi-Max and UWB**

Therefore, the authors developed an irregular-shape MPA [33] and rectangular MPA [34] for UWB

application where PSO is applied to estimate design parameters under unknown dimensional specifications. Nonetheless, a differentia-evolution (DE) based estimation method is recommended to design a square monopole antenna, but the overall dimensions of the antenna are not justified [35]. However, a miniaturized stepped-triangular MPA design parameter is tuned for WB application with the PSO technique [36]. Thus, various types of printed MPA using PSO [37] and using hybrid GA-PSO [38] can be designed for UWB applications.

Similar to PSO, authors [39] proposed a GSA function based on the concept of gravitational force and mass interaction. Detailed analysis revealed that GSA can be superior to classical GA, PSO, etc. based approaches. A study conducted by the authors [40], which compared GSA against a PSO-based optimization technique that was designed for the synthesis of ring array MPA (RA-MPA), revealed that GSA is more effective, particularly in terms of fitness values and time. In [41], the authors designed a reconfigurable RA-MPA using GSA, while the same method [42] was applied to estimate the dimensional features of the rectangular MPA. However, it exists a simple antenna design employing an equivalent transmission line model without significant attention to WB application demands and operating environment. The most of above reported literature focused on design optimization of MPA. Therefore, a novel optimization concept based on GSA and QPSO for the modeling of irregular-shaped patch antenna with desired BW is proposed in this paper. Brief comparisons in terms of antenna surface area dimensions, operating frequency band, optimization techniques, and their applications are also reported in Table 1 with some recently stated literature.

3. PROPOSED MODEL AND METHODS

This section primarily discusses the overall proposed system and its implementation.

In sync with the overall research intend, where the key emphasis has been made on designing a state-of-art new heuristic-based irregular-shaped MPA design for WB applications. In this research, we focused on designing an irregular-shaped MPA with multiple cones to reduce design complexity, while retaining wider BW.

Unlike classical heuristics algorithms such as GA, PSO, and even GSA, etc., which are often criticized for their local minima and convergence limitations, in this paper, a new hybrid evolutionary computing concept is proposed by applying QPSO and GSA to perform irregular-shaped antenna design optimization. The proposed GSA-QPSO model with a state of art new weighted Average Personal Best Position (APBP) and Adaptive Local Attractor (ALA) not only intends to alleviate the at-hand convergence and local optima problem but also exploits efficient local search ability (by GSA) to ensure highly accurate and computationally efficient optimization solution. In other words, in the suggested GSA-QPSO model, QPSO (Note: we applied Quantum based Delta PSO model with Levy's flight-based particle positioning concept using Mantegna algorithm) at one hand exploits optimal "social thinking" ability to estimate global (gbest) values, while GSA helps to strengthen its "local search ability". Being a co-evolutionary approach [PSOGSA], the proposed algorithm performs in parallel, where each contributes to estimating the optimal design parameters of the targeted irregular-shaped MPA design with higher bandwidth performance while maintaining lower (-10 dB) S_{11} output. The proposed GSA-QPSO model has been integrated with the finite element method (FEM) using HFSS to obtain optimal design parameters of the irregular radiator for wider BW.

Before discussing the system implementation, a snippet of the proposed GSA-QPSO model is given in the subsequent sections.

3.1. GRAVITATIONAL SEARCH ALGORITHM (GSA)

GSA is a recently evolved heuristic search algorithm that uses the conception of Newton's law of gravity and object motion [43]. Similar to the other models, GSA at first initializes the population where it considers the initial random locus of the N agents. The random N agents are deployed as (1).

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots, x_i^D\right) \quad \text{for } i = 1,2, \ldots., N \tag{1}$$

Where D signifies the search space's dimension while x_i^d refers to the i^{th} agent present in the d^{th} dimension. Thus, each participating agent is considered an object and the mass of each participating agent is obtained in the form of the fitness value of the current population. Mathematically, the mass of each participating object or agent is estimated using the fitness value, as defined in (2-3). Where, $q_i(t)$ and $M_i(t)$ state the fitness value and the mass of the i^{th} agent, correspondingly. Noticeably, these values are obtained over the at-hand t^{th} iteration.

$$q_i(t) = \frac{fit_i(t) - Worst(t)}{best(t) - Worst(t)} \tag{2}$$

$$M_i(t) = \frac{q_i(t)}{\sum_{j=1}^{N} q_j(t)} \tag{3}$$

Considering the minimization issue, the best solution and worst solution are defined as best(t) and worst(t), using (4) and (5), respectively.

$$best(t) = \min_{j \in \{1,\ldots,N\}} fit_j(t) \tag{4}$$

$$worst(t) = \max_{j \in \{1,\ldots,N\}} fit_j(t) \tag{5}$$

Now, considering Newton's law of gravitation, the cumulative force active on an i^{th} agent from another j^{th} agent can be estimated using (6). The parameter $R_{i,j}$ signifies the Euclidian distance existing between i^{th} and j^{th} agents, while ϵ represents a constant value.

$$f_{ij}^d(t) = G(t) \frac{M_i(t) \times M_j(t)}{R_{i,j+\epsilon}} \left(x_j^d(t) - x_i^d(t)\right) \tag{6}$$

As defined in (7), G(t) describes a function for an iteration time t, that drops exponentially over the period. G0 is the inception value, while the reduction factor and total amount of iterations are denoted by α and T, respectively.

$$G(t) = G_0 e^{-\alpha \frac{t}{T}} \tag{7}$$

So, the overall force acting on an agent i owing to the overall current population is estimated as per (8).

$$f_i^d(t) = \sum_{j \in Kbest, j \neq i} rand_j F_{ij}^d(t) \tag{8}$$

In (8), K_{best} presents the set of the initial K agents possessing the highest mass value, which decreases linearly over time-period t, and thus executing over iterations in result a single agent in K_{best} with the highest fitness value. The component $rand_j$ refers to a linearly deployed arbitrary number placed at the interval of [0, 1]. Noticeably, it is employed to assure the stochastic nature of the detection mechanism. Now, employing the perception of Newton's second rule of motion, the acceleration of the i^{th} agent can be attained by (9). Where $F_i^d(t)$ is the entire force acting on ith agent and $M_i(t)$ denotes the mass component of the agent i at time t.

$$a_i^d(t) = \frac{F_i^d(t)}{M_i(t)} \tag{9}$$

Now, the agent's velocity is updated by applying equations (10-11).

$$v_i^d(t+1) = rand_i \times v_i^d(t) + a_i^d(t) \tag{10}$$

$$x_i^d(t+1) = x_i^d(t) + v_i^d(t+1) \tag{11}$$

Thus, employing the above-derived equations, the acceleration can be estimated using (12).

$$a_i^d(t) = \sum_{j \in Kbest, j \neq i} rand_j \times M_j(t) \frac{G(t)}{R_{ij}(t) + \epsilon} \left(x_j^d(t) - x_i^d(t)\right) \tag{12}$$

Accordingly, the positions of masses are estimated as per (11), and once reaching the maximum number of iterations, the anticipated model stops, and the optimal solution is considered the best sub-solution to the at-hand problem.

3.2. QUANTUM-BASED DELTA PSO

Similar to the other heuristic model, PSO is also a kind of stochastic population-based approach, motivated by the cumulative behavior of bird flocks. In this process, each participating particle is considered as a solution, which possesses distinct fitness values and velocities. These possible solutions or candidates (say, particles) can fly across the multidimensional search space by learning from the previous traces or historical information. Noticeably, the key historical information encompasses the memories of the participating particles towards their individual best positions and the global best position in the groups over the search period. Unlike GA systems, PSO is easier to implement due to lower tuning parameters and allied adjustment [44-45]. Despite its superiority, it often undergoes adverse performance such as the local optima, and premature convergence [46]. Numerous efforts have been made to enhance PSO; however, a recent innovation in the form of quantum mechanics and trajectory analysis enabled it to perform more superior [47]. Unlike classical PSO algorithm, QPSO avoids any additional need for a velocity estimator and distinct velocity vector for particles. Moreover, it requires fewer parameters to tune which makes it computationally more efficient. Considering such robustness, in this paper, the hybrid search concept is applied by using the GA-QPSO model to achieve an appropriate convergence rate and global optima. However, achieving it using classical QPSO is highly complex and hence requires enhancing

global search capacity and optimal acceleration rate, simultaneously. In the native QPSO model [47], both the average personal best position and local attractor have a decisive impact on overall performance. Noticeably, the average best position signifies the average of the personal best positions of the participating particles and hence doesn't considers the difference of the impact of the different particles possessing different fitness values to search for the optimum solution. It marks questions on the generalization of the classical QPSO model. Though, to achieve better performance authors have made efforts like Gaussian distributed local attractor point-based QPSO (GAQPSO) algorithm [48]. In the GAQPSO model, the local attractor was subjected to the Gaussian distribution whose average value was the original local attractor, as defined in classical QPSO. Unlike [48], Jia et al. [49] developed a weighted sum of particle personal and global best positions as the local attractor to improve the performance. However, it was unable to monitor the population diversity over the computation period and hence is limited to at hand antenna design optimization problem. Additionally, these approaches require ensuring a diversity of solutions (being population-based stochastic optimization problems) [50]. In this reference, nursing the diversity of QPSO to obtain local attractors towards particle optimization can improve the overall performance. Therefore, to balance the local as well as global search ability, in this research paper we applied a set of weighted coefficients which could differentiate the fitness of particles to estimate the "average personal best position" (APBP) even at reduced computational overhead. Here, we applied the GSA algorithm to estimate the set of weighted coefficients which were later used to update the APBP to further improve adaptive local attractor (ALA) for better (design) optimization.

3.2.1. Particle Swarm Optimization (PSO)

In PSO, the movement of the participating particles is directed by their corresponding best-known position (pbest) and the best-known position of the entire particles (gbest). In this manner, the location and velocity of the i^{th} participating particle can be estimated as per (13) and (14), respectively.

$$V_i^{t+1} = wV_i^t + c_1 r_1 (pbest_i - X_i^t) + c_2 r_2 (gbest - X_i^t) \quad (13)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1} \quad (14)$$

Where, $X_i^t=(x^t_{i1},\ldots,x^t_d,\ldots,x^t_{iD})$ represent the position vector, while $V_i^t=(v^t_{i1},\ldots,v^t_d,\ldots,v^t_{iD})$ corresponds to the velocity vector for i^{th} particle at t^{th} iteration. c_1 and c_2 represent the positive constants controlling the impact of pbesti and gbest on the search process. The other parameters, r_1 and r_2 represent the arbitrary values existing from 0 to 1. In (13), w presents the inertia weight that helps in balancing the algorithm's global and local search ability. Here, the fitness value of each particle's position is estimated as per a certain fitness function. Thus, PSO is executed iteratively to estimate the values of (13) and (14), till it reaches the stopping criteria or the predefined number of iterations. Noticeably, reaching the stopping criteria, the velocity update turns out to be near zero.

3.2.2. Quantum-based PSO (QPSO)

Recalling the suggestions in [44-50], which stated that the convergence of PSO can be accomplished only when the particle converges to its local attractor (for $1 \le d \le D$), $LA^t_i=(la^t_{i1},\ldots, la^t_{id},\ldots, la^t_{iD})$. It can be denoted mathematically in the equations (15-16).

$$la_{id}^t = \frac{c_1 r_1 pbest_{id}^t + c_2 r_2 pbest_{id}^t}{c_1 r_1 + c_2 r_2} \quad (15)$$

$$la_{id}^t = \varphi pbest_{id}^t + (1-\varphi) gbest_d^t,\ \varphi \sim \mathbb{U}(0,1) \quad (16)$$

Where the parameter t signifies the number of iterations, while j presents the arbitrary number distributed uniformly over (0, 1), that is $\varphi \sim U(0,1)$. Here, $pbest_i$ presents the best earlier position observed by the i^{th} paticle, while the current global solution i.e. global best position is indicated by gbest. Being motivated by the quantum mechanism and trajectory assessment methods of PSO, the authors [47-48] proposed two models named QPSO and delta potential well model (QDPSO). In QDPSO, the location of the participating particle i, over t^{th} iteration is obtained as per (17), by using Levy's flight method.

$$x_{id}^{t+1} = \begin{cases} la_{id}^t + \alpha |la_{id}^t - x_{id}^t| \ln\left(\frac{1}{u}\right), & \text{if rand} \ge 0.5 \\ la_{id}^t - \alpha |la_{id}^t - x_{id}^t| \ln\left(\frac{1}{u}\right), & \text{Otherwise} \end{cases} \quad (17)$$

In (17), the parameters u and rand indicate random numbers distributed arbitrarily over (0, 1), while a indicates the positive real number, here we call the contraction-expansion coefficient (CEC), which is defined as a=0.5+0.5 (T-t)/T to balance local as well as global search capability of QDPSO. Where, t and T are iterations and the maximum number of iterations, correspondingly.

$$mbest^t = \frac{1}{s}\left(\sum_{i=1}^{S} pbest_{i1}^t, \ldots, \sum_{i=1}^{S} pbest_{id}^t, \ldots, \sum_{i=1}^{S} pbest_{iD}^t\right) \quad d = 1,2,\ldots D \quad (18)$$

In this method, a global point indicating $mbest=(mbest_1,\ldots,mbest_d,\ldots,mbest_D)$ and stated as the APBP, which is applied to improve the global search skill. The global point of the t^{th} iteration is obtained using (18). In (18), S presents the number of particles. Therefore, the position of the i^{th} particle in t^{th} iteration is updated as per (19). In this case, mbesttd presents the APBP of the swarm for the d^{th} dimension at the t^{th} iteration.

$$x_{id}^{t+1} = \begin{cases} la_{id}^t + \alpha |mbest_d^t - x_{id}^t| \ln\left(\frac{1}{u}\right), & \text{if rand} \ge 0.5 \\ la_{id}^t - \alpha |mbest_d^t - x_{id}^t| \ln\left(\frac{1}{u}\right), & \text{Otherwise} \end{cases} \quad (19)$$

3.2.3. Population Diversity

Typically, the swarm's participating candidates or the population diversity is vital towards estimating and tuning its optimal path estimation. Population diversity can be estimated as per (20).

$$\sigma^2(t) = \sum_{i=1}^{S}\left(\frac{f_i^{(t)} - f_{avg}^{(t)}}{F}\right), f_{avg}^{(t)} = \frac{1}{S}\sum_{i=1}^{S} f_i^{(t)} \quad (20)$$

Where $\sigma^2(t)$ represents the sum of squared deviations of the particles' fitness values, S stands for the swarm size, $f_i^{(t)}$ is the fitness of the i^{th} particle at the t^{th} iteration, $f_{avg}^{(t)}$ is the average fitness of the swarm at the t^{th} iteration and F is the normalized calibration factor to confine $\sigma^2(t)$. Mathematically, it is denoted as (21).

$$F = \begin{cases} \max\left|f_i^{(t)} - f_{avg}^{(t)}\right|, & \text{if } \max\left|f_i^{(t)} - f_{avg}^{(t)}\right| > 1 \\ 1 & \text{otherwise} \end{cases} \quad (21)$$

3.2.3.1. Weighted APBP and ALA-based QPSO

In QPSO, mbest of the population is tracked by i^{th} particle during the search process. It applies coefficients of the similar weights to form the linear combination of each particle's best position and is unable to differentiate the variance in the impact of the particles possessing different fitness values on guiding particles i to identify the global solution. Classical PSO models often undergo loss of the significant information hidden inside the particles' personal best positions (information). Being a minimization problem, the elite particle would have the minimum objective function value. In other words, the smaller objective function value of a particle would signify a corresponding better fitness value. In this approach, the elite would enable a better solution and therefore it is applied towards APBP estimation by assigning higher weights to the particles possessing better fitness while assigning smaller weights to those with relatively lower fitness values. In the offered model, the weighted APBP values are estimated as per (22) and (23). Consequently, based on the feedback of the fitness values of the particles can be estimated for guiding particles i to attain the global optima solution.

$$r_i(t) = \begin{cases} \frac{1}{S-1}\left(1 - \frac{f_{Obj}^{i}(t)}{\sum_{k=1}^{S} f_{Obj}^{i}(t)}\right), & \text{if } \sum_{k=1}^{S} f_{Obj}^{k}(t) \neq 0 \\ \frac{1}{S} & \text{Otherwise} \end{cases} \quad (22)$$

$$\begin{aligned} mpbest^t &= \sum_i^S r_1(t) pbest_i^t \\ &= \left(\sum_{i=1}^{S} r_1(t) pbest_{i1}^t, \ldots, \sum_{i=1}^{S} r_i(t) pbest_{iD}^t\right) \end{aligned} \quad (23)$$

In (22-23), t presents the at-hand iteration number, while the objective function value is defined as $f_{Obj}^{\ i}(t)$. The number of particles in the considered swarm population or particles is S, while $r_i(t)$ refers to the coefficient of the best position which is employed to construct the weighted APBP. Observing (22), one can find that the sum of $r_i(t)$ is 1, where $r_i(t)$ exists between 0 and 1 over the iteration t. If the sum of the objective function values of the all-participating agents will be 0, then coefficient $r_i(t)$ would be 1/S. Otherwise, the smaller $f_{Obj}^{\ i}(t)$ value leads to a larger $r_i(t)$ value. Summarily, when estimating a weighted APBP to guide the particles in a swarm over a trajectory to get the optimal solution, the larger fitness value of a particle would yield a more significant and near-optimal best position. In this manner, the proposed QPSO model is capable to distinguish the impact of the particles even with the varied fitness values.

3.2.3.2. Adaptive Local Attractor (ALA)

Considering an existing study [47-48] [51], where authors found that each participating agent or the particle in PSO intends to converge towards its local attractor (LA). Observing (15) or (16), the LA function amalgamates pbesti and gbest. And therefore, it becomes pertinent to estimate an optimal way to amalgamate the relevant information embedded in the above-stated two best-known positions. Being a population-based stochastic prediction and optimization method; it is expected to encourage the initial population to wander across the search space without converging around local optimal. During the later phase, it becomes necessary to improve convergence towards the global optimum, to realize the best solution. As a result, population diversity is significant in population-based optimization problems, as it impacts their performance. Although high-diversity results in better solution retrieval, particularly in the initial iterations, in the later phase, it is important to utilize a small area of the search space to retain computationally efficient solutions without premature convergence or time exhaustion. Similarly, the experience of each particle has a larger impact on particles after updating their position at the beginning of the next iteration. In contrast, other particles' experience has a higher impact on particles when updating their subsequent position at the later stage of iterations. Observing (20), one can find that $\sigma^2(t)$ exists between 0 and S. In case all particles are found at the same position, $\sigma^2(t)$ turns out to be zero, signifying the strongest (swarm) aggregation degree. On the contrary, $\sigma^2(t)$ changes to be S when all absolute discrepancies between the current fitness values of whole particles and their average fitness values are equal to one. In this manner, the sum of squared deviations of the particles' fitness values illustrates a reducing pattern with the increase in the number of generations. In this reference, a new approach is formulated to estimate the LA using (24). The prime motive of this approach was to improve the global search within a short span or the early part of the optimization to alleviate the problem of convergence and to accomplish the global optima at the end. In this manner, the place of the particle or the agent i over the iteration t is updated as per (25).

$$Al_{id}^t = \varphi\frac{\sigma^2(t)}{S} pbest_{id}^t + (1-\varphi)\left(1 - \frac{\sigma^2(t)}{S}\right) gbest_d^t \quad (24)$$

$$x_{id}^{t+1} = \begin{cases} Al_{id}^t + \alpha\left|mpbest_d^t - x_{id}^t\right| \ln\left(\frac{1}{u}\right), & \text{if rand } \geq 0.5 \\ Al_{id}^t - \alpha\left|mpbest_d^t - x_{id}^t\right| \ln\left(\frac{1}{u}\right), & \text{otherwise} \end{cases} \quad (25)$$

When implementing the proposed QPSO model with weighted APBP and ALA, the position of the particle i can be determined as per (25). Where, $\sigma^2(t)$ refers to the sum of squared deviations of the participating par-

ticle's fitness values at t^{th} iteration, over S swarm size, and j is a random number distributed uniformly in the range of (0, 1). In equation (25), the parameter a and u possesses the similar significance as shown in (17), while $mpbest^t_d$ refers to the weighted APBP for the d^{th} dimension at the t^{th} iteration. The ALA is given as Al^t_{id} for i^{th} particle over dth dimension at t^{th} iteration. Thus, the proposed weighted APBP and ALA have been implemented using (24), which estimates the optimal set of parameters for non-linear radiator design.

The overall anticipated model employs the subsequent approach to achieve design parameter optimization.

- Step-1: Initialize population with swarm size S and Max_Iteration count T.
- Step-2: Deploy particles or agents across the swarm with arbitrary position vectors.
- Step-3: Estimate pbest for every agent and gbest for the complete population or swarm.
- Step-4: Estimate population diversity (J) using (20).
- Step-5: Update weighted APBP mpbest as per (23).
- Step-6: Update ALA for each agent using (24).
- Step-7: Update the position of each agent as per (25).
- Step-8: If stopping criteria are not met, go to step 3.

Recalling the fact that being a low-level co-evolutionary concept, both GSA and QPSO algorithms are applied in parallel to estimate the most optimal design parameters for an irregular polyline antenna design with higher BW performance, with minimal S_{11} outputs. A snippet of the integration of the GSA and QPSO model to yield the above-stated result is given as follows:

In order to amalgamate both GSA and QPSO models, the subsequent approach has been applied.

$$V_i(t+1) = w \times V_i(t) + c'_1 \times rand \times ac_i(t) + c'_2 \times rand \times \left(mgbest - x^t_{id}(t)\right) \quad (26)$$

The parameter $V_i(t)$ presents the velocity of the i^{th} agent at t^{th} iteration, while c'_j refers to the weighting factor, and w is a reduction function. Here, rand represents a random number existing from 0 to 1, while $ac_i(t)$ presents the acceleration of the i^{th} agent over t^{th} iteration. The best solution realized so far is given by gbest. Using the GSA-QPSO model, the place of the participating particles is updated iteratively over each iteration using the following equation (27).

$$x^{t+1}_{id}(t+1) = x^t_{id}(t) + V_i(t+1) \quad (27)$$

Functionally, in the GSA-QPSO model, initially, all participating particles or the agents are arbitrarily initialized, where each solution is considered as a candidate solution. Once initializing the model, the proposed model estimates gravitational force, gravitational constant, and consequent forces amongst particles. Subsequently, it estimates the acceleration of the particles, followed by pbest update and acceleration estimation. By doing so, the results show the most optimal velocity and position of the particles, indicating the optimal solution. The details of the hybridization concept can be found in [51].

4. SYSTEM IMPLEMENTATION

A snippet of the targeted irregular-shaped patch antenna is given in Fig. 1. Fig. 1 presents the initial radiator shape, which is supposed to be processed for optimization to yield an optimal design with expected performance (higher bandwidth and frequency operation while maintaining a lower reflection coefficient (S_{11}). The initial parameters of MPA are established based on the stated concept and calculation [5-7], [37-38]. The suggested MPA model encompasses an irregular-shaped radiator fed by a 50 Ω microstrip line with a predefined length (L_f) and width (W_f). Here, the measurement of L_f is considered as 12.5 mm, while W_f is fixed at 3 mm. As the base material for fabrication, FR-4 substrate is considered with a surface area of 30×30 mm^2 and 1.6 mm thickness. The considered material possessed a relative permittivity of 4.4, while the loss tangent is considered as 0.02.

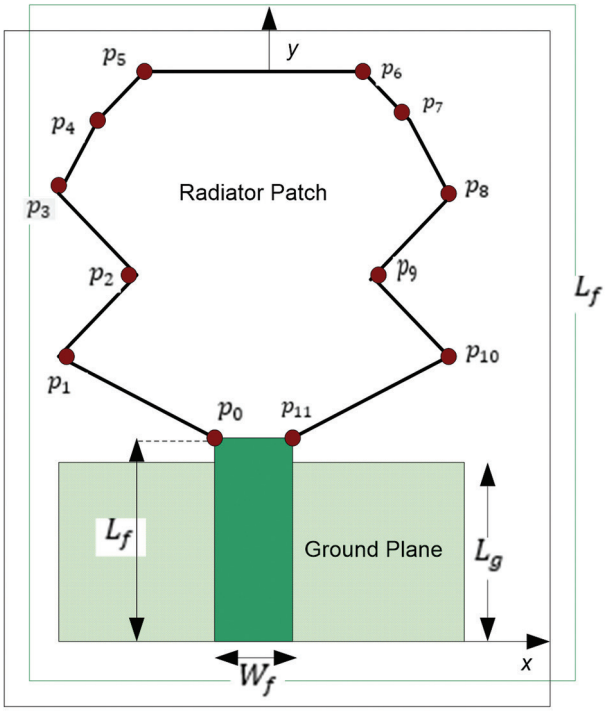


Fig. 1. Initial target irregular-shaped MPA design.

Unlike classical patch antenna designs, ground metallization under the radiator has been avoided to enable proper functionality; however, it required a partial ground length (L_g) of 11.5 mm. The proposed patch antenna design encompassed an irregular-radiator shape that changes shapes with non-linear steps in both x as well as y directions. In the x-direction, there are non-linear steps between x_1 and x_5, with a lower limit of 0.01 mm and an upper limit of 14.5 mm. Δy_1 to Δy_5 represent the varying steps in the y-direction with the lower limit

of 0.01 mm and the higher limit of 5 mm. In this model, the left half of the radiator shape is considered as a set of polylines, while another half (i.e., the right half) represents the mirror image. Where maintained the initial coordinate point (p_0) as the output value of $x_0=W_f/2$, along with $y_0=L_f$ (Eq. (28)). Now, the other coordinates of the antenna radiators (i.e., p_1, p_2, p_3, p_4, and p_5) are obtained as per equations (28-39), which are found using the GSA-QPSO algorithm.

$$p_0 = (x_0, y_0,) \quad (28)$$

$$p_1 = (x_1, y_0 + \Delta y_1) \quad (29)$$

$$p_2 = (x_2, y_0 + \Delta y_1 + \Delta y_2) \quad (30)$$

$$p_3 = (x_3, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3) \quad (31)$$

$$p_4 = (x_4, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3 + \Delta y_4) \quad (32)$$

$$p_5 = (x_5, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3 + \Delta y_4 + \Delta y_5) \quad (33)$$

$$p_6 = \left(-x_5, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3 + \Delta y_{4+\Delta y_5}\right) \quad (34)$$

$$p_7 = (-x_4, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3 + \Delta y_4) \quad (35)$$

$$p_8 = (-x_3, y_0 + \Delta y_1 + \Delta y_2 + \Delta y_3) \quad (36)$$

$$p_9 = (-x_2, y_0 + \Delta y_1 + \Delta y_2) \quad (37)$$

$$p_{10} = (-x_1, y_0 + \Delta y_1) \quad (38)$$

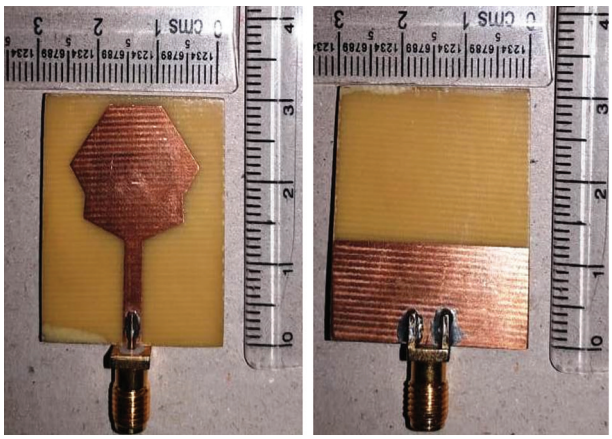
$$p_{11} = (-x_0, y_0,) \quad (39)$$

Though, MATLAB is applied as the key development tool; where HFSS has been utilized to validate the performance of reflection coefficients and radiation characteristics. Additionally, Microsoft VB-script has been used to interface (GSA-QPSO) MATLAB program with HFSS. The simulation environment is designed in such a manner that inputting initial model parameters in Microsoft VB-script, it applied HFSS to estimate the S_{11} parameter, which is subsequently passed to the MATLAB program to assess fitness function and update design parameters, iteratively. Eventually, once updating the coordinate parameters of the target model (Fig. 1), the updated design parameters were fed back to the HFSS through VB-script, and thus it generates the updated S_{11} parameters. This process continued until the obtained S_{11} parameter reached the expected goal. Noticeably, the maximum iteration has been assigned as 500. The proposed research intended to maintain S_{11} parameter below -10 dB within the frequency range of 2.4 to 10.4 GHz. The frequency interval of 0.1 GHz has been considered throughout the targeted WB frequency range.

$$\text{Fitness} = \min\left(\sum_{2.40\ \text{GHz}}^{10.4\ \text{GHz}} S_{11}\right) \quad (40)$$

Thus, maintaining the minimum fitness as per (40), the optimized antenna coordinates have been obtained which resulted in an optimal patch antenna design, serving optimal radiation over 2.4 GHz to 10.4 GHz.

The final optimized coordinates are considered as a proposed antenna. The fabricated model of the MPA design is shown in Fig. 2. The overall simulation and measured results as well as allied inferences are given in the subsequent sections.



(a) Top (b) Bottom

Fig. 2. Dimensions of the fabricated MPA, (a) Patch view and, (b) Ground view

5. RESULTS AND DISCUSSION

This paper focused on improving the computational environment as well as the design aspects of irregular-shaped (polyline) MPA to achieve wider bandwidth. In the proposed model, the efficacy of GSA and QPSO (Delta model with Levy's flight-based particle positioning) has been exploited. Here, the GSA is applied to improve local search efficacy while QPSO helped to achieve global best (gbest) solution. The APBP with the ALA model is applied to enhance the performance with higher fitness value and convergence problem. Accordingly, the obtained results helped to improve the accuracy or the optimality of the design parameters. Further, both GSA and QPSO model have been applied in parallel to solve the aforesaid optimization problem, where GSA helped QPSO to obtain global value (gbest) and velocity optimally. In order to execute the program, 500 generations have been considered as the stopping criteria, while the search space is modeled with six dimensions. This polyline (irregular) shaped patch antenna (Fig.1) has 11 coordinates for identifying a population of 25 agents or particles. In other words, the proposed model is designed in such a manner that it estimates five coordinates with 10 parameters (x1 to x5 and Δy1 to Δy5) to result in an irregular-shaped patch antenna with higher bandwidth. Noticeably, the GSA-QPSO model is regarded as a constrained optimization model, where the limits of x and y are predefined to avoid inappropriate results probability. During simulation, the lower and higher threshold values are maintained for $x_{i=1,2,3,4,5}$ as 0.01 and 14.5, respectively. Similarly, the lower and upper thresholds are kept for $\Delta y_{i=1,2,3,4,5}$ as 0.01 and 5, correspondingly. The target is to achieve a polyline irregular-shaped MPA with a 50 Ω microstrip feed line, where the values of L_f and W_f are fixed which is designed over FR4 substrate with relative permittivity of 4.4. To enable optimal radiation performance, the ground metallization is escaped below the radiator; however, it required a L_g of 11.5 mm.

In order to simulate the performance, the sweep size or interval band of 0.01 GHz has been taken into account. Recalling the overall research intend where the key emphasis has been made on achieving higher BW under the assign frequency range. In this case, the lower and upper bound of the frequency is maintained at 2.4 GHz and 10.4 GHz, respectively. Thus, the optimal values of the above-stated irregular radiator coordinates are obtained by processing the GSA-QPSO model. Noticeably, to perform coordinate optimization, at first, the GSA-QPSO algorithm is applied which obtains the set of coordinate values (i.e., $x_{i=1,2,3,4,5}$ and $\Delta y_{i=1,2,3,4,5}$) for each iteration. Once obtaining the updated co-ordinate values, it is fed as input to the HFSS through a VB-script which acts as an interface between MATLAB and HFSS. Thus, for each set of radiator's coordinates, HFSS examined corresponding S_{11} values and this process continued till GSA-QPSO reached the stopping criteria. Considering the proposed model to be a population-based stochastic prediction (because both GSA and QPSO use initial populations to estimate sub-optimal solutions), it has been simulated for different test cases. This generated different coordinates, based on which the performance is obtained. To assess relative performance, the proposed model is simulated multiple times and noted the corresponding performance outcomes in terms of S_{11} parameter. The overall proposed model is simulated using MATLAB 2020b with HFSS 18. A few test results for coordinate values by the proposed GSA-QPSO model are given in Table 2.

Table 2. GSA-QPSO optimized MPA design parameters.

Design Constraints		L_f= 12.5 mm W_f=3 mm L_g=11.5 mm $\in_r$=4.4		
Variables	Threshold (mm)	Test Case-1	Test Case-2	Test Case-3
x_1	Upper Value= 14.5 Lower Value=0.01	7.089	7.319	7.617
x_2		7.479	7.100	6.748
x_3		8.528	9.680	8.678
x_4		7.391	6.538	5.719
x_5		8.239	6.689	3.594
Δy_1	Upper Value= 5 Lower Value=0.01	0.728	0.408	0.532
Δy_2		1.931	4.613	4.159
Δy_3		2.390	2.702	3.931
Δy_4		2.679	3.171	3.249
Δy_5		4.378	2.828	2.747

Fig. 3 presents the reflection coefficient performance of optimized MPA for three different cases. Because an ideal patch antenna requires maintaining S_{11}<-10dB, perceiving the outcome, it is observed that the anticipated MPA retains S_{11} lower than the aforesaid acceptance range. This pattern can be operated within the assessment frequency band from 2.4 GHz to 10.4 GHz. For instance, the optimized antenna provides the impedance BW in case (1) 8 GHz, ranging from 2.4 GHz to 10.4 GHz, case (2) 8.03 GHz, ranging from 2.38 GHz to 10.41 GHz, and case (3) 8.1 GHz, ranging from 2.34 GHz to 10.44 GHz. In other words, we can say that our proposed MPA can cover the WB or UWB frequencies from 2.4 to 10.4 GHz in these cases.

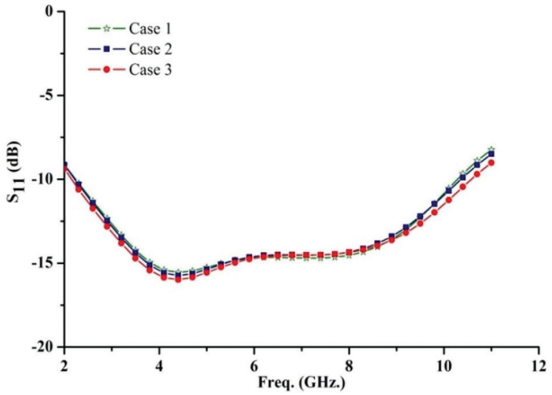


Fig. 3. Reflection coefficient of optimized MPA for different cases.

In the case (3), the optimized MPA provided a wider bandwidth, so this is considered for fabrication and measurement assessment. The final optimized (case 3) MPA is fabricated by a CNC machine, and Agilent's Vector Network Analyzer (VNA) N5247A is used to measure its performance. The fabricated model of MPA is depicted in Fig. 2.

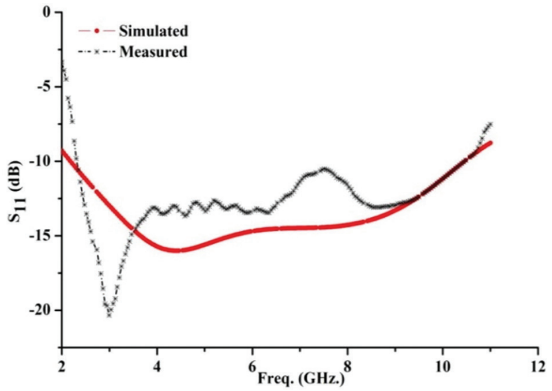


Fig. 4. Reflection coefficient of the final optimized MPA

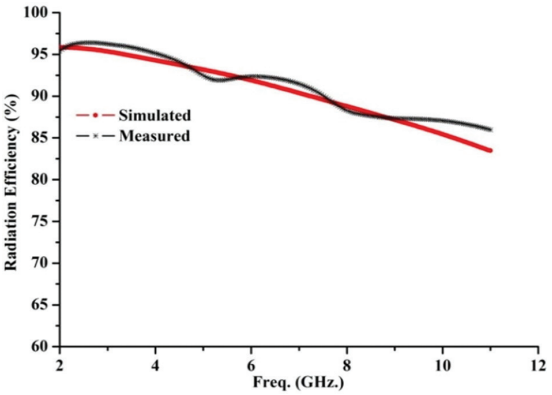


Fig. 5. Radiation efficiency of the final optimized MPA

Simulation and measurement of the final optimized MPA are correlated in Fig. 4 and Fig. 5. Across the entire simulation and measurement frequency range at 2.34 GHz - 10.44 GHz and 2.42 GHz -10.38 GHz, S_{11} is sustained with less than -10 dB. An antenna's radiation efficiency (%) can be compared using Fig. 5 and the corresponding simulations.

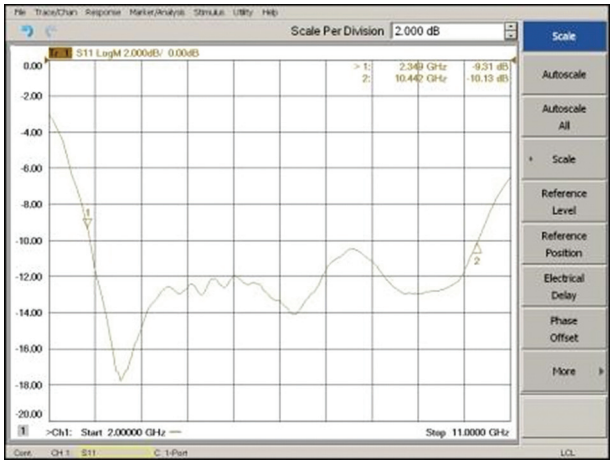


Fig. 6. Reflection coefficient on VNA of the final optimized MPA

The radiation efficiency of an antenna can be determined by comparing the radiated power with the power delivered at its terminals. In this case, the distance between transmitter and receiver is 1.5 meters. Horn antenna (AMkom) is used as a reference antenna, which has a broad frequency range from 1-18 GHz. The details of the experimental method for determining the radiation efficiency of the antenna are given in [52]. As shown in Fig. 5, the measured and simulated radiation efficiency is more than 84% for the whole operational band. When frequencies rise, dielectric loss increases and reduces efficiency. Modeling and simulating over a large BW have led to slightly different outcomes due to fabrication tolerances and software restrictions. The measured outcome of S_{11} on the VNA is presented in Fig. 6. The print of the VNA and anechoic chamber for the projected antenna can be seen in Fig. 7 (a) and (b).

Fig. 7(a). VNA setup

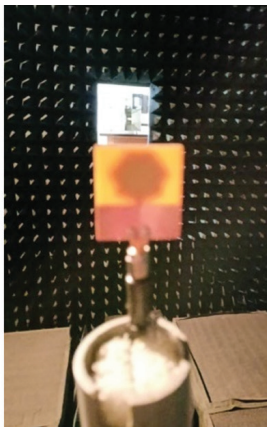
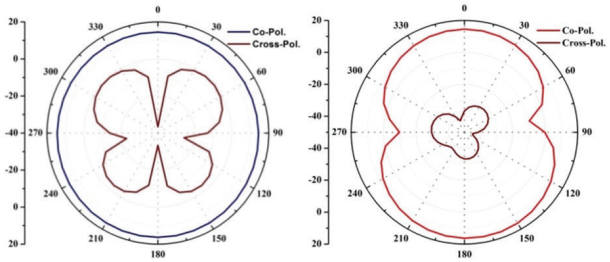
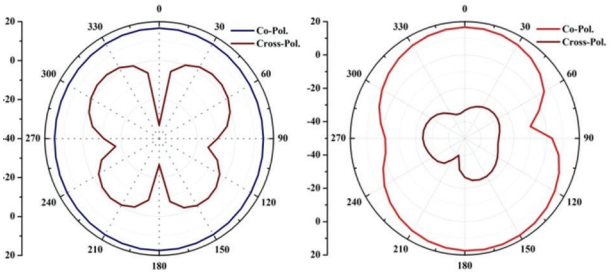


Fig. 7(b). Anechoic chamber

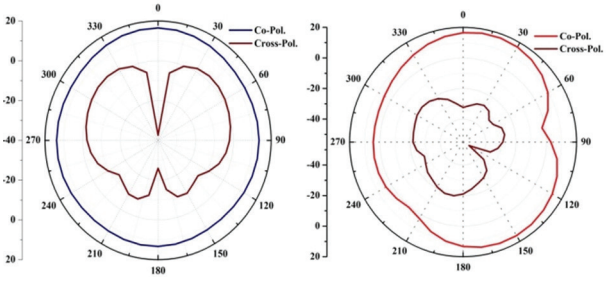
Fig. 8 (a), (b), and (c) displays the simulated two-dimensional radiation profiles of a final projected antenna. The co-polarization and cross polarization for the E and H plane at different operating frequencies of 4.4 GHz, 5.6 GHz, and 8.6 GHz, correspondingly is presented. Co-polarization is the preferred polarization of the wave to be transmitted by the radiator, while cross-polarization is the symmetrical radiation of the preferred polarization of the wave. Essentially, cross polarization is a loss of a signal at the recipient end. Likewise, it is a noise as far as detection is concerned. To reduce the obstruction of the waves, cross-polarization should be less than co-polarization. Generally, 15-20 dB down is adequate, except if the recipient has explicit prerequisites. The undesirable signal can be made adequate until it doesn't influence the detection.



(a) 4.4 GHz



(b) 5.6 GHz



(c) 8.6 GHz

Fig. 8. The radiation pattern of the final optimized MPA at varied resonant frequencies (Left and right side indicate E-plane H-plane, respectively).

In contrast to co-polarization intensities, the cross-polarization intensities are lesser than 14-17 dB and 30-34 dB in the E and H planes. The performance of the radiation pattern of the anticipated MPA is the same as those of a normally printed monopole, so the proposed MPA is appropriate for wide-band applications [2] [5].

6. CONCLUSION

In this paper, the emphasis is made on optimizing both the computational environment as well as the research goal, which is a broader band compatible patch antenna design. This research aimed to optimize the computational environment by implementing a hybrid heuristic concept, named the GSA-QPSO algorithm. Unlike classical heuristic algorithms, in the proposed GSA-QPSO model, QPSO with Levy's flight-based positioning on the one hand enabled optimal estimation, while GSA helped in achieving optimal local search activity. As a result, the GSA-QPSO model assisted in the estimation of appropriate MPA parameters. Noticeably, to further improve accuracy and computational efficiency, the proposed model employed QPSO with APBP and ALA functions. This approach enabled optimal estimation with the suitable set of design parameters while maintaining local minima and convergence avoidance and global optima accomplishment. This mechanism facilitated optimal coordinate parameter estimation for irregular-shaped patch antenna design. The proposed GSA-QPSO model optimized the MPA coordinate parameters in such a manner that it retained higher BW. The proposed model is simulated for the different test cases. This generated different coordinates, regarding which we obtained the performance. In terms of reflection coefficient, three different cases are observed to get optimum coordinate parameters of anticipated MPA. Finally, for case 3, the optimized MPA provided 126.6 % impedance BW with more than 84 % radiation efficiency over the entire operating frequency range from 2.34 to 10.44 GHz. The simulated results are also compared with measured results, which are closer to each other. The performance of the radiation patterns of the anticipated MPA is nearly omnidirectional. It revealed that the proposed MPA can be used for real-world wideband communication drives.

7. REFERENCES:

[1] First report and order, "Revision of Part 15 of the Commission's Rules regarding Ultra-Wideband Transmission Systems", Washington, DC, Federal Communications Commission FCC 02-48, 2002.

[2] S. Baudha, M. V. Yadav, "A Novel Design of a Planar Antenna with Modified Patch and Defective Ground Plane for Ultra-Wideband Applications", Microwave and Optical Technology Letters, Vol. 61, No. 5, 2019, pp. 1320-1327.

[3] M. Gupta, K. K. Mutai, V. Mathur, D. Bhatnagar, "A Novel Elliptical Ring Microstrip Patch Antenna for Ultra-Wideband Applications", Wireless Personal Communications, Vol. 114, 2020, pp.3017-3029.

[4] N. Sharma, S. S. Bhatia, "Performance Enhancement of Nested Hexagonal Ring-shaped Compact Multiband Integrated Wideband Fractal Antennas for Wireless Applications", International Journal of RF and Microwave Computer-Aided Engineering, Vol. 30, No. 3, 2020, pp.1-19.

[5] M. Kumar, J. A. Ansari, A. K. Saroj, R. Saxena, Devesh, "A Novel Microstrip Fed L-Shaped Arm Slot and Notch Loaded RMPA with Mended Ground Plane for Bandwidth Improvement", Progress in Electromagnetics Research C, Vol. 95, 2019, pp. 47-57.

[6] C. Mbinack, B. Bodo, J-S. A. E. Fouda, E. Tonye, "Inset-fed Rectangular Microstrip Patch Antenna Bandwidth Enhancement", Microwave and Optical Technology Letters, Vol. 61, No. 2, 2019, pp. 562-567.

[7] L. Tao, J. Xu, H. Li, Y. Hao, S. Huang, M. Lei, K. Bi, "Bandwidth Enhancement of Microstrip Patch Antenna Using Complementary Rhombus Resonator", Wireless Communications and Mobile Computing, Vol. 2018, 2018, pp. 1-8.

[8] M. T. Islam, M. Cho, M. Samsuzzaman, S. Kibria, "Compact Antenna for Small Satellite Applications", IEEE Antennas and Propagation Magazine, Vol. 57, No. 2, 2015, pp.30-36.

[9] M. G. Al-Halawani, M. Al-Najjar, M. K. Abdelazeez, "Microstrip Antenna Design for UWB Applications", Proceedings of the IEEE International Symposium on Antennas and Propagation, Fajardo, PR, USA, 26 June - 1 July 2016, pp. 43-44.

[10] R. N. Tiwari, P. Singh, B. K. Kanaujia, "A Modified Microstrip Line Fed Compact UWB Antenna for WiMAX/ISM/WLAN and Wireless Communications", International Journal of Electronics and Communications, Vol. 104, 2019, pp. 58-65.

[11] I. H. Hasan, M. N. Hamidon, A. Ismail, I. Ismail, A. S. Mekki, M. A. M. Kusaimi, S. Azhari, R. Osman, "YIG Thick Film as Substrate Overlay for Bandwidth Enhancement of Microstrip Patch Antenna", IEEE Access, Vol. 6, 2018, pp. 32601-32611.

[12] M. Alibakhshikenari, B. S. Virdee, C. H. See, R. A. Alhameed, A. Ali, F. Falcone and E. Limiti, "Wideband Printed Monopole Antenna for Application in Wireless Communication Systems", IET Microwaves, Antennas and Propagation, Vol. 12, No. 7, 2018, pp. 1222-1230.

[13] K. Suvarna, N. R. Murty, D. V. Vardhan, "A Miniature Rectangular Patch Antenna Using Defected Ground Structure for WLAN Applications", Progress in Electromagnetics Research C, Vol. 95, 2019, pp.131-140.

[14] P. Kala, R. Saxena, M. kumar, A. Kumar, R. Pant, "Design of Rectangular Patch Antenna using MLP Artificial Neural Network", Journal of Global Research in Computer Sciences, Vol. 3, No. 5, 2012, pp. 11-14.

[15] T. Khan, A. De, "A Generalized ANN Model for Analyzing and Synthesizing Rectangular, Circular, and Triangular Microstrip Antennas", Chinese Journal of Engineering, Vol. 2013, 2013, pp. 1-9.

[16] R. Saxena, M. Kumar, S. Aslam, "Evolutionary Computing based Neuron-Computational Model for Microstrip Patch Antenna Design Optimization", International Journal of Computer Networks and Communications, Vol. 13, No. 3, 2021, pp. 15-40.

[17] L. Y. Xiao, W. Shao, F. L. Jin, B. Z Wang, "Multiparameter Modeling with ANN for Antenna Design", IEEE Transactions on Antennas and Propagation, Vol. 66, No. 6, 2018, pp. 3718-3723.

[18] S. S. Sran, J. S. Sivia, "ANN and IFS based Wearable Hybrid Fractal Antenna with DGS for S, C and X band Application", AEU-International Journal of Electronics and Communications, Vol. 127, 2020, pp. 1-12.

[19] M. V.V.P. Kantipudi, S. Vemuri, S. S. Kashyap, R. Aluvalu, Y. S. Kumar, "Modeling of Microstrip Patch Antenna using Artificial Neural Network Algorithms", Proceedings of the 4th International Conference on Advanced Informatics for Computing Research, Gurugram, Haryana, India, 26-27 December 2020, pp. 27-36.

[20] V. S. Kushwah, A. P. S. Kushwah, "ANN Modeling of Microstrip Patch Antenna for WLAN Application using FR-4 Material", Materials Today: Proceedings, Vol. 47, No. 19, 2021, pp. 6647-6651.

[21] J. S. Sivia, A. P. S. Pharwaha, T. S. Kamal, "Neuro-computational Models for Parameter Estimation of Circular Microstrip Patch Antennas", Procedia Computer Science, Vol. 85, 2016, pp. 393-400.

[22] K. Guney, N. Sarikaya, "Concurrent Neuro-Fuzzy Systems for Resonant Frequency Computation of Rectangular, Circular, and Triangular Microstrip Antennas", Progress in Electromagnetics Research, Vol. 84, 2008, pp. 253-277.

[23] V. K. Singh, A. Lala, A. K. Singh, "Novel Inset Feed Circular Slotted Microstrip Antenna using Multilayer Feed-Forward Back-Propagation and Radial Basis Function Neural Network", National Academy Science Letters, Vol. 43, 2020, pp. 343-345.

[24] V. Thakare, P. K. Singhal, "Bandwidth Analysis by Introducing Slots in Microstrip Antenna Design using ANN", Progress In Electromagnetic Research M, Vol. 9, 2009, pp. 107-122.

[25] Z. Wang, S. Fang, "ANN Synthesis Model of Single-Feed Corner-Truncated Circularly Polarized Microstrip Antenna with an Air Gap for Wideband Applications", International Journal of Antennas and Propagation, Vol. 2014, 2014, pp. 1-7.

[26] D. Sarkar, T. Khan, F. A. Talukdar, "Forward and Reverse Neural Network Modelling of Beveled Stepped Rectangular UWB Antennas", Advances in Intelligent Systems and Computing, Vol. 1392, 2021, pp. 103-113.

[27] R. G. Mishra, R. Mishra, P. Kuchhal, N. P. Kumari, "Optimization and Analysis of High Gain Wideband Microstrip Patch Antenna using Genetic Algorithm", International Journal of Engineering and Technology, Vol. 7, No. 1.5, 2018, pp. 176-179.

[28] S. Sun, Y. Lu, J. Zhang, F. Ruan, "Genetic Algorithm Optimization of Broadband Microstrip Antenna", Frontiers of Electrical and Electronic Engineering, Vol. 5, 2010, pp. 185-187.

[29] C. R. M. Silva, H. W. C. Lins, S. R. Martins, E. L. F. Barreto, A. G. D'Assuncao, "A Multiobjective Optimization of a UWB Antenna using a Self Organizing Genetic Algorithm", Microwave and Optical Technology Letters, Vol. 54, No. 8, 2012, pp. 1824-1828.

[30] D. Sarkar, T. Khan, F. A. Talukdar, " Multi-adaptive Neuro-Fuzzy Inference System Modelling for Prediction of Band-notched Behaviour of Slotted-UWB Antennas Optimised using Evolutionary Algorithms", IET Microwaves, Antennas and Propagation, Vol. 14, No. 12, 2020, pp. 1396-1403.

[31] F. Mir, L. Kouhalvandi, L. Matekovits, E. O. Gunes, "Automated Optimization for Broadband Flat-

Gain Antenna Designs with Artificial Neural Network", IET Microwaves, Antennas and Propagation, Vol. 15, No. 12, 2021, pp. 1537-1544.

[32] R. K. Verma, D. K. Srivastava, "Optimization and Parametric Analysis of Slotted Microstrip Antenna using Particle Swarm Optimization and Curve Fitting", International Journal of Circuit Theory and Applications, Vol. 49, No. 7, 2021, pp. 1868-1883.

[33] W.-C. Weng, "Optimal Design of an Ultra-Wideband Antenna with the Irregular Shape on Radiator using Particle Swarm Optimization", The Applied Computational Electromagnetics Society Journal, Vol. 27, No. 5, 2012, pp. 427-434.

[34] P. Elechi, S. Orike, C. E. Ikpo, "Performance Analysis of Patch Antenna for Ultra-Wideband using Particle Swarm Optimization", Journal of Telecommunication, Electronic and Computer Engineering, Vol. 13, No. 3, 2021, pp. 53-59.

[35] S. Shabnam, S. Manna, U. Sharma, P. Mukherjee, "Optimization of Ultra Wide-Band Printed Monopole Square Antenna using Differential Evolution Algorithm", Proceedings of the 2nd International conference on Information Systems Design and Intelligent Applications, Kalyani, West Bengal, India, 8-9 January 2015, pp. 81-89.

[36] M. A. Trimukhe, B. G. Hogade, "Design of the Compact Ultra-Wideband (UWB) Antenna Bandwidth Optimization using Particle Swarm Optimization Algorithm", Iranian Journal of Electrical and Electronic Engineering, Vol. 15, No. 2, 2019, pp. 195-202.

[37] M. Zubair M. Moinuddin, "Joint Optimization of Microstrip Patch Antennas using Particle Swarm Optimization for UWB Systems", International Journal of Antennas and Propagation, Vol. 2013, 2013, pp. 1-8.

[38] M. Zubair, J. Ahmad, S. S. H. Rizvi, Miniaturization of Monopole Patch Antenna with Extended UWB Spectrum via Novel Hybrid Heuristic Approach", Wireless Personal Communications", Vol. 109, 2019, pp. 539–562.

[39] K. R. Mahmoud, "UWB Antenna Design using Gravitational Search Algorithm", JES-Journal of Engineering Sciences, Vol. 41, No. 5, 2013, pp. 1890-1903.

[40] A. Chatterjee, G. Mahanti, N. N. Pathak, "Comparative Performance of Gravitational Search Algorithm and Modified Particle Swarm Optimization Algorithm for Synthesis of Thinned Scanned Concentric Ring Array Antenna", Progress in Electromagnetics Research B, Vol. 25, 2010, pp. 331-348.

[41] A. Chatterjee, G. Mahanti, P. R. S. Mahapatra, "Design of Fully Digital Controlled Reconfigurable Dual-beam Concentric Ring Array Antenna using Gravitational Search Algorithm", Progress in Electromagnetics Research C, Vol. 18, 2011, pp. 59-72.

[42] O.T. Altinoz, A. E. Yilmaz, "Calculation of Optimized Parameters of Rectangular Patch Antenna using Gravitational Search Algorithm", Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15-18 June 2011, pp. 349-353.

[43] E. Rashedi, H. N. Pour, S. Saryazdi, "GSA: A Gravitational Search Algorithm", Information Sciences, Vol. 179, No. 13, 2009, pp. 2232-2248.

[44] Y. Zhang, S. Wang, G. Ji, "A Comprehensive Survey on Particle Swarm Optimization Algorithm and its Applications", Mathematical Problems in Engineering, Vol. 2015, 2015, pp. 1-38.

[45] Y. Zhang, S. Wang, Y. Sui, M. Yang, B. Liu, H. Cheng, J. Sun, W. Jia, P. Phillips, J. M. Gorriz, "Multivariate Approach for Alzheimer's Disease Detection using Stationary Wavelet Entropy and Predator-Prey Particle Swarm Optimization", Journal of Alzheimer's Disease, Vol. 65, No. 3, 2018, pp. 855-869.

[46] D. Tian, Z. Shi, "MPSO: Modified Particle Swarm Optimization and its Applications", Swarm and Evolutionary Computation, Vol. 41, 2018, pp. 49-68.

[47] J. Sun, B. Feng, W. B. Xu, "Particle Swam Optimization with Particles having Quantum Behavior", Proceedings of the IEEE Congress on Evolutionary Computation, Portland, OR, USA, 19-23 June 2004, pp. 326-331.

[48] J. Sun, W. Fang, V. Palade, X. Wu, W. Xu, "Quantum-behaved Particle Swarm Optimization with Gaussian Distributed Local Attractor Point", Applied Mathematics and Computation, Vol. 218, No. 7, 2011, pp. 3763-3775.

[49] P. Jia, S. Duan, J. Yan, "An Enhanced Quantum-behaved Particle Swarm Optimization Based on a Novel Computing way of Local Attractor", Journal of information science and technology, Vol. 6, No. 4, 2015, pp. 633-649.

[50] F. Han, Q. Liu, "A Diversity-guided Hybrid Particle Swarm Optimization Algorithm Based on Gradient Search", Neurocomputing, Vol. 137, 2014, pp. 234-240.

[51] A. A. Nagra, F. Han, Q. Ling, S. Mehta, "An Improved Hybrid Method Combining Gravitational Search Algorithm with Dynamic Multi Swarm Particle Swarm Optimization", IEEE Access, Vol. 7, 2019, pp. 50388-50399.

[52] A. Duzdar, A. Fuchs, C. Thiam, "Radiation Efficiency Measurements of a Microstrip Antenna Designed for the Reception of XM Satellite Radio Signals", Journal of Passenger Cars: Electronic and Electrical Systems, Vol. 115, No. 7, 2006, pp. 685-689.

A Hybrid Modified Ant Colony Optimization - Particle Swarm Optimization Algorithm for Optimal Node Positioning and Routing in Wireless Sensor Networks

Original Scientific Paper

Shaik Imam Saheb
JNTUA College of Engineering,
Research Scholar,Department of Computer Science and Engineering
Ananthapuramu, Andhra Pradesh, 515002, India
shaikimamsa@gmail.com

Khaleel Ur Rahman Khan
ACE Engineering College,
Professor & Dean Academics,Department of Computer Science and Engineering
Hyderabad, 501301, India, khaleelrkhan@gmail.com

C. Shoba Bindu
JNTUA College of Engineering,
Professor,Department of Computer Science and Engineering
Ananthapuramu, Andhra Pradesh, 515002, India, shobabindhu@gmail.com

Abstract *– Wireless Sensor Networks (WSNs) have been widely deployed in hostile locations for environmental monitoring. Sensor placement and energy management are the two main factors that should be focused due to certain limitations in WSNs. The nodes in a sensor network might not stay charged when energy draining takes place; therefore, increasing the operational lifespan of the network is the primary purpose of energy management. Recently, major research interest in WSN has been focused with the essential aspect of localization. Several types of research have also taken place on the challenges of node localization of wireless sensor networks with the inclusion of range-free and range-based localization algorithms. In this work, the optimal positions of Sensor Nodes (SNs) are determined by proposing a novel Hybrid M-ACO – PSO (HMAP) algorithm. In the HMAP method, the improved PSO utilizes learning strategies for estimating the relay nodes' optimal positions. The M-ACO assures the data conveyance. A route discovers when it relates to the ideal route irrespective of the possibility of a system that includes the nodes with various transmission ranges, and the network lifetime improves. The proposed strategy is executed based on the energy, throughput, delivery ratio, overhead, and delay of the information packets.*

Keywords*: wireless network, PSO, modified ACO, HMAP algorithm, node placement, relay node selection*

1. INTRODUCTION

WSNs are structured networks with many SNs, where SNs sense, compute, and transmit data within small distance ranges [1]. SNs have low powers and cost very less. The applications of WSNs includedifferent areas, like enemy monitoring and tracking, forest fire detection, battlefield surveillance, and disaster management. The data is collected and sent back towards Base stations (BSs)or sink nodes by SNs deployed over a region in the applications of WSNs. SNs have very low communication ranges and are driven by batteries resulting in issues of coverages, energy consumptions, network lifespans, and costs during deployments of WSNs in regular/irregular terrains. For improving the WSN's performance, an effective resolution is required. In this efficient method, the optimum location for SNs is to be positioned as the NP-hard issue [2]. The energy consumption, operative lifespan, and sensing coverage are affected by the sensor node's locations [3]. Therefore, there is an essential needfor the careful placement of SNs. The trade-off exists between SNs' energy consumption and network coverage, [4]. However, the network coverage will turn smaller, and the energy consumption isreduced using nearby SNs.

As a result of its low battery life, a main problem in the WSNs are the nodes' energy consumption. Since power consumptions and transmission distances are directly proportional, data transmissions are the primary reasons for energy depletion. A solution to this problem can be the additions of expensive high-powerrelay nodes or CHs (cluster heads) that extend networklifespans, enhance network proficiencies, and minimize data transmission distances based on connections and fault tolerances [5]. The network lifespan is prolonged, and therefore by minimizing the transmission distance to its equivalent relay nodes which act as CHs for clusters of SNs transmitting data. Relay nodes are similar to SNs in that they are high-powered nodes with tiny batteries and might point toward failures. Various factors like environmental problems, external destructions, and hardware failures make networks inefficient, and similarly, relay nodes may become idle or get damaged. Therefore, BSs cannot receive sensed data of SNs that are combined towards relay nodes [6]. Thus, in the circumstance of a relay node failure, placing an appropriate quantity of relay nodes linked to a sensor is necessarily such that it is still attached to an additional relay node. By considering the connectivity problem, the number of relay nodes is minimized since the relay nodes are costly. Since both are contrary to one another, the minimization of the relay node and the connectivity are opposing ideas. Therefore, employing a method that mutually considers problems is highly significant.

WSNs are cheaper in terms of costs and require little or no maintenance costs once installed hence the need for using WSNs for various applications [7]. Routing protocols of WSNs map paths between sources and destinations. They are routing algorithms that split the network into more manageable chunks and provide mechanisms for exchanging information amongst neighbors initially followed by coverage of entire networks [8-9]. For WSNs applications to be efficient and reliable there is a need to design a routing optimization to manage the communication of WSNs in energy-aware and also traffic and distance-aware mechanisms. The focus of this research work is to optimize routing in WSNs. Based on the energy of SNs, network traffics, and the distances between source and destination SNs, this study aims to select the best ideal paths for SNs to deliver information to BSs. Hence in this work, the optimal positions of SNs (Sensor Nodes) are determined by proposing a novel hybrid algorithm HMAP algorithm for estimating the relay nodes' optimal positions that includes the nodes with various transmission ranges, and the network lifetime improves.

The technical work is organized as given. Section 2 analyzes the different research approaches, which are presented to attain the Optimal Node Positioning and Routing in WSNs. Section 3 discusses the proposed research approach in detail with appropriate diagrams and examples. Section 4 discusses the suggested research approach's performance examination based on the numerical evaluation. Lastly in section 5, the research study's conclusion is studied depending on the achieved findings.

2. LITERATURE SURVEY

Network performances, the energy efficiency of Media Access Controls (MACs), topology controls, reduced energy routings, enhanced TCP, and domain-based schemes have been studied in WSNs [10]. These studies imply issues of battery powers, the density of SNs, and limitations in preferred statistical information in WSNs which are dissimilar when compared to other networks [11]. For the supply of energy, energy-limited small batteries are used by the Sensor nodes [12-13]. Hence, to extend the network operation lifetime, power consumption is the primary task. For reducing the power consumption and transmission range using the appropriate protocol design and the method of an advanced hardware application, different techniques have been proposed by increasing SNs density[14]. The development of algorithms is achieved to determine the disjoint paths of minimum energy in an all-wireless network. SEAD was proposed in [15] to reduce energy usage in both creations and dissemination of trees for delivering data to BSs. A few studies look at how the placement of sensors or coupled nodes affects the performance of WSNs.

The authors in [16] used Neuro-Fuzzy Rule-based clustering for WSNs with Internet of Things (IoTs). The study enhanced network lifespans by adopting cluster-based routing where MLTs (Machine learning techniques) and fuzzy rules updated weights to accomplish energy modeling. The study's comparative performance results with LEACH, FLCP and HEED procedures showed its superior performances.

Routings in the study [17] were based on adaptive ranks where CHs were selected by SN's ranks which were in turn based on energy residues and geographical positions.

The study in [18] suggested that ECRP-energy coverage ratio protocols were better alternatives to LEACH protocols for reducing network energy usage. The study found ideal cluster counts and CHs based on the least energy usage and maximum coverage area. Network longevities were enhanced by replacing CHs with low energy residues and high usages. Catalina A. S. and Mihaela C. extended their previous work of mobile BSs on Spatio-temporal event identifications and reports in [19].

In designing sensor networks, evolutionary algorithm grounded methods, namely Gas, EA, GP, and so on, are used effectively by several investigators [20].

Shaik Imam Saheb et al. [21] have proposed a novel and efficient self-deployment strategy, i.e. IPONP algorithm, to restrict the relay node placement issue. Two different parameters like relay nodes' deployed quantity and movement price minimization are concerned to provide the maximum coverage area.

Mao Li and Feng Jiang et al. [22] proposed two-dimensional topologies based on Optimal Transmission Distance Algorithms (OTDAs) to lower energy usage and extend the lives of networks.

Belal Al-Fuhaidi et al. [23] suggested heterogeneous sensor network deployments using Harmony Search Algorithms (HSAs) and probabilistic sensing models (PSMs) for improving maximum coverage and increasing probable coverage without interfering with each other.

Table 1. Comparison of the existing Approaches

Author	Approaches	Results	Disadvantages
Thangaramya et al (2019)	Algorithmic Energy aware clustering and neuro-fuzzy-based routings	Energy utilization, PDR, and network lifetime	All SNs were considered trustworthy, an impossible condition
Chithaluru et al (2019)	AREOR– Adaptive ranks based on energy-efficient opportunistic routing schemes	Better Message success rates, reduced energy consumption, minimized end-to-end delays, and better PDRs	Time and computational complexity
Mengjia Zeng et al (2019)	Heterogeneous Energy-based Clustering Protocol for WSNs	Enhanced Network lifetimes, reduced load balancing, and improved overall energy consumption	Applicable for heterogeneous networks only
Aranzazu-Suescun et al (2019)	Anchor-based routing protocols where dynamic clusters were used ring	Reduced energy consumptions	Computational complexity
Shaik Imam Saheb et al. (2019)	IPONP algorithm	The algorithm has achieved the best energy consumption, and delayed performance	Reduced lifetime
Mao Li and Feng Jiang et al. (2020)	optimal transmission distance (OTDA)	Minimizing energy usages and maximizing network life spans	Very meagre improvements in terms of first node death

3. PROPOSED SYSTEM

In this section, the proposed HMAP scheme is explained in detail. In the HMAP scheme, the optimal position of SNs is determined by a novel improved PSO algorithm that utilizes the learning strategies to increase the particle population diversity and enhance the capability to escape from local optima. The M-ACO assured the optimal data transmission, and it discovers the ideal route irrespective of the system that includes the nodes with various transmission ranges.

3.1 LOCATION ESTIMATION USING IMPROVED PSO

PSOs are an enhanced population-based stochastic optimization approach inspired by fish schooling and bird flocking. A swarm of S potential solutions is contained in the basic PSO, and they refer to particles through the problem space of D-dimension when searching for the optimum global position. Thus, the best fitness values of an objective function are generated.PSOs can be understood better by knowing elements that makeup PSOs

Particles may be defined as Pi.

Fitness Functions-These functions determine the best solutions and are generally objective functions.

Local Bests-they signifies the particle's best locations in the swarm between locations visited so far.

Global Bests-The best locations of particles based on the best fitness values amongst all particles.

Updates to Velocities—-Velocities are vectors that determine a particle's speeds and directions.

Positional Updates-Particles attempt to get into ideal positions for maximum fitness. Particles in PSOs update their locations regularly based on global optima values. The flowchart of the proposed system is shown in Figure 1.

Primarily, a position Xid is assigned by each particle i randomly, where i = 1, 2,…D and a velocity Vid(i = 1, 2, ….S). The best position of pbesti and the global best gbesti can be tracked by each particle. The particles' velocity and position are updated as follows:

$$V_{id}(t+1) = w * V_{id}(t) + L_1 * R_1 * \left(pbest_i(t) - X_{id}(t)\right) + L_2 * R_2 * \left(Gbest_i(t) - X_{id}(t)\right) \quad (1)$$

$$X_{id}(t+1) = X_{id}(t) + V_{id}(t+1) \quad (2)$$

Where w refers to the weight factor that controls the particle's velocity, L1 and L2 are the learning factors, and r1 and r2 refer to the random variables between 0 and 1.

Design of Fitness Function:

The particle position merits are evaluated, the Fitness function guides the particle selection direction, and the calculation is given below:

$$fitness\ (i) = 1/M \sum_{j=1}^{M} \sqrt{(xi - xj)^2 + (yi - yj)^2} \quad (3)$$

Where, (xj, yj) refers to the location coordinates of the anchor node j, the particle i's position coordinates (xi, yi), and its fitness value is fitness i.

In PSO, the aggregation of particles contributes towards the global best position and its finest position. The population causes convergence effects and impulsive convergence, and later search stagnates in the iterative phase. Resulting in low search accurateness of PSO algorithm, it is challenging to escape from local optimal.

A learning strategy using probability selection is proposed to increase the particle population diversity and enhance the capability to escape from local optimal. In the initial iterations, the algorithm's search efficiency and convergence speed are enhanced with the help of a learning strategy. The local optimum becomes favorable, and the quality of the candidates is improved in later iterations by using a learning strategy. Every measurement that is to be chosen by the learning strategy is diverse because of various particle position vectors' qualities. Therefore, self-determining learning approaches are employed by each dimension.

3.2 The pseudo-code learning process of a particle i

The particle i's position vector is xi, the search space dimension is D, the intermediate vector is xx, the minimum and maximum particle population's j-dimensional components are max(x:,j) and min(x: j), the fitness function is f(x), and the jth dimensional component of the particle population average is mean (x:j). Here, the fitness value reciprocal average is considered as weight.

For j = 1:D

xx = xi;

If rand<P1,k

Xxj = 2*gbestj – xxj; //learning strategy 1

Else If rand < P1j+P2j

Xxj = 2*pbesti,j – xxj; // learning strategy 2

Else If rand < P1,j + P2,j + P3,k

Xxj = max (x:,j) + min(x:,j) – xxj; // learning strategy 3

Else

Xxj = 2*mean(x:,j) – xxj; //learning strategy 4

End If

If f(xx) < f(xi)

Xi= xx;

f(xi) = f(xx);

End If

End For

Formulas of mean (x:,j), P1, j~P4,j in the Pseudo-code are as follows:

$$mean(x,j) = \frac{\sum_{i=1}^{N} \frac{1}{f(xi)} Xxi,j}{\sum_{i=1}^{N} \frac{1}{f(xi)}} \quad (4)$$

Where N represents the number of particles in the population, the particle i's j-dimensional component in the position vector is j, and f(xi) is the particle i's fitness value.

The local development low capacity can improve the accuracy and convergence of the optimal solution and other shortcomings in the later stage and disturbances to the current position as the PSO algorithm falls into the local optimum. The mutation of any dimension can be done as the current position's each dimension is not the best. For mutating the current best value, the step of Levy distribution is used based on the mutation formula as follows:

$$g_{best} = g_{best} + \alpha * s \quad (5)$$

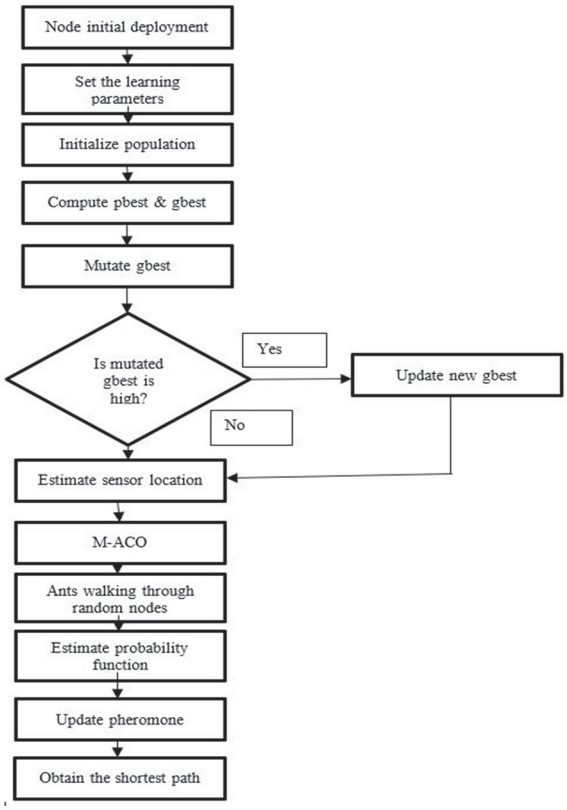


Fig. 1. Flowchart of the proposed system

Where, a represents the step size factor, whose value is 0.1 and s refers to the step subject to Levy's distribution.

Algorithm process: The process involves below

Step 1: Several SNs are placed at random in target regions, and distances between SNs are computed.

Step 2: Learning factors are set along with maximum iterations counts Tmax and maximum and lowest counts of particles.

Step 3: The population is initialized. The particle i's initial position xi = (xi1, xi2, …, xiD), best position is pbesti = xi, and speed is vi = (vi1, vi2, …viD). The swarm fitness value of a particle is evaluated based on equation (3). The best initial position of particle population is the global best position gbest.

Step 4: The particle speed v and position x are updated using the formula (1) & (2)

Step 5: The learning strategies implemented for particles based on pseudo-code

Step 6: The global best position and best position of particles are updated.

Step 7: The optimal position gbest mutates based on equation (5), and gbest is updated when the position of mutation is better

Step 8: The steps are repeated until all nodes' estimated location is determined.

3.3 ACO-BASED OPTIMAL ROUTE SELECTION

To determine the surmised answers for optimization challenges, the ACO could be used because it is a populace-based meta-heuristic algorithm. From the concept of ants' conduct, the ACO's vital thought has been considered. In a self-assertive way, each ant would be crossed the region while researching the synthetic substance in the first place. The chemical is called a pheromone, and it is the preface details of a neighborhood at each node. The pheromone quantity would be deposited based on the number of ants on the path and its length. The measure of higher pheromone will get by the shortest path. The ants with more increased pheromone obsession would be taken away and brace on the path they have considered. For handling various combinatorial issues, it is utilized as a masses-based scan practice. The network directions would imitate the way followed by ants. As a part of the position of ants, ant packets could be used. Each node's likelihood could be supplanted by the synthetic substance 'pheromone'. In ACO, pheromone trails would be helpful as dispersed and numerical data. The ants create the answers probabilistically for the comprehended challenge, and the ants have been adjusted to implement the algorithm for mirroring the pursuit encounter. At each ant packet's entry, the pheromone ought to be refreshed or modified as it is a volatile substance.

ACO-based algorithm: The assumption of random circulation of SNs in a rectangular locale is considered in the network model to detect the purpose of using the algorithm. For applying the principle behind the fundamental ACO, the outlines would be given by the fragment. Each node in the graphical issue would represent a point or vertex. The nodes or vertices that are joining the line are called edges.

Step 1: Random Deployment

The source over the framework would communicate the ant packets. In addition to the pheromone concentration, each path distribution is done randomly.

Step 2: Solution set generation

Based on the previous studies, the random deployment of the incalculable number of ants would be commenced, and the random path is strolled along these lines that constructed each set of solutions. Based on the provided constraints, each solution would be generated.

Step 3: Node selection

From the present node, the following node's selection would have relied on the probability function. Along the network edges, the associated pheromone would be considered while selecting the nodes. As a part of the Markov chain, each move of the ant could consider that the possibility of the move will be relying on the present value and not on the initial value.

Step 4: Probability estimation for the selection of a node

The equation below provides the possibility of selecting the next node 'j' from the present node 'i'.

$$\text{Pij} = \frac{[Tij]^{\alpha} * [nij]^{\beta}}{\sum k \in Ni\ [\{Tik\}^{\alpha} * \{nik\}^{\beta}]} \quad (6)$$

Here, α & β are control constraints Tij would represent the pheromone concentration along with the edges nij is going to signify the information of heuristic. It is equivalent to 1/dij, where dij is the distance between 2 nodes, Pij characterizes the node j probability to be selected from node i, and Ni would be represented the nodes' set.

Step 5: Pheromone update

Any node receives each ant packet and updates the pheromone while each node travels over all ants formerly using the below equation,

$$Tij(t+1) = (1-p)Tij(t) + \sum_{k=1}^{m} Tijk(i,j) \quad (7)$$

Where ρ is the pheromone rate of evaporation for avoiding the accumulation of pheromone, which would represent the pheromone quantity that would require be adding or subtracting to the path traveled by the ant k. As these results are made as particles, these k particles are selected randomly from the particle population and the k particles' appropriate positions are compared. Rather than choosing the global position for guiding the particle motion, it selects its individual best position. Based on the iteration, the k value increases gradually. The high-grade solution is made worse by a smaller k value that performs better towards the global search, and the population diversity rises. A larger k value attains high quality.

4. RESULTS AND DISCUSSION

Network simulator-2 was used to perform simulations based on the AODV routing protocol that focuses on the technique of node deployment. The network's energy consumption is discussed through the simulation results after carrying out the routing packets via AODV [24]. The network lifetime is increased by covering the maximum area using the proposed algorithm that saves the network energy. For nodes' deployment, the results are plotted using different parameters [25]. The proposed H-MAP system was compared to the Intersection Point Based Optimal Node Placement (IP-ONP) Algorithm, optimal transmission distance (OTDA) Algorithm, and harmony search algorithm using the network simulator NS2 (HSA). End-to-end latency, en-

ergy usage, and throughput characteristics were utilized to compare and appraise this example. The simulation parameters are shown in Table 2.

Table 2. Simulation parameters

Simulation Parameters	Values
Simulation Tool	NS-2.35
No. of nodes	50
Simulation time	100 sec
Simulation area	1500*1500 m
Pause time	2-20 S
Mobility model	Random waypoint model
Routing protocol	HMAP
Packet rate	1000 bytes/0.1ms
Transmission Protocols	UDP, NULL
Channel type	Wireless
Mac layer	802.11, SMAC
Traffic type	constant Bit Rate (CBR)
Antenna type	Antenna/Omni antenna
Initial energy	100 j

End-to-end delays: Average time taken by packets to get transferred from network's sources to destinations and based on Equation (8).

$$End - to - \text{end}delay = \frac{\sum_{i=1}^{n}(t_{ri} - t_{si})}{n}$$

Where t_{ri} – ith packet delivery time, t_{si} – a time when an ith packet was sent, n – number of packets.

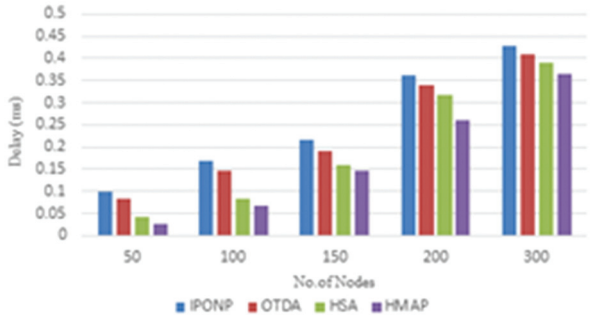


Fig. 2. End-to-end delay vs number of nodes

The simulation results in Figure 2 show the evaluation of the proposed HMAP method's end-to-end delay time. Table 2 shows how effective node placement using a learning technique and optimum relay selection using ACO, which guarantees that only high-quality nodes are chosen as relays, helped to reduce end-to-end latency. The proposed HMAP approach achieved the shortest average latency in the network 0.25ms for 50 numbers of nodes, whereas the previous methods like IPONP, OTDA, and HSA are 0.98ms, 0.84ms, and 0.41 which are higher delays than the proposed method.

Table 2. Comaprision of end to end delay

Nodes	IPONP (ms)	OTDA (ms)	HSA (ms)	HMAP(ms)
50	0.098	0.084	0.041	0.025
100	0.169	0.146	0.084	0.068
150	0.217	0.191	0.160	0.148
200	0.362	0.338	0.316	0.259
300	0.429	0.409	0.390	0.365

Energy consumptions: These refer to average energies required for transmitting packets to nodes within particular time frames in Equation (9)

$$Energy\ (e) = [(2 * pi - 1)(e_t + e_r)d \qquad (9)$$

Where pi – data packet, e_t - packet i's source energy, e_r – energy needed for the receipt of packet i, d – distance among source and destination nodes.

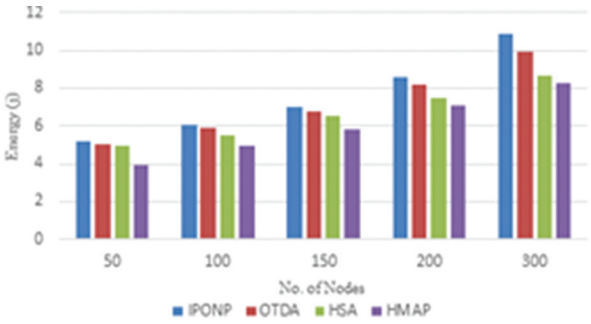


Fig. 3. Energy consumption against counts of SNs

The simulation results in Figure 3 and Table 3 show that the proposed approach saves a significant amount of energy compared to the prior existing methods. The proposed HMAP method attains minimum energy consumption obtained in the network was 3.920 for 50 numbers of nodes, whereas the previous methods like IPONP,OTDA, and HSA are 5.23, 5.01, and 4.954 which are higher values of energy consumption when compared to the proposed method.

Table 3. Comparison of energy consumption

Nodes	IPONP (j)	OTDA (j)	HAS (j)	HMAP (j)
50	5.23	5.01	4.954	3.920
100	6.071	5.93	5.480	4.933
150	7.027	6.80	6.540	5.811
200	8.619	8.20	7.510	7.067
300	10.8544	9.90	8.70	8.296

Network lifetime: network lifespan can be expressed in eqn (10)

$$Lifetime\mathbb{E}[L] = \frac{\varepsilon_0 - \mathbb{E}[E_w]}{P + \lambda\mathbb{E}[E_r]} \qquad (10)$$

here P - constant network power consumption and continuous, ε_0 - total non-rechargeable initial energy, λ- average sensor reporting rate, $\mathbb{E}[E_w]$ – expected energy wastage or non-utilized energy till the death of

the network, and $\mathbb{E}[E_r]$ –reported energy consumption of nodes.

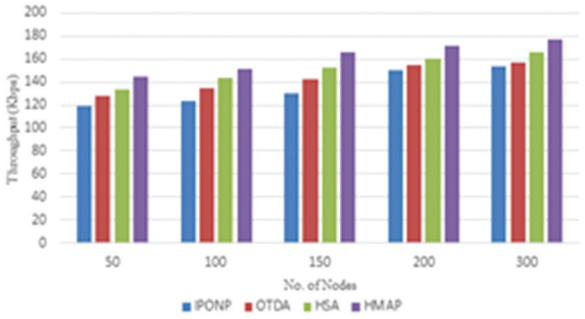


Fig. 4. Network performance vs number of nodes

Throughput metric refer to the amount of data transmitted between SNs. High throughputs ensure higher amounts of data deliveries. Table 4 and figure 4 show that, when compared to existing approaches, the suggested method has a high throughput rate. The high throughput rate was due to the efficient node placement strategy and optimal relay node selection, which always selects the interference-free paths. In the experiment, the suggested approach kept the average throughput rate at up to 160kbps, whereas current methods kept it at less than that and maintained low throughputs than the results of the proposed method.

Table 4. Comparison of network performance

Nodes	IPONP (Kbps)	OTDA (Kbps)	HSA (Kbps)	HMAP (Kbps)
50	119.34	127.93	133.42	144.13
100	122.88	133.98	143.17	151.42
150	129.88	142.88	152.46	165.75
200	149.59	154.68	160.42	171.01
300	153.58	157.37	165.64	177.39

PDR: indicates the proportion of total lost packets to overall sent packets can be expressed in eqn (11)

$$\text{Packet loss ratio} = \frac{N^{tx} - N^{rx}}{N^{tx}} \times 100\% \qquad (11)$$

Where N^{tx} - transmitted packets, N^{rx} - received packets. This evaluation was carried out through the extraction of all real-time packet sizes, which are sent and obtained.

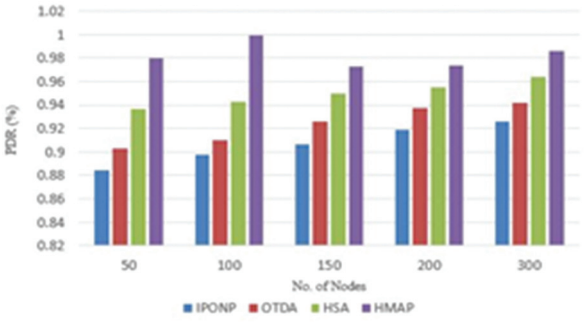


Fig. 5. Packet delivery ratio vs number of nodes

PDRs: They refer to the ratio of successful delivery of packets to intended destinations.

PDRs describe the network's data transmission quality. The successful delivery of packets was aided by the selection of reliable relay nodes and the optimal placement of relay nodes using learning algorithms. The proposed approach obtained a maximum PDR of 0.97 percent, whereas previous methods averaged 0.93 percent. The packet delivery ratio is shown in Figure 5, and a comparison of existing approaches is shown in Table 5.

Table 5. Comparision of Packet delivery ratio

Nodes	IPONP (%)	OTDA(%)	HSA (%)	HMAP (%)
50	0.8847	0.9026	0.9369	0.9795
100	0.8973	0.9103	0.9430	0.9991
150	0.9065	0.9256	0.9497	0.9726
200	0.9186	0.9370	0.9548	0.9739
300	0.9255	0.9417	0.9635	0.9861

5. CONCLUSIONS

In this work, the optimal position of SNs is determined by proposing a novel hybrid algorithm HMAP algorithm. In the HMAP method, the improved PSO utilizes learning strategies for estimating the relay nodes' optimal position. The M-ACO assures the data conveyance, and a route is discovered that is very close to the ideal route irrespective of the system with nodes of various transmission ranges. The probability-based selection scheme used in ACO improves the self-node selection strategy. The experimental results prove the effectiveness of the HMAP scheme over the IPONP, OTDA, and HSA schemes. The proposed HMAP method attains minimum energy consumption experienced in the network was 3.920 for 50 numbers of nodes, whereas the previous methods like IPONP, OTDA, and HSA are 5.23, 5.01, and 4.954 which are the higher value of energy consumption when compared to the proposed method. The suggested approach appears to have a good chance of being implemented in a static WSNs system. Future research should look at the possibilities of implementing the proposed method in a dynamic WSNs system, so that its full potential may be used to solve real-world challenges like sensor lifetime and geographical conditions.

6. REFERENCES:

[1] M. K. Singh, S. I. Amin, S. A. Imam, V. K. Sachan, A. Choudhary, "A Survey of Wireless Sensor Network and Its Types", Proceeding of the 2018 International Conference on Advances in Computing, Communication Control and Networking, Greater Noida, India, 2018, pp. 326-330.

[2] I. Tomić, J. A. McCann, "A Survey of Potential Security Issues in Existing Wireless Sensor Network

Protocols", IEEE Internet of Things Journal, Vol. 4, No. 6, 2017, pp. 1910-1923.

[3] J. N. Al-Karaki, A. Gawanmeh, "The Optimal Deployment, Coverage, and Connectivity Problems in Wireless Sensor Networks: Revisited", IEEE Access, Vol. 5, 2017, pp. 18051-18065.

[4] M. Ram, S. Kumar, V. Kumar, A. Sikandar, R. Kharel, "Enabling Green Wireless Sensor Networks: Energy Efficient T-Mac Using Markov Chain Based Optimization", Electronics, Vol. 8, No. 5, 2019, p. 534.

[5] A. K. M. Al-Qurabat, C. Abou Jaoude, A. K. Idrees, "Two Tier Data Reduction Technique for Reducing Data Transmission in Iot Sensors", Proceeding of the 15th International Wireless Communications & Mobile Computing Conference, Tangier, Morocco, 2019, pp. 168-173.

[6] H. Yetgin, K. T. K. Cheung, M. El-Hajjar, L. H. Hanzo, "A Survey of Network Lifetime Maximization Techniques in Wireless Sensor Networks", IEEE Communications Surveys & Tutorials, Vol. 19, No. 2, 2017, pp. 828-854.

[7] W. Heinzelman, J. Kulik, H. Balakrishnan, "Adaptive protocols for information dissemination in wireless sensor networks", Proceedings of 5th Annual Joint ACM/IEEE International Conference on Mobile Computing and Networking; Seattle, WA, USA, 1999; pp. 174-185.

[8] D. Braginsky, D. Estrin, "Rumor routing algorithm for sensor networks", Proceedings of 1st ACM International Workshop on Wireless Sensor Networks and Applications, Atlanta, GA, USA. October 2002; pp. 22-31.

[9] C. Schurgers, M. B. Srivastava, "Energy efficient routing in wireless sensor networks", Proceedings of Military Communications Conference on Communications for Network-Centric Operations: Creating the Information Force; USA. 2001.

[10] M. Kodialam, T. Nandagopal, "Characterizing Achievable Rates in Multi-Hop Wireless Networks: The Joint Routing and Scheduling Problem", Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, San Diego, California, 2003, pp. 42-54.

[11] K. Sundaresan, V. Anantharaman, Hung-Yun Hsieh and A. R. Sivakumar, "ATP: A Reliable Transport Protocol for Ad Hoc Networks", IEEE Transactions on Mobile Computing, Vol. 4, No. 6, 2005, pp. 588-603.

[12] A. Ahilan, G. Manogaran, C. Raja, S. Kadry, S. N. Kumar, C. A. Kumar, N. S. Murugan, "Segmentation by Fractional Order Darwinian Particle Swarm Optimization Based Multilevel Thresholding and Improved Lossless Prediction Based Compression Algorithm for Medical Images", IEEE, Vol. 7, 2019, pp. 89570-89580.

[13] P. Du, A. Nakao, "Application Specific Mobile Edge Computing Through Network Softwarization", Proceeding of the 2016 5th IEEE International Conference on Cloud Networking, Pisa, Italy, 2016, pp. 130-135.

[14] V. K. Sachan, S. A. Imam, M. T. Beg, "Energy-Efficient Communication Methods in Wireless Sensor Networks: A Critical Review", International Journal of Computer Applications, Vol. 39, No. 17, 2012, pp. 35-48.

[15] H. S. Kim, T. F. Abdelzaher, W. H. Kwon, "Minimum-Energy Asynchronous Dissemination to Mobile Sinks in Wireless Sensor Networks", Proceedings of the 1st International Conference on Embedded Networked Sensor Systems, San Diego, CA, USA, 2003, pp. 193-204.

[16] K. Thangaramya, K. Kulothungan, R. Logambigai, M. Selvi, S. Ganapathy, A. Kannan, "Energy aware cluster and neuro fuzzy based routing algorithm for wireless sensor networks in IoT", Computer Networks, Vol. 151, 2019, pp.211-223.

[17] P. Chithaluru, R. Tiwari, K. Kumar, "AREOR–Adaptive ranking based energy efficient opportunistic routing scheme in Wireless Sensor Network", Computer Networks, Vol. 162, 2019, p. 106863.

[18] M. Zeng, X. Huang, B. Zheng, X. Fan, "A heterogeneous energy wireless sensor network clustering protocol", Wireless Communications and Mobile Computing, 2019, p. 7367281.

[19] C. Aranzazu-Suescun, M. Cardei, "Anchor-based routing protocol with dynamic clustering for Internet of Things WSNs", EURASIP Journal on Wireless Communications and Networking, Vol. 1, 2019, p. 130.

[20] B. A. Özdemir, A. Attea, Ö. A. Khalil, "Multi-Objective Evolutionary Algorithm Based on Decomposition for Energy Efficient Coverage in Wireless Sensor Networks", Wireless Personal Communications, Vol. 71, No. 1, 2013, pp. 195-215.

[21] S. I. Saheb, K. U. R. Khan, C. ShobaBindu, "An Intersection Point-Based Optimal Node Placement Algorithm for Wireless Sensor Networks", International Journal of Innovative Technology and Exploring Engineering, Vol. 8, No. 8, 2019, pp. 3371-3380.

[22] M. Li, F. Jiang, "Relay Node Placement Based on Optimal Transmission Distance in Two-Tiered Sensor Network", IEEE Access, Vol. 8, 2020, pp. 110438-110445.

[23] B. Al-Fuhaidi, A. M. Mohsen, A. Ghazi, W. M. Yousef, "An Efficient Deployment Model for Maximizing Coverage of Heterogeneous Wireless Sensor Network Based on Harmony Search Algorithm", Journal of Sensors, Vol. 2020, 2020, pp. 1-18.

[24] R. Dhaya, R. Kanthavel, A. Ahilan, "Developing an Energy-Efficient Ubiquitous Agriculture Mobile Sensor Network-Based Threshold Built-In Mac Routing Protocol", Soft Computing, Vol. 25, No. 18, 2021, pp. 12333-12342.

[25] A. J. G. Malar, C. A. Kumar, A. G. Saravanan, "Iot based Sustainable Wind Green Energy for Smart Cites Using Fuzzy Logic Based Fractional Order Darwinian Particle Swarm Optimization", Measurement, Vol. 166, 2020, p. 108208.

Secure and Energy Aware Cluster based Routing using Trust Centric – Multiobjective Black Widow Optimization for large scale WSN

Original Scientific Paper

Sampath Reddy Chada
Sree Chaitanya Institute of Technological Sciences,
Department of Computer Science and Engineering,
Karimnagar, Telangana, India
Sampath553@gmail.com

Narsimha Gugulothu
JNTUH College of Engineering,
Department of Computer Science and Engineering, Sultanpur, Telangana, India
Narsimha06@jntuh.ac.in

Abstract *– Wireless Sensor Network (WSN) is a promising approach that is developed for a wide range of applications due to its low installation cost. However, the nodes in the WSN are susceptible to different security threats, because these nodes are located in hostile or harsh environments. Moreover, an inappropriate selection of routing path affects the data delivery of the WSN. The important goal of this paper is to obtain secure data transmission while minimizing energy consumption. In this paper, Trust Centric - Multiobjective Black Widow Optimization (TC-MBWO) is proposed for selection of Secure Cluster Head (SCH) from the large-scale WSN. Moreover, the secure routing path is generated by using the TC-MBWO, in which the factors considered for the cost function are: residual energy, distance, trust and node degree. Therefore, the secured clustering and routing achieved by using TC-MBWO, provides the resistance against malicious nodes and simultaneously the energy consumption is also minimized by identifying the shortest path. The proposed TC-MBWO method is analyzed in terms of alive nodes, dead nodes, energy consumption, throughput, and network lifetime. Here, the TC-MBWO method is compared with different existing methods such as Low Energy Adaptive Clustering Hierarchy (LEACH), Particle Swarm Optimization - Grey Wolf Optimizer (PSO-GWO), Particle-Water Wave Optimization (P-WWO) and Particle-based Spider Monkey Optimization (P-SMO). The alive nodes of the TC-MBWO are 70 for 2800 rounds which are higher in number when compared to the PSO-GWO, P-WWO and P-SMO.*

Keywords*: Cluster head, Energy Consumption, Secure Clustering and Routing process, Trust Centric- Multiobjective Black Widow Optimization, Wireless Sensor Networks.*

1. INTRODUCTION

WSN contains a huge amount of sensors for observing environmental situations such as sound, humidity, temperature, etc. [1]. WSNs are used in various applications comprising security systems, disaster management, agricultural areas, medical domains, weather forecasting and military applications wherein, the WSN gathers data to perform an appropriate analysis [2] [3]. The sensors in the network have a power supply, communication unit, and microcontroller. The sensor unit analyzes the environment, gathers the data, processes it and then transfers the processed information to other sensors over the communication medium. But, the sensor faces certain issues related to memory, computation and energy [4]. Security is considered as an important issue when broadcasting sensitive information in WSN. Broadcasting the information through the multi-hop route with a higher distance, leads to intrusion of different malicious attacks [5] [6] [7]. Moreover, energy preservation is also a main issue in the WSN. The major prevalent approach i.e., clustering of sensor nodes is accomplished for solving the issue of energy consumption [8] [9] [10].

The clustered routing protocol efficiently deals with the requirements of large-scale applications for hierarchical WSN. But the selection of SCH and secure routing is difficult during the clustering and routing phase respectively [11] [12]. Clustering is generally an energy efficient approach wherein the sensors are divided into numerous clusters. Accordingly, the Cluster Members

(CM) observe the surroundings and broadcast the information to the Cluster Head (CH). Next, the CH eliminates the unwanted data from the aggregated data. Since, the CH is closer to the BS, it rapidly exhausts its energy over the network [13]. An optimal path is identified by the routing algorithm and is used to broadcast the observed data over the discovered path which helps to increase lifetime and minimize energy consumption [14]. Moreover, the issue of energy consumption also persists when the sensors are involved in malicious behaviors. Hence, the node's energy is preserved by avoiding the malicious nodes [15]. Therefore the main issues of WSN are energy efficiency and security. Because, the existence of malicious nodes in the network causes packet drop and unwanted energy consumption. These issues of WSN are the main motivations of this research, therefore the TC-MBWO based secure clustering and routing are developed to ensure the reliable communication.

The major contributions of the research paper are given below:

- An SCH and routing path selection is achieved by using the TC-MBWO with distinct cost parameters. Here, the MBWO is taken for selecting the SCH and routes, due to its efficient global search process.
- Therefore, a secure and energy aware routing is developed for achieving reliable communication. This kind of communication minimizes energy consumption while improving the throughput.

This research paper is arranged as follows: Section 2 provides the related work about the secure data transmission performed in the WSN. A detailed explanation of the TC-MBWO is given in Section 3. Section 4 delivers the outcomes of the TC-MBWO method. The conclusion is made in Section 5.

2. RELATED WORK

Hu et al. [16] provided security against the attacks by developing a Trust-aware Secure Routing Protocol (TSRP). Here, the node's trust value was calculated using the residual energy, volatilization factor, direct trust value and indirect trust value. Next, the hop count and link quality were used to identify the optimal path. However, the developed TSRP failed to perform analysis in large scale WSNs.

Shi et al. [17] implemented the information-aware secure routing for a network wherein cost functions such as trust metric and each node's status, were considered during the secure route identification. The distance and residual energy were included in the node's status. The node's energy consumption was minimized by detecting the path with a lesser distance. Sometimes, the packet loss was huge because of the energy exhaustion in the sensor.

Sefati et al. [18] presented the optimized black hole algorithm to detect appropriate CHs and Ant Colony Optimization (ACO) for route detection. The parameters used to optimize the selection of CH were distance, node's free buffer and residual energy. This work considered both the single and multi-hop data transmission to transmit the data, but it had not considered the trust values to improve the security.

Prithi and Sumathi, [19] developed a hybrid PSO-GWO for effective usage of energy and secure broadcast of data. The environment's dynamic role was learned by developing the Learning Dynamic Deterministic Finite Automata (LD2FA) which was used for providing the learned data to PSO-GWO. This work failed to properly utilize the advantages of the fitness function used in PSO-GWO, which is an important requirement alongside optimization, in an effective research.

Kumar and Vimala [20] developed energy and trust based routing by using Exponentially-Ant Lion Whale Optimization (E-ALWO). This E-ALWO was the combination of the exponentially weighted moving average with ant lion and whale optimizations. The designed E-ALWO provided less delay while transmitting the data packets. The E-ALWO selected the CH only based on the energy and delay.

Khot and Naik, [21] presented the P-WWO for routing the data in the optimal secure path. The P-WWO was the integration of Particle Swam Optimization (PSO) and water wave optimization. Here, the PSO selected the CHs according to their fitness which included maintainability factor, consistency factor, trust, energy and delay. Moreover, the routing path with less delay and distance was chosen as an optimal path. However, the distance measure was not considered in the selection of CH which caused higher energy consumption.

Khot and Naik [22] developed the P-SMO which is the combination of PSO and spider monkey optimization. The developed P-SMO was used to perform the secure data transmission through the CH whereas the secure routing was accomplished by considering trust, consistency factor, energy and delay. However, the packets received by the BS were not analyzed in this P-SMO.

The drawbacks found from the related works are mentioned as follows: high amount of packet loss due to node failure, higher energy consumption and inappropriate cost function selection. To overcome the aforementioned issues, the secure and energy aware routing is developed by using the TC-MBWO. In this TC-MBWO, the malicious nodes are avoided while broadcasting the data packets, which results in lesser energy consumption and reduced packet drop.

3. TC-MBWO METHOD

In this TC-MBWO, a secure and energy aware cluster based routing is developed to improve the network lifetime and packet delivery. The important processes ac-

complished in the TC-MBWO are sensor deployment, SCH selection, clustering and routing path generation. Here, the malicious nodes that exist in the network are avoided during SCH selection and routing, by considering the trust value of the nodes. Accordingly, the energy consumption of the nodes are minimized in the network. The block diagram for the TC-MBWO is shown in Figure 1.

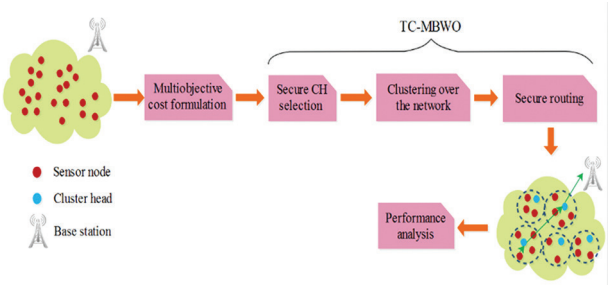


Fig. 1. Block diagram of the TC-MBWO method

3.1 INITIALIZATION OF SENSORS

At first, the sensor nodes are randomly deployed in the large-scale network area. Here, two sinks are considered to create the multi sink large scale WSN environment. The SCH and route generation using TC-MBWO are explained in the following section.

3.2 SCH SELECTION USING TC-MBWO

In this phase, secure cluster heads are selected to enhance the security and to lessen the energy consumption of the network. This SCH selection is used to avoid the malicious nodes during the communication. In general, the Black Widow Optimization (BWO) is operated on the idea of reproduction style and cannibalism of black widows [21].

3.2.1. Representation and Initialization

The potential solution of TC-MBWO is denoted as spider population, in which it specifies the candidate sensors that can be selected as SCHs. In this phase, the candidate solutions are referred to as spiders that specify the nodes which can be chosen as SCHs. The widow's dimension is equal to the amount of SCHs. Here, a random node ID from 1 to N is used to initialize the position of each widow, wherein the total nodes in the WSN is represented as N. The i^{th} widow initialized in the TC-MBWO is expressed in Equation (1).

$$x_i = (x_{i,1}, x_{i,2}, .., x_{i,NCH}) \quad (1)$$

Wherein, the $x_{i,d}$ defines the widow's position and the candidate nodes between the total nodes is represented as $1 \le d \le NCH$.

3.2.2. Iterative process of SCH selection using TC-MBWO

The iterative process of the TC-MBWO involves the movement and pheromone update that are detailed as follows:

3.2.2.1. Movement

Equation (2) shows the spider's motion which is exhibited in liner and spiral manner.

$$\vec{x}_i(t+1) = \begin{cases} \vec{x}_*(t) - m\vec{x}_{r1}(t) & if\ rand() \le 0.3, \\ \vec{x}_*(t) - \cos(2\pi\beta)\,\vec{x}_i(t) & in\ other\ case \end{cases} \quad (2)$$

Wherein, the $\vec{x}_i(t+1)$ defines the spider's new position which denotes the motion of the spider; $\vec{x}_*(t)$ specifies the best spider identified from the whole population; m is the floating value created between [0.4,0.9]; $r1$ defines the random number generated between 1 and the total search agent size; $\vec{x}_{r1}$ is the chosen $r1$ search agent where $i \neq r1$; β denotes the random float number created between [- 1.0,1.0] and $\vec{x}_i(t)$ denotes the current search agent.

3.2.2.2. Pheromone update

In this searching process, the emitted pheromones from the spider are important for the courtship-mating process. Here, the male spider provides a high response to the sex pheromones received from the healthy females which have a high fertile possibility. Moreover, this kind of activity is utilized to avoid the dangerous mating attempt with hungry cannibal females. In this TC-MBWO, the male black widow chooses the high fertile females rather than the female spider with cannibalism. Therefore, a male black widow chooses only the female spider with high pheromone. Equation (3) shows the computation of spider's pheromone rate.

$$Pheromone = \frac{Cost_{max} - Cost\ (i)}{Cost_{max} - Cost_{min}} \quad (3)$$

Where, the finest and worst costs in the recent population are denoted as $Cost_{max}$ and Cos_{tmin} respectively; the i^{th} spider's current cost is specified as $Cost\ (i)$. The female black widow is specified as cannibal when it has less pheromone and the corresponding female is interchanged with another spider as shown in the equation (4).

$$\vec{x}_i(t) = \vec{x}_*(t) + \frac{1}{2}[\vec{x}_{r1}(t) - (-1)^{\sigma} \times \vec{x}_{r2}(t)] \quad (4)$$

Wherein, $\vec{x}_i(t)$ refers to females with less pheromone; $r1$ and $r2$ are the random values generated from 1 and the total black widow population ($r1 \neq r2$) and σ is the random binary number. The cost function that is used to measure the pheromone rate is formulated in the following section.

3.3 MULTIOBJECTIVE COST FORMULATION FOR SCH SELECTION

The cost functions considered in the TC-MBWO for selecting optimal SCHs are trust (f_1), residual energy (f_2), intracluster distance (f_3), distance from the SCH to BS (f_4) and node degree (f_5). These cost functions are converted into a single objective as shown in equation (5).

$$Cost = \gamma_1 \times f_1 + \gamma_2 \times f_2 + \gamma_3 \times f_3 + \gamma_4 \times f_4 + \gamma_5 \times f_5 \quad (5)$$

Where, γ_1, γ_2, γ_3, γ_4 and γ_5 denotes the weighted parameters allocated to each cost parameter.

- The primary cost value considered in this TC-MBWO is the trust value of each node, in which two distinct trust values are considered, named as direct and indirect trust values. The direct trust (*DT*) value is the ratio between the received packets and broadcasted packets from the source node which is expressed in equation (6). On the other hand, indirect trust (*IDT*) is calculated according to the direct trust measured from the target node which is expressed in equation (7). Accordingly, the calculation of final trust value is shown in equation (8).

$$DT = \frac{R_{a,b}(t)}{S_{a,b}(t)} \tag{6}$$

$$IDT = \frac{1}{NN}\sum_{u=1}^{U} DT_{u,s} \tag{7}$$

$$f_1 = \sum_{i=1}^{P} (DT + IDT)/i \tag{8}$$

Wherein the received and sent packets between the nodes a and b at time t are represented as $R_{a,b}(t)$ and $S_{a,b}(t)$; NN denotes the number of nodes adjacent to the node s and P specifies the total amount of participating nodes

- During the communication, the energy utilization of SCH becomes high as it performs various tasks such as packet receiving, aggregation and broadcasting over the network. Hence, the sensor with higher residual energy is preferred as SCH and the residual energy is expressed in equation (9).

$$f_2 = \sum_{i=1}^{NCH} \frac{1}{E_{CH_i}} \tag{9}$$

Wherein, the E_{CH_i} is the residual energy of the i^{th} SCH.

- Two different distances known as; i) intracluster distance and ii) distance from the SCH to BS, are considered in the cost, because the energy consumption of the node mainly depends on the transmission distance over the network. Hence, the node with less transmission distance is preferred to minimize the energy consumption. Equation (10) and (11) expresses the intra-cluster distance and distance from the SCH to BS.

$$f_3 = \sum_{j=1}^{M} \left(\sum_{i=1}^{I_j} dis(N_i, CH_j)/I_j \right) \tag{10}$$

$$f_4 = \sum_{i=1}^{M} dis(CH_i, BS) \tag{11}$$

Where, distance from i^{th} node to j^{th} SCH and distance from i^{th} SCH to BS are represented as $dis(N_i, CH_j)$ and $dis(CH_i, BS)$ respectively; The amount of normal sensors in the cluster *j* is specified as I_j.

- The amount of normal nodes belonging to the next hop node is node degree which is expressed in equation (12). The node consumes less energy, when it has less node degree in the network.

$$f_5 = \sum_{i=1}^{M} I_j \tag{12}$$

The selection of optimal SCH is done by using the derived cost function. The malicious nodes are avoided using trust value while choosing the SCHs, because the malicious nodes cause packet losses and unwanted energy consumption. The energy used in the cost is used to avoid the node failure which results in high packet delivery, as well as minimal distance, which is used to minimize the energy consumption. Further, the node degree is used to minimize the energy distribution. Therefore, the proposed TC-MBWO selects the optimal SCH to achieve reliable transmission.

3.4 CLUSTER FORMATION

In this phase, the CMs are assigned to chosen SCHs, in which the clusters are created based on the distance and residual energy. The potential function to form the clusters in the network is expressed in equation (13).

$$Potential\ of\ sensor\ (N_i) = \frac{E_{CH}}{dis(N_i, CH)} \tag{13}$$

The derived function is used to allocate the CM to the SCH with less distance and high residual energy.

3.5. ROUTING PATH GENERATION USING TC-MBWO

The TC-MBWO method was also used to discover the secure routing path. In this multi-sink scenario, the sink which is near the source node is taken as the final destination. The steps processed in this routing stage are mentioned as follows:

1. The possible routes between the source SCH and BS are initialized in the spiders whereas the dimension of each spider is equal to the amount of relay nodes.
2. Subsequently, location and pheromone updates are accomplished based on the cost computed for the route.
3. The cost usage while generating the transmission path, involves considering residual energy, the distance from SCH to BS and node degree. Equation (14) shows the cost used in the TC-MBWO based route generation.

$$Cost = \varphi_1 \times \sum_{i=1}^{P} \frac{(DT+IDT)}{i} + \varphi_2 \times \sum_{i=1}^{NCH} \frac{1}{E_{CH_i}} + \varphi_3 \times \sum_{i=1}^{M} dis(CH_i, BS) + \varphi_4 \times \sum_{i=1}^{M} I_j \tag{14}$$

Where, φ_1, φ_2, φ_3 and φ_4 are weighted parameters assigned to each objective of route generation.

The overall flowchart of the TC-MBWO method is shown in the Figure 2. As illustrated in the Figure 2, initially the nodes are deployed randomly in the network area. Subsequently, the cost formulation of TC-MBWO for CH selection takes place as shown in the section 3.3. The formulated cost value is used to choose an appropriate SCH, followed by the clusters, formed as shown in section 3.4. Fi-

nally, the secure path over the network is identified using the TC-MBWO. For a better analysis of the TC-MBWO, the simulation is executed until the dead nodes of the WSN is equal to the total number of initialized nodes. Therefore, the proposed TC-MBWO helps to identify the secure path with higher residual energy, lesser transmission distance and lesser node degree. Hence, the energy consumption of the nodes are minimized by using the TC-MBWO-based routing which helps to improve the network lifetime.

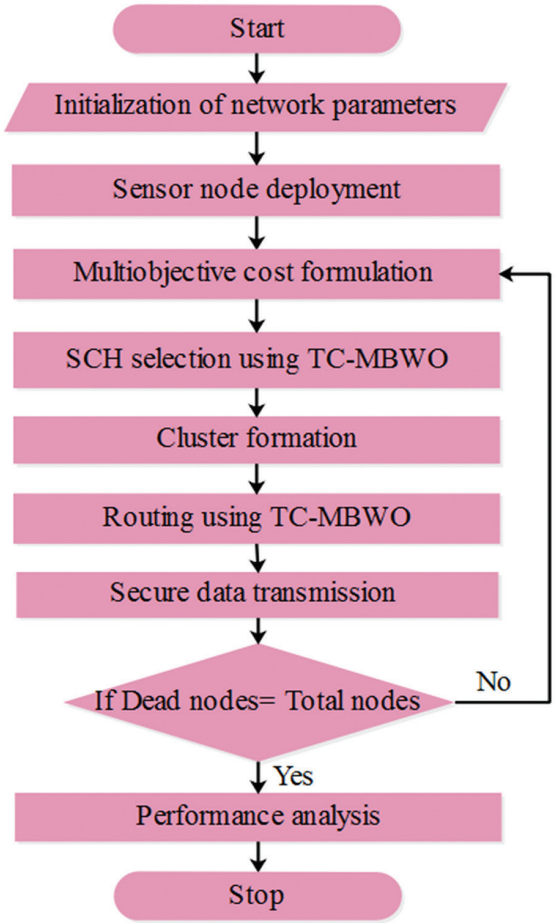


Fig. 2. Flowchart of the TC-MBWO method

4. RESULTS AND DISCUSSION

The design and implementation of reliable transmission using TC-MBWO are done using MATLAB R2018a. The system used in the analysis is operated with i5 processor having 6GB of RAM. The main objective of the TC-MBWO method is to achieve an improved security and energy efficiency for large scale WSN. The simulation parameters of the TC-MBWO are mentioned in Table 1.

Table 1. Simulation parameters

Parameters	Value
Area	500m×500m
Nodes	100
Location of sink	(500, 500) & (250, 250)
Packet size	4000 bits
Initial energy	0.5 J

4.1 PERFORMANCE ANALYSIS

The performance of the TC-MBWO is analyzed by means of alive nodes, dead nodes, energy consumption, throughput, and network lifetime. Here, the TC-MBWO's performances are evaluated with one classical approach i.e., LEACH in which the implementation is done with the same specifications as in Table 1.

4.1.1. Alive nodes and dead nodes

Alive nodes are defined as the nodes with enough residual energy to transmit the data packets to the sink. On the contrary, the dead nodes are inversely proportional to the alive nodes of the network. Specifically, the node is declared as dead when it exhausts its energy while transmitting the packets. Figures 3 and 4 respectively show the alive node and dead node comparison, for the TC-MBWO and LEACH. From the analysis, it is concluded that the TC-MBWO achieves higher alive nodes and lesser dead nodes than the LEACH. In general, the malicious nodes that exist in the network cause higher energy consumption. But, the TC-MBWO avoids the malicious nodes during the SCH selection and routing, therefore the energy consumption of the nodes is minimized. This helps in increasing the alive nodes of the TC-MBWO.

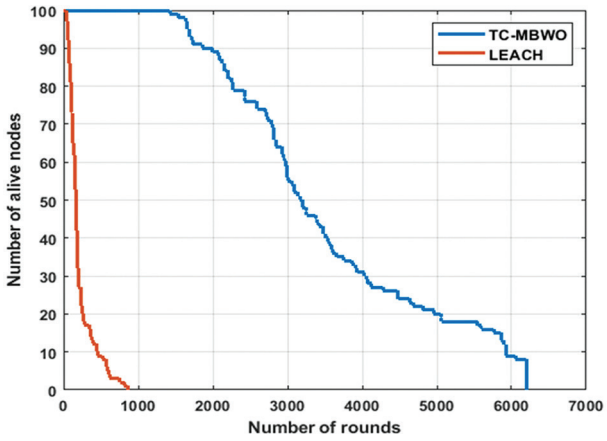


Fig. 3. Alive nodes Vs. rounds

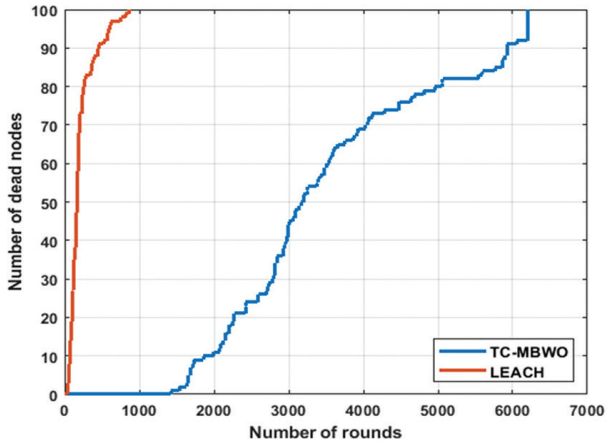


Fig. 4. Dead nodes Vs. rounds

4.1.2 Energy consumption

Energy consumption of the network is defined as the amount of energy consumed while receiving and broadcasting the data packets. The energy consumption comparison for the TC-MBWO and LEACH is shown in Figure 5. From Figure 5, it is concluded that the energy consumption of the TC-MBWO is lesser when compared to LEACH, which achieves higher energy consumption because it fails to mitigate the malicious nodes and also performs single hop transmission. Moreover, TC-MBWO achieves higher energy efficiency because of the mitigation of malicious nodes using trust and the generation of the shortest route.

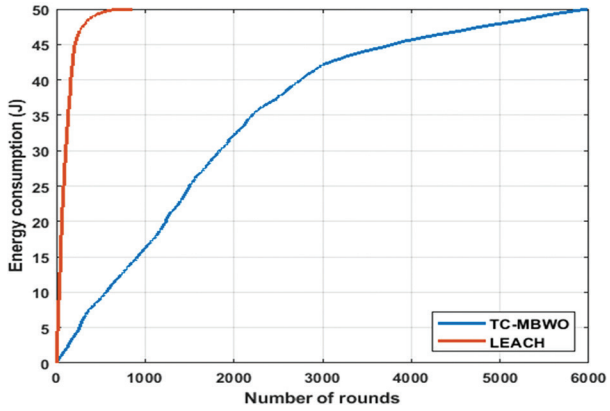


Fig. 5. Energy consumption Vs. rounds

4.1.3. Throughput

Throughput is defined as the amount of packets successfully received at the sink, and is analyzed in bits per second. Figure 6 shows the throughput comparison for the TC-MBWO and LEACH. The throughput of TC-MBWO is greatly increased than the LEACH, because of the secure data transmission. Specifically, the data delivery of the TC-MBWO is improved by avoiding node/link failure and malicious nodes while broadcasting the data packets.

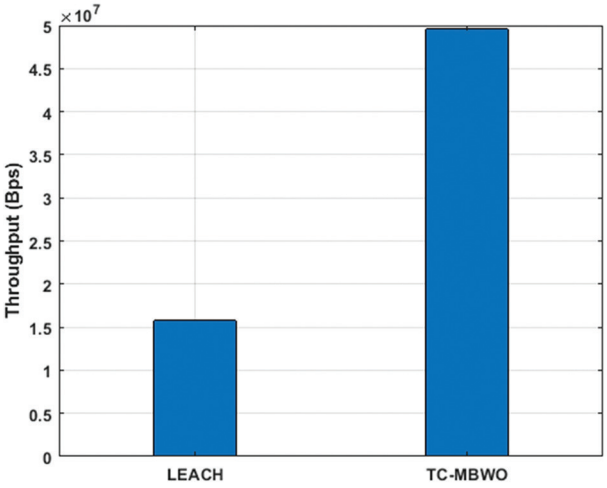


Fig. 6. Throughput Vs. rounds

4.1.3. Network lifetime

Network lifetime is defined as the round where all the nodes exhaust their energy over the large-scale WSN. Here, the lifetime is analyzed by using three metrics: First Node Die (FND), Half Node Die (HND) and Last Node Die (LND). The lifetime comparison for the TC-MBWO and LEACH is shown in Figure 7. From Figure 7, it is inferred that the lifetime of the TC-MBWO is high when compared to the LEACH. For example, the LND of the TC-MBWO is 6204 whereas the LND of the LEACH is 864. The LEACH results in lesser lifetime due to the presence of malicious attacks and single hop transmission. But, the proposed TC-MBWO achieves a higher lifetime due to its appropriate cost function.

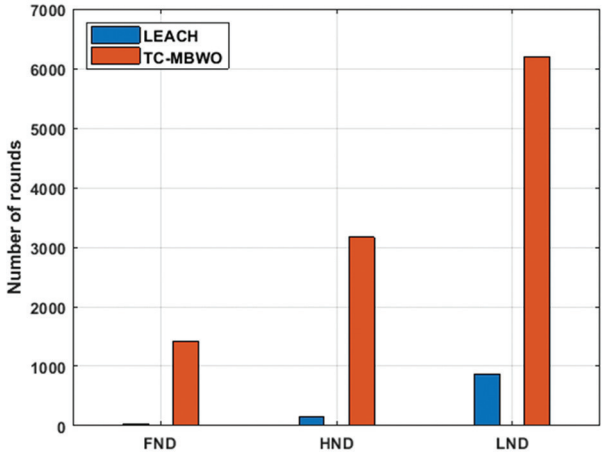


Fig. 7. Network lifetime Vs. rounds

4.2. COMPARATIVE ANALYSIS

The existing research PSO-GWO [19], P-WWO [21] and P-SMO [22] are used to evaluate the efficiency of the TC-MBWO. Table 2 provides the comparative analysis of the PSO-GWO [19], P-WWO [21], P-SMO [22] and TC-MBWO. Additionally, the graphical comparison of alive nodes is shown in Figure 8.

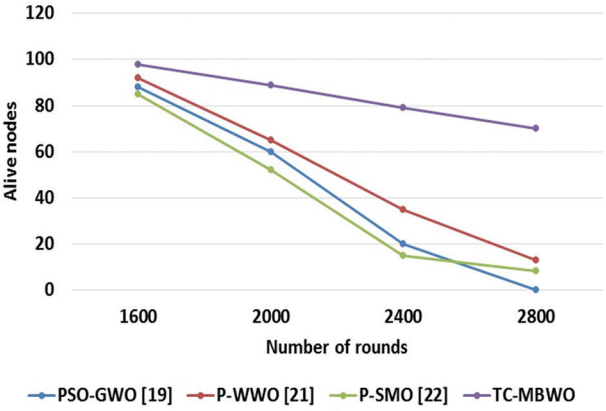


Fig. 8. Graphical comparison of alive nodes

From Table 2 and Figure 8, it is concluded that the TC-MBWO achieves better performance than the PSO-GWO [19], P-WWO [21], and P-SMO [22], because of its

optimal cost function selection. The derived multiobjective cost function is used to achieve a secure and energy-aware data transmission over the large scale WSN. The energy consumption of the nodes are minimized by avoiding malicious nodes and discovering shortest path using the TC-MBWO. The lesser energy consumption of the nodes increases the network lifetime, therefore the alive nodes of the TC-MBWO are higher in number in the WSN. Moreover, the throughput of the TC-MBWO is improved by avoiding the malicious nodes during the selection of SCH and route.

Table 2. Comparative analysis of TC-MBWO

Performance measures	Methods	Number of rounds			
		1600	2000	2400	2800
Alive nodes	PSO-GWO [19]	88	60	20	0
	P-WWO [21]	92	65	35	13
	P-SMO [22]	85	52	15	8
	TC-MBWO	98	89	79	70
Energy consumption	PSO-GWO [19]	920	1150	1400	1625
	TC-MBWO	26.5968	32.2538	36.7779	40.3968
Throughput	PSO-GWO [19]	15.8×104	16×104	16×104	16×104
	TC-MBWO	12.8×106	16×106	19.2×106	22.4×106

5. CONCLUSION

A secured clustering and a routing path are developed using the TC-MBWO algorithm to obtain secure data transmission between the nodes. The node's energy depletion is minimized by using the TC-MBWO-based optimal SCH selection. Next, the secure route between the desired nodes is generated by using the TC-MBWO. The SCH selection and secure routing path generation done by TC-MBWO are improved by using distinct cost parameters such as trust, residual energy, distance, and node degree. The trust considered in the TC-MBWO helps to mitigate malicious attacks during data transmission. From the obtained results, it is concluded that the TC-MBWO method outperforms the PSO-GWO, P-WWO and P-SMO in the comparative analysis. The number of alive nodes in TC-MBWO is 70 for 2800 rounds which is a much higher result, when compared to PSO-GWO, P-WWO and P-SMO. In the future, a novel optimization algorithm can be used for improving the performance of the WSN.

6. REFERENCES:

[1] V. Kavidha, S. Ananthakumaran, "Novel energy-efficient secure routing protocol for wireless sensor networks with Mobile sink", Peer-to-Peer Networking and Applications, Vol. 12, No. 4, 2019, pp. 881-892.

[2] M. Selvi, S. V. N. Santhosh Kumar, S. Ganapathy, A. Ayyanar, H. Khanna Nehemiah, A. Kannan, "An energy efficient clustered gravitational and fuzzy based routing algorithm in WSNs", Wireless Personal Communications, Vol. 116, No. 1, 2021, pp. 61-90.

[3] C. Deepa, B. Latha, "HHSRP: a cluster based hybrid hierarchical secure routing protocol for wireless sensor networks", Cluster Computing, Vol. 22, No. 5, 2019, pp. 10449-10465.

[4] P. Rodrigues, J. John, "Joint trust: An approach for trust-aware routing in WSN", Wireless Networks, Vol. 26, No. 5, 2020, pp. 3553-3568.

[5] G. Dhand, S. S. Tyagi, "SMEER: secure multi-tier energy efficient routing protocol for hierarchical wireless sensor networks", Wireless Personal Communications, Vol. 105, No. 1, 2019, 17-35.

[6] M. Revanesh, V. Sridhar, J. M. Acken, "Secure coronas based zone clustering and routing model for distributed wireless sensor networks", Wireless Personal Communications, Vol. 112, No. 3, 2020, pp. 1829-1857.

[7] H. Zhou, Y. Wu, L. Feng, D. Liu, "A security mechanism for cluster-based WSN against selective forwarding", Sensors, Vol. 16, No. 9, 2016, p. 1537.

[8] W. Ke, O. Yangrui, J. Hong, Z. Heli, L. Xi, "Energy aware hierarchical cluster-based routing protocol for WSNs", The Journal of China Universities of Posts and Telecommunications, Vol. 23, No. 4, 2016, pp. 46-52.

[9] M. Pavani, P. T. Rao, "Adaptive PSO with optimised firefly algorithms for secure cluster-based routing in wireless sensor networks", IET Wireless Sensor Systems, Vol. 9, No. 5, 2019, pp. 274-283.

[10] V. Vijayalakshmi, A. Senthilkumar, "USCDRP: unequal secure cluster-based distributed routing protocol for wireless sensor networks", The Journal of Supercomputing, Vol. 76, No. 2, 2020, pp. 989-1004.

[11] W. Fang, W. Zhang, W. Chen, J. Liu, Y. Ni, Y. Yang, "MSCR: multidimensional secure clustered routing scheme in hierarchical wireless sensor networks", EURASIP Journal on Wireless Communications and Networking, Vol. 2021, No. 1, 2021, pp. 1-20.

[12] T. Kalidoss, L. Rajasekaran, K. Kanagasabai, G. Sannasi, A. Kannan, "QoS aware trust based routing algorithm for wireless sensor networks", Wireless Personal Communications, Vol. 110, No. 4, 2020, pp. 1637-1658.

[13] M. Maheswari, R. A. Karthika, "A novel QoS based secure unequal clustering protocol with intrusion detection system in wireless sensor networks", Wireless Personal Communications, Vol. 118, No. 2, 2021, pp. 1535-1557.

[14] Y. U. Xiu-Wu, Y. U. Hao, L. Yong, X. Ren-rong, "A clustering routing algorithm based on wolf pack algorithm for heterogeneous wireless sensor networks", Computer Networks, Vol. 167, 2020, pp. 106994.

[15] M. Selvi, K. Thangaramya, S. Ganapathy, K. Kulothungan, H. Khannah Nehemiah, A. Kannan, "An energy aware trust based secure routing algorithm for effective communication in wireless sensor networks", Wireless Personal Communications, Vol. 105, No. 4, 2019, pp. 1475-1490.

[16] H. Hu, Y. Han, H. Wang, M. Yao, C. Wang, "Trust-aware secure routing protocol for wireless sensor networks", ETRI Journal, Vol. 43, No. 4, 2021, pp. 674-683

[17] Q. Shi, L. Qin, Y. Ding, V. Xie, J. Zheng, L. Song, "Information-aware secure routing in wireless sensor networks", Sensors, Vol. 20, No. 1, 2020, p. 165.

[18] S. Sefati, M. Abdi, A. Ghaffari, "Cluster-based data transmission scheme in wireless sensor networks using black hole and ant colony algorithms", International Journal of Communication Systems, Vol. 34, No. 9, 2021, p. e4768.

[19] S. Prithi, S. Sumathi, "Automata based hybrid PSO–GWO algorithm for secured energy efficient optimal routing in wireless sensor network", Wireless personal communications, Vol. 117, No. 2, 2021, pp. 545-559.

[20] K. SureshKumar, P. Vimala, "Energy efficient routing protocol using exponentially-ant lion whale optimization algorithm in wireless sensor networks", Computer Networks, Vol. 197, 2021, p. 108250.

[21] P. S. Khot, U. Naik, "Particle-Water Wave Optimization for Secure Routing in Wireless Sensor Network Using Cluster Head Selection", Wireless Personal Communications, Vol. 119, No. 3, 2021, pp. 2405-2429.

[22] P. S. Khot, U. L. Naik, "Cellular automata-based optimised routing for secure data transmission in wireless sensor networks", Journal of Experimental & Theoretical Artificial Intelligence, 2021, pp. 1-19.

Digital Signature Method to Overcome Sniffing Attacks on LoRaWAN Network

Original Scientific Paper

Rahayu Indah Lestari
Telkom University, School of Computing
Jl. Telekomunikasi no. 1, Bandung, Indonesia
rahayuindahlestarii@student.telkomuniversity.ac.id

Vera Suryani*
Telkom University, School of Computing
Jl. Telekomunikasi no. 1, Bandung, Indonesia
verasuryani@telkomuniversity.ac.id

Aulia Arif Wardhana
Telkom University, School of Computing
Jl. Telekomunikasi no. 1, Bandung, Indonesia
auliawardan@telkomuniversity.ac.id

Abstract *– LoRa or Long Range with LoRaWAN technology is a protocol for low-power wireless networks. The absence of an encryption process on the data payload becomes a challenge for the LoRaWAN network. When the process of sending messages is running inter devices, sniffing might occur, thereby reducing the confidentiality aspect of the data communication process. This paper optimized the digital signature method to secure messages sent by LoRaWAN network devices, along with Advanced Encryption Standard (AES) algorithm and Ed25519 algorithm. AES was used for message encryption, while Ed25519 was used for signature purposes. The aim of applying digital signatures in this paper was to verify that the payload data sent was original and not changed during the transmission process and to ensure data confidentiality. The addition of security mechanisms to the LoRaWAN network, such as the process of encryption, decryption, and verification results, has caused some overheads. The overhead caused by the usage of a digital signature is also analyzed to ensure that the digital signature is feasible to be implemented in LoRa devices. Based on the experimental results, it was found that there was an increase in the size of memory usage and some additional processing delay during the deployment of digital signatures for LoRa devices. The overall overhead caused by implementing digital signatures on the LoRa devices was relatively low, making it possible to implement it on the LoRa network widely.*

Keywords*: LoRaWAN, sniffing, digital signature, AES, overhead*

1. INTRODUCTION

Internet of Things (IoT) refers to a communication paradigm to build the interactions between machines without any human interference [1]. IoT networks can be classified based on their physical radio layer, achievable bit rate, and power consumption or communication range. A network that operates remotely, use low power, and is able to tolerate low bit rates tend to use network technology such as LoRa [2].

LoRaWAN is a network infrastructure based on the Long Range (LoRa) radio modulation technology with some security flaws [3]. Payload data is not protected by encryption, making it subject to the sniffer.

Sniffing is the process of snooping the data packet on a network system. Some of which can monitor and capture all passing network traffic, regardless of who will receive the packet. During the sniffing process, it is potential to emerge an attack on LoRaWAN devices when receiving data from other devices.

The lack of security protection on LoRaWAN networks, which makes sniffing activities susceptible to attacks, is the main research problem of this paper. The digital signature is utilized to anticipate any attacks that sniffing operations may induce. This method aims to enhance the authentication aspect during data transmission of LoRaWAN communication. Furthermore, the purpose

of this digital signature is to ascertain that the content being transmitted does not change until they reach the recipient; thus, the receiver may be confident that the message received is genuinely original from the sender [4]. The digital signature is not a new method, but it is an alternative solution to encountering sniffing attacks on LoRa networks. Digital signatures are well suited to identifying valid users involved in the LoRa network's data communication process. Digital signatures are frequently used in software distribution, financial transactions, and other situations where modification or forgery must be detected.

The encryption algorithm of the digital signature implemented in this paper Advanced Encryption Standard (AES), and Ed25519 algorithm. The AES algorithm was used considering that it is lightweight and efficient in both software and hardware, and it can be applied to the digital signature method [5]. For encryption and decryption purposes, the AES variant applied were AES 128 and AES 256. The Ed25519 algorithm was selected for the signature process because it applies the Curve25519 algorithm in which the algorithm is compatible and found more efficient to be applied to the digital signature method. The overhead computation is the performance parameter of the proposed method in this research. The computed overhead consists of the change of the data payload size, the memory usage of the device executing the application, the RAM utilization, and the response processing time between the sender and receiver.

This study was conducted to investigate the usage of digital signature to prevent data sniffing in LoRa devices during data communication. The addition of a security mechanism to the network will certainly produce computational overhead on the system. This computational overhead was computed to determine how many additional resources are required when a digital signature is utilized. The parameters of the system's overhead analysis are payload length, memory usage, RAM usage, and response processing time [6]. This evaluation aims to determine the feasibility of applying digital signatures to LoRa devices.

The remaining sections of the paper are structured as follows. Section 2 provides a brief summary of relevant works or the current state of the art about other techniques for addressing security vulnerabilities in the Lora Network. In Section 3, the architecture of the proposed approaches is explained. In Section 4, the authors assess the proposed solution and show the experiment results. Section 5 concludes with a summary of the investigation's findings.

2. RELATED WORKS

Many studies have been conducted using digital signature and other methods in preventing the attack on the LoRaWAN network devices caused by the sniffing process. Table 1 depicts the comparison of related studies regarding this problem. Paper [7] compares traditional and new methods to deal with selective jamming attacks. The new methods suggested are game-theoretic approaches and the usage of machine learning. These two new methods significantly impact the detection of selective jamming attacks rather than the traditional ones.

Due to the growing usage of LoRa and the expansion of IoT devices, the paper [4] explains how to avoid sniffer activities on wireless sensor networks, particularly in LoRa networks. The AES and MAC algorithms are implemented in the LoRa network to protect data during transmission. The overhead analysis of IoT constrained devices class 0, and class 2 was also explored in this paper, with the findings indicating that these two algorithms could be applied to these devices. The network architecture in this research was still a local network. As a result, it is believed that additional study would enable this technology to be implemented on the LoRaWAN network, allowing data to be accessible over the internet.

Paper [8] analysis of LoRaWAN and its future directions focused on the threat of LoRaWAN, such as physical capture of end devices, sniffing gateways, and self-replay processes. These threats required particular attention from developers and organizations implementing LoRa networks. The problems that occurred were about the comprehensive security risks of the protocol and the way to find solutions to these security risks. Hence, the results and advantages obtained are the creation of a threat catalog for LoRaWAN by conducting discussions and analysis from the perspective of scale, impact, possibilities of each threat, and the drawbacks that may have an impact on several relevant network device security threats.

Paper [9] entitled Onboarding and Software Update Architecture for IoT was focused on Ed25519 as a derivative of the signature of EdDSA scheme. The Ed25519 algorithm applied a symmetric key using SHA- 512, a member of the SHA-2 family in the hashing process. The result showed that EdDSA provided attack resistance equal to 128-bit symmetric ciphers, using a 16-byte public key and a signature key of 64 bytes for the Ed25519 algorithm. This paper is used as a reference for designing the Lora system in this research.

Meanwhile, paper [10] discusses experimental tests focusing on the energy efficiency and security of LoRaWAN end devices (WisNode RAK811 and Seeeduino SX1301). In the security aspect, the experiment depicted that Activation By Personalization (ABP) mode is a more energy-efficient solution that comes at the sacrifice of security. Due to the lack of a join method, the ABP mode exchanges fewer messages. ABP offers an additional security risk because it relies on counter values maintained in memory and is unable to renew session keys. The end device will go into an out-of-sync condition and become useless if there is a problem retaining or reading these settings. WisNode devices are more vulnerable to physical memory assaults due

to this security feature. An OTAA system that provides secure session keys to safeguard communication is an option to secure these devices.

The authors of [11] stated that a dual-blockchain structure could be used to secure a LoRa-based information system. The algorithms used in this research are decentralized to reduce the dependency on a centralized server. Blockchain is also utilized for securely updating IoT device firmware using LoRa as a communication protocol [12].

Encrypt then Sign was the digital signature approach used in this research because when communication is exploited by a third party, the sender and receiver of the message can determine who is exploiting the message. Due to the fact that the signature key retrieved no longer belongs to the message's sender but rather to the sniffer party, the application of the Encrypt then Sign approach drastically reduces the likelihood of message exploitation. Using the Sign then Encrypt approach, when a sniffer exploits a message from the sender and forwards it to the receiver, the received message still has the sender's signature key. This is because the sniffing party only modifies the message from the decryption process in plaintext and not the signature key of the message's sender. Consequently, the sender and receiver cannot identify the sniffing party who compromised the message.

Table 1. Comparison with previous methods

Reference	Attack/Vulnerability Type	Techniques	Security Aspect
[7]	Selective jamming	Game-theoretic approaches and reinforcement machine learning methods	Integrity
[4]	Sniffing	Advanced Encryption Standard (AES) and Message Authentication Code (MAC)	Confidentiality, Integrity
[8]	Device Cloning, Self-Replay, Rogue End-Device	Tamper-resistant hardware, Public Certificate, End-to-End Encryption	Confidentiality, Integrity
[9]	Software update, MiTM	Elliptic curve cryptography (Curve25519), authenticated key establishment, and a public key encryption	Authentication
[10]	Remote access of IoT device	Over- The-Air Activation (OTAA) and exchanging keys	Authentication
[11]	Information asymmetry	Blockchain-based LoRa-IS combined with contract theory	Authentication
[12]	Software update, MiTM	Blockchain-based system for securely updating IoT device firmware	Privacy, Authentication
This paper	Sniffing	Digital signature using Advanced Encryption Standard (AES) algorithm	Authentication, Integrity

3. PROPOSED SYSTEM

This research used two 868MHz LoRa Shield Module devices and two Arduino Mega 2560 devices as node 1 sender and node 2 receiver. A Raspberry Pi device for LoRaWAN acted as a sniffer. The programming language used was C++ with the data type sent as string data. The attack scenario was conducted by testing a man-in-the-middle attack for the sniffing process. This attack was tested before the encryption algorithm was deployed and after the signature process was implemented. The goal is to determine whether or not the payload data sent has been modified. AES 128 with 128-bit key length and AES-256 with 256-bit key length were implemented in the experiment. Moreover, Ed25519 algorithm was deployed for the signature implementation. Table 2 shows the hardware specifications and scenarios used in the experiment. Fig. 1 illustrates the overall system architecture, where node 2 as the receiver would only react if the received data contained the same ID found in node 1 as the sender. If node 2 successfully received the message sent, then the node 2 device as the receiver would send an acknowledgment to node 1 informing that the message received was valid.

Fig. 2 illustrates the flowchart of sensor node 1 as a sender. The first process was to connect the sender to node 2. The plaintext message would be added with ID and message digest when the connection was established. This plain text was then encrypted to produce cipher text. The subsequent step would be checking the key used for the signing process. If the signing process is successful, the signature key and ciphertext message will be merged and delivered to the node 2 receiver.

Fig. 3 portrays the flowchart of sensor node 2 as the receiver. First, node 2 must be connected to sensor node 1 as the message sender device. If successfully connected, then node 2 would check the messages. Furthermore, the checking process was conducted for a total length of 80 bytes message, where 64 bytes was the length of the signature key, and 16 bytes was the length of the ciphertex message. If the value of message length was equal, i.e. 80 bytes; the following process is splitting the signature key and ciphertext message. The decryption procedure then required for the inversion of ciphertext into plaintext. Before beginning the decryption procedure, node 2 would match the signature key of node 1.

Table 2. Hardware and Scenario Specifications

No	Scenario	Hardware
1	Device 1 (communicate with device 2)	Arduino Mega 2560 Rev3 with Dragino LoRa Shield
2	Device 2 (communicate with device 1)	Arduino Mega 2560 Rev3 with Dragino LoRa Shield
3	Sniffing Device	Raspberry Pi 3 model B with Dragino LoRa Hat

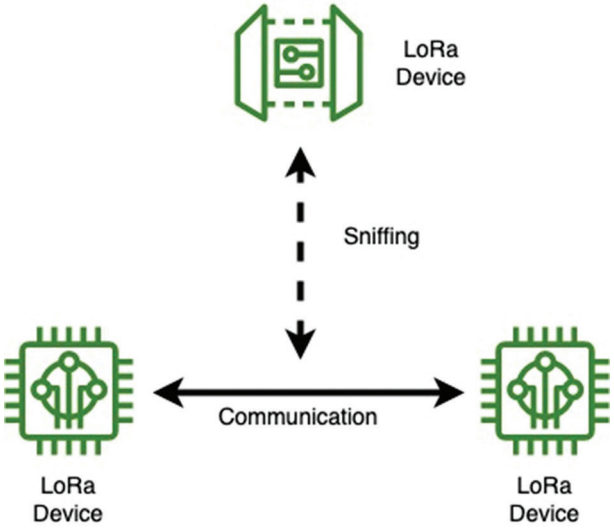


Fig. 1. System architecture

After discovering that the key is identical to the one used by node 1 during the signing procedure, decryption would be performed. The results obtained from the decryption process are sender ID, message digest, and plaintext; therefore, it is crucial to separate these three results. Following the process of splitting, the three values are stored. After the results have been saved, the receiver will verify that the sender is a legitimate user, not an adversary.

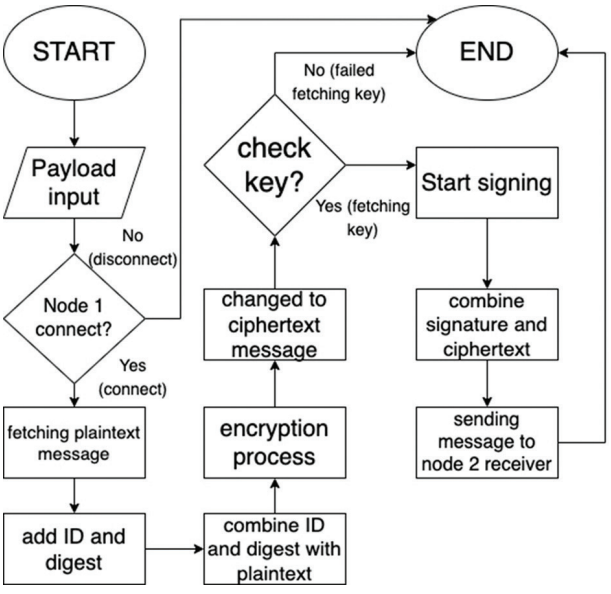


Fig. 2. Sender Flowchart

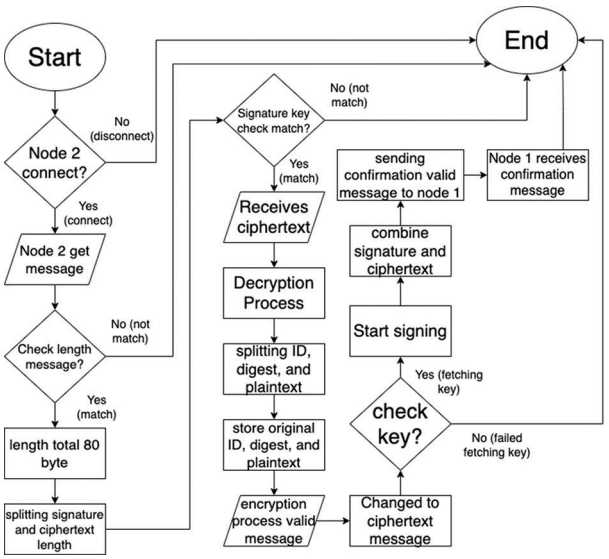


Fig. 3. Receiver Flowchart

4. RESULT AND DISCUSSION

Man-in-the-middle attack testing was applied for the sniffing process during the experiment. Three modes were set up on LoRa devices:

- mode 0 for "MODE_NON_SIGNATURE"
- mode 1 for "MODE_SIGNATURE_AES128"
- mode 2 for "MODE_SIGNATURE_AES256"

The experiment started with mode 0, followed by mode 1 and mode 2. Detail procedures are depicted on flowcharts in Fig. 2 and Fig. 3. Furthermore, Fig. 4 shows the test results of the sniffing process using mode 0 "MODE_NON_SIGNATURE". Mode 1 "MODE_SIGNATURE_AES128" is shown in Fig. 5 and mode 2 "MODE_SIGNATURE_AES256" is shown in Fig. 6.

After conducting the man-in-the-middle test for the sniffing procedure, the following test observed the system's overhead values. Similar 3 modes were utilized to evaluate the sniffing process for calculating the overhead values.

```
There are data has been sniffed.
Received:
30 31 74 65 73 74 69 6e 67
31 32 33 34 00 ab ac 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00
Packet RSSI: -50, RSSI: -106, SNR: 9, Length: 84
current time: 20-07-01 23:07:08:392
There are data has been sniffed.
Received:
52 45 43 45 49 56 45 52 5f
53 52 56 30 31 ba bc 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00 00 00 00 00 00 00
Packet RSSI: -58, RSSI: -106, SNR: 10, Length: 84
current time: 20-07-01 23:07:08:585
```

Fig. 4. MODE NON-SIGNATURE

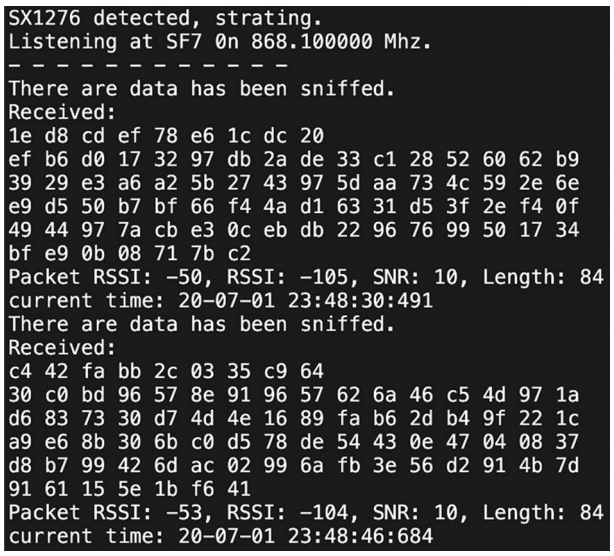


Fig. 5. MODE SIGNATURE AES 128

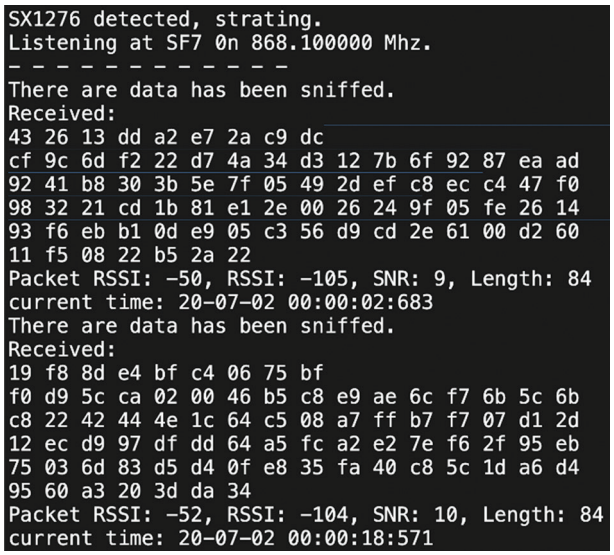


Fig. 6. MODE SIGNATURE AES 256

Overhead testing was done by sending string data from node 1 sender to node 2 receiver. The goal of the overhead analysis was to investigate the payload length, memory utilization, RAM usage, and response processing time characteristics. Table 3 shows the results of the overhead comparison of the three modes used.

As depicted in Fig. 4, the test result in mode 0 showed that the device which acted as a sniffer knew all the original payloads data of the two communicating node devices. LoRaWAN devices are susceptible to attacks and message alterations if the payload data is not encrypted. Meanwhile, for testing mode 1 in Fig. 5 and mode 2 in Fig. 6, it can be seen that the devices acted as the sniffer also knowing all communications between devices. However, the payload data obtained have been encrypted and signed so that the authenticity of the payload data could be well maintained.

This evidence shows that using digital signatures can reduce the potential for attacks due to its encryption process. Furthermore, Table 3 depicted the overhead testing outcomes for each sender and recipient. The first overhead analysis was the analysis of the length of the payload data. Based on the experiment findings shown in Table 3, Table 4 provides a more detailed description of the payload length test value. It can be seen that mode 1 and mode 2 used in the test had additional header data. In mode 0 the size of the payload length used was only 16 bytes, 4 bytes of which were the additional data consisting of 2 bytes of ID sender and 2 bytes digest, and the rest 12 bytes are considered as actual data.

Meanwhile, in mode 1 and mode 2 the payload length used was 80 bytes with 64 bytes were the additional header data, i.e., the signature key and 2 bytes of IDsender, 2 bytes of digest, and 12 bytes of actual data.

Mode	Additional data headers			Real Data	Payload
	Sender ID	digest	signature key		
Mode 0	2	2	0	12	16 bytes
Mode 1	2	2	64	12	80 bytes
Mode 2	2	2	64	12	80 bytes

Table 4. Detail Overhead Test Results

Parameters	Mode 0		Mode 1		Mode 2	
	S	R	S	R	S	R
Payload length	16 bytes	16 bytes	80 bytes	80 bytes	80 bytes	80 bytes
Memory usage	13506 bytes (5%)	12304 bytes (4%)	36770 bytes (14%)	35680 bytes (14%)	37080 bytes (14%)	35998 bytes (14%)
RAM usage	1754 bytes (21%)	1374 bytes (16%)	2564 bytes (31%)	2091 bytes (25%)	2644 bytes (32%)	2171 bytes (26%)
Delay (S to R)	174.64 ms	34.68 ms	16063.53 ms	9884.53 ms	15763.2 ms	9584.58 ms
Delay (R to S)	42.14 ms	169.07 ms	9683.87 ms	6187.79 ms	10004.4 ms	6187.77 ms

S = sender; R = receiver

The message length in mode 0 corresponded to the encryption key used in the AES algorithm, where the total key length was 32 bits. Hence, each AES algorithm was divided by 8 bits, so 128 bits = 16 bytes, and 256 bits = 32 bytes. Using AES 128 or AES 256, the size of the encrypted plaintext was only 16 bytes, independent of the AES algorithm library being used [13]. It is obvious that adding a header to modes 1 and 2 would result in a payload length that was 4 times bigger than it was in mode 0, which only used 12 bytes of actual data and 4 bytes of extra data. Consequently, the resulting payload length was indeed higher. Even though separate digital signature algorithms are utilized for modes 1 and 2, the payload length results produced for both modes are equal. Therefore, memory and RAM utilization on the system rose since the more program functionalities that were implemented, the greater memory and RAM consumption was required.

The fourth overhead analysis was a delay from sender to receiver. Table 4 shows that mode 1 and mode 2 produced a longer response time when the sender sent the payload data to the receiver, with the encryption process and added digital signing. After the message from the sender is successfully received, the receiver will carry out the signature key validation process and provide a key validation response to the sender. The fifth overhead analysis is the delay or response processing time from receiver to sender. The response time used in mode 1 and mode 2 was also longer than mode 0. This is because after the payload data were received, the receiver would confirm to the sender that the payload data received was a valid message. However, before the confirmation process was sent, the data payload must be converted into encrypted form, and the signing process was carried out first. Therefore, from the two results of response processing time testing on LoRaWAN devices, after applying the digital signature method, the transmission time was increased because the device was charged for several extra operations. Based on prior study, if the work cycle was applied to 1%; a node was allowed to send only for 36 seconds/hour or about 36 ms [14]. After establishing a security system which resulting a work cycle of 14%, the highest response processing time in this experiment was 16063.53 ms. Nevertheless, the system's utilization grows more secure.

5. CONCLUSION

Based on the experiment on the sniffing process and overhead calculations that have been conducted, it can be concluded that the digital signature method could secure the message sent between LoRa devices.

The results of the overhead analysis in this study showed that the use of digital signatures produced a high overhead value compared to those without implementation. The experiment results also revealed that this system was more efficiently applied to the AES 128 than AES 256 encryption algorithm. For future work, different security methods can be applied to improve the confidentiality, integrity, and availability aspects of LoRa network. The security method algorithm should be lightweight to be compatible with the characteristics of LoRa devices having limited resources.

6. REFERENCES:

[1] A. A. Laghari, K. Wu, R. A. Laghari, M. Ali, A. A. Khan, "A Review and State of Art of Internet of Things (IoT)", Archives of Computational Methods in Engineering, Vol. 29, No. 3, 2022, pp. 1395-1413.

[2] E. Aras, N. Small, G. S. Ramachandran, S. Delbruel, W. Joosen, D. Hughes, "Selective jamming of LoRaWAN using commodity hardware", Proceedings of the 14th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, November 2017, pp. 363-372.

[3] N. Hayati, K. Ramli, S. Windarta, M. Suryanegara, "A Novel Secure Root Key Updating Scheme for LoRaWANs Based on AES DRBG 128", IEEE Access, Vol. 10, 2022, pp. 18807-18819.

[4] P. A. Windya, V. Suryani, A. A. Wardana, "Sniffing Prevention in LoRa Network Using Combination of Advanced Encryption Standard (AES) and Message Authentication Code (MAC)", Proceedings of the International Conference Advancement in Data Science, E-learning and Information Systems, Bali, Indonesia, 13-14 October 2021.

[5] M. A. Mughal, X. Luo, A. Ullah, S. Ullah, Z. Mahmood, "A lightweight digital signature-based security scheme for human-centered internet of things", IEEE Access, Vol. 6, 2018, pp. 31630-31643.

[6] H. Hidayat, P. Sukarno, A. A. Wardana, "Overhead Analysis on the Use of Digital Signature in MQTT Protocol", Proceedings of the International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, 9-10 July 2019, pp. 87-92.

[7] D. Basu, T. Gu, P. Mohapatra, "Security Issues of Low Power Wide Area Networks in the Context of LoRa Networks", arXiv:2006.16554v1, 2020.

[8] I. Butun, N. Pereira, M. Gidlund, "Security risk analysis of LoRaWAN and future directions", Future Internet, Vol. 11, No. 1, 2018, pp. 1-22.

[9] H. Gupta, P. C. Van Oorschot, "Onboarding and Software Update Architecture for IoT Devices", Proceedings of the 17th International Conference on Privacy, Security and Trust, Fredericton, NB, Canada, 26-28 August 2019.

[10] M. Mehic, M. Duliman, N. Selimovic, M. Voznak, "LoRaWAN End Nodes: Security and Energy Efficiency Analysis", Alexandria Engineering Journal, Vol. 61, No. 11, 2022, pp. 8997-9009.

[11] G. Yu et al. "A novel Dual-Blockchained structure for contract-theoretic LoRa-based information systems", Information Processing & Management, Vol. 58, No. 3, 2021, pp. 1-23.

[12] A. Anastasiou, P. Christodoulou, K. Christodoulou, V. Vassiliou, Z. Zinonos, "IoT Device Firmware

Update over LoRa: The Blockchain Solution", Proceedings of the 16th International Conference on Distributed Computing in Sensor Systems, Marina del Rey, CA, USA, 25-27 May 2020, pp. 404-411.

[13] L. A. F. Fernandes, M. M. Oliveira, "Handling Uncertain Data in Subspace Detection", Pattern Recognition, Vol. 47, No. 10, 2014, pp. 3225-3241.

[14] D. Zorbas, K. Abdelfadeel, P. Kotzanikolaou, D. Pesch, "TS-LoRa: Time-slotted LoRaWAN for the Industrial Internet of Things", Computer Communications, Vol. 153, No. October 2019, 2020, pp. 1-10.

Deep Learning Algorithms for Diagnosing Covid 19 Based on X-Ray and CT Images

Original Scientific Paper

M. Shanthi
Manonmaniam Sundaranar University,
Department of Computer Science, Nesamony Memorial Christian College,
Marthandam, Tamilnadu, India
shanthim104@gmail.com

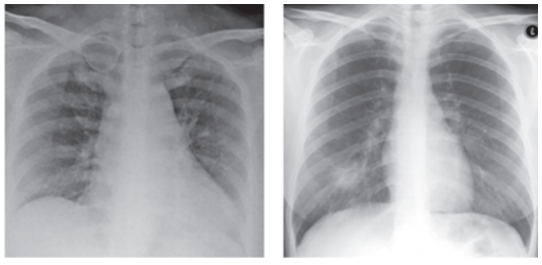
C. H. Arun
Manonmaniam Sundaranar University,
Department of Computer Science, Nesamony Memorial Christian College,
Marthandam, Tamilnadu, India
arunch394@gmail.com

Abstract *– An outbreak of a highly pathogenic coronavirus, which can cause chronic respiratory illness and high mortality rates. It takes a considerable amount of time to perform the polymerase chain reaction (PCR) used in COVID tests. Its accuracy ranges from 30% to 70%. In contrast, CT and chest X-ray diagnostics are 98% and 80% accurate in detecting COVID, respectively. A deep learning algorithms was applied to CT and X-ray images to enable rapid and accurately diagnosis of COVID-19 within seconds. In this survey, we revised all state-of-the-art studies of COVID-19 based on CT and X-ray images. Also, we analysed multiple deep learning networks and compared the performance of each technique. The result of the comparison shows that the baseline neural network has better efficiency in the recognition of COVID-19. The detection accuracy of baseline networks ranges between 93% and 98.7%. This shows the efficiency of deep learning techniques in identifying COVID-19.*

Keywords*: COVID-19, chest X-rays, CT scans, deep learning networks*

1. INTRODUCTION

The onset of the once-in-a-century pandemic coronavirus or SARS-CoV in Wuhan, China, spread its roots within a split second and triggered the global issue. According to Global Statistics, over 2.3 billion pathological cases and 4.8 million deaths were recorded across 188 nations and territories as of October 1, 2021. COVID-19 is a infectious disease that is spread by the physical contact of an infected person, saliva droplets, and sometimes airborne [1]. It suppresses the immune system and causes serious respiratory disorders. The standard Polymerase chain reaction (PCR) test for the diagnosis of COVID-19 has only 30 to 70% efficiency and it consumes time. Modern diagnostic tools such as computed tomography (CT) and Chest X-ray also have effective results on Covid-19 [2]. Early detection of the syndrome and infection level prevents the severity, increases the recovery phase and particularly reduces the further spread of infection. Implementation of the deep learning technique in medical diagnosis allows the creation of approaches from beginning to end without the need for manual intervention, yielding predictable outputs from input data [3-5]. For recognizing suspected instances of the coronavirus, computer vision studies use DL techniques like convolution neural network (CNN), recurrent neural network (RNN), and supervised and semi-supervised models to classify the CT images or chest X-rays of the chest as normal or abnormal. Herein, CNN models such as ResNet, Mobile Net, LeNet, and some other backbone techniques such as VGG, inception, and Xception structures have shown extremely accurate performance in the areas of image identification and computer vision detection [6]. They are often used for computer vision tasks. CNN [7, 8], COVID Screen [9], and COVINet [10] are some of the deep learning networks that have been designed for identifying COVID-19 occurrences. Fig. 1. illustrates the difference between positive and negative COVID-19 chest X-ray images.



(a) (b)

Fig. 1. Chest X-Ray images of Covid-19 (a) positive, and (b) negative

Deep learning networks require adequate data to train, process, and classify the input image. In the field of medical image processing, in which there is only a finite amount of information available, that results in an insufficient training database. To solve this issue, the transfer learning methodology was established. [11, 12], employed transfer learning to train the model with minimal trials by transferring data from a previously trained network to a newer model for evaluation. Integration of AI with chest X-ray technology has latterly been proven to be more successful in detecting this problem. The models used for the analysis were ResNet, InceptionV3, InceptionResNetV2, VGG16, Mobile Net, Xception, and DenseNet121, which were fine-tuned through a new batch of layers substituted by the network's head. But most of the deep learning networks fail in the execution process. The pitfalls of deep learning and its execution failures are discussed in.

The main aspect of this work is to analyze all the up-to-date DL techniques based on the COVID-19 diagnosis. All the recent deep learning studies on the Covid-19 using chest X-rays and CT scans are reviewed with a detailed view of existing techniques along with their drawbacks. The key focus of the review is elaborated as follows:

- To study all advanced deep learning networks on predicting COVID-19 infection, particularly from CT and chest X-ray images.
- To present a detailed view of different data sources reliable on Covid-19.
- To discuss the challenges faced by the recent learning structure both in the training and testing process.
- To provide a future guideline for overcoming existing limitations and designing an effective system for detecting COVID-19 infection.

The remainder of this study is arranged as follows; section II explains a detailed study of COVID-19 datasets. Section III narrates a detailed description of deep architecture. Section IV comprises the comparison study and tabulation of up-to-date deep learning techniques on COVID-19, especially using CT and chest X-ray images on Covid infection. Section V explains the discussion of the survey, and the conclusion part is stated in section VI.

2. DATASET

A detailed study on the COVID-19 dataset is described in this section. Data harmonisation is the process of combining data from various sources such as CT scans and X-rays into a single cohesive data set by modifying data formats. It covers most of the reliable datasets on COVID-19. The COVIDx-CT dataset is significantly large. The data is obtained from the CNCB. It is limited to data from China's various provinces, implying that COVID-19 symptoms in CT imaging may not be appropriate for instances outside of China. The GitHub dataset comprises about 1140 normal and abnormal images of Covid-19 infection. The dataset includes bacterial, viral, Chlamydophila, E. coli, fungal, COVID, Influenza, Klebsiella, Legionella, Lipoid, MERS, Mycoplasma, No Finding, Pneumocystis, pneumonia, SARS, Strecoccus, Varicella, and viral infection images. There are two perspectives for COVID images: PA and AP views. The normal images were collected from the Kaggle website's Pneumonia dataset. There are more than 500 images, but they must be counted in the same way as COVID images.

Table 1. Dataset description

Datasets	Total images	Attributes (Patients / Age)	Classes	Severity level (positive class)
COVIDx-CT	201,103	3,745/ 18-80	Negative (100548) Positive (100555)	Normal-PCR+:9568 Mild:25137 Moderate:50274 Severe:15568
RSNA pneumonia CXR challenge	30,227	1546/ 3-35	Negative (15115) Positive (15112)	Normal-PCR+:945 Mild:3778 Moderate:7556 Severe:2833
Chest X-ray8	32,717	5428 / 20-45	Negative (16360) Positive (16357)	Normal-PCR+:661 Mild:4564 Moderate:8178 Severe:2954
MIMIC-CXR	377,110	65,379/ 25-60	Negative (188,555) Positive (188,555)	Normal-PCR+:19064 Mild:53569 Moderate:80138 Severe:35784
PadChest	160,000	67,000/ 18-55	Negative (15115) Positive (15112)	Normal-PCR+:871 Mild:3778 Moderate:7523 Severe:2940

COVID-19 ImageData Collection is the primary source for the COVID-19 class. It has 76 good and 26 negative PA perspectives. Most research use CXR from one or more public pulmonary illness data sets to create non-COVID classes. The following are some examples of these repositories: On Kaggle, one can find the RSNA Pneumonia CXR challenge dataset, Dataset ChestX-ray8, MIMIC-CXR dataset, PadChest dataset. Multiple kinds of research have demonstrated that better data leads to better models. All pre-processing procedures that raise the value and validity of data are referred to as smart data. These tactics include noise reduction, data augmentation, and data transformation.

3. MATERIALS AND METHODS

In this section, we have discussed the existing deep learning techniques developed during the past two years for the detection of Covid-19 infection and the diagnosis process. The analysis discusses the architectural modification and structural implementation of neural networks in the case of covid-19 diagnosis. The following networks were some of the recently developed deep learning architectures which show remarkable efficiency in covid-19 prediction.

3.1 DEEP LEARNING

Deep learning (DL) and machine learning (ML), two important AI disciplines, has recently sparked a lot of interest in medical applications. Some of the DL systems are designed using a pre-trained model that employs transfer learning, while others utilize custom networks. Machine learning and data science are also frequently employed in the disciplines of coronary diagnosis, prognosis, prediction, and epidemic forecasting. The intensity of the epidemic has also been reduced because of computer vision [29]. A DL model that is both dependable and accurate could be utilized as a triage tool and to assist in clinical decision-making. A widening number of recent studies assert to have attained remarkable sensitivity levels of > 95%, considerably higher than competent radiologists. In the modern environment, CNN is the main factor in exhibiting remarkable performance in the field of medical image analysis. ResNet, Xception, Inception, DenseNet, GoogleNet, and other CNNs are among the most useful. By extracting characteristics from the CXR images, the DL technique was able to distinguish between normal, pneumonia, and COVID-19. Because it is lightweight, the equipment required for this test is less cumbersome and portable. This sort of resource is more often accessible than necessary for RT-PCR and CT-scan testing. Furthermore, a patient's chest X-ray takes only 15 seconds, making it one of the most cost-effective and time-efficient evaluation techniques. The CNN is a deep learning network with ResNet, Mobile Net, LeNet and some other backbone techniques such as VGG, inception and Xception structures has developed multiple neural networks to detect the Covid infection and yields effective progress. The basic architecture of the CNN model is depicted in Fig. 2. Internet of Things (IoT), big data, and smart technologies are also effective in combating the spread of COVID 19.

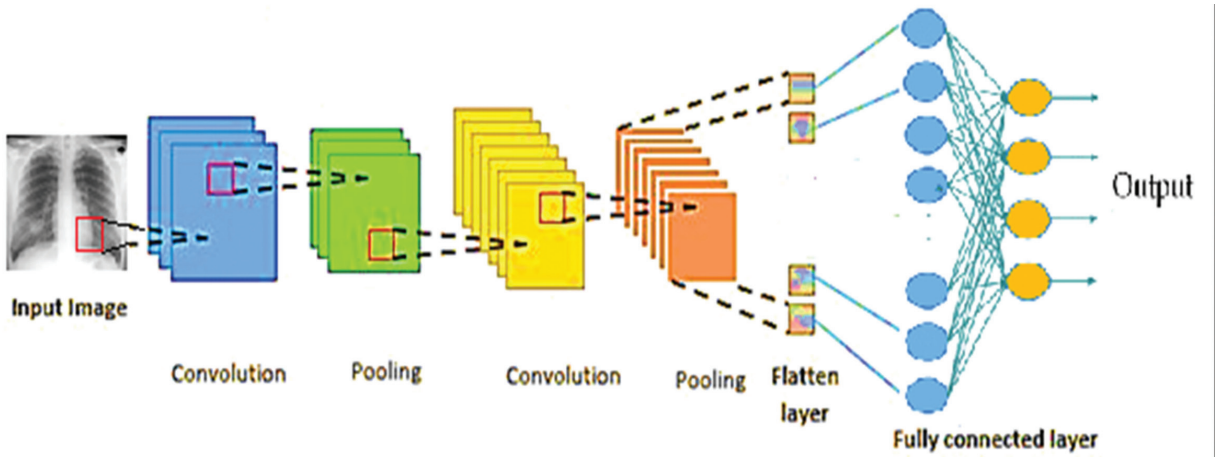


Fig. 2. Basic architecture of deep learning network

CNN can learn hierarchic characteristics, an essential trait, automatically. The initial several CNN layers frequently identify fundamental characteristics like horizontal, vertical, and diagonal borders. The output of these layers is transmitted to the intermediate layers, which extract more advanced characteristics like corners and edges. This implies that the characteristics calculated by the initial layers are generic and may be used for a range of issues, while the characteristics calculated by the latter layers are particular to the given dataset and tasks. CNN offers an important benefit from the need for a reduced number of neurons and hyperparameters compared with conventional feed-in neural networks. Several CNN baseline designs have been created and effectively used to solve complicated visual imaging tasks for image recognition applications. In this work, we opt to construct suggested models with pre-trained models like VGG16, InceptionV3, and Xception.

VGG 16 Net

A standard object-cognition model with up to 16 layers is the VGG16 Object-cognizing model. ImageNet performs VGG16, a deep CNN, on a variety of tasks and datasets. VGG16 is one of the most commonly used imaging models today. We thus suggest that this model be used.

Inception V3

InceptionV3 was one of the earliest batch standardization models. It also used the factorization method for more efficient calculations. InceptionV3 will parallel operations and execute convolutions and batch normalization in parallel before the results are concatenated instead of linearly performing processes, with additional parameters and complexities. InceptionV3 enables advanced treatment with directed acyclic graphs.

Xception

Xception is a CNN completely made of convolutionary layers, which are profoundly separate. Xception combines ResNets with a profoundly separable convolution to produce a light yet highly powerful network. Xception achieved greater precision than InceptionV3 with the ImageNet dataset. Table 2. shows the performance analysis of a pretrained deep learning networks on the Covid dataset.

Table 2. Performance analysis of pretrained deep learning model on Covid Chest X-ray 8 dataset.

Models	Labels	Precision	Sensitivity	Specificity	F1-score	Accuracy (%)
DenseNet121	Normal Pneumonia COVID	0.93 1 1	1 0.92 1	1 0.96 1	0.96 0.96 1	97
Xception	Normal Pneumonia COVID	0.91 1 1	1 0.90 1	1 0.95 1	0.95 0.95 1	96
MobileNetv2	Normal Pneumonia COVID	0.91 1 0.95	1 0.86 1	1 0.93 1	0.95 0.92 0.98	95
ResNet50v2	Normal Pneumonia COVID	0.86 0.98 1	1 0.84 0.98	1 0.92 0.99	0.93 0.90 0.99	94
NASNetMobile	Normal Pneumonia COVID	0.86 0.95 1	0.98 0.84 0.98	0.99 0.92 0.99	0.92 0.89 0.99	93
VGG19	Normal Pneumonia COVID	0.83 0.96 1	0.98 0.86 0.90	0.98 0.93 0.96	0.90 0.91 0.95	92

4. RESULT AND DISCUSSION

This section compares the available Covid infection approaches using CT and X-ray imaging in detail.

4.1 RESULTS OF PROPOSED MODELS

The functionality of the pretrained CNN models generated in this research is assessed. The experiments have been carried out with the tuned hyper-parameters are depicted in table.3, which produced the better results during training. Table 3 mention the comparison details of the existing baseline technique based on model size, training and inference time.

Table 3. Comparison of baseline technique based on model size, training time and FPS

Model	Parameters	Model Size	Training Time	FPS
U Net	31.07M	118.24MB	51min	1.88
E Net	343.7K	1.33MB	15min	4.03
U Net ++	9.16M	34.95MB	58min	1.81
Attention U Net	34.87M	133.05MB	63min	1.75
ANAM-Net	4.47M	17.21MB	27min	2.76

In terms of sensitivity, accuracy, and specificity. Anam-Net outperformed the better outcomes. On the other hand, UNet++ has a vast dense connection, which results in hierarchical encoder-decoder modules that enable efficient feature propagation for accurate segmentation. When compared to other models, the proposed Anam-Net with fewer parameters was able to provide reliable segmentation results. In the cross-data set assessment, Anam-Net did quite well (second best), as shown in Fig. 3, and was similar to the highest performing technique.

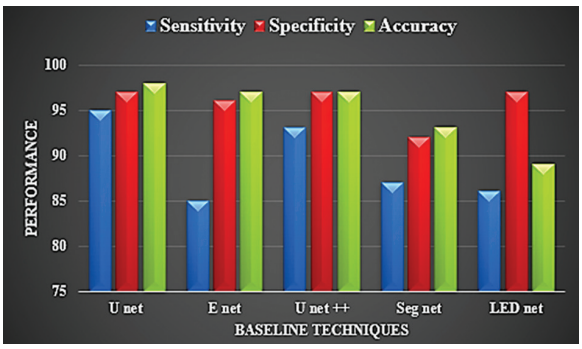


Fig. 3. Comparison of various baseline techniques based on sensitivity, specificity and accuracy based on COVIDx-CT dataset

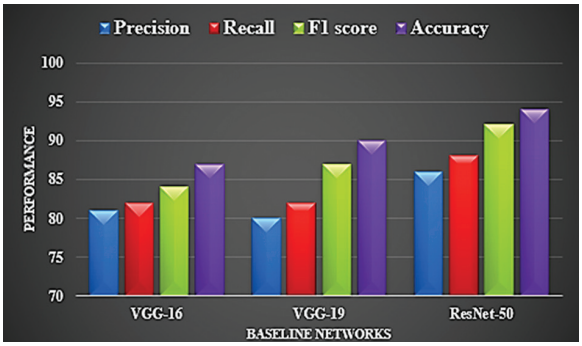


Fig. 4. Comparison of baseline technique based on precision, recall, F1 score and accuracy based on COVIDx-CT dataset

Necessity of DL in CXR

To diagnose COVID-19, many researchers and practitioners have relied on simple radiographic imaging or X-rays. Nonetheless, these images lack the requisite resolution and precision to diagnose COVID infection in the initial stage and they have some drawbacks in this regard. As a result, AI researchers went to the aid of medical specialists and deployed DL as a potent tool to progress the accuracy rate of COVID-19 detection using X-rays. The below figure 5. depicts the rate of using different radiological techniques for diagnosis and detection of COVID-19.

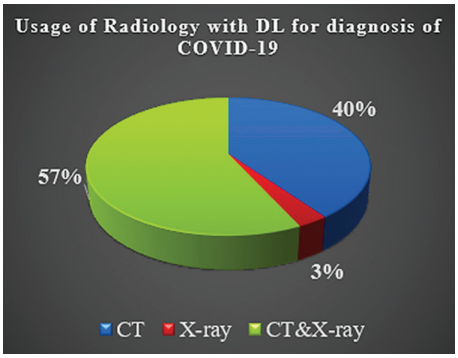


Fig. 5. Rate of using DL accompanied radiology

This showed that a greater proportion of persons will need to be examined in short durations by a few pro-

fessionals with limited resources. The uniform database encompasses all severity levels, from normal with Positive RT-PCR to Mild, Medium, and Extreme. The performance measure for SD-Net on various Chest x-ray databases is listed in Table 4. And Table.5 represents the efficiency of the hybrid network and baseline network in Covid-19 classes based on CXR images respectively. Fig.6. represents the efficiency rate of CNN models in COVID-19 images.

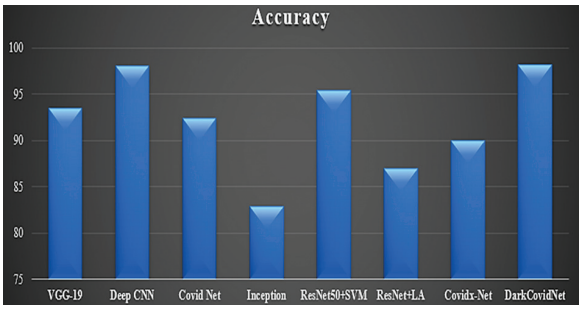


Fig. 6. Accuracy rate of CNN architectures of radiology modality images of COVID-19.

Table 4. Comparative analysis of hybrid Model based on CXR images

Metric	Sensitivity	Specificity	Precision	Accuracy
COVID Net-CXR	69.12	88.23	4.56	72.1
COVID-CAOS	74.89	65.12	68.56	69.56
RES Net without segmentation	68.63	80.23	75.23	79.25
RES Net with segmentation	2.01	79.34	78.52	75.89
FuCiT Net	78.94	81.56	76.01	77.23
COVID –SD Net	80.04	89.56	79.74	80.12

Table 5. Performance evaluation of five-fold cross-validation for Covid -19 classes based on CXR images

Model	Sensitivity	Specificity	Precision	F1-score
COVID-CAPS	0.953	0.351	0.710	0.822
COVID-CAPS (focal)	0.962	0.231	0.664	0.781
COVID-CAPS scaled	0.921	0.542	0.763	0.831
COVID-CAPS scaled (focal)	0.994	0.363	0.674	0.812
POCOVID-Net	0.882	0.764	0.852	0.873
POCOVID-Net (focal)	0.881	0.792	0.861	0.873
Mini- COVIDNet	0.923	0.684	0.821	0.863
Mini-COVIDNet (focal)	0.921	0.712	0.832	0.874
MobileNet-v2	0.964	0.612	0.813	0.881
MobileNet-v2 (focal)	0.912	0.643	0.813	0.860
NASNetMobile	0.921	0.494	0.734	0.822
NASNetMobile (focal)	0.952	0.521	0.751	0.842
ResNet50	0.841	0.572	0.761	0.801
ResNet50 (focal)	0.952	0.472	0.742	0.832

Table 6. Comparison of different frameworks for COVID-19 prediction on the X-RAY images

Author & year	Dataset	Data description	Preprocessing method	Network model	Partitioning	Validation results
Dong, S., et al [14](2021)	Public CXR dataset (COVIDx)	15134 (Normal-8851Pneumonia-6045COVID-19-238)	-	RCoNet	5-fold cross validation	Accuracy-92.98% Sensitivity-93.39% Specificity-93.51%
Tabik, S.,[15] (2020)	COVIDGR-1.0	Normal-426 COVID-19-426	Segmentation-based cropping, class-inherent transformation network (to increase the discrimination capacity)	COVID-SDNet	Traning-80% Testing-20%	N- specificity-80.79±6.98, N- precision-74.74±3.89, N- FI-76.94±2.82 p- sensitivity-72.59±6.77, p- precision-78.67±4.70, p- FI-75.71±3.35 Accuracy-76.18±2.70
Wang, L et al., [16](2020)	COVIDx 1.0, RSNA pneumonia detection challenge dataset	13800(Normal-8066,Pneumonia-5538,COVID-358)	-	COVID-Net	Training-98% Testing-2%	Sensitivity-80% Accuracy-92.6% Precision-88.9%
Ozturk, T et [17] al.,(2020)	chestX-ray8 data base,covid-19 X-ray image	Normal -500Pneumonia-500Covid-19-127	-	Dark covidnet	5-fold cross validation	Accuracy-95.13% Precision-98.03% Specificity-95.3% Sensitivity-85.35% FI-score- 96.51%
Ohata, E.F et al.,[18](2020)	Kaggle COVID-19 in X rays	Healthy-194 Covid-19-194	Transfer learning	Mobile net+SVM Densenet201+MLP	10-fold cross validation	mobilenetAccuracy&FI-score-98.5% densenetAccuracy&FIscore-95.6%

Author & year	Dataset	Data description	Preprocessing method	Network model	Partitioning	Validation results
Oh, Y., et al [19] (2020)	CXR dataset, JSRT,NLM	502(normal-191, bacterial-54, tuberculosis-57, viral-20, COVID_19-180)	Data augmentation via batchextraction	Densenet-103 Resnet-18	80%-training 20%-testing	Accuracy-91.9%,
Horry et al.,[20](2020)	COVID-19 X-ray image database, NH chest X-ray	400(COVID-19-100, pneumonia-100, normal-200	Equalization of sampling bias, segmentation-based noise reduction, data augmentation	VGG16, VGG19, Resnet50, Inception v3, Xception	Training 80% Tesing-20%	Sensitivity=80 Precision-83 FI-score-80
Panwar H et al., (2020)[21]	Chest xray data set	337(covid-19-192)	Resize the image, data augumentation	nCOVnet	70% - training 30%- testing	Accuracy-97%
Toğaçar, M [22] (2020)	Joseph Paul Cohen dataset, Kaggle	458(Covid-19-295, normal-65, pneumonia-98)	Fuzzy color technique	MobileNetV2, SqueezeNet	5-fold cross validation	Accuracy=98.25% FI-score-93.48
Hussain E [23] (2021)	Covid-R	7390 (Covid-19-2843Normal-3108, Pneumonia-1439)	Generating dataset	CoroDet	5-fold cross validation	2-class acuuracy-91.1% 3-class accuracy-94.2% 4-class accuracy-91. %
Jain, R (2021) [24]	Kaggle	6432 (583-normal, 576-covid-19,4273-pneumonia)	-	Inception V3, Xception Net,ResNeXt	90%testing, 10%validation	Accuracy-97.97%
Haghanifar, A (2020)[25]	NH CXR-14	CAP-4600, Normal-5000, COVID-19-780	Image augmentation, enhancement algorithm	COVID-CXNet	-	Accuracy-96.72%
Hemdan, E.E.D et al (2020)[26]	Public dataset of X-ray	50 (25-normal, 25-COVID-19	One-hot encoding	COVIDX-Net	80%-20%	Inception v3-50%, VGG19 &DenseNer201-90% MobileNetV2-60%
Basu, S.,et al(2020) [27]	NIH Chest X-ray Dataset,	Data A (normal-350, pneumonia-322, other-disease-300, Covid-305) Data B (57560, diseased-50819)	Domain Extensive transfer learning	CNN	5-fold cross validation	Accuracy-90.13%±0.14
Ouchicha, C(2020)[28]	Kaggle's COVID-19 Radiography Database	Pneumonia-1345, COVID-19-219, normal-1341	Cropping and resizing	CVDNet	K fold cross validation	Accuracy-97.02%
Gupta, A2021[29]	Kaggle, Chest X-ray dataset	COVID-19-316, normal-1341, pneumonia-1345	Fuzzy color image enhancement, stacking	InstaCovNet-19	80%-20%	3-class Accuracy-99.08%
Rajaraman, S (2020)[30]	Pediatric CXR dataset, RSNA CXR dataset, Twitter COVID-19 CXR dataset, Montreal covid-19 car dataset	-	Lung segmentation, Median filtering, rescaling	VGG-16, VGG-19, Inception-V3	90%-10%	Accuracy-99.01% AUC-95%

Table 7. Comparison of different approaches for COVID-19 detection on the CT images

Author& year	Dataset	Data description	Preprocessing method	Network model	Partitioning	Validation results
Farid et al [31] (2020)	Kaggle benchmark dataset	102(COVID-19=51, SARS=51)	Feature extraction	CNN	10-fold cross validation	Accuracy=94.11% Precision-99.4% FI-score-94% AUC=99.4%
Gunraj, H.,[32] (2020)	COVIDx-CT Dataset	Total-104009	Data augumentation	COVIDNet-CT	-	Accuracy-99.1% Sensitivity-98.76% Specificity-99.53%
Yazdani, S et al., (2020) [33]	COV-2 CT scan dataset	2482 (COVID-19-1252, normal-1230)	Data augmentation	COVID CT-Net	-	For 0.9 threshold Sensitivity- 0.850±0.002 Specificity = 0.962 ± 0.001 FI score = 0.900 ± 0.001

Author& year	Dataset	Data description	Preprocessing method	Network model	Partitioning	Validation results
Zheng et al (2020)[34]	Three different hospitals (Union hospital, Tongji Medical college, Huazhong University of science and technology)	Total-630	Lung segmentation, data augmentation	DeCoVNet	Traning-80% Testing-20%	Accuracy=90.1% Sensitivity=90.7 Specificity=91.1 Precision=84 NPC=98.2 AUC=93
Wang, S., etal(2021)[35]	3 different hospitals	259 (pneumonia-180,79-SARS-COV-2)	ROI separate, background area filling, reverse color, grayscale binarization	M- inception	Traning-80%, Testing-20%	Accuracy-89.5% Sensitivity-88% Specificity-87%
Fan, D.P., etal (2020)[36]	COVID-SemiSeg	638 (285 -normal, 353 –COVID-19)	Data augmentation	INF-Net	Training-50%, Testing-50%	Sensitivity-87% Specificity-97% Precision-50%
Zhou, T., (2021) [37]	previous publications, authoritative media reports, and public databases	2933	Resize	Ensemble CNN	5-fold cross-validation	Alexenet accuracy-98.16 Alexnet accuracy-98.16 Googlenet accuracy-98.25
Ni, Q., (2020) [38]	Individual patient,	12291 (COVID-193854, Pneumonia-6871, Normal-8566)		MVP-Net, 3D U-Net		FI score-97% Sensitivity-95%

4. 2 DISCUSSIONS

The aim of the study is to review and present various prominent COVID-19 diagnostic techniques based on CT and chest X-ray images. Although many of the characteristics described in the related works are emphasised here, some limits still required to be considered in future studies. Initially, the COVID-19 deep-learning diagnostic systems have been presented, but no underlying knowledge descriptions of deep-learning methods highlighting mathematical evaluations are provided. This work takes on a degree of domain expertise. Secondly, several features of the studied neural networks are not addressed here, especially for custom designs, such as layer numbers, layer speciation, study rate, epoch number, lots size, dropdown layer, optimization, and loss function. Third, while this study examines diagnosis COVID-19 from a computer viewpoint, this paper does not give qualitative diagnostic results in CT scans with chest X-rays. The study aims to evaluate and provide many well-known diagnostic methods for COVID-19 based on CT and chest X-ray images. Although many of the features were discussed in this study are stressed, some restrictions still need to be considered in future studies. First of all, COVID-19 deep-learning diagnostic tools are described, but no basic explanations of deep-learning processes are offered that emphasise mathematical representations.

5. CONCLUSION

As mentioned above, early detection and the DL method for COVID-19 are the fundamental stages in avoiding sickness and reducing the complexity of the infection. The addition of DL algorithms to radiological equipment increases computation speed with a low cost and efficient diagnosis. Implementation of these efficient tools reduces human error and predicts accurate results in critical cases. This review supports the concept that DL algorithms are a potential method in which diagnostic and therapeutic processes might be optimised and improved. Despite of deep learning is the most effective computational tools for pneumonia diagnosis, in particular COVID-19, the development of COVID-19 DL diagnostic techniqu4s should be careful to prevent overfitting and optimise the generalisation and utility of deep learning models.

6. REFERENCES

[1] M. M. Islam, F. Karray, R. Alhajj, J. Zeng, "A Review on Deep Learning Techniques for the Diagnosis of Novel Coronavirus (covid-19)", IEEE Access, Vol. 9, 2021, pp. 30551-30572.

[2] A. Abbas, M. M. Abdelsamea, M. M. Gaber, "Classification of COVID-19 in Chest X-ray Images using DeTraC Deep Convolutional Neural Network", Applied Intelligence, Vol. 51, No. 2, 2021, pp. 854-864.

[3] A. Sedik, M. Hammad, F. E. Abd El-Samie, B. B. Gupta, A. A. Abd El-Latif, "Efficient Deep Learning Approach for Augmented Detection of Coronavirus Disease", Neural Computing and Applications, 2021, pp. 1-18.

[4] K. C. Kamal, Z. Yin, M. Wu, Z. Wu, "Evaluation of Deep Learning-Based Approaches for COVID-19 Classification based on Chest X-ray Images", Signal, Image and Video Processing, Vol. 15, 2021, pp. 1-8.

[5] R. K. Singh, R. Pandey, R. N. Babu, "COVIDScreen: Explainable deep learning framework for differential diagnosis of COVID-19 using chest X-Rays", Neural Computing and Applications, Vol. 33, 2021, pp. 1-22.

[6] Y. Nan et al. "Data Harmonisation for Information Fusion in Digital Healthcare: A State-of-the-Art Systematic Review, Meta-Analysis and Future Research Directions", Information Fusion, Vol. 82, 2022, pp. 9-122.

[7] W. Zhao, W. Jiang, X. Qiu, "Deep Learning for COVID-19 Detection Based on CT Images", Scientific Reports, Vol. 11, No. 1, 2021, pp. 1-12.

[8] S. Aggarwal, S. Gupta, A. Alhudhaif, D. Koundal, R. Gupta, K. Polat, "Automated COVID-19 Detection in Chest X-ray Images Using Fine-Tuned Deep Learning Architectures", Expert Systems, p. e12749.

[9] M.E. Chowdhury et al. "Can AI help in screening viral and COVID-19 pneumonia?", IEEE Access, Vol. 8, 2020, pp. 132665-132676.

[10] A. Bustos, A. Pertusa, J. M. Salinas, M. de la Iglesia-Vayá, "Padchest: A Large Chest X-Ray Image Dataset with Multi-Label Annotated Reports", Medical Image Analysis, Vol. 66, 2020, pp. 101797.

[11] M. Rey-Area, E. Guirado, S. Tabik, J. Ruiz-Hidalgo, "FuCiTNet: Improving the Generalization of Deep Learning Networks by the Fusion of Learned Class-inherent Transformations", Information Fusion, Vol. 63, 2020, pp. 188-195.

[12] Y. Liu et al. "Exploring Uncertainty Measures In Bayesian Deep Attentive Neural Networks For Prostate Zonal Segmentation", IEEE Access, 8, 2020, pp. 151817-151828.

[13] N. Paluru, A. Dayal, H. B. Jenssen, T. Sakinis, L. R. Cenkeramaddi, J. Prakash, P. K. Yalavarthy, "Anam-Net: Anamorphic Depth Embedding-based Lightweight CNN for Segmentation of Anomalies in COVID-19 Chest CT Images", IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 3, 2021, pp. 932-946.

[14] S. Dong, Q. Yang, Y. Fu, M. Tian, C. Zhuo, "RCoNet: Deformable Mutual Information Maximization and High-order Uncertainty-aware Learning for Robust COVID-19 Detection", IEEE Transactions on Neural Networks and Learning Systems, Vol. 32, No. 8, 2021, pp. 3401-3411.

[15] S. Tabik et al. "COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 based on Chest X-Ray Images", IEEE Journal of Biomedical and Health Informatics, Vol. 24, No. 12, 2020, pp. 3595-3605.

[16] L. Wang, Z. Q. Lin, 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, 2020, pp. 1-12.

[17] T. Ozturk, M. Talo, E. A. Yildirim, U. B. Baloglu, O. Yildirim, U. R. Acharya, "Automated Detection of COVID-19 Cases using Deep Neural Networks with X-ray Images", Computers in Biology and Medicine, Vol. 121, 2020, p. 103792.

[18] U. R. Ohata, G. M. Bezerra, J. V. S. das Chagas, A. V. L. Neto, A. B. Albuquerque, V. H. C. de Albuquerque, P. P. Reboucas Filho, "Automatic Detection of COVID-19 Infection using Chest X-ray Images through Transfer Learning", IEEE/CAA Journal of Automatica Sinica, Vol. 8, No. 1, 2020, pp. 239-248.

[19] Y. Oh, S. Park, J. C. Ye, "Deep Learning covid-19 Features on CXR using Limited Training Data Sets", IEEE Transactions on Medical Imaging, Vol. 39, No. 8, 2020, pp. 2688-2700.

[20] M.J. Horry, S. Chakraborty, M. Paul, A. Ulhaq, B. Pradhan, M. Saha, N. Shukla, ''X-Ray Image Based COVID-19 Detection Using Pre-Trained Deep Learning Models", 2020, https://doi.org/10.31224/osf.io/wx89s (accessed: 2022)

[21] H. Panwar, P.K. Gupta, M.K. Siddiqui, R. Morales-Menendez, V. Singh, "Application of Deep Learning For Fast Detection of Covid-19 in X-Rays Using nCOVnet", Chaos, Solitons & Fractals, Vol. 138, 2020, p.109944.

[22] M. Toğaçar, B. Ergen, Z. Cömert, "COVID-19 Detection using Deep Learning Models to Exploit Social Mimic Optimization and Structured Chest X-ray Images using Fuzzy Color and Stacking Approaches", Computers in Biology and Medicine, Vol. 121, 2020, p. 103805.

[23] E. Hussain, M. Hasan, M. A. Rahman, I. Lee, T. Tamanna, M. Z. Parvez, "CoroDet: A Deep Learning Based Classification for COVID-19 Detection using Chest X-ray Images", Chaos, Solitons & Fractals, Vol. 142, 2021, p. 110495.

[24] R. Jain, M. Gupta, S. Taneja, D. J. Hemanth, "Deep Learning Based Detection and Analysis of COVID-19 on Chest X-ray Images", Applied Intelligence, Vol. 51, No. 3, 2021, pp. 1690-1700.

[25] A. Haghanifar, M. M. Majdabadi, Y. Choi, S. Deivalakshmi, S. Ko, "Covid-cxnet: Detecting Covid-19 in Frontal Chest X-ray Images using Deep Learning", arXiv preprint arXiv:2006.13807 (accessed: 2020)

[26] E. E. D. Hemdan, M. A. Shouman, M. E. Karar, "Covidx-net: A Framework of Deep Learning Classifiers to Diagnose Covid-19 in X-ray Images", arXiv preprint arXiv:2003.11055 (accessed: 2020)

[27] S. Basu, S. Mitra, N. Saha, "Deep Learning for Screening Covid-19 using Chest X-ray Images", Proceedings of the IEEE Symposium Series on Computational Intelligence, Canberra, Australia, 1-4 December 2020, pp. 2521-2527.

[28] C. Ouchicha, O. Ammor, M. Meknassi, "CVDNet: A Novel Deep Learning Architecture for Detection of Coronavirus (Covid-19) from Chest X-ray Images", Chaos, Solitons & Fractals, Vol. 140, 2020, p. 110245.

[29] A. Gupta, S. Gupta, R. Katarya, "InstaCovNet-19: A Deep Learning Classification Model for the Detection of COVID-19 Patients using Chest X-ray", Applied Soft Computing, Vol. 99, 2021, p. 106859.

[30] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, S.K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-rays", IEEE Access, Vol. 8, 2020, pp. 115041-115050.

[31] A. A. Farid, G. I. Selim, H. A. A. Khater, "A novel approach of CT images feature analysis and prediction to screen for corona virus disease (COVID-19), Imaging Informatics and Artificial Intelligence, Vol. 11, No. 3, 2020, pp. 1-10.

[32] H. Gunraj, L. Wang, A. Wong, "Covidnet-ct: A Tailored Deep Convolutional Neural Network Design for Detection of Covid-19 Cases from Chest CT Images", Frontiers in Medicine, Vol. 7, 2020, p. 608525.

[33] S. Yazdani, S. Minaee, R. Kafieh, N. Saeedizadeh, M. Sonka, "Covid ct-net: Predicting Covid-19 From Chest CT Images Using Attentional Convolutional Network", arXiv preprint arXiv:2009.05096 (accessed: 2020)

[34] C. Zheng et al. "Deep Learning-Based Detection for Covid-19 from Chest CT using Weak Label", IEEE Transactions on Medical Imaging, Vol. 39, No. 8, 2020, pp. 2615-2625.

[35] S. Wang et al. "A Deep Learning Algorithm Using CT Images to Screen for Corona Virus Disease (COVID-19)", European Radiology, Vol. 31, 2021, pp. 1-9.

[36] D. P. Fan et al. "Inf-net: Automatic Covid-19 Lung Infection Segmentation from CT Images", IEEE Transactions on Medical Imaging, Vol. 39, No. 8, 2020, pp. 2626-2637.

[37] T. Zhou, H. Lu, Z. Yang, S. Qiu, B. Huo, Y. Dong, "The Ensemble Deep Learning Model for Novel COVID-19 on CT Images", Applied Soft Computing, Vol. 98, 2021, p. 106885.

[38] Q. Ni et al. "A Deep Learning Approach to Characterize 2019 Coronavirus Disease (COVID-19) Pneumonia in Chest CT Images", European Radiology, Vol. 30, No. 12, 2020, pp. 6517-6527.

Deep learning approach for Touchless Palmprint Recognition based on Alexnet and Fuzzy Support Vector Machine

Original Scientific Paper

John Prakash Veigas
Department of Information Science and Engineering,
A J Institute of Engineering and Technology,
Kottara Chowki, Mangaluru, India
john.veigas@gmail.com

Sharmila Kumari M
Department of Computer Science Engineering,
P A College of Engineering,
Mangaluru, India
sharmilabp@gmail.com

Abstract *– Due to stable and discriminative features, palmprint-based biometrics has been gaining popularity in recent years. Most of the traditional palmprint recognition systems are designed with a group of hand-crafted features that ignores some additional features. For tackling the problem described above, a Convolution Neural Network (CNN) model inspired by Alex-net that learns the features from the ROI images and classifies using a fuzzy support vector machine is proposed. The output of the CNN is fed as input to the fuzzy Support vector machine. The CNN's receptive field aids in extracting the most discriminative features from the palmprint images, and Fuzzy SVM results in a robust classification. The experiments are conducted on popular contactless datasets such as IITD, POLYU2, Tongji, and CASIA databases. Results demonstrate our approach outperformers several state-of-art techniques for palmprint recognition. Using this approach, we obtain 99.98% testing accuracy for the Tongji dataset and 99.76 % for the POLYU-II datasets.*

Keywords*: Palmprint Recognition, Deep learning, Support Vector Machine, Fuzzy*

1. INTRODUCTION

Palmprint recognition has recently become a study of interest in image processing, artificial intelligence, and pattern recognition as a type of biometric technology. It is a popular biometric with several advantages, including stable line features, which can handle low-resolution imaging, and cheaper capturing devices with ease of use. Palmprint images have rich and discriminative features that allow for reliable person identification. Existing palmprint recognition methods may be categorized into strategies based on structuring, texture, subspace, and statistics. Structure-based techniques are utilized to obtain relevant line and point features [1][2]. However, the recognition accuracy in structure-based approaches is low. Further, the features demand higher storage space. In texture-based techniques, rich texture information from palmprints will be extracted [3][4][5]. These methods have better classification capabilities with higher recognition accuracy. In these methods, the coding of palmprint features is carried out. Therefore they could be influenced by image translation and rotation. In subspace-based techniques, images are transformed and mapped from their higher dimensional representation to lower-dimensional vector space [6][7].

These approaches have higher accuracy and recognition speed. Additionally, most of these features are created by a biometric specialist and are hand-crafted to accomplish better performance with a specific type of biometric dataset. Conventional image-based palmprint recognition systems have drawbacks such as preprocessing, parameter settings, and hand-crafted features that need to be carried out by biometric specialists. Several techniques and applications have recently included deep learning for biometric identification. A variety of patterns are being used to train the deep network. Once the deep learning model has learned the dataset's unique characteristics, it can be incorporated to identify similar patterns. Deep learning techniques have primarily been utilized to acquire features for palmprint recognition [8][9][10].

Additionally, deep learning can be highly effective for classification and clustering tasks. The system classifies the input examples according to their associated class labels during the classification task. In contrast, when performing a clustering task, the instances are clustered according to their similarity without reference to class labels. Numerous methodologies discussed below are built on deep learning techniques for recognizing palmprints. Within the deep learning framework, the input images are fed to the CNNs, determining the optimal way to merge the pixels to obtain maximum recognition accuracy. Wang et al. [11] used two-dimensional Gabor wavelets to decompose the palmprint images. In this work, PCNN (Pulse-Coupled Neural Network) is employed to simulate the perceptive function of creatural vision and break down every Gabor sub-band together into a sequence of binary images. Entropies are computed for the binary images and are considered features. For classification, the SVM-based classifier is used.

To create a better genuine score distribution of touchless palmprint datasets, Svoboda et al. [12] suggested a CNN using the AlexNet model, which is trained by optimizing a loss function linked to the d-prime index. Minaee et al. [13] devised a palmprint recognition system based on a deep scattering CNN with two layers. Then, Principal Component Analysis(PCA) is employed for dimensionality reduction of the data. A multiclass Support Vector Machine and nearest-neighbor classifiers are used to perform the classification task. Meraoumia et al. [14] proposed a model for deep learning called PCANet for feature extraction of the palmprint. They experimented with Random Forest Transform(RFT), KNN, Radial Basis Function(RBF), and SVM classifiers for the multispectral datasets. Cheng et al. [15] proposed a technique called DCFSH that extracts palmprint convolutional features using CNN-F architecture, then learned binary coding from distilled features. A multispectral palmprint database is used to analyze DCFSH. The Hamming distance is used for matching. In [16], the palmprint images are preprocessed using a fuzzy enhancement algorithm and then trained using Alexnet, which results in higher accuracy than some conventional techniques. Zhong et al. [17] suggested a novel approach for end-to-end palmprint identification through a Siamese network. Two parameter-sharing VGG-16 networks are employed in the network to retrieve convolutional features from two input palmprint images. The top network obtained similarity in two input palmprints directly from convolutional features. Khaled Bensid et al. [18] proposed a discrete cosine transform network (DCTNet) deep feature extraction algorithm for palmprint recognition for multispectral datasets. Genovese et al. [19] introduced PalmNet. This convolutional network employs Gabor responses and PCA filters in an unsupervised approach on several touchless palmprint databases and performs classification using a 1-NN classifier based on Euclidean distance. Gong et al. [20] used Alexnet with the PRelu activation function for palmprint recognition. In [21], a pre-trained MobileNet V2 neural network is applied to learn the palmprint features, and linear SVM is used for the classification to obtain higher accuracy. Zhao et al. [22] developed a Joint Constrained Least-Square Regression model using the CNN model to address the under-sampling classification task by extracting various deep local convolutional features from multiple patches from the same palmprint image. Veigas et al. [23] proposed a genetic-based 2D Gabor filter with CNN for palmprint recognition. The filters are tuned using a genetic algorithm, and Gabor features are extracted.

Liu et al. [24] proposed SMHNet that extracts features of the palmprint at structure and pixel levels. Recently, many biometric initiatives are increasingly being explored using deep learning techniques because of the capability to extract features from noisy data and adjust to biometric data samples captured with various devices and achieve good recognition in less-constrained environments. Numerous CNNs, such as AlexNet, VGG-Net, Inception-V3, and ResNet, perform better at image recognition and classification [25].

Although CNN-based techniques efficiently capture perceptual and biometric information extracted from the input images, applying this approach to verify palmprints poses some challenges. Firstly, the sample size is a constraint in existing palmprint databases, as most CNN approaches require a substantial quantity of input data during the training phase. Secondly, the performance of CNNs is highly dependent on the underlying architecture. Due to the above-stated limitations, utilizing a CNN architecture in small datasets may result in overfitting. It is observed that data augmentation strategies are only marginally effective in reducing overfitting due to the low intra-class variability of palmprint images.

Although the pre-trained CNN model may be faster, there is another approach called Transfer learning [26]. Transfer learning is typically used to solve problems where the datasets include insufficient data to train a full-scale model from the start. Transfer learning is a process that adopts previously trained CNN, removing fully connected layers and also training the remaining layers from the required dataset. A CNN may obtain discriminative features of the image by freezing the weights of CNN layers and fully connected layers for classifying palmprint images. These addresses the challenges associated with CNN approaches that include substantial computational cost during the learning phase and overfitting induced by tiny palmprint databases.

As per the above literature survey, Alexnet has been widely used and yielded better results than other deep networks. Besides, the Alexnet has a simple architecture and can be trained with a few epochs. Therefore, Alexnet has been chosen for this work.

SVM is a discriminant method that analytically solves the problem of convex optimization and gives some optimal hyperplane parameters, unlike perceptron which are mostly used in machine learning for classification. In the case of perceptrons, the solutions depend on the criteria of initialization and termination. To address the highly nonlinear problem, kernels like RBF(Radial Basis Function) are used. Although kernel SVM solves the nonlinear problem, it cannot optimally give a solution for the hard boundary conditions [27]. Hence, fuzzy SVM have been chosen for classifying the palmprint and solving hard boundary condition.

The main contribution of this work is as follows:

- Proposed a fuzzy SVM classification approach for the palmprint recognition that provides better classification accuracy.
- Proposed a framework using transfer learning and fine-tuning the Alexnet model for feature extraction.
- Extensive experimentation is performed on four openly accessible palmprint datasets: PolyU-II, CASIA, Tongji, and IITD Contactless databases.
- Systematic analysis is performed by comparing the proposed approach with eight different state-of-the-art schemes such as CR-CompCode, LLDP, HOL, LDP, LBP, AlexNet, VGG-16, and VGG-19

The remainder of the paper is discussed as follows: Section-2 demonstrates the proposed methodology. Section-3 presents the implementation and results. Lastly, the conclusion of the work.

2. METHODOLOGY

2.1 IMAGE PREPROCESSING

After capturing the image, pre-processing is the most crucial step in developing any biometric system. First, the Region of Interest is extracted using the techniques mentioned in [23]. In the pre-processing step, the noise or any other artifacts are removed. The image is enhanced using a fuzzy enhancement algorithm [28], using the membership function with a fuzzy enhancement operator built up of piecewise continuous function. The fuzzy membership function is given by Eqn. (1).

$$P_{ij} = F(X_{ij}) = \begin{Bmatrix} s_1\, tg^2\left(\frac{\pi X_{ij}}{4(L-1)}\right) \ldots 0 \le X_{ij} \le X_T \\ 1 - s_2\left(1 - tg\frac{\pi X_{ij}}{4(L-1)}\right)^2 \ldots X_T < X_{ij} \le L-1 \end{Bmatrix} \quad (1)$$

The general idea behind the fuzzy enhancement algorithm is to perform weakening and strengthening the operations in low grey scale and high grey scale regions, respectively. Thereby the pixel's grey levels will decrease in the low scale region and increase in the high grey scale region. The enhanced ROI is given as an input to train the convolutional neural network.

2.2 CONVOLUTION NEURAL NETWORK(CNN)

CNN is a well-studied and widely applied branch of deep learning. It is a multi-layered network model, which is improved from back propagation neural network. The network uses forward propagation to compute output values and back propagation to fine-tune weights and biases. In contrast to the traditional recognition algorithm, the CNN repetitively performs convolution and pooling on the original input to produce progressively complex feature vectors and delivers the output directly via the fully connected neural network. It consists of five layers: the input layer, the convolution layer, the pooling layer, the full connection layer, and finally, the output layer. Convolutional neural networks are made use in extracting features from the image.

2.3 SUPPORT VECTOR MACHINE(SVM)

SVM is one of the successful computational mathematical models for solving classification problems. The SVM algorithm is capable of classifying both linear and nonlinear data. Support vector machines are algorithms that create a hyperplane or a series of hyperplanes by transforming the training data into multi-dimensional or infinite-dimensional space. These hyperplanes are referred to as decision planes or decision boundaries.

For the binary linear classification problem, the hyperplane is defined using Vapnik's theory [29]. Given input training dataset of the form (x_i, y_i), where x_i belongs to the class for which $y_i \in \{-1,1\}$. It is required to obtain a hyperplane that separates the classes such that

$$w.\varphi(x_i) + b = 0 \quad (2)$$

where z is a vector and b is a scalar that separates the points in the class x_i. The two sides of the hyper plane meet the inequality function criteria which is given by

$$w.\varphi(x_i) + b \ge 1 \ , where\ if\ y_i = 1 \quad (3)$$

$$w.\varphi(x_i) + b < -1 \ , where\ if\ y_i = -1 \quad (4)$$

The smallest perpendicular distance from the hyperplane to the data point is called the margin. The decision plane with the largest margin is known as the maximum marginal hyperplane. The maximum boundary separating hyperplane is chosen by SVM. SVM's maximum marginal hyperplane selection improves classification accuracy and reduces the likelihood of misclassification.

In Non-linear SVM, separation is obtained by mapping the n-dimensional input feature vector x into to the k-dimensional feature vector using the nonlinear vector function φ(x). We then construct the decision function f(x) that distinguishes data from between two different classes in the feature vector.

$$f(x) = w^T \varphi(x) + b \quad (5)$$

where w and b are the k-dimensional feature vector and biased term, respectively.

The L1-SVM is then expressed in its primitive form given as follows:

$$P(w,b,\varepsilon) = \frac{1}{2}w^T w + R\sum_{i=1}^{m}\varepsilon_i \quad (6)$$

which is constrained to $y_i f(x_i) \geq 1-\varepsilon_i,\ \forall_i = \{1,2,\dots m\}$ where R is the boundary parameter which regulates the trade-off involving training error and generalization ability, xi is a set of M n-dimensional training inputs that belongs to either Class1 or Class2, and the corresponding labels are y_i=1 and y_i=-1 for both the classes, respectively, given that the slack variable $\varepsilon_i \geq 0 \quad \forall_i = \{1,2,\dots m\}$.

Multi Class SVM

Multiclass SVM attempts to assign labels to a set of multiple items which depend on a set of either linear or nonlinear basic SVMs. In the literature [30][32], a common way to accomplish this is to divide the single multiclass problem into numerous binary class problems. There are two approaches:

One-vs-rest (OVR) approach is the simplest way to extend SVMs for multiclass problems. It involves breaking down the multiclass dataset into several binary classification problems. Here, n linear SVMs are trained separately, while data from other classes become negative cases. A binary classifier is trained on each binary classification problem, and predictions are made using the most confident model. Initially, the binary classifier is trained for a given class using the training samples. These samples are separated from the rest of the class samples. In the classification, x is classified into a multi-label class which is given as follows:

$$L_c = \{k \mid f(x) > 0 \ \ for\ k = 1,\dots m\} \quad (7)$$

where m represents the number of classes. This approach is called binary relevance method or OVR. This is an augmentation of single-label one-vs-rest classification.

One-vs-one (OVO) approach: It creates M*(M-1)/2 binary classifiers by forming the combination of binary pair-wise possibilities of the M classes.

In[31], it is observed that OVR is superior as compared to OVO approach.

Fuzzy support vector machines

Given a M class problem with class labels 1 to M, we specify distinct class labels from M+1 to N to the target training dataset, in which number of newly created classes are N-M. These classes are called multi-label classes. Here, class 1 to class m is single-labeled class The remaining k number of multi-label classes contains single-labeled classes k_1 to k_e, where $L_{k_e} = \{k_1, k_2 \dots k_e\} \in L_m = \{1,2 \dots, m\}$.

Given an optimal boundary fi(x) = 0 which separates the training samples of class i from the rest of the other class samples, then the convex region is given by

$$R_k = \left\{x \middle| \begin{Bmatrix} f_i(x) > 0, & i \in L_{k_e} \\ f_i(x) < 0, & i \in L_m - L_{k_e} \end{Bmatrix}\right\} for\ k = \{1,2\dots m\} \quad (8)$$

Where R_k represents the region for class k which contains all the training samples for class k, wherein if class k denotes a single-labeled class, then 1≤ k ≤ n, k_1=k_e=k.

Suppose we want to classify training data x into one of m distinct classes. It is accomplished by defining a membership decision function for x within the region Rk where k= {1...m}.

By introducing a fuzzy membership function [31] $fz_{ij}(x_i)$ on the direction perpendicular to decision function $f_{ij}(x)$=0 as

$$fz_{ij}(x) = \begin{Bmatrix} 1 & for\ f_{ij}(x) \geq 1 \\ f_{ij}(x) & otherwise \end{Bmatrix} \quad (9)$$

Given $fz_{ij}(x)$ (i≠j, j=1, 2... m) the membership function of x belonging to a class i is given as

$$fz_i(x) = min_{j=1,2..m} f_{ij}(x) \quad (10)$$

Which is also expressed as

$$fz_i(x) = min\left(1, min_{i\neq j, j=1,\dots m} f_{ij}(x)\right) \quad (11)$$

Finally, a given unknown sample x is classified using

$$arg\ max_{i=1,2,\dots m} fz_i(x) \quad (12)$$

Feature Extraction and Classification

Figure 1 shows the block diagram of the proposed work. The input required for the model is 224x224x3 pixel size. Hence, ROI is resized to fit to the network requirements. The model used here is Alexnet, which is consists of five convolutional layers and three fully connected ones. The convolutional layers are given as follows:

96 kernels of size 55 x 55 x 3 in the first layer, 256 kernels of size 5 x 5 x 64 in the second layer, and 256 kernels of size 3x3x256 in the following three layers. They are followed by two fully connected (FC)layers with 4096 neurons. In the final layer, a fuzzy SVM classifier is used.

3. IMPLEMENTATION AND RESULT

This section shows the implementation and results obtained for the various standard databases. The experiments are implemented using Intel(R) Xeon(R) 2.30 GHz with NVIDIA Tesla K80 GPU. The final result is evaluated using accuracy, EER, and Receiver Operating Characteristics Curves.

3.1 DATABASES USED

The presented methodology is evaluated using four openly available databases: CASIA[33], PolyU-II[34], IITD[35], and Tongji[36]. The CASIA Palmprint Database contains 5502 palmprint images from 312 different

persons. The palmprint database at the IITD has 2601 images from 460 palms associated to 230 individuals. PolyU-II database has 7752 grayscale palmprint images in the collection containing 386 palmprint images. In the Tongji dataset, there are 6000 images belonging to 600 people.

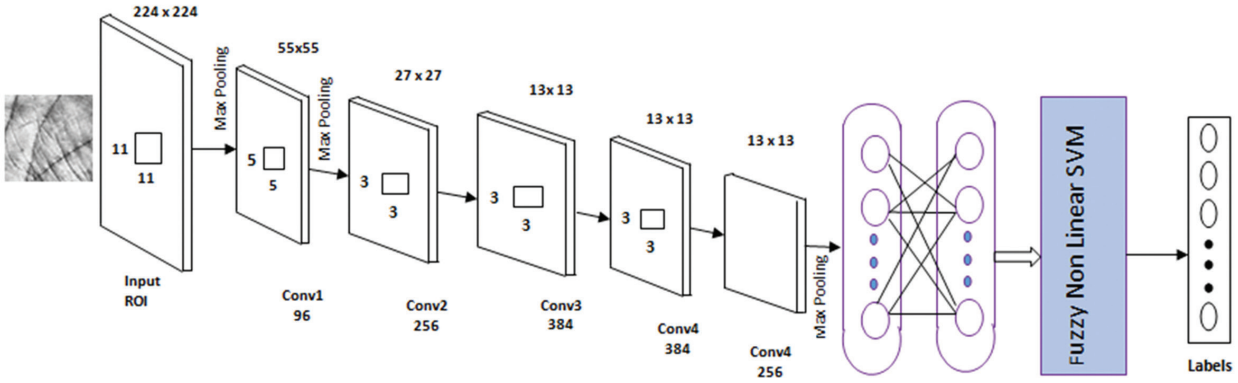


Fig. 1. Proposed Block diagram

3.2 EXPERIMENTAL RESULTS

The proposed model is assessed based in terms of accuracy using the equation-13. The accuracy of the classifier is given by ratio of correct image classification to the total number of images.

$$accuracy = \frac{(T\_P + T\_N)}{S} \tag{13}$$

Where T_P, T_N, S are true positive, true negative, and total number of sample images to be classified respectively. To perform the experimentation, we split the dataset into two parts: training dataset and testing dataset in the ratio of 80% and 20 % respectively.

Table 1. Comparison of Validation accuracy

Sl no	Method[Ref]	Classification accuracy (%)			
		IITD	POLYU-II	Tongji	CASIA
1	LBP[37]	92.81	98.2	98.9	97.3
2	LDP[38]	85.17	95.18	99.28	98.91
3	HOL[39]	95.90	98.3	99.21	98.96
4	CR-Compcode[40]	94.44	98.88	99.11	96.33
5	PCANet[14]	98.63	99.45	99.78	98.37
6	VGG-16[41]	92.56	94.28	97.14	92.14
7	VGG-19[41]	92.25	94.22	96.04	92.16
8	AlexNet[20]	96.1	98.88	99.32	96.73
	proposed	98.78	99.76	99.98	98.93

Table 1shows the accuracy of the proposed method with the already existing methods in palmprint identification experiments, all of which were evaluated on the PolyU-II, IITD, CASIA, and Tongji databases. To configure the methods mentioned above, we use the parameters made available by the researchers. Our technique is compared to the newly published techniques known in the literature. To make comparisons against methods built on local-based texture descriptors, we have used the CR-CompCode, LLDP, HOL, LDP, and LBP approaches. We compared PCANet with the pre-trained AlexNet, VGG-16, and VGG-19 CNNs as deep learning approaches. The experimental results show that the classification accuracy of the proposed method for the IITD database has an improvement of 2.68 %,6.53%, and 6.22% compared to the pre-trained Alexnet, VGG-16, and VGG-19, respectively, as shown in Table 1. The accuracy of PCAnet is on par with the proposed approach. But the amount of training time taken in PCAnet is more than the proposed approach. The proposed method has an improvement of 0.31%,5.48%,5.54%, and 0.88% when compared with the planet, VGG-16, VGG-19, and Alexnet, respectively, for the PolyU-II dataset. In the case of the Tongji database, the proposed approach has an improvement of 0.2%,2.84%,3.94%, and 0.66% for PCAnet, VGG-16, VGG-19, and Alexnet, respectively.

Traditional approaches based on local texture descriptors [37-40] reveal performance variations on different databases. However, the proposed CNNs have consistent accuracy across all databases experimented with. The recognition accuracy is highest in the Tongji dataset compared with the other datasets considered for this experiment. The data analysis shows the proposed method's classification accuracy, which is consistent across all the datasets with accuracy >98%. The advantage of fuzzy SVM is that it classifies the palmprint and solves hard boundary conditions.

The results of palmprint recognition are studied in terms of Equal Error Rate (EER), where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are the same. Table-2 shows the comparison of EER% for the different methods for the palmprint recognition for the considered databases. It has been observed that the proposed approach produces the least EER in the Tongi database with the EER % of 0.16 and performs consistently better than other approaches for other datasets.

Table 2. Comparison of EER%

Sl no	Method[Ref]	Classification accuracy (%)			
		IITD	POLYU-II	Tongji	CASIA
1	LBP[37]	10.79	3.62	1.70	4.37
2	LDP[38]	18.87	4.82	2.44	4.84
3	HOL[39]	6.7	0.31	0.41	4.62
4	CR-Compcode[40]	4.65	0.89	0.47	3.67
5	PCANet[14]	1.37	0.43	0.20	1.63
6	VGG-16[41]	7.44	5.72	2.86	7.86
7	VGG-19[41]	7.75	5.8	3.96	7.84
8	AlexNet[20]	3.90	2.01	0.68	3.22
	proposed	1.22	0.84	0.16	2.42

Figure 2 depicts the training accuracy and validation accuracy versus Epochs. The graph shows that with almost 10 Epochs, the accuracy is reached nearly above 95%. The learning rate is kept at 0.001.

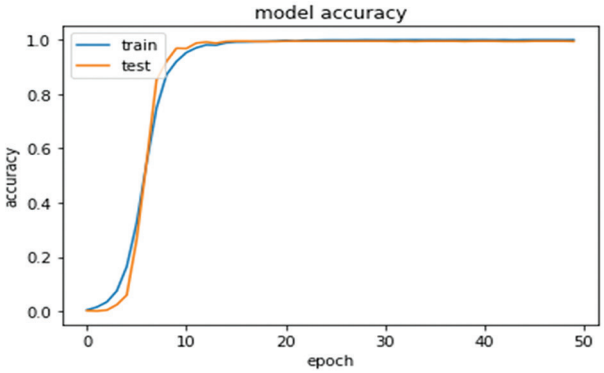


Fig. 2. Accuracy vs Epochs

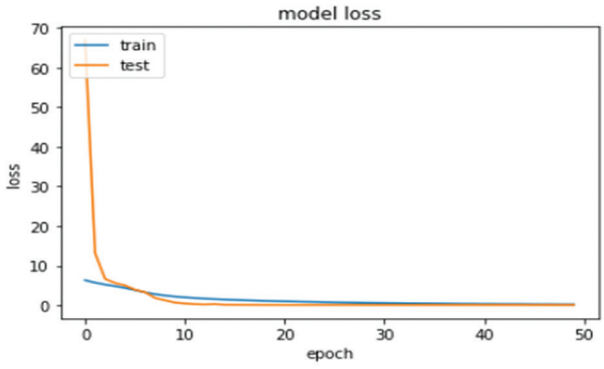


Fig. 3. Loss vs Epochs

Receiver Operating Characteristic:

The receiver operating characteristic (ROC) curve is a graphical illustration of the trade-off between genuine acceptance rate (GAR) and false acceptance rate (FAR) represented in the y and x coordinates respectively. The GAR and FAR equations are as follows:

$$GAR = \frac{T\_P}{T\_P + F\_N'} \quad (14)$$

$$FAR = \frac{F\_P}{T\_P + F\_N'} \quad (15)$$

Measuring the area under the ROC curves is a reliable approach to comparing the performance of the different classifiers. Figure 4-7 shows the ROC curves of various techniques and multiple datasets. The proposed work has achieved high GAR and low EER with the different datasets and various benchmark methods. It is observed from the graph that the ROC for the approximate coefficients is close to the optimal ROC curve. AUC lies between the values 0.95 to 1.0 which shows that the classifier is very efficient.

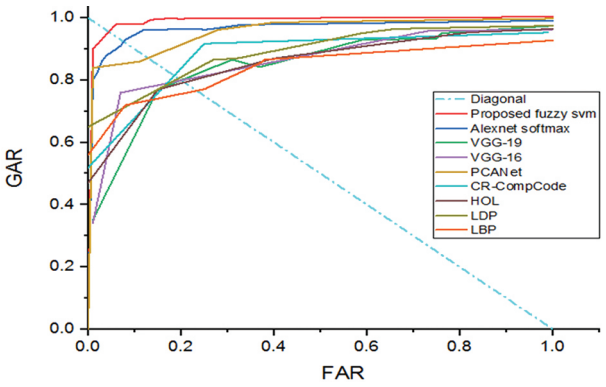


Fig. 4. ROC curve analysis for the proposed method using Tongji DataSet

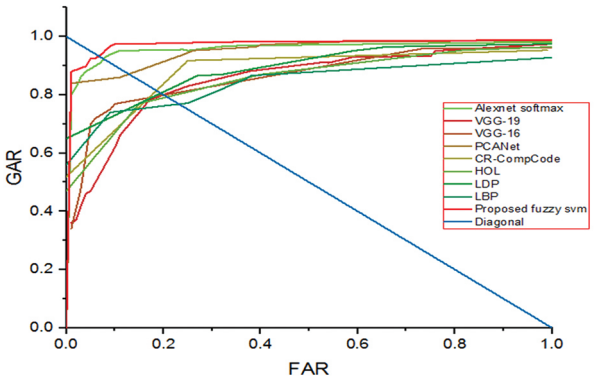


Fig. 5. ROC curve analysis for the proposed model with other methods for IITD dataset

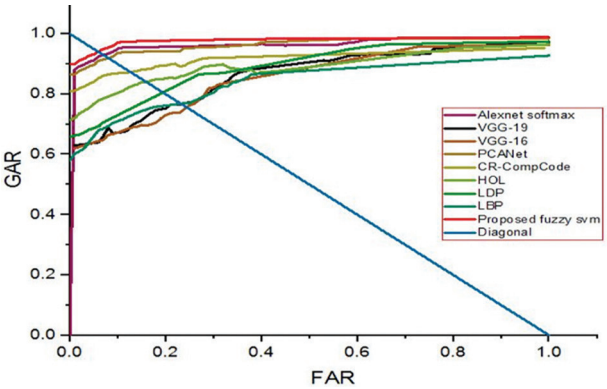


Fig. 6. ROC curve analysis for the proposed model with other methods for POLYU2 dataset

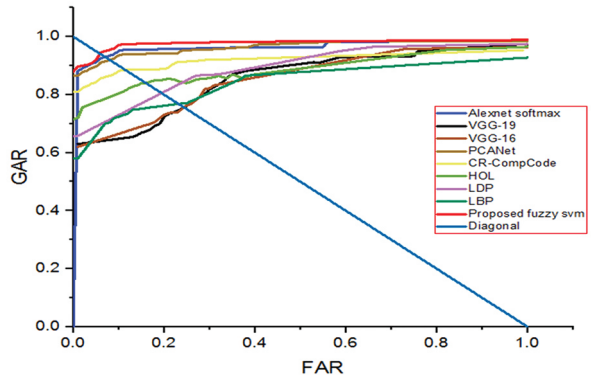


Fig. 7. ROC curve analysis for the proposed model with other methods for CASIA dataset

4. CONCLUSION AND FUTURE WORK

The presented paper proposes Convolution Neural Network (CNN) inspired by Alex-net to learn the features, and a fuzzy support vector machine is used for classification. The output of the CNN is fed as input for the support vector machine. The CNN's receptive field aids in extracting the most discriminative features from the palmprint images, and Fuzzy SVM results in a robust classification. The experiments are conducted on popular contactless datasets such as IITD, POLYU2, Tongji, and CASIA databases. Results demonstrate our approach outperformers several state-of-art techniques for palmprint recognition. Results show that the proposed method is efficient with good accuracy and very low EER values compared to the several state-of-art methods. Analysis of the ROC curves demonstrates that the proposed techniques' have higher accuracy on all databases evaluated based on Genuine Acceptance Rate and Acceptance Rate values. The fuzzy SVM is not feasible for large overlapping class labels, and one can overcome this disadvantage by using evolutionary or quantum computing techniques.

5. REFERENCES

[1] D. Zhang, W. Shu, "Two novel characteristics in palmprint verification: Datum point invariance and line feature matching", Pattern Recognition, Vol. 32, No. 4, 1999, pp. 691-702.

[2] N. Duta, A. K. Jain, K. V. Mardia, "Matching of palmprints", Pattern Recognition Letters, Vol. 23, No. 4, 2002, pp. 477-485.

[3] D. Zhang, W. K. Kong, J. You, M. Wong, "Online palmprint identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 25, No. 9, 2003, pp. 1041-1050.

[4] A. W. K. Kong, D. Zhang, "Competitive coding scheme for palmprint verification", Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004, pp. 520-523.

[5] W. Jia, D. S. Huang, D. Zhang, "Palmprint verification based on robust line orientation code", Pattern Recognition, Vol. 41, No. 5, 2008, pp. 1504-1513.

[6] G. Lu, D. Zhang, K. Wang, "Palmprint recognition using eigenpalms features", Pattern Recognition Letters, Vol. 24, No. 9-10, 2003, pp. 1463-1467.

[7] S. Zhang, Y. K. Lei, Y. H. Wu, "Semi-supervised locally discriminant projection for classification and recognition", Knowledge-Based Systems, Vol. 24, No. 2, 2011, pp. 341-346.

[8] Dexing Zhong, Xuefeng Du, Kuncai Zhong, Decade progress of palmprint recognition: A brief survey, Neurocomputing, Vol. 328, 2019, pp. 16-28.

[9] D. Samai, K. Bensid, A. Meraoumia, A. Taleb-Ahmed, M. Bedda, "2D and 3D Palmprint Recognition using Deep Learning Method", Proceedings of the 3rd International Conference on Pattern Analysis and Intelligent Systems, Tebessa, Algeria, 24-25 October 2018, pp. 1-6.

[10] A. Nalamothu, J. Vijaya, "A review on Palmprint Recognition system using Machine learning and Deep learning methods", Proceedings of the International Conference on Technological Advancements and Innovations, Tashkent, Uzbekistan, 10-12 November 2021, pp. 434-440.

[11] X. Wang, L. Lei, M. Wang, "Palmprint Verification Based on 2D - Gabor Wavelet and Pulse-Coupled Neural Network", Knowledge-Based Systems, Vol. 27, 2012, pp. 451-55.

[12] J. Svoboda, J. Masci, M. M. Bronstein, "Palmprint recognition via discriminative index learning", Proceedings of the International Conference on Pattern Recognition, Cancun, Mexico, 4-8 December 2016, pp. 4232-4237.

[13] S. Minaee, Y. Wang, "Palmprint recognition using deep scattering network", Proceedings of the IEEE International Symposium on Circuits and Systems, Baltimore, MD, USA, 28-31 May 2017, pp. 1-4.

[14] A. Meraoumia, F. Kadri, H. Bendjenna, S. Chitroub, A. Bouridane, "Improving Biometric Identification Performance Using PCANet Deep Learning and Multispectral Palmprint", Biometric Security and Privacy. Signal Processing for Security Technologies. Springer, 2017.

[15] J. Cheng, Q. Sun, J. Zhang, Q. Zhang, "Supervised Hashing with Deep Convolutional Features for Palmprint Recognition", Biometric Recognition. ,Springer, 2017.

[16] S. Dian, Liu, Dongmei, "Contactless Palmprint Recognition Based On Convolutional Neural Network", Proceedings of the Signal Processing Proceedings, Chengdu, China, 6-10 November 2016, pp 1363-1367

[17] D. Zhong, Y. Yang, X. Du, "Palmprint recognition using siamese network", Biometric Recognition, Springer, 2018, pp. 48-55.

[18] K. Bensid, D. Samai, F. Z. Laallam, A. Meraoumia, "Deep learning feature extraction for multispectral palmprint identification", Journal of Electronic Imaging, Vol. 27, No. 3, 2018.

[19] A. Genovese, V. Piuri, K. N. Plataniotis, F. Scotti, "PalmNet: Gabor-PCA convolutional networks for touchless palmprint recognition", IEEE Transactions on Information Forensics and Security, Vol. 14, No. 12, 2019, pp. 3160-3174.

[20] W. Gong, X. Zhang, B. Deng, X. Xu, "Palmprint Recognition Based on Convolutional Neural Network-Alexnet", Proceedings of the Federated Conference on Computer Science and Information Systems, Leipzig, Germany, 1-4 September 2019, pp. 313-316.

[21] A. Michele, V. Colin, D. D. Santika, "Mobilenet convolutional neural networks and support vector machines for palmprint recognition", Procedia Computer Science, Vol. 157, 2019, pp. 110-117.

[22] S. Zhao, B. Zhang, "Joint constrained least-square regression with deep convolutional feature for palmprint recognition", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 52, No. 1, 2022, pp. 511-522.

[23] John. P. Veigas, M. S. Kumari, G. S. Satapathi, "Genetic Algorithm Based Gabor CNN For Palmprint Recognition", International Journal of Recent Technology and Engineering, Vol. 8, No. 6, 2020, pp. 4895-4899.

[24] C. Liu, D. Zhong, H. Shao, "Few-shot palmprint recognition based on similarity metric hashing network", Neurocomputing, Vol. 456, 2021, pp. 540-549.

[25] W. Jia, J. Gao, G. Xia, Y. Zhao, H. Min, J. Lu, "A performance evaluation of classic convolutional neural networks for 2D and 3D palmprint and palm vein recognition", Machine Intelligence Research, Vol. 18, 2021, pp. 18-44.

[26] M. Korichi, D. Samai, A. Meraoumia, A. Benlamoudi, "Towards Effective 2D and 3D Palmprint Recognition Using Transfer Learning Deep Features and ReliefF method", Proceedings of the International Conference on Recent Advances in Mathematics and Informatics, Tebessa, Algeria, 21-22 September 2021, pp. 1-6.

[27] J. Shao, X. Liu, W. He, "Kernel Based Data-AdaptiveSupport Vector Machines for Multi-Class Classification", Mathematics, Vol. 9, No. 9, 2021; p. 936.

[28] X. Liu, "An Improved Image Enhancement Algorithm Based on Fuzzy Set", Physics Procedia, Vol. 33, 2012, pp. 790-797.

[29] C. Cortes, V. Vapnik, "Support-Vector Networks", Machine Learning, Vol. 7, No. 20, 1995, pp. 273-297.

[30] C. W. Hsu, C. J. Lin, "A comparison of methods for multiclass support vector machines", IEEE Transactions on Neural Networks, Vol. 13, No. 2, 2002, pp. 415-425.

[31] S. Abe, "Fuzzy support vector machines for multilabel classification", Pattern Recognition, Vol. 48, No. 6, 2015, pp. 2110-2117.

[32] M. Awad, R. Khanna, "Support Vector Machines for Classification", Efficient Learning Machines, Apress, 2012.

[33] CASIA database, Chinese Academy of Sciences Institute of Automation, Biometrics Ideal Test, Institute of Automation, Chinese Academy of Sciences, http://biometrics.idealtest.org/ dbDetailForUser.do?id=5 (accessed: 2022)

[34] IITD Touchless Palmprint Database, Version 1.0, https://www4.comp.polyu.edu.hk/~csajaykr/IITD/Database_Palm.htm (accessed: 2022)

[35] PolyU Palmprint Database, The Biometric Research Centre at The Hong Kong Polytechnic University, http://www4.comp.polyu.edu.hk/~biometrics/ (accessed: 2022)

[36] L. Zhang, L. Li, A. Yang, Y. Shen, M. Yang "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach", Pattern Recognition, Vol. 69, 2017, pp. 199-212

[37] X. Wang, H. Gong, H. Zhang, B. Li, Z. Zhuang, "Palmprint identification using boosting local binary pattern", Proceedings of the International Conference on Pattern Recognition, Hong Kong, China, 20-24 August 2006, pp. 503-506.

[38] T. Jabid, M. H. Kabir, O. Chae, "Robust facial expression recognition based on local directional pattern", ETRI Journal, Vol. 32, No. 5, 2010, pp. 784-794.

[39] W. Jia, R. Hu, Y. Lei, "Histogram of Oriented Lines for Palmprint Recognition", Vol. 44, No. 3, 2014, pp. 385-395.

[40] L. Zhang, L. Li, A. Yang, Y. Shen, M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach", Pattern Recognition, Vol. 69, 2017, pp. 199-212.

[41] A. S. Tarawneh, D. Chetverikov, A. B. Hassanat, "Pilot Comparative Study of Different Deep Features for Palmprint Identification in Low-Quality Images", Proceedings of the Ninth Hungarian Conference on Computer Graphics and Geometry, 2018, pp. 3-8.

Task level disentanglement learning in robotics using βVAE

Original Scientific Paper

Midhun M S
Department of Electronics,
Cochin University of Science and Technology, Kerala, India
midhunms@cusat.ac.in

James Kurian
Department of Electronics,
Cochin University of Science and Technology, Kerala, India
james@cusat.ac.in

Abstract *– Humans observe and infer things in a disentanglement way. Instead of remembering all pixel by pixel, learn things with factors like shape, scale, colour etc. Robot task learning is an open problem in the field of robotics. The task planning in the robot workspace with many constraints makes it even more challenging. In this work, a disentanglement learning of robot tasks with Convolutional Variational Autoencoder is learned, effectively capturing the underlying variations in the data. A robot dataset for disentanglement evaluation is generated with the Selective Compliance Assembly Robot Arm. The disentanglement score of the proposed model is increased to 0.206 with a robot path position accuracy of 0.055, while the state-of-the-art model (VAE) score was 0.015, and the corresponding path position accuracy is 0.053. The proposed algorithm is developed in Python and validated on the simulated robot model in Gazebo interfaced with Robot Operating System.*

Keywords*: Machine Learning, Robotics, Neural Networks, Variational Autoencoder, beta-VAE*

1. INTRODUCTION

With the emergence of Artificial Intelligence (AI), trajectory planning in robotics has been solved for many scenarios with different methods [1]. However, task planning with generative models is still because of the complex nature of the joint trajectory, joint constraints, self-collision and collision with the workspace objects. Numerous problems in robotics are solved based on reinforcement learning, established in real-time feedback from sensors. Serial manipulator robots possess multiple joints and links where each joint is controlled by one or many actuators using link actuator signals. Numerous models propose modelling the lower-dimensional joint values with sensory feedback. This approach models are an open-loop higher-dimensional abstraction of the lower-dimensional joint values.

Generative models are best suited for generating data from the same distribution. Generative Adversarial Network (GAN) [2] and Variational Autoencoder (VAE) [3, 4] are the two most common forms of generative deep learning networks. VAEs were picked because their training is more stable than GAN (no mode collapse). VAE has two models, namely encoder and decoder. The encoder maps the input into a higher-dimensional latent space, and the decoder rebuilds it.

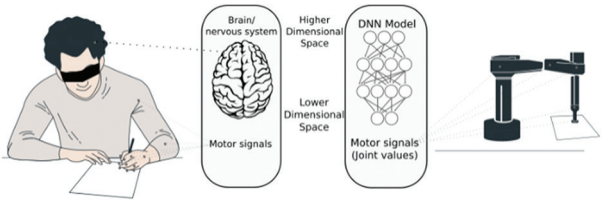


Fig. 1. The Proprioception intelligence model for human and robots. Higher and lower-dimensional space exists in the nervous system and DNN for human and robots. Lower dimensional signals directly control the joints using the motor signals.

Proprioception is the ability of a human to sense position, orientation, joint angle etc. If a robot model has these features? Fig. 1 shows the model of a human drawing an image on paper blindfolded and a robot without any visual feedback - both project the higher dimensional planning to lower-dimensional action.

A change in an independent factor in a higher dimension only affects a single factor in output is called disentanglement [5]. Disentanglement representation of data has been getting more critical in the machine learning community in recent years. The human brain coded each object based on colour, shape, size, etc.;

similarly, if the robot can learn the task's underlying nature, human interpretability, predictive performance and compressed representation will benefit.

In the proposed disentanglement robot model, the encoder generates the mapping function from lower dimensional raw trajectory data to higher dimensional representation, and all the interpretable higher-dimensional vectors generate the mapping function from higher-dimensional representation to lower-dimensional motor signals in the decoder, which encapsulates all the kinematics complexities, sequence, and task information. The main contributions in the work are as follows

- The model generates a generative model for robot task planning
- Human interpretable, disentangled latent space is learned by the model, which is an effective way to make new data from the underlying factors of variations.

The rest of the paper is organised as follows. Section 2 explains related works in the field; Section 3 describes the system implementation for disentanglement representation. Section 4 presents the simulation setup. Section 5 discusses the results obtained, and Section 6 explains the conclusions.

2. RELATED WORK

Disentanglement models generate data from the independent factors of variation. Since *βVAE* [6], disentanglement learning is getting strong community attention. Disentanglement learning is divided into supervised and unsupervised. The supervised method needs the dataset to contain all the factors of variations. Supervised and unsupervised disentanglement models gain much attention in the image, audio and video domains. Most real-world datasets do not have variation factors, so the proposed work implements the unsupervised model.

Disentanglement in robotics usually processes the input image and learns the disentanglement on those. Y. Hristov et al. [7] presented the robot learns from demonstrations from the captured scene. Mobile robot path planning and execution are demonstrated with disentanglement scene representation by V.A.K.T Rajan et al. [8]. Learning the changing surroundings by mobile robots in [9, 10] uses image-based disentanglement. M.Wulfmeier et al. [9] represent an improved reinforcement learning approach for better perception and exploration with the help of disentanglement. J. Pajarinen et al. [11] presented a probabilistic approach to disentangle the objects from an image and waste sorting using the state of the art machine learning algorithms with a robotic arm. M. Zolotas et al. [12] presented a robotic wheelchair that uses a disentangled Variational sequence encoder for trajectory planning and execution with a joystick and laser scanner inputs. Robot disentanglement models produce closed-loop control systems with cameras, sensors, user inputs, or combinations. The proposed implementation uses an open-loop control system, which learns and produces the task without feedback.

3. SYSTEM IMPLEMENTATION

The robot trajectory contains the sequence of moving joint values. Each channel in the data includes a single joint motion and collectively moves to achieve a particular goal. The data is recorded in a simulator/emulator environment and used as the dataset.

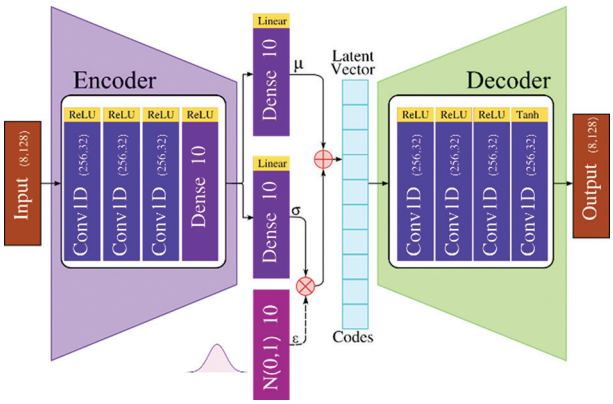


Fig. 2. The Variational Autoencoder model architecture for 1-D robot task sequence data.

3.1. THE PROPOSED NETWORK ARCHITECTURE

Autoencoder (AE) network consists of two networks named encoder and decoder. One dimensional Convolution layer captures the hidden features in the data and generates the model. The architecture is portrayed in Fig. 2. CNN learns features from the raw data while training, sparsely connected layers make it more efficient to learn large networks than the densely connected Multi-Layer Perceptron (MLP). Also, they have low computational requirements and are immune to small changes in translation, scaling and distortion in the input. Hence 1-D CNNs are used for learning the underlying data structures of robot tasks with non-linear Rectifier Linear Unit (ReLU) activation functions ($max(0, x)$). The initial layers of the encoder network learn the simple joint-trajectory features in its kernels, and the higher layers model the complex task level representation.

Table 1. DH parameters of SCARA robot.

Parameter	Link1	Link2	Link3	Link4
Link length (a)	0.45 m	0.45 m	0	0
Link Twist(α)	0	π	0	0
Joint distance(d)	0	0	d_3	0
Joint Angle(θ)	θ_1	θ_2	0	θ_4

The decoder network/generative network uses the transposed convolution layers to model the latent dimensional vector into the robot task data. The probabilistic nature of the model makes it a generative network for robot-task data. The model is trained using a variant Stochastic Gradient Descent optimiser called Adam, which is relatively computationally efficient and less prone to noise.

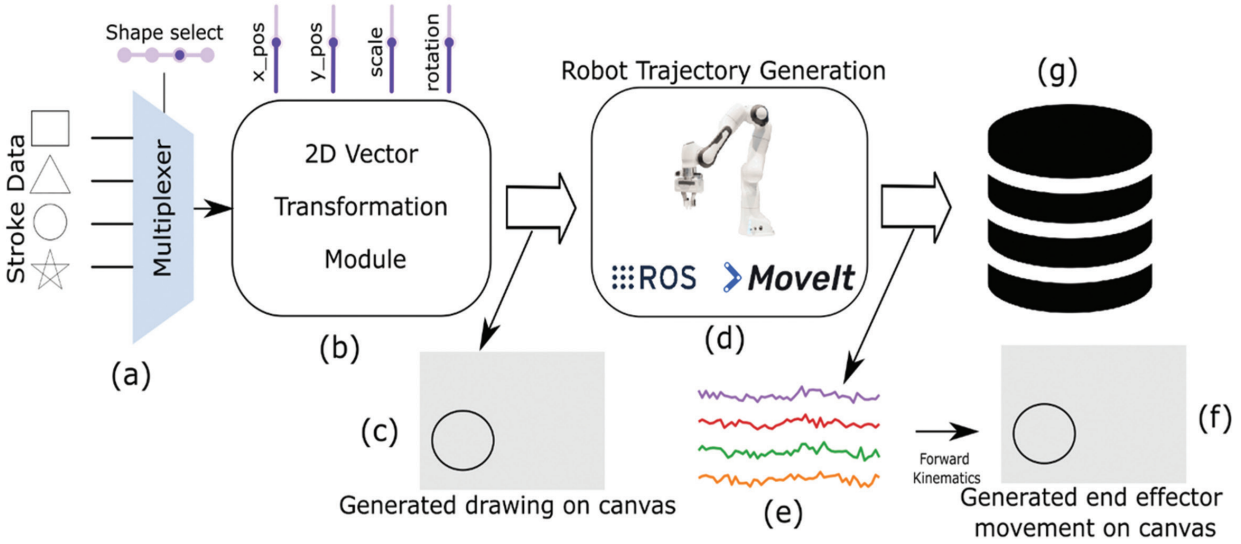


Fig. 3. The disentanglement robot task dataset generation. a) Input path generated using interpolation of points and selected using a multiplexer module, b) 2D vector transformation module with translation, rotation and scaling operations c) The generated trajectory is shown - grey colour represents the canvas d) Robot trajectory generation module e) The generated joint values f) Task space path is plotted by applying forward kinematics to the joint-values g) stored in a dataset.

Table 2. Factors of variation

Factor	Values	Count
Shape	Circle, Square, Triangle, Star	4
Scaling	0.5, 0.6, 0.7, 0.8, 0.9, 1.	6
Orientation	0, 9, 18 … 351	40
Position_x	-70.0, -68.5 … 70	32
Position_y	-70.0, -68.5 … 70	32
Total configurations		**983040**

3.2. SIMULATION SYSTEM DESIGN

Selective Compliance Assembly Robot Arm (SCARA) [13] is a 4-degree of freedom (DOF) serial manipulator robot used in the proposed work. The simulated model of the SCARA robot is developed as a physical linkage system with a Unified Robotics Description Format (URDF) file based on Denavit–Hartenberg (DH) [14] parameters described in Table 1. Robot Operating System (ROS)[15] interfaced with Gazebo simulator with Open Dynamics Engine (ODE) physics engine is used for simulating the model with joint, link, visual and collision parameters in the URDF file. Since the simulator is computationally complex, a low-footprint kinematics model is also developed with the robotics toolbox for Python [16] for evaluating the model performance in the evaluation phase.

Considering $X \in \{x_i\}$ as input and $Z \in \{z_i\}$ as the latent space vector in the network, Evidence Lower Bound (ELBO) [3] in VAE is defined as

$$\log p(X) \geq E[\log p(X|Z)] - D_{KL}[q(Z|X)||p(Z)] \quad (1)$$

where $p(X/Z)$ and $p(Z/X)$ are two probability distributions. The term $E[logp(X/Z)]$ represents the reconstruction and $D_{KL}[q(Z/X)||p(Z)]$ represents the similarity between the two probability distribution. The goal of the network is to maximise the ELBO, i.e., maximise the similarity while keeping the prior and posterior distribution closure as possible.

The model ϕ and θ represent the weights and biases of the probabilistic encoder and decoder, respectively, and its corresponding distribution is defined as $q_\phi(z/x)$ and $p_\theta(x/z)$. The final objective is to minimise the loss as

$$L_{VAE}(\theta, \phi) = -E_{z\, q_\phi(z|x)} \log p_\theta(x|z) + D_{KL}(q_\phi(z|x)||p_\theta(z)) \quad (2)$$

The gradients cannot backpropagate since the sampling operation exists in the model. Re-parameterisation trick is used to model the z as

$$z = \mu + \sigma \odot \epsilon \quad (3)$$

where fully connected layers model the mean (μ) and variance (σ) of the prior representation $p_\theta(z)$ and a sampling layer with a sampling normal vector ($\epsilon \sim N(0,1)$) is utilised. ($\odot$ represents element-wise product) The latent vector is called codes in disentanglement representation. The compressed human interpretable vector is learned - each robotic task generated with varying human interpretable factors like position, orientation, constraints etc.

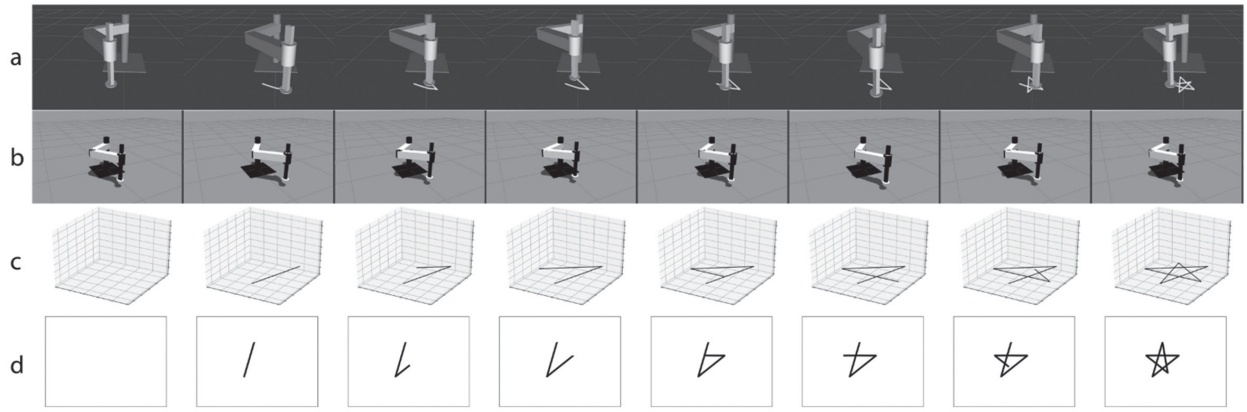


Fig. 4. The disentanglement robot task dataset generation (equidistant samples) of a sample (drawing a star). a) ROS robot visualisation (Rviz) with end-effector movement is shown, b) Gazebo simulated robot, c) The generated trajectory in three-dimensional space is shown d) Generated plot in canvas.

Adding more importance to the KL loss term in equation 2 with a new-hyper-parameter $\beta(> 1)$ will enhance the divergence factor and improves the disentanglement performance. The network is called βVAE, and the new loss is given by,

$$\mathrm{L}_{\beta VAE}(\theta, \phi, \beta) = -\mathrm{E}_{\mathrm{z}\, \mathrm{q}_{\phi}(\mathrm{z}|\mathrm{x})} \log \mathrm{p}_{\theta}(\mathrm{x}|\mathrm{z}) + \beta \mathrm{D}_{\mathrm{KL}}(\mathrm{q}_{\phi}(\mathrm{z}|\mathrm{x})||\mathrm{p}_{\theta}(\mathrm{z})) \quad (4)$$

As the β increases, the representation becomes more suitable, but the reconstruction loss increases, leading to lower precision in robotic tasks, which is not advisable.

4. SIMULATION SETUP

4.1 DATASET

Disentanglement testing Sprites dataset (dSprites) [15] is a popular dataset with images and its underlying factors of variations. The robot version of the dSprites-like dataset is developed using the SCARA robot with a straightforward task - "draw a shape on a canvas", as shown in Fig. 3. Four different shapes were picked - box, circle, triangle and star. Each shape creates multiple instances by varying the independent factors – Position (x, y), scaling (s), and rotation (θ). The transformation is accomplished by using a 2D vector algebra equation.

The dataset preparation is carried out in two steps. The task space trajectory is generated using the transformation metric in the first phase and the generation of the joint space trajectory in the second phase. Four basic shapes are selected in the first phase - circle, square, triangle and star. The vector drawing of each shape is generated using linear algebra equations. Then interpolate the points in the shape and create a sequence in the task space of the robot canvas. Each point is transformed using the equation

$$\begin{bmatrix} x_o \\ y_o \\ z_o \\ s_o \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & x_{offset} \\ 0 & 1 & 0 & y_{offset} \\ 0 & 0 & 1 & z_{offset} \\ 0 & 0 & 0 & s_{canvas} \end{bmatrix} \begin{bmatrix} cos\theta & sin\theta & 0 & x \\ -sin\theta & cos\theta & 0 & y \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & s \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 0 \\ 1 \end{bmatrix} \quad (5)$$

where x, y, θ and s represent independent factors of variations, x_i, y_i represents the position points in the generated trajectory.

The x_{offset}, y_{offset}, z_{offset} and s_{canvas} projects the points into the robot workspace and the points on the robot task space obtained as $(x_o/s_o, y_o/s_o, z_o/s_o)$.

The values represent the 2D representation of the task as plotted on a canvas, as shown in Fig. 4d. Then the values are translated into the robot workspace, which will be the 3D representation shown in Fig. 4c. The robot trajectory points in Robot visualization (Rviz) and gazebo simulator are shown in Fig. 4a and Fig. 4b.

All possible combinations of the factors of variations are generated, as shown in Table 2. The total configurations are estimated as

$$C_{total} = \prod_{i=1}^{5} factor_i \quad (6)$$

The Robot Trajectory generation uses Cartesian planners available in the Moveit planning library [18]. The generated trajectory is post-processed to remove outliers, and the dataset is created for the training.

Each configuration $c_i \in C_{total}$ is taken and generated, the task space path using a 2D vector transformation block with equation 5. The robot Trajectory generation module plans the trajectory and appends it to a dataset. The dataset is normalized based on the corresponding joint limits in the post-processing phase.

The generated task is stored with the corresponding factors and metadata in the database. The data filtering and outlier removal were done by using Python scripts.

4.2 ROBOTICS METRICS

Accuracy and repeatability are the two most common evaluation metrics for robot performance defined in ISO 9283:1998 [19]. Each is calculated by generating n data points. Path position accuracy and repeatability are computed as the maximum pose position accuracy and repeatability value.

4.3 DISENTANGLEMENT METRICS

Each independent element in the labelled data is called a factor, and the varying independent variables in the latent space are called codes. There is no proper way to measure true disentanglement, completeness and informativeness, but the literature suggests many metrics to rely on. βVAE score is one of the first metrics to evaluate the disentanglement's performance, also called the z-min variance score. Later many methods were suggested with different advantages.

βVAE [6] and FactorVAE [20] are some of the initial disentanglement representation methods which measure the variance in codes. Mutual information Gap (MIG) [5] is used to evaluate the disentanglement by using the mutual information between the true underlying factors and generated codes, which uses the difference between the most prominent two variations. *jemmig* is introduced in [21], which measures the modified *MIG score*, including all the factors of variation instead of top2 as in *MIG score*.

Later disentanglement is represented using three terms disentanglement, completeness and informativeness (DCI) [22]. DCI run k linear repressor and evaluates the metrics. Disentanglement represents the amount of disentangling the underlying data variations. Completeness measures the amount of data a single variable captures, and informativeness represents how informative the latent vector is. Disentanglement library functions are used for evaluating the metrics [23].

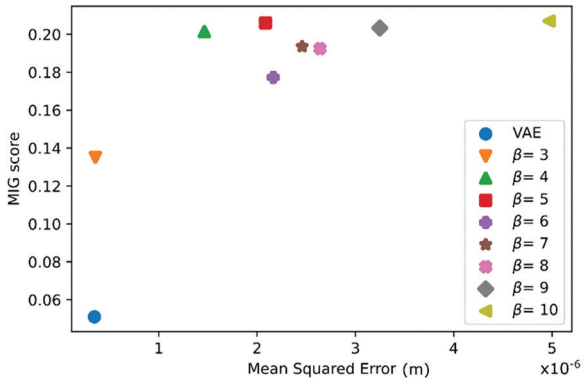


Fig. 5. The disentanglement score (MIG) vs. reconstruction error (MSE) plot for different models.

5. RESULTS AND DISCUSSIONS

In order to evaluate the performance, each model is trained for 100 epochs with a batch size of 128. The resulting z (codes) is evaluated against the factors in the dataset and different metrics calculated for measuring various disentanglement metrics. Fig. 5 represents the reconstruction loss (mean squared error) with a MIG disentanglement score of VAE and βVAE models with different beta values. βVAE models provide better disentanglement by compromising reconstruction quality. The work aims to find the trade-off between reconstruction loss and disentanglement. The VAE model achieves a minimum reconstruction metric (3.4x10-5 m), with a better reconstruction loss but poor disengagement (MIG score = 0.015). $\beta VAE_{(\beta=5)}$ model has a better disentanglement score of 0.206 and a reconstruction error of 2x10^{-4} m. The figure shows that the $\beta VAE_{\beta=10}$ model has a slightly greater MIG score (0.207) than the $\beta VAE_{\beta=5}$ model, but has a higher reconstruction error of 4.9x10^{-4} m. For the performance evaluations, models in VAE, $\beta VAE_{\beta=5}$ and $\beta VAE_{\beta=10}$ are considered.

The evolution of disentanglement metric and reconstruction error over epoch are shown in Fig. 6. MIG metric shows a massive improvement over traditional VAE models but will affect the reconstruction performance. As MIG is not directly linked with the loss function, it does not monotonically increase. Fig. 7 shows the reconstruction performance of models, and the corresponding evaluation metrics are shown in Table 3.

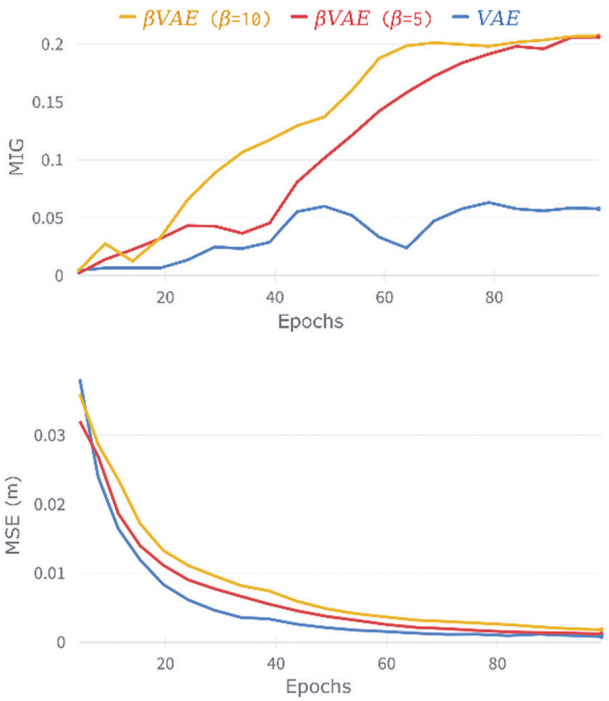


Fig. 6. The evolution of MIG score metric and reconstruction loss (MSE) for different models

Fig. 7. The reconstruction performance of different models - input, VAE, βVAE (β=5), βVAE (β=10).

In order to evaluate the repeatability, input and reconstructed images are plotted in Fig. 8 in the first and second rows, respectively. Joint space reconstruction error in VAE, $\beta VAE_{\beta=5}$ and $\beta VAE_{\beta=10}$ are 4.4 x 10^{-4} ± 1.7 x 10^{-4} m, 5.4 x 10^{-4} ± 4.9 x 10^{-5} m and 3.5 x 10^{-3} ± 2.1 x 10^{-4} m respectively.

And the corresponding robot task space loss (computed by applying forward kinematics) is 8.9 x 10^{-4} ± 2.6 x 10^{-4} m, 3.3 x 10^{-4} ± 5.7 x 10^{-5} m and 2.3 x 10^{-3} ± 2.1 x 10^{-4} m respectively. The joint space loss is less in the case of VAE model, but the tasks space loss is lower in $\beta VAE_{\beta=5}$ model. It is because of the non-linear forward kinematics conversions. The latent dimensional transversal representation of the best-performing model is depicted in Fig. 8, rows 3-12.

Each row shows the transversal of only one code, and the decoded the plot is shown. Fig. 8a, Fig. 8b and Fig. 8c show the transversal in VAE, $\beta VAE_{\beta=5}$ and $\beta VAE_{\beta=10}$ models respectively with Gaussian reconstruction loss. While considering the VAE model in Fig. 8, the 5th and 6th rows show y position transversal and the 5th and 9th rows show x position transversal. Rows 10th and 11th produce similar instances over the changes in the corresponding code.

(a) $\boldsymbol{VAE}$ (b) $\boldsymbol{\beta VAE_{\beta=5}}$ (c) $\boldsymbol{\beta VAE_{\beta=10}}$

Fig. 8. The first row shows the input image, and the second row shows the corresponding reconstructed images for computing the repeatability of the model. Rows 3-12 show the latent transversal performance of the model. The latent transversal performance of each model, each row shows the code, and columns show the transverse in that particular code with all other codes kept constant. The fifth column in each figure shows the reconstructed instance and their transversals generated based on this.

Other rows produce some noise outputs. The $\beta VAE_{\beta=5}$ in Fig. 8b shows x position, y position and orientation transversal in the 5th, 6th and 7th rows, respectively. Scaling is embedded in code in the 3rd and 7th rows. Rows 4th and 8th produce shape transversal, and code variation in 9th -12th rows does not produce much difference. The $\beta VAE_{\beta=10}$ in Fig. 8c produces position x, y and scale transversals in the 5th, 6th and 4th rows, respectively. The 3rd and 8th rows encode code for shape, and rows 7, 9-12 do not produce a visible output difference.

While analyzing the variations, the 5th row in the VAE model changes x and y positions and not all factors of variations are not encoded, while $\beta VAE_{\beta=5}$ and $\beta VAE_{\beta=10}$ encodes the codes and has achieved a higher disentanglement score. It can be observed that the reconstruction performance is quite prominent in lower β values.

Generative models produce samples from a sample distribution, so each time it generates a new sample, it belongs to the same distribution, but some variations exist. It is the property of VAE which causes the variation in standard deviation. Table 3 shows the precision of different models. Robot tasks need to be precise and accurate. The higher value of standard deviation in Reconstruction loss, accuracy and repeatability are due to the generative nature of the model. Accuracy and repeatability are calculated in task space and the Reconstruction loss in the joint space of the robot. The non-linear kinematics operation produces variations between the reconstruction values and the robotics metrics. The model is executed 100 times and generates the plot sown in Fig. 9.

Table 3. Metrics considered.
(↑ Means higher is better).

Model	VAE	βVAE (β=5)	βVAE (β=10)
Rec_loss ↓(x10^{-3} m)	0.346 ± 0.345	2.085 ± 2.714	4.970 ± 7.019
Accuracy ↓ (m)	0.053 ± 0.013	0.055 ± 0.039	0.080 ± 0.065
Repeatability ↓ (m)	0.017 ± 0.002	0.030 ± 0.007	0.039 ± 0.010
MIG [14] ↑	0.051	0.206	0.207
Disentanglement [37] ↑	0.5	0.914	0.786
Completeness [37] ↑	0.128	0.216	0.302
Informativeness [37] ↑	0.122	0.213	0.21
jemmig [36] ↑	0.192	0.297	0.292
βVAE score [15] ↑	0.546	0.617	0.612
FactorVAE [16] ↑	0.611	0.647	0.759

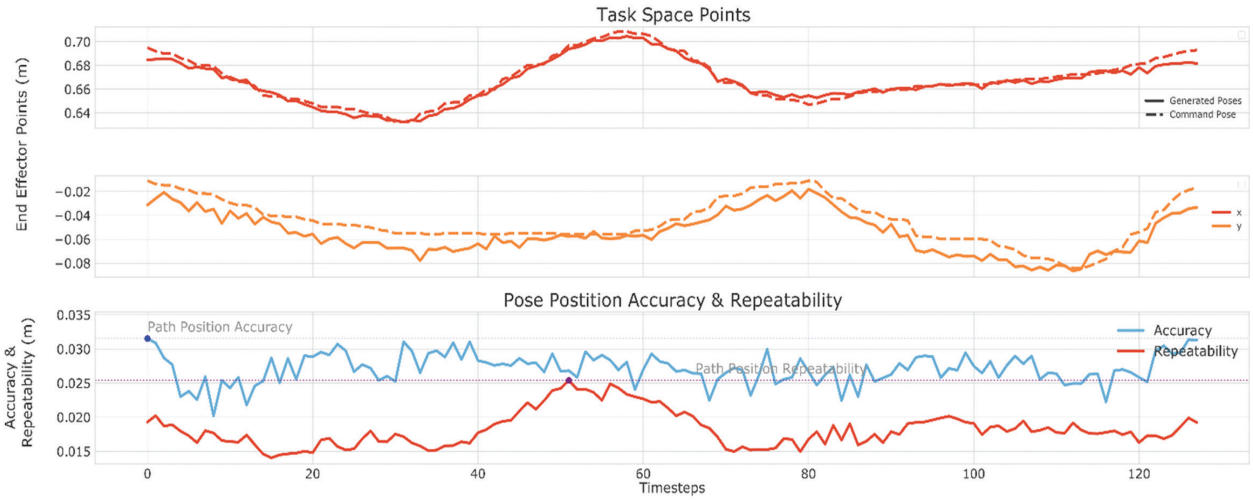


Fig. 9. $\beta VAE_{\beta=5}$ model task space input and output representation in x and y axis is represented (top) and its corresponding accuracy and repeatability is plotted (bottom)

The performance analysis of different disentanglement representations, losses and robot precision are listed in Table 3. Joint space reconstruction loss is lower in the VAE model. As the β value increases, reconstruction performance decreases. However, there is an allowed limit for each task's precision and accuracy range. Optimization of β based on task nature and disentanglement required can be achieved by hyperparameter tuning. Literature shows that the βVAE and FactorVAE scores do not provide practically feasible metrics. This work uses the MIG score as the primary metric for evaluating disentanglement.

6. CONCLUSION

In this work, CNN-based Variational Autoencoder models have been utilized for disentanglement representation of robot tasks. All the models have been trained on the robot disentanglement dataset proposed. Popular disentanglement metrics such as MIG score, DCI, jemmig, VAE and FactorVAE scores are used to evaluate the model performance as well as robot accuracy and precision metrics. From various disentanglement metrics, it has been found that the underlying factors of variation in tasks learn better with disentanglement losses. The model has been found to generate disentanglement representation with a path position accuracy of 0.055, close to the VAE model (0.053) and better disentanglement of 0.206, which is far better than the state-of-the-art VAE model.

The disentanglement generative models can be used as a supervised data generator for training deep learning models and can be directly used in robot task generation applications (e.g. Painting tasks).

7. ACKNOWLEDGEMENTS

The authors would like to acknowledge University Grants Commission (UGC), India for providing financial assistance to carry out this research work.

8. REFERENCES

[1] H. Demir, F. Sari, M. R. Tolun, "Review for path planning for robots and its applications", ACADEMIC STUDIES, 1st Edition, 2019, pp. 196-211.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial networks", Proceedings of the Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 8-13 December 2014, pp. 2672-2680.

[3] D. P. Kingma, W. Max, "Auto-encoding variational bayes", arXiv:1312.6114, 2013.

[4] D. J. Rezende, S. Mohamed, D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models", Proceedings of the 31st International Conference on Machine Learning, Bejing, China, 22-24 June 2014, pp. 1278-1286.

[5] R. T. Chen, X. Li, R. Grosse, D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders", arXiv:1802.04942, 2013.

[6] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, A. Lerchner, "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework", Proceedings of the 5th International Conference on Learning Representations, Toulon, France, April 2017, pp. 24-26.

[7] Y. Hristov, D. Angelov, M. Burke, A. Lascarides, S. Ramamoorthy, "Disentangled Relational Rep-

resentations for Explaining and Learning from Demonstration", Proceedings of the Conference on Robot Learning, Osaka, Japan, 30 October - 1 November 2019, pp. 870-884.

[8] V. A. Kumar, T. Rajan, A. Nagendran, A. Dehghani-Sanij, R. C. Richardson, "Tether monitoring for entanglement detection, disentanglement and localisation of autonomous robots", Robotica, Vol. 34, No. 3, 2016, pp. 527-548.

[9] M. Wulfmeier, A. Byravan, T. Hertweck, I. Higgins, A. Gupta, T. Kulkarni, M. Reynolds, D. Teplyashin, R. Hafner, T. Lampe, M. Riedmiller, "Representation matters: Improving perception and exploration for robotics", Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May - 5 June 2021, pp. 6512-6519.

[10] C. Qin, Y. Zhang, Y. Liu, S. Coleman, D. Kerr, G. Lv, "Appearance invariant place recognition by adversarially learning disentangled representation", Robotics and Autonomous Systems, Vol. 131, No. 9, 2020, p. 103561.

[11] J. Pajarinen, O. Arenz, J. Peters, G. Neumann, "Probabilistic approach to physical object disentangling", IEEE Robotics and Automation Letters, Vol. 5, No. 4, 2020, pp. 5510-5517.

[12] M. Zolotas, Y. Demiris, "Disentangled sequence clustering for human intention inference", arXiv:2101.09500.

[13] M. T. Das, L. C. Dülger, "Mathematical modelling, simulation and experimental verification of a SCARA robot", Simulation Modelling Practice and Theory, Vol. 13, No. 3, 2005, pp. 257-271.

[14] J. Denavit, R. S. Hartenberg., "A kinematic notation for lower-pair mechanisms based on matrices", Journal of Applied Mechanics, Vol. 22, No. 2, 1955, pp. 215-221.

[15] M. Quigley, et al. "ROS: an open-source Robot Operating System", Proceedings of the ICRA workshop on open source software, Kobe, Japan, 12-17 May 2009, p. 5.

[16] P. Corke, J. Haviland, "Not your grandmother's toolbox–the robotics toolbox reinvented for python", Proceedings of the IEEE International Conference on Robotics and Automation, Xi'an, China, 30 May -5 June 2021, pp. 11357-11363.

[17] L. Matthey, I. Higgins, D. Hassabis, A. Lerchner, dsprites: "Disentanglement testing sprites dataset", https://github.com/deepmind/dsprites-dataset/ (accessed: 2021)

[18] S. Chitta, I. Sucan, S. Cousins, "Moveit![ros topics]", IEEE Robotics & Automation Magazine, Vol. 19, No. 1, 2012, pp. 18-19.

[19] ISO 9283:1998(en), "Manipulating industrial robots-Performance criteria and related test methods", International Organization for Standardization, Geneva, Switzerland, Technical Report, 1998.

[20] H. Kim, A. Mnih, "Disentangling by factorising", Proceedings of the 35th International Conference on Machine Learning, Stockholm Sweden, 10-15 June 2018, pp. 2649-2658.

[21] K. Do, T. Tran, "Theory and evaluation metrics for learning disentangled representations", Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26 April - 1 May, 2020.

[22] C. Eastwood, C. K. I. Williams, "A framework for the quantitative evaluation of disentangled representations", Proceedings of the International Conference on Learning Representations, Vancouver, Canada, 30 April - 3 May 2018.

[23] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations", Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 10-15 June 2019, pp. 4114-412.

Using Attribute-Based Access Control, Efficient Data Access in the Cloud with Authorized Search

Original Scientific Paper

K. S. Saraswathy
Manonmaniam Sundaranar university,
Department of Computer Science,
Abishekapatti, Tirunelveli – 627012.
789saraswathy@gmail.com

S. S. Sujatha
Manonmaniam Sundarnar university,
S.T.Hindu College,
Abishekapatti, Tirunelveli – 627012.

Abstract *– The security and privacy issues regarding outsourcing data have risen significantly as cloud computing has grown in demand. Consequently, since data management has been delegated to an untrusted cloud server in the data outsourcing phase, data access control has been identified as a major problem in cloud storage systems. To overcome this problem, in this paper, the access control of cloud storage using an Attribute-Based Access Control (ABAC) approach is utilized. First, the data must be stored in the cloud and security must be strong for the user to access the data. This model takes into consideration some of the attributes of the cloud data stored in the authentication process that the database uses to maintain data around the recorded collections with the user's saved keys. The clusters have registry message permission codes, usernames, and group names, each with its own set of benefits. In advance, the data should be encrypted and transferred to the service provider as it establishes that the data is still secure. But in some cases, the supplier's security measures are disrupting. This result analysis the various parameters such as encryption time, decryption time, key generation time, and also time consumption. In cloud storage, the access control may verify the various existing method such as Ciphertext Policy Attribute-Based Encryption (CP-ABE) and Nth Truncated Ring Units (NTRU). The encryption time is 15% decreased by NTRU and 31% reduced by CP-ABE. The decryption time of the proposed method is 7.64% and 14% reduced by the existing method.*

Keywords*: Cloud computing, data access control, Nth Truncated Ring Units, Ciphertext Policy Attribute-Based Encryption, database authentication.*

1. INTRODUCTION

The term "cloud computing" represents the supply of computational services on demand, primarily the collection of information and processing capacity [1]. This concept is commonly used to identify data centers that are accessible to multiple users on the Internet, without the user actively managing them [2]. Data is being transferred by an increasing number of businesses and individuals, personal data, and vast archive systems to cloud-based storage services because they provide a range of attractive services, such as limitless space, straightforward costs, and longstanding services. [3]. Consumers can also access applications and services without location limitations. However, according to several recent reports, 88% of cloud users are disturbed by the confidentiality of their information, and protection is frequently cited as the primary reason for using cloud-based storage solutions [4]. Cloud computing is a knowledge that allows Cloud storage service providers (CSP) to offer applications, calculate, and collection of information to customers located all over the world [5].

It has lately piqued the attention of both IT firms and academic organizations. In cloud computing, there are three major service delivery models: (PaaS) Platform as a Service, (IaaS) Infrastructure as a Service, and (SaaS) Software as a Service [6]. Private cloud, public cloud, community cloud, and hybrid cloud are the four types of cloud. Agility, flexibility, scalability, pay-per-use, and resiliency are only a few of the benefits of cloud computing. Scaling, low information technology costs, reliability, market stability, and almost unlimited efficiency are the advantages of cloud computing [7]. It has two

major data protection and access control issues, with data security weakening when reviewing its web-based services [8]. An access management model is a method for a user to gain access to data stored on cloud servers [9]. With the exponential development of big data technology and cloud computing, a growing number of enterprises and organizations have opted to automate their information to the server [10]. The majority of cloud data, like confidential medical history and business internal data, are extremely vulnerable [11].

The information would be maintained on the public cloud throughout the context of ciphertext in particular to provide data confidentiality and user privacy [12]. The encryption technique can be thought of as a protection assurance for gaining data access control. However, controlling access to encoded information is a significant problem [13]. Through the increasing adoption of cloud computing, increasingly consumers are opting to offload both the high responsibility for data processing and the complexity of computing to the public cloud [14]. About the benefits of cloud storage, secure information access management maintains among the most challenging obstacles, since the private cloud is not completely accepted via the data owner, and data collected in the cloud may contain sensitive data [15]. As a consequence, since distributing information to the server, the data owner should encrypt the message to preserve the safety of the customer and maintain secure communications. Here, Attribute-based access control is utilized to access the data in the cloud storage [16]. The remaining part of the paper contains section 2 explains the related work in various techniques and problems, and section 3 provides the proposed methodology and the step-by-step procedure of ABAC. Section 4 explains the result part and section 5 contains the conclusion parts.

2. LITERATURE REVIEW

There has been a lot of research on the different access controls in cloud storage. This section includes a discussion of the relevant work on access control.

A well-organized EACAS (attribute-based access control with an authorized search scheme) has been established by Jialu Hao *et al(2019)* [17] for the cloud storage access control. In the intended strategy, EACAS enables data users to customize search strategy with a focus on their data access and accumulate the respective trapdoor by using a private key conferred by the cloud provider to extract their valuable research by incorporating the key delegation methodology into AKP-ABE. But the limitation includes further modulation of the proposed methodology with supple exchanges of information with confined retaining of data in the cloud.

In 2019, Wang, S., et al, [18] analyzed a secure cloud storage framework. In this article, Ethereum blockchain architecture was used to construct a modern secure cloud storage framework including authentication, which was a mixture of Ethereum blockchain and CP-ABE. There was no trustworthy third party in the cloud computing system because it was decentralized. It has three features: it was built using Ethereum blockchain technologies, the storage operator can establish legitimate information usage times, and it can be preserved in the blockchain.

In 2018, Xu, Q., et al. defined that In a multi-authority cloud storage system, PMDAC-ABSC is a privacy-preserving shared data management mechanism based on Ciphertext-Policy ABSC that offers fine-grained control mechanisms and attributes privacy security at the same time [19]. The overhead decryption for users has been substantially reduced via outsourcing the unnecessary bilinear pairing to the cloud server without damaging the privacy of the attributes. The standard model is robust and can include anonymity, unforgeability, confidential authentication, and public verifiability. Their architecture would match protection goals towards practical computational efficiency, as demonstrated by the protection strategy asymptotic complexity comparison and execution outcomes.

In 2017, Liu, H., et. al,[20] implemented a logical secret sharing reward exchange mechanism, and a fair information access control system for data storage. The scheme produces a huge amount of fake keys. When a consumer deviates from the specified scheme during a share exchange, he or she must first send his or her shares. This discourages users from being narcissistic and encourages them to use the shared data as a community. According to mathematical research, the suggested scheme's Nash equilibrium is that both users still give their shares, enabling them to reconstruct the decoding key fairly. Furthermore, extensive research shows that the proposal will successfully control access control policies.

H-KCABE in data storage with fine-grained access control was developed by Sangeetha, M., et al [21]. In the HABE model, they propose an H-KCABE encryption algorithm with a few minor changes to improve performance through the re-encryption process. The HABE model helps the users to access information hierarchically through generating traffic, and the KCABE methodology improves efficiency by decreasing data transmission time in a fine-grained authentication method. They can easily improve efficiency by reducing time with the KCABE algorithm then the HABE model, which allows them to access information in a hierarchical manner without creating any traffic between users.

A new approach that resolves the essential encryption issue while also allowing for quick user voiding retraction has been employed by Zhihua, Liangao, and Dandan (2016) [22]. First, an access regulator is added to the current strategy, and so the attribute authority and authorization controllers create encryption data on a corporate level. Second, a version key that enables forward and reversible security is used to provide a convenient revocation process. The proposed method is simple and reliable in terms of user authorization and

revocation, according to the assessment. But lack of accuracy in terms of encryption of cloud storage

Saravanan, N., and Umamakeswari, D. A. [23] suggested a layered method to protecting client information that includes lattice-based encryption strategies It has been shown that by combining an access management architecture with a double authentication strategy, cloud data can be better protected. Users will be able to store their vast quantities of personal data in the cloud without fear of security threats thanks to this strong protection technique. The RSA and AES algorithms prevent the operator from guessing the key and encrypted text. Intruders' intelligence was almost irrelevant in terms of the hybrid paradigm. Bell and LaPadula (BLP) and lattice versions add user-level authentication as well.

In 2020 Challagidad, P. S., & Birje, M. N. [24] proposed an effective multi-authority intrusion detection system that enables efficient, fine-grained user authentication utilizing an attribute-based encryption scheme. For information storage anonymity, multi-authority access management, and fine-grained accessibility to encrypted information, the scheme uses HAS algorithm and a single RHA. The (RHA) Role Hierarchy Algorithm separates cloud users into groups depending on their assigned attributes. The (HAS) Hierarchy Access Structure assists in determining the authorization process for fine-grained and multi-authority cloud resource access management. In comparison to current works, analysis findings indicate that the RHA, HAS, was successful. Because more information is deposited on the cloud computing server, the scheme's advantages are growing increasingly obvious.

In cloud services, revocable server identity-based encryption for secure shard data was developed by Vurukonda, N., et al [25]. This paper explains revocable storage Identity-Based Encryption, a device that manages authenticated text back-and-forth authentication through disabled user revocation and software maintenance authentication functionality. Furthermore, the revocable storage IBE was compared to previous IBE approaches, demonstrating the reliability and sufficiency of the enables.

In 2019 Prabhu kavin, B., & Ganapathy, S. [26] proposed the latest data management method built on the Chinese Remainder Theorem (CRT) for safely processing user data in a cloud database. In addition, CRT was used to build a new community key management scheme for accessing encrypted data from the cloud database. In CRT-based secure processing systems, two encryption techniques were introduced using new methods for first and second authentication, as well as the formula for data storage authentication. In comparison, during the group key generation process formula for obtaining authenticated cloud data from a database server on a cloud server was introduced. By evaluating the experimental effects, the safeguards models' performance level has been assessed. Finally, the data protection model is superior to other current models.

3. PROPOSED METHOD

This section describes the access control in cloud storage using attribute-based access control. First, the data must be stored in the cloud and security must be strong for the user to access the data. This paradigm takes into account some of the features of cloud data seen in the database's security process for storing information about registered groups and the user's stored keys. Clusters, the registry's message identification code, and usernames and party names, each with their package of benefits. Initially, information must be secured and delivered to the service provider; this indicates that the data is protected if the supplier's security procedures are disrupted in some cases. The overall diagram of the design is given below.

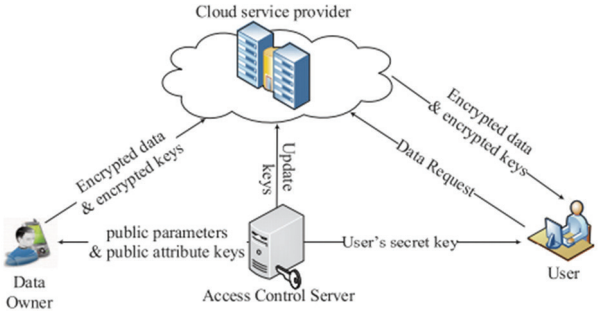


Fig. 1. Overall diagram of the proposed method

Data owner (DO): After transferring information to the server, DO must be authenticated with data based on its features, which expands user access to information based on their computer locations or passwords rather than data characteristics. DO has total trust in our method and is in charge of main development.

Data user (DU): The user is allowed to decode ciphertexts whose characteristics comply with DU's information system. It can also set a tighter search policy than his entry policy, and only his hidden key can be used to create the trapdoor. DU uses the cloud server's trapdoor to request the relevant data to extract the ciphertext that matches the search strategy. It is untrustworthy, and they can band together to procure data information outside their access rights. They're still curious about the data's attribute detail.

Cloud server (CS): CS is believed to have a lot of storage and processing power and is still available to help. The CS contains two parts: the (CSS) cloud storage server and the (DSS) delegated search server, with CSS supporting, DO in storing their information and DSS conducting data searches on behalf of DU and returning the related data to Data user. Cloud server is semi-honest, which ensures it would diligently comply with DO and DU's demands, but it is interested in data details, such as data content and attribute privacy. The Additional Private Key DSS is used to ensure that those without a private key are unable to guess the attribute values in the dropout by guessing offline.

A following objectives should be contacted when managing access to cloud storage.

Fine-grained access control: An information stored in the CSS is authenticated using its attributes, which can be decrypted via Data user if the ciphertext attributes obey the access policy. The access control should be built into the decoding mechanism rather than being handled by CS. Consequently, any threshold gate with an articulate information system should be enabled to ensure fine-grained network access.

Flexible and authorized search: DU must be allowed to obtain the information ciphertext whose attributes fulfill the selection policy using DSS. DU, on the other hand, can only scan the information inside the limits of his security authorization which ensures it must be allowed to provide a trapdoor with an exploration strategy that is more stringent than his information system. At the same time, the selection strategy must be expressed in a way that allows for an agile search.

Attribute privacy preservation: In ciphertext and trapdoor, the default attribute name is visible, but the associated component attributes should be concealed to secure sensitive information and privacy protection. Attribute values found in the ciphertext cannot be deduced by an attacker. Furthermore, any attackers who do not have the DSS private key are not exposed to attribute values in the search policy by the trapdoor.

Practical implementation: For functional implementations, device processes can be performed with lower computing and processing expenses.

3.1 OVERVIEW OF ABAC

ABAC is used to describe descriptive security policies for the DU and to explicitly encrypt attribute values in ciphertext which allows for better and more privately controlled access to outsourced data. The secret attribute knowledge, on the other hand, makes data search a difficult issue. ABAC's key delegates adopt a strategy that allows the DU to identify a more stringent search policy than the access system and use encryption data to create the next dropout to solve the issue. Figure 2 represents the overview of the ABAC method.

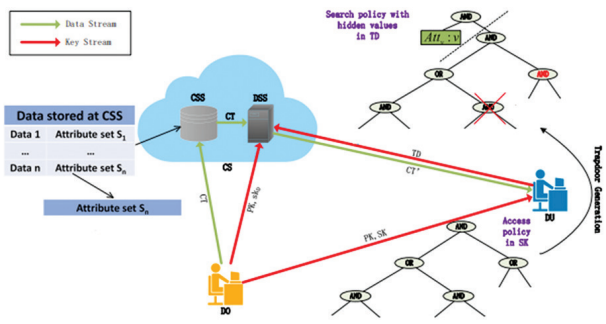


Fig. 2. Comprehensive operating procedures of ABAC

To protect the attribute information, the attribute values in the trapdoor are also concealed. A synthetic attribute on both the ciphertext then the trapdoor prevents DSS from accessing the data content. In detail, the ciphertext is made up of two sections: (1) The actual attribute set is used to encrypt the initial data; (2) a meaningless data "1" encoded with the synthetic attribute inserted through the original attribute set. The virtual characteristic is connected to the search tree root node with an AND gate while in the trapdoor, allowing it a prerequisite for successful matching. DSS will decrypt the ciphertext which descriptor set supports the search policy by deciding whether the trivial data "1" can be obtained by checking the ciphertext, but it cannot decrypt the ciphertext of the original data that is encrypted without the virtual attribute. As a consequence, information security will achieve fine-grained access control with an agreed-upon search on information outsourced to the cloud while maintaining data integrity and attribute privacy.

3.2 Step by step procedure of ABAC2

ABAC consists of six phases: data encryption, system setup, data decryption key generation, trapdoor generation, and data search.

3.2.1 System Setup

To produce *PK* (Public key) and *MSK (master secret key)*, *DO* choose a security limitation ξ and call the Setup (ξ) algorithm. The Setup algorithm is similar to ABE, with the exception that the public key includes a virtual attribute V_a containing the value *v* which is the value of real attributes, and additio nal public and private key pair (pk_D, sk_D) for DSS is created as $pk_D = g^y$, and $sk_D = \gamma$, where γ is a random value in Z^*_p. The system's public key is then made available as,

$$PK = \langle g, u, h, w, e(g,g)^{\alpha}, g_1, g_2, g_3, g_4, [V_a : v], pk_D{}^1 \rangle \quad (1)$$

DO maintains the machine master secret key as $MKS = \langle \alpha, \tau_1, \tau_2, \tau_3, \tau_4 \rangle$. DO also passes the private key $sk_D = \gamma$ to the DSS.

3.2.2 Key generation

The *DSS* public key pk_D is used in the machine public key *PK* for convenience. *DO* creates an access policy *AP* for *DU* based on his position and distributes the hidden key $SK = \langle AP, \{D_x, D_{x,0}, D_{x,1}, D_{x,2}, D_{x,3}, D_{x,4}\}_{n_x \in atts(\tau)} \rangle$ created by the *KeyGen(PK, MSK, AP)* algorithm to *DU* when he enters the framework. The *ABE* and *KeyGen* algorithms are the same to generate keys for public and private keys. *DU* will decode the ciphertext whose attribute collection satisfies *AP* using *KeyGen* and the secret key *SK*

Data Encryption

DO creates a characteristic set *S* based on the information specifications before transferring the data *M* to *CS*, and then uses the *Encrypt(PK, M, S).* process to generate the ciphertext *CT*. *DO* computes $E=M.e(g,g)^{as}$, $E=g^s$,

$E' = g^{s'}$, two random values s and s'. Then take specific datatypes $s(x,1)$, $s(x,2)$, and $s_{x,1}, s_{x,2}, z_x$ from z^*_p for each component in S are chosen and computed.

$$E_{x,0} = w^{-s}(u^{sx}h)^{zx}, E'_{x,0} = w^{-s'}(u^{sz}h)^{zx} \quad (2)$$

$$E_{x,1} = g_1{}^{zx-8x,1}, E_{x,2} = g_2{}^{sx,1}, \quad (3)$$

$$E_{x,3} = g_3{}^{zx-8x,2}, E_{x,4} = g_4{}^{s_{x,2}} \quad (4)$$

DO selects a random value $r_v \in Z^*_p$ for the virtual attribute v_a and quantifies $E_{v,0} = w^{s'}(u^v h)rv$, $E_{v,1} = g^{rv}$. Lastly, the ciphertext that will be uploaded to the cloud is developed as follows:

$$CT = \langle N_S, \tilde{E}, \widetilde{E'}, E, E', E_{v,0}, E_{v,1}, \{E_{x,0}, E'_{x,1}, E_{x,1}, E_{x,2}, E_{x,3}, E_{x,4}\}_{x \in S} \rangle \quad (5)$$

3.2.3 Trapdoor generation

DU establishes a *SP* development scheme based on user access policy, in which search policy *(SP)* will have the same expression style as access policy *(AP)*, the search tree architecture in search policy is extremely strict than the request tree architecture in *AP*, and the meaning of the element related to the attribute name cannot be altered. Then, using his hidden key *SK* identified through the data access access policy, *DU* uses the *TrapGen(PK, SK, SP)* algorithm to produce the trapdoor TD identified with the search policy search policy. The *TrapGen* algorithm uses the key delegation method, in which a sequence of simple operations is carried out to transform the hidden key *SK* for the efficient key access policy to the trapdoor for the search policy search policy. The *TrapGen* algorithm includes the corresponding three steps in particular. The first step is to manipulate the current gates to convert the actual private key to a different encryption key, the second option is to prevent DSS from decoding the information to the encrypted message via adding an AND gate to the root node, then the final step is to protect the related data in the trapdoor against disconnected manipulation attacks by attackers who do not have access to the DSS private key sk_D.

3.2.3 Data decryption

DU uses the *Descrypt (CT', SK)* algorithm to retrieve the received information since obtaining the encrypted message from DSS. With increasing attribute name $n_x \in \widetilde{\mathcal{N}_l}$ in the Decrypt algorithm computes,

$$P_x = e(E, D_x)e(E_{x,0}, D_{x,0}).(Q_x)^{1/\delta_y} = e(g,g)^{px(0)s} \quad (6)$$

In the *ABE.Decrypt* algorithm, the term $(g,g)^{\alpha s}$ can be improved and M can be determined concluded $E/_{e(g,g)^{\alpha s}}$ The encrypting data the product supplier is assigned to assures that the data is still secure in case the supplier's security measures are violated. After that, the data can be accessed by the user using an encrypted key.

4. RESULT AND DISCUSSION

In this section, the access control in cloud storage is analyzed using ABAC. The proposed methodology is applied in the JAVA programming language with JDK 1.7.0. This proposed concept is mainly used in the health care system. The experimental used datasets are collected from different sources.

4.1 COMPARATIVE ANALYSIS

A proposed method is analyzed via several methods like key generation, encryption time, time consumption and decryption time. A current method is investigated against the existing methods are NTRU and CP-ABE. The proposed method comparative analysis against the existing technique is given below,

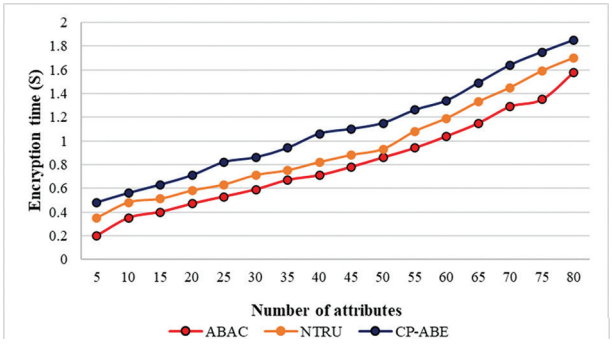


Fig. 3. Comparative analysis of encryption time

In Figure 3, represents the comparative analysis of ABAC, NTRU, and CP-ABE with encryption time and several attributes. The encryption time is the amount of time it takes for an encryption algorithm to generate a ciphertext from plaintext and it is used to measure the performance of the encryption scheme. Each attribute in the encryption process (ABAC, NTRU, and CP-ABE) can begin at the same attribute value, but they can vary by changing the values of the time representation. The encryption time may increase which indicates the speed of the encryption process. The encryption time of the proposed method is 15% decreased by the existing method NTRU and 31% of the encryption time is reduced by the existing method CP-ABE. The graphical representation of the key generation is given below,

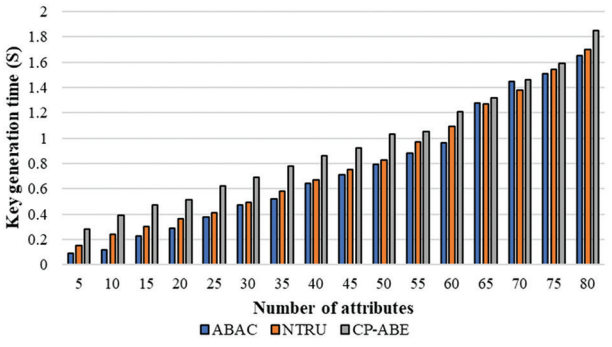


Fig. 4. Comparative analysis for Key Generation Time

In Figure 4, represents the comparative analysis of ABAC, NTRU, and CP-ABE with key generation time and a number of the attribute. The key generation time can also increase the variance of attributes, the starting stage of attribute values may same but the variations of keys may differ. The ABAC is lower compared to other graphical representations, NTRU may be slightly higher compared to ABAC and CP-ABE are higher values in key generation performance. The proposed method key generation is 6.16% reduced by the existing method NTRU and 22% decreased by CP-ABE. The decryption time graph is given below,

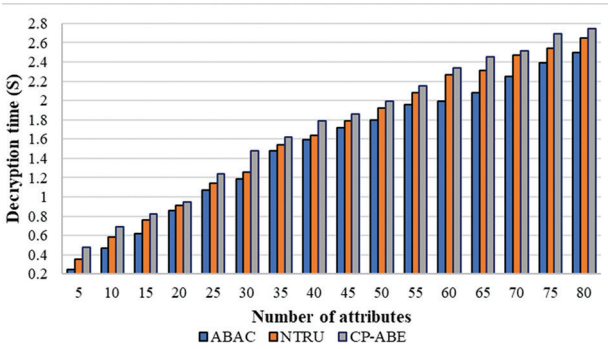


Fig. 5. Comparative Analysis for decryption time

In figure 5, represents the comparative analysis of ABAC, NTRU, and CP-ABE with decryption time and number of the attribute. In a decryption time process, the ABAC attributes may be very less compared to other attribute representations and the other attribute value may increase step by step. The decryption time of the proposed method is 7.64% and 14% reduced by the existing method. The graphical representation of time consumption is given below,

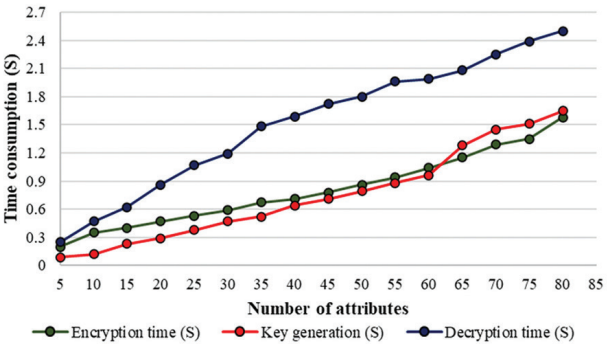


Fig. 6. Comparative analysis for time consumption with encryption time, key generation, decryption time

In figure 6, represents the comparison analysis of time consumption with encryption time, key generation, and decryption time. The encryption time may increase slightly throughout the key generation, the key generation may also increase but it can decrease towards the encryption process. The encryption time and key generation are mixed while increases neither decrease. The decryption time can increase highly by comparing the other two comparisons.

5. CONCLUSION

In this section, we have introduced the access control in cloud storage data using ABAC. First, the data can be stored in the cloud and security must be strong for the user to access the data. This model considers some of the characteristics of the cloud data contained in the authentication mechanism that the database uses to retain data around groups that have been registered, as well as the user's saved keys. User names and party names, as well as groups and the database message encryption method all, have unique benefits. Encrypting the data before sending it to the network operator means that, it remains encrypted despite the supplier's protection protocols being breached. The suggested method's experiment results are assessed utilizing a variety of metrics, including encryption time, decryption time, key generation time, and time usage. The encryption time of the proposed method is 15% decreased by the existing method NTRU and 31% of the encryption time is reduced by the existing method CP-ABE. The decryption time of the proposed method is 7.64% and 14% reduced by the existing method. The key generation of the proposed method is 6.16% reduced by the existing method NTRU and 22% decreased by CP-ABE. By comparing the time consumption, the key generation time is reduced.

6. REFERENCES:

[1] J. Li, W. Yao, Y. Zhang, H. Qian, J. Han, "Flexible and Fine-Grained Attribute-Based Data Storage in Cloud Computing", IEEE Transactions on Services Computing, Vol. 10, No. 5, 2017, pp. 785–796.

[2] J. Shi, J. Lai, Y. Li, R. H. Deng, J. Weng, "Authorized Keyword Search on Encrypted Data", Proceedings of the European Symposium on Research in Computer Security, Wroclaw, Poland, 7-11 September 2014, pp. 419–435.

[3] P. Jiang, Y. Mu, F. Guo, Q. Wen, "Public Key Encryption with Authorized Keyword Search", Proceedings of the Australasian Conference on Information Security and Privacy, 2016, pp. 170–186.

[4] H. Cui, Z. Wan, R. H. Deng, G. Wang, Y. Li, "Efficient and Expressive Keyword Search Over Encrypted Data in The Cloud", IEEE Transactions on Dependable and Secure Computing, Vol. 15, No. 3, 2016, pp. 409–422.

[5] H. Cheng, C. Rong, K. Hwang, W. Wang, Y. Li, "Secure Big Data Storage and Sharing Scheme For Cloud Tenants", China Communications, Vol. 12, No. 6, 2015, pp. 106–115.

[6] J. Baek, Q. H. Vu, J. K. Liu, X. Huang, Y. Xiang, "A Secure Cloud Computing-Based Framework for Big

Data Information Management of Smart Grid", IEEE Transactions on Cloud Computing, Vol. 3, No. 2, 2015, pp.233–244.

[7] G. Zhuo, Q. Jia, L. Guo, M. Li, P. Li, "Privacy-Preserving Verifiable Set Operation in Big Data for Cloud-Assisted Mobile Crowdsourcing", IEEE Internet of Things Journal, Vol. 4, No. 2, 2016, pp. 572–582

[8] L. Guo, Y. Fang, M. Li, P. Li, "Verifiable Privacy-Preserving Monitoring for Cloud-Assisted M-health Systems", Proceedings of the IEEE Conference on Computer Communications, 26 April - 1 May 2015, pp. 1026–1034.

[9] L.-Y. Yeh, P.-Y. Chiang, Y.-L. Tsai, J.-L. Huang, "Cloud-Based Fine Grained Health Information Access Control Framework for Lightweight IoT Devices with Dynamic Auditing and Attribute Revocation", IEEE Transactions on Cloud Computing, Vol. 6, No. 2, 2018, pp. 532–544.

[10] Z. Yan, X. Li, M. Wang, A. V. Vasilakos, "Flexible Data Access Control Based On Trust And Reputation In Cloud Computing", IEEE Transactions On Cloud Computing, Vol. 5, No. 3, 2017, pp. 485–498.

[11] K. Yang, K. Zhang, X. Jia, M. A. Hasan, X. Shen, "Privacy-Preserving Attribute-Keyword Based Data Publish-Subscribe Service on Cloud Platforms", Information Sciences, Vol. 387, 2017, pp. 116–131.

[12] J. Hao, C. Huang, J. Ni, H. Rong, M. Xian, X. Shen, "Fine-Grained Data Access Control with Attribute-Hiding Policy for Cloud-Based IoT", Computer Networks, Vol. 153, 2019, pp. 1-10.

[13] H. Cui, R. H. Deng, G. Wu, J. Lai, "An Efficient and Expressive Ciphertext-Policy Attribute-Based Encryption Scheme with Partially Hidden Access Structures", Proceedings of the International Conference on Provable Security, 2016, pp. 19–38.

[14] R. Fernando, R. Ranchal, B. An, L. Othmane, B. Bhargava, "Consumer Oriented Privacy Preserving Access Control of Electronic Health Records in The Cloud", Proceedings of the IEEE 9th International Conference on Cloud Computing, San Francisco, CA, USA, 27 June- 2 July 2016, pp. 608–615.

[15] R. Ranchal, B. Bhargava, R. Fernando, H. Lei, Z. Jin, "Privacy Preserving Access Control in Service-Oriented Architecture," Proceedings of the IEEE International Conference on Web Services, San Francisco, CA, USA, 27 June - 2 July 2016, pp. 412–419.

[16] Z. Wang, D. Huang, Y. Zhu, B. Li, C.-J. Chung, "Efficient Attribute-Based Comparable Data Access Control", IEEE Transactions on Computers, Vol. 64, No. 12, 2015, pp. 3430–3443.

[17] J. Hao, J. Liu, H. Wang, L. Liu, M. Xian, X. Shen, "Efficient Attribute-Based Access Control with Authorized Search in Cloud Storage", IEEE Access, Vol. 7, 2019, pp.182772–182783.

[18] S. Wang, X. Wang, Y. Zhang, "A Secure Cloud Storage Framework with Access Control based on Blockchain", IEEE Access, Vol. 7, 2019, pp. 112713–112725.

[19] Q. Xu, C. Tan, Z. Fan, W. Zhu, Y. Xiao, F. Cheng, "Secure Multi-Authority Data Access Control Scheme in Cloud Storage System Based on Attribute-Based Signcryption", IEEE Access, Vol. 6, 2018, pp. 34051–34074.

[20] H. Liu, X. Li, M. Xu, R. Mo, J. Ma, "A Fair Data Access Control Towards Rational Users In Cloud Storage", Information Sciences, Vol. 418, 2017, pp. 258–271.

[21] M. Sangeetha, P. Vijayakarthik, S. Dhanasekaran, B. S. Murugan, "Fine Grained Access Control Using H-KCABE in Cloud Storage", Materials Today: Proceedings, Vol. 37, 2021, pp. 2735–2737

[22] Z. Xia, L. Zhang, D. Liu, "Attribute-based Access Control Scheme With Efficient Revocation In Cloud Computing", China Communications, Vol. 13, No. 7, 2016, pp.92–99.

[23] N. Saravanan, D. A. Umamakeswari, "Lattice Based Access Control for Protecting User Data in Cloud Environments with Hybrid Security", Computers & Security, Vol. 100, 2021, p.102074.

[24] P. S. Challagidad, M. N. Birje, "Efficient Multi-authority Access Control using Attribute-based Encryption in Cloud Storage", Procedia Computer Science, Vol. 167, 2020, pp. 840–849.

[25] N. Vurukonda, M. T. Basu, V. Velde, K. Enumula, "Revocable Storage Identity-Based Encryption For Protected Shared Data In Cloud Computing", Material Today: Proceedings, 2020

[26] B. P. Kavin, S. Ganapathy, "A Secured Storage and Privacy-Preserving Model Using CRT for Providing Security on Cloud and IoT-Based Applications," Computer Networks, Vol. 151, 2019, pp.181–190.

An Optimal Virtual Machine Placement Method in Cloud Computing Environment

Original Scientific Paper

Ashalatha Ramegowda
Faculty, Department of Computer Science
Gulbarga University, Kalaburagi
Karnataka, India
ashalatha.dsce@gmail.com

Abstract *– Cloud computing is formally known as an Internet-centered computing technique used for computing purposes in the cloud network. It must compute on a system where an application may simultaneously run on many connected computers. Cloud computing uses computing resources to achieve the efficiency of data centres using the virtualization concept in the cloud. The load balancers consistently allocate the workloads to all the virtual machines in the cloud to avoid an overload situation. The virtualization process implements the instances from the physical state machines to fully utilize servers. Then the dynamic data centres encompass a stochastic modelling approach for resource optimization for high performance in a cloud computing environment. This paper defines the virtualization process for obtaining energy productivity in cloud data centres. The algorithm proposed involves a stochastic modelling approach in cloud data centres for resource optimization. The load balancing method is applied in the cloud data centres to obtain the appropriate efficiency.*

Keywords*: Cloud computing, Virtualization, Stochastic modeling, Energy efficiency, Cloud service provider, Resource optimization*

1. INTRODUCTION

Cloud is a large server for storing different services and data of the users. Cloud is a concept of using services not stored on your computer. The virtualization process consolidates many workloads into smaller physical servers in the data centres of the cloud to meet the Service Level Agreement (SLA) standards using Virtual Machines (VMs) [1]. The Virtualization process allows multiple users to share a physical server. Major companies include VMware, Hyper-V, HP, F5, Nuage, Nicira, etc. The virtualization technology uses a Virtual Machine Monitor (VMM) at the software level for abstraction purposes [2]. The Cloud Service Providers (CSPs) make the infrastructure as a Service (IaaS) for cloud consumers. They use VMs and Virtual Clusters (VCs) for computing cloud resources. Primary cloud providers include Amazon EC2 and IBM [3].

The user, service, and cloud computing infrastructure are significant entities involved in the cloud environment. Primary attacks in the cloud can be service attacks, including browser attacks, phishing attacks, and SSL certificate spoofing attacks. User attacks can be accomplished during spoofing or the cloud infrastructure services. Significant Quality of Service (QoS) matter has service availability considered a cloud environment. Therefore, the virtual data centre is becoming popular because of providing IaaS in the field of the cloud [4]. IaaS is the most crucial delivery model developed in cloud computing. IaaS offers virtual infrastructure like servers and data storage. Cloud providers can use virtualization techniques for virtual data centres in cloud computing. Amazon is an excellent example of providing a cloud platform that offers infrastructure at an affordable price to its customers [5].

A VM is a single computer with a dedicated platform environment with limited resources. The resource virtualization process is the simplified version of traditional resource management. The virtualization system encourages replication procedures in the cloud to perform elasticity processes within a given system. They run on a hardware platform controlled by VMM. Each VM runs under VMM, which can change the virtual status from one data centre to another. A VMM is hypervisor software that divides the computation resources into the number of VMs through the guest operating system [6]. VMM has various operating systems to execute particular hardware simultaneously for resource isolation. The significant benefits of using VMM include high-security performance using multiple services simultaneously. VMM uses a Live VMM scheme to transfer the cloud server from one hardware platform to another using system modification. Mobile devices consume less energy to achieve resource optimization in the cloud network. The parameter metrics to consider are data size and delay constraints for optimal

solutions in the mobile cloud. Different energy optimal models for mobile devices include:

- The mobile execution models.
- The execution model for the cloud.
- The optimal application execution policy model.

This research aims to observe the energy efficiency of cloud data centers using a stochastic modelling approach. The principal enrichment of this paper is as follows.

- Increasing energy productivity in cloud data centers using a resource optimization approach.
- Load balancing scheme using resource virtualization method.
- A stochastic modeling method for energy efficiency in cloud data centers.

The remaining part of this work is as follows. Section 2 presents the overall literature survey part. Section 3 depicts a system model with architecture. Section 4 details the methodology part. Section 5 provides the performance analysis, and section 6 concludes the proposed work.

2. RELATED WORK

An enormous amount of literature has been reviewed for energy efficiency for achieving high performance in cloud computing. The stochastic models accomplish service requests and maintenance of the server. Zhang et al. provide a Markov chain process scheme for optimum policy schedule and scaling problems of cloud servers [7]. CSPs must reduce the energy usage in cloud data centres, and a reliable system can produce with less cost for operation purposes.

Xia et al. have presented energy efficiency in the cloud data centre using the VM migration process's stochastic method for high performance [8]. Han et al. have used the VMM policy in cloud data centres to achieve high performance and robustness. A dedicated stochastic process model has been used for energy efficiency with high production [9].

Ait Salaht et al. have used a technique called the hysteresis queuing model, which is used for cloud data centres. The stochastic bounding models provide performance analysis in the cloud [10]. Anastasopoulos et al. have operated the optical networks and cloud infrastructures considered for service provision in the cloud. A stochastic linear-based programming approach has been used for resource provisioning in the cloud to evaluate and use renewable sources [11].

Ghosh et al. have used cloud host services to reduce costs in the data centres. The VMs provide IaaS cloud service, which shares the instances of Physical Machines (PMs) instances within cloud data centres. An optimal PM has been chosen to minimize operational costs with excellent infrastructure in the cloud. The stochastic model has been used to analyze value and optimize the framework within the IaaS cloud [12].

Zhou et al. use cloud computing with a sophisticated infrastructure and comprehensive data-sharing service. The stochastic process with high-quality evaluation and modelling has been proposed in work. The assessment of the IaaS cloud performs quality metrics based on the criteria such as completion time of user requests, system overhead rate and rejection time probability [13].

Maguire et al. provide the stochastic analysis model, which uses the load balancing process and the VM and is a scheduled concept of cloud. In Cloud Computing Cluster (CCC) theory, every job uses VMs under a stochastic process. A dedicated algorithm for load balancing and VM scheduling is analyzed for high capacity within the system [15]. Chen et al. say a scalable and flexible approach is designed for high-scale cloud computing. The multi-data centre model provides substantial data processing for large applications and top computing resources.

High performance attained by using workload scheduler under VM in cloud data centres. The QoS-based approach model is designed for energy efficiency and resource optimization [16]. Vasileios et al. have presented an intelligent city framework using the Internet of Things [17]. Modern computing techniques are adopted to increase usability and are also helpful in preserving confidential details [18]. Tahar et al. use a VM placement scheme for cloud data centres. The integer linear model saves energy and hence increases QoS [19].

Every VM is decided through a scheduling strategy to achieve budgetary deadlines. Resource provisioning is provided for accessing the tasks with execution time. The stochastic-based scheme uses multi-objective scheduling criteria for making energy-efficient in a cloud [20]. A stochastic approach is used for energy and cost minimization purposes. LP and LDPP schemes are used for cost-savings sake. CCDF method has been proposed for high performance. A stochastic optimization model is given for green data centres [21]. Stochastic Petri nets are used for QoS and any time system availability. The VM placement strategy achieves good accuracy in the cloud. The SRN model performs VM migration and placement strategy using proposed algorithms [22].

3. SYSTEM MODEL

Cloud system network varies from traditional network distributed systems. They are characterized by many resources that can span different administrative domains. Various clouds appropriate to one particular or different organizations can dynamically join each other to achieve a common objective, usually represented using the optimization of cloud resources. This method is known as cloud federation [23]. The system architecture comprises data users, owners, trusted authority, cloud proxy, and Third-Party Auditor (TPA). The cloud users are connected to share multimedia files

through the wireless access point to the cloud proxy server. The cloud data owner uses data file ciphertext to store the contents of the cloud data centres. TPA is used in cloud data centres to achieve additional security purposes.

The trusted authority uses privilege data management requests and privilege update requests by using public and attributes keys in cloud data centres. TPA is used in cloud data centres to achieve additional security purposes. The trusted authority uses privilege data management requests and privilege update requests by using public and attributes keys in cloud data centres. The system security model for public auditing scheme cloud servers, data users, and TPA [24]. The proposed system model represents computer systems composed of many resources, making it possible to describe physical and virtual resources. Each system's general stochastic data model comprises M = m X N VMs running parallel. Here, m refers to the number of virtual machines in the cloud.

M is the maximum number of VMs running in a particular method of a cloud data centre for resource optimization. The stochastic process maintains three primary servers in the cloud. They include images, video and document servers [25].

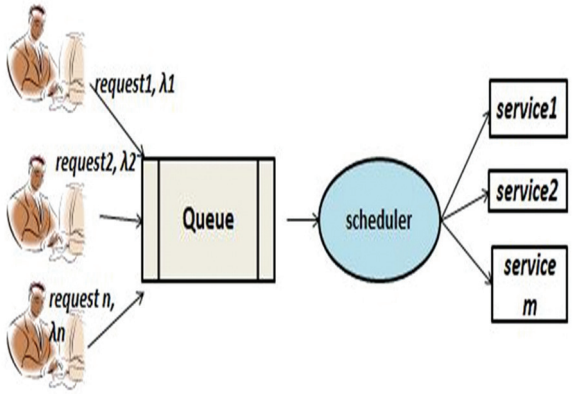


Fig. 1. Basic system model

Figure 1 depicts the overall cloud system security model for various services and user requests. The significant entities included in the system architecture are cloud users, input queue, resource scheduler, and many services to be computed in the cloud server [26]. It depicts the queuing performance model in the cloud for the service requests to be performed. Figure 2 represents the proposed system storage model in cloud systems. Primary operations involved in the model include cloud users, data owners, trusted authority, cloud proxy servers and TPA.

A data owner is responsible for generating the encrypted files and uploading the files to the cloud server [27]. Trusted authority checks for the incoming requests and sends the required key to the data owner. The proxy server requests the needed key, gets the essential key, selects the file to encrypt, decrypts the confidential data, downloads the received files and shares the files with the cloud users [28]. TPA can view all the files on the server and perform the auditing process for cloud security. The overall system architecture has the following stages.

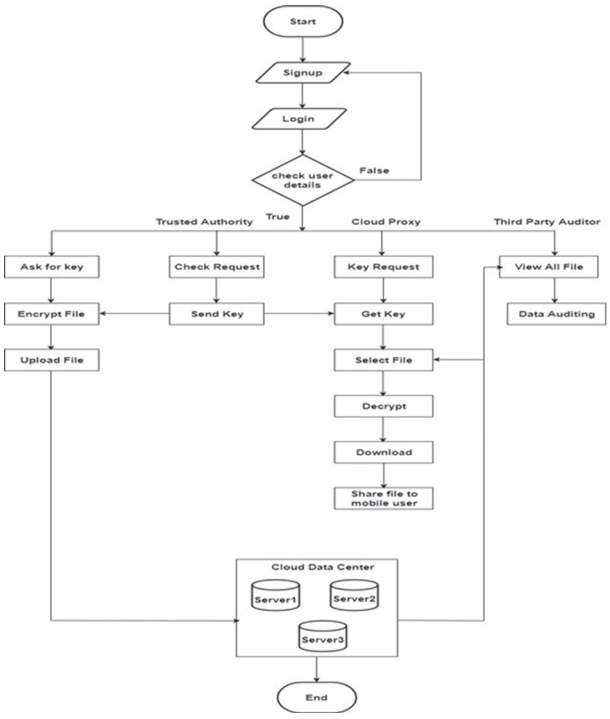


Fig. 2. System storage model

Figure 2 represents the system storage model that includes the following steps.

i. Data owner who uploads the data files to the cloud data centre in an encrypted manner.
ii. Cloud proxy who shares files from cloud server for cloud users.
iii. A third-party auditor who is used for maintaining the cloud data files
iv. A trusted authority that provides the attribute key for cloud proxy for file sharing.
v. Data users could be mobile phones, laptops, personal computers or Tablet PCs.

4. METHODOLOGY

The dynamic load balancing (DLB) approach uses only the present machine state for balancing the current workloads to achieve high-performance satisfaction and complete deployment of cloud resources. The load balancing scheme for data centres improves the energy efficiency of the cloud resources in the cloud. DLB is a method that distributes scalable workloads evenly among all the system nodes in the cloud, used to create new instances [29]. The massive amount of energy from the industry and companies leads to high-cost cloud data centres. Thus, the cloud data centres must change according to the energy used to gain energy efficiency using the virtualization process. A

stochastic model that uses the queuing theory concept is used to achieve high performance and energy consumption. Dynamic right-sizing of the data centres can be gained using stochastic modelling in cloud data centres [30]. The cloud maintains the central storage of data in its remote data centres. Data management has been made easier through cloud storage. The queuing model concept applies to managing and storing applications in the cloud [31].

4.1 SYSTEM ARCHITECTURE

The following section describes the overall system architecture for the stochastic system process. Figure 3 represents the stochastic system model. It includes a load balancing process, resource optimization and a data centre management module. VM scheduler works on VM section strategy using placement process. The cloud data centres use hypervisors for VM management [32].

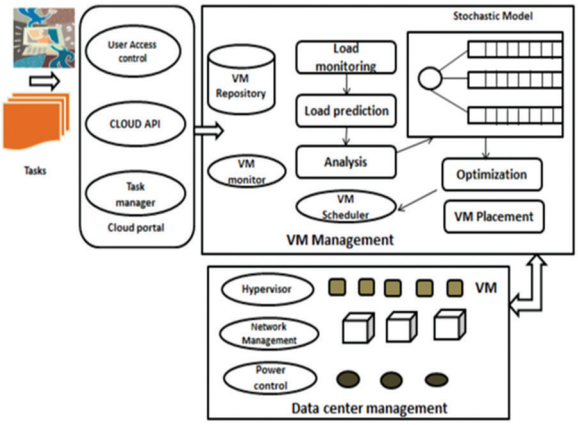


Fig 3. System architecture

4.2 MODEL ANALYSIS AND SOLUTION

The steady-state includes the probability distribution analysis method, which accomplishes essential information about the given data centre model.

Let J (t) is given as number of VMs for the system at time t.

Let L (t) is given as the number of working servers for the time t, then

L(t) and J(t) provides the input with a process with the given state space.

$$\Omega = \{(i,j): 0 \le j \le M-1, 0 \le i \le j\} \cup \{(i,j): j \ge M, 0 \le i \le N\} \quad (1)$$

The stochastic process solution includes the following entities.

Let X(*t*) be the stochastic process where t ≥ 0

S: finite state space

R: Reward function

π_i: Steady-state probability of state S_i

The stochastic modelling is a process of evolution, where $r_{X(t)}$ gives the stochastic process using system reward rate at time t.

$r(r:S \rightarrow R)$ is given as reward function

Probability π_i allows $i \in S$ as steady state and *r(i)* denotes reward and is written as r_i

The model solution has $t(\pi_i(t))$ and $t_i(t)$ presents the expected reward rate given in equation 2.

$$E[r_x] = \sum_{i \in S} r_i . \pi_i \quad (2)$$

4.3 MARKOVIAN CHAIN PROCESS

The following equation gives the markovian continuous model.

$$w, P1, P2, \ldots, Pm, V1, V2, \ldots, Vm, F1, F2, \ldots, Fm \quad (3)$$

Where, w is the number of VM requests in the waiting queue

P_i is the position of the virtual machine level

V_i is the number of virtual machines

F_i=0 means the virtual machine is alive in the process else failure.

The queuing theory scheme can be used for modelling the data and applications in the cloud. The model solutions are built using an analytic method using the probability vectors. The model analysis provides a key by using an M/M/N queuing model using a Markovian chain process in the cloud environment [33]. This model helps develop the algorithm for a specific VM during the execution time of the cloud resources.

4.4 STEADY STATE VECTOR

Here, πij is the probability vector stating that there are j VMs on the system and i working servers. The steady state vector is given as π and can be considered as:

$$\begin{matrix} \{\Pi Q = 0 \\ \sum_{j=0}^{\infty} \pi j e = 1\} \end{matrix} \quad (4)$$

The steady state vector is given by:

$$\begin{matrix} \Pi = [\Pi_0 \Pi_1 \Pi_2 - - - - \Pi_N] \\ \Pi_j = [\Pi_0 \Pi_1 \Pi_2 - - - - \Pi_{jj}] \text{ for j < N+1} \\ \Pi_j = [\Pi_{0j} \Pi_{1j} \Pi_{2j} - - - - \Pi_{Nj}] \text{for j > N} \end{matrix} \quad (5)$$

4.5 THE STOCHASTIC MODEL

Each system's general stochastic data model comprises M = m X N VMs running parallel. Here, m refers to the number of virtual machines in the cloud. M is the maximum number of VMs running in a particular method of a cloud data centre for resource optimization [34]. The stochastic process maintains three central servers in the cloud. They include images, video and document servers [35]. Here, each server can consider one of the following states.

i. serving the VMs. The service rate always gives the energy spent in this specific state. Suppose

there is m' VMs present where m' ≤ m is the number of cloud servers running in the system, the CPU utilization can be given as μm'. The energy consumption can be provided by:

$$\mu m' P(on) \quad (6)$$

ii. *OFF state*: Here, the server is not serving any VMs. The image, video and document servers are made off as there are insufficient VMs. Hence, the energy consumption for any given server is 0.

ii. *Setup state*: The state, which goes from off state to on form, is a setup state. Here, all the servers are made for file accessing purposes. Therefore, the energy consumption for the given server is given as P_{on}.

ii. *Failed state*: The server has failed due to some erroneous failure in the server or the data centres. The system can crash due to some catastrophic losses or damage due to disasters, and hence, the server goes to a failed state. Here, energy consumption can be given as P_{fail}.

The performance efficiency can be achieved using scalable analytic models for cloud resources. The optimization algorithms are used in cloud data centres to achieve optimal values such as service rate μ, optimized servers N and the performance function F (μ, N). The dynamic workflow uses a critical path VM selection strategy for optimization sequence. The resource optimization algorithm is defined in algorithm 4.1.

Algorithm 4.1: Resource optimization algorithm

Step 1: Procedure measure1

Step 2: Input: performance metric function F (μ, N), N, λ, θ, with an initial point μ_0 and a positive tolerance Δ.

Step 3: Output: Approximate solution μ*

Step 4: Calculate unit matrix $H=I_n$ (7)

Step 5: Compute the gradient
$g_0=\nabla F(\mu_0,N)$ at point $x_0=\mu_0$

Step 6: Set k to 0

Step 7: While $|\nabla F(x_k+1)\leq\Delta|$ do (8)

Step 8:Generate the search direction
d_k=-H_k g_k (9)

Step 9: Search along d_k from point X_k, find the step-length $\propto_k$ by satisfying
$F(X_k+\propto_k d_{-k})=\min\{F(X_k+\propto_k d_k)\}$ (10)

Step 10: let $g_{k+1}=\nabla F(X_k+1)$, $p_k=X_{k+1}-X_k$, $q_k=g_{k+1}-g_k$ (11)

Step 11: $H_{k+1}=H_k+1+p_k^T q_k/q_k^T H_k q_k)$ (12)

Step 12: $X_{k+1}=X_k+\alpha_k d_k$ (13)

Step 13: $k=k+1$

Step 14: end while

Step 15: return X_k

Step 16: end Procedure

Step 17: Stop

The optimization procedures are used in algorithms using various measures. The optimal solution can be found using an optimization procedure to see the complexity of the optimization algorithms. The main factors used are variables, validity and effectiveness in our optimization algorithm. The resource optimization algorithm for stage 2 is given in the algorithm.

Algorithm 4.2: Optimization algorithm

Step 1: Procedure measure2

Step 2: Input: F (μ, N), μ*, λ, θ and M, where M is a sufficient large number

Step 3: Output: Approximate solution (μ*, N*)

Step 4: F*= ∞

Step 5: for (N=1;N<M; N++) do

Step 6: Use algorithm 1 to calculate F^* (mu^*, N)

Step 7: if ($F^*(mu^*, N) < F^*$ then

Step 8: $F^*=F^*(mu^*, N)$

Step 9: S= (mu^*, N)

Step 10: end if

Step 11: end for

Step 12: return S

Step 13: end procedure

Step 14: Stop

The resource optimization algorithm is used to gain optimal values of service rate μ, and optimal server N. The process uses a minimum performance metric function to achieve efficiency.

4.6 SYSTEM MODULES

The system design comprises the following system modules in cloud data centres.

i. *Load monitoring module*: The load monitoring module is used for calculating the statistics of λ, θ and σ for the resource optimization in the cloud system.

ii. *Load prediction module*: The load prediction module predicts the actual load by using λ (i + 1) for the next optimization period Δ.

iii. *Analysis and optimization*: These modules are used to implement the mathematical model and solution for resource optimization. The mathematical system model for this system design has been presented in Table 1.

Table 1: Mathematical model

μ	Service rate
m'	Virtual machines
N	Optimal web servers
E (W)	Execution time
E (P)	Energy consumption
F(μ, N)	Metric function
μ*	Optimal value

5. PERFORMANCE ANALYSIS

The efficiency of the proposed work is shown through performance metrics which are defined as follows.

i. *Accessing data files*

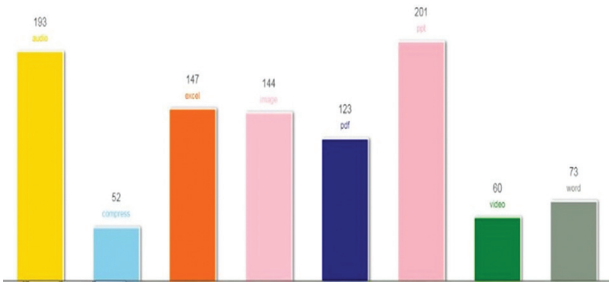


Fig. 4. Proxy accessing cloud

Figure 4 depicts the number of users accessing a proxy server for obtaining various types of files in the cloud. Cloud proxy can view the uploaded files and request for the file to download. Once the 'key' is requested, the request is sent to the trusted authority. Data files can be accessed more efficiently using a proxy server than the actual cloud. A cloud proxy server can view the uploaded files and request for the file to download. Once the key is requested, the request is sent to the trusted authority. Various multimedia servers involve video, word, pdf, ppt, image, excel, compress, audio servers etc.

ii. *Downloading multimedia files*

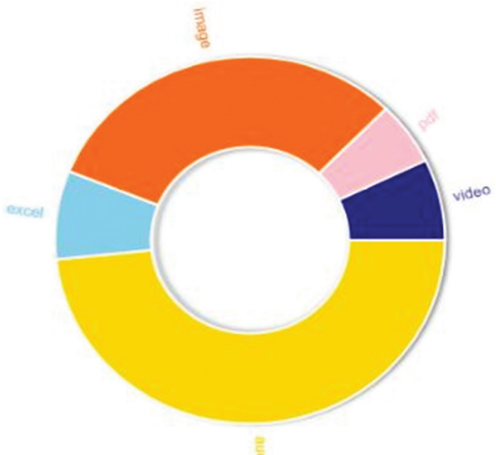


Fig. 5. Users accessing proxy server

Figure 5 depicts the number of users using a proxy server for downloading the data. The data owner can upload the data in a given format such as video, audio, image or document format. The process of downloading multimedia files can be made through a cloud proxy. The pie chart represents various multimedia files used for downloading from the cloud proxy server, including documents, video, audio, image servers, etc.

iii. *Resource access time with several cloud users*

In the following figure, the resource access audit time from the cloud proxy server is compared with access time from the cloud server. Bandwidth has been used broader, and user access has now improved through the proposed method. The audit access time of the cloud data can be reduced relatively by using the proxy server than the cloud server, as shown in Figure 6.

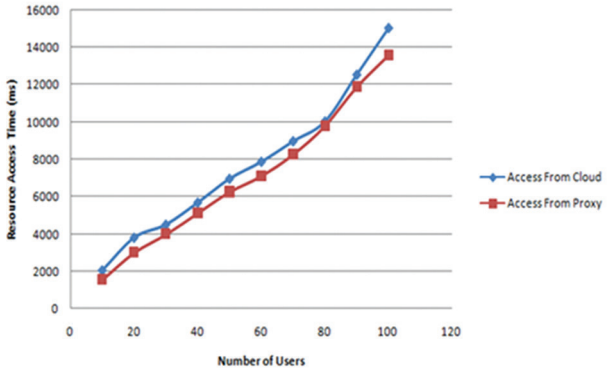


Fig. 6 Resource access auditing time

In the comparison graph of Figure 6, the resource access time from the cloud proxy server is compared to that of access time from the cloud server in milliseconds. In this approach, bandwidth has been used higher, and user access is improved through the proposed method. The trusted authority approves the file request from the proxy cloud. The cloud proxy can receive the mail with the token and attribute key to decrypt in registered mail id. Once the cloud proxy gets the mail, it can solve and view the file. And cloud proxy has additional work, which means they must upload the file to access the end mobile user. The Android mobile app is created to view and access this detailed data for accessing the file.

iv. *Virtualization graph*

The VMM policy can be used to transfer the VMs taken from one DC to another in the distributed systems. To evaluate the system behaviour in multiple data centres, the analysis of VMM time and balancing of the incoming load is required. The virtualization graph specifies resource type and the number of files in the cloud. It analyses which resource type occupies more on cloud and proxy. If the resources are stored in the proxy cloud, the speed will automatically increase. Here, the proxy cloud acts as a virtual server. The consolidation ratio denotes the virtual servers running independently on the host machine.

The number of multimedia resources, their count, and specific resource type are mentioned here. Figure 7 depicts the virtualization process within the cloud data centres. The graph shows the number of users accessing proxy servers for accessing various types of files in the cloud. Cloud proxy can view the uploaded files and request for the file to download. The demand for the specific file is sent to the trusted authority by the cloud proxy server.

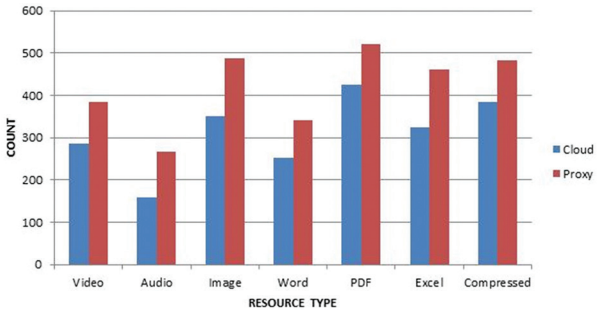


Fig 7. Virtualization graph

v. *Energy consumption*

The energy usage in various DCs can be reduced using the dynamic optimization scheme. The energy consumption graph is given between the resource name and the energy taken to access the resource. The chart emulates the cloud's energy consumption and the cloud proxy server to access the resource.

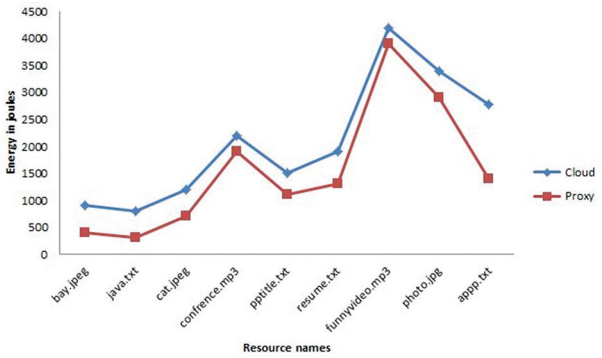


Fig. 8. Energy consumption

The stochastic type programming models reduce the energy consumption in large DCs in terms of joules. Figure 8 represents the energy consumption in the cloud centres for various files and applications. The numbers of resource names and the energy used by the cloud server are depicted in the figure.

vi. *Internet load monitoring*

Load Monitoring analyses resource size and time taken to access the particular resource in that size. It gives a more precise picture of how long it will take to obtain the specific volume.

Figure 9 gives the load monitoring process with the file size in kilobytes. The time taken in milliseconds from each file in the cloud is given. A large number of resource types with size, along with the time taken to load the files onto a cloud server, are shown in the graph.

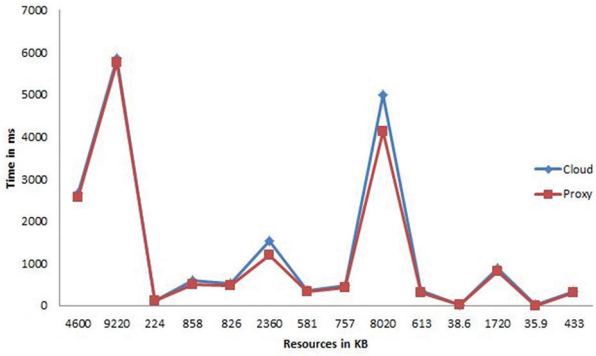


Fig. 9. Load monitoring

viii. *Load balancing*

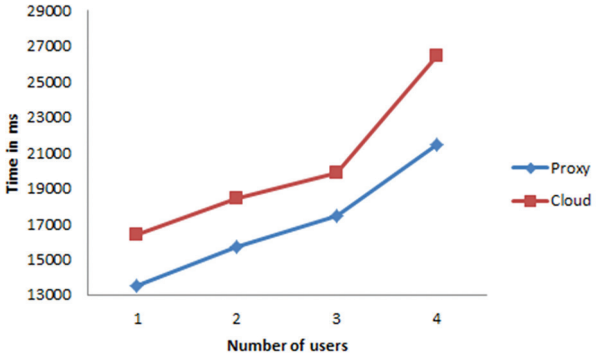


Fig. 10. Load balancing

Figure 10 depicts the load balancing feature involving several VMs in the cloud and proxy server. Time has been calculated in milliseconds for different VMs. The result shows that the proxy server takes less time than the cloud server.

ix. *Resource utilization comparison*

Figure 11 compares the resource utilization for both proxy and cloud servers using a virtualization system. The result shows that a proxy server utilizes less time for considering the cloud resources than a cloud in milliseconds.

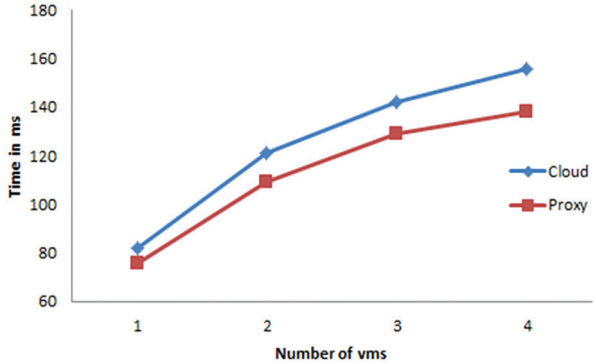


Fig. 11. Resource utilization

6. CONCLUSION

Cloud security has become a significant concern these days in the computing world. Computing and communication security has been taken into consideration by substantial researchers. The availability of cloud data services may fail anytime due to power failures or the occurrence of any catastrophic failures. A third-party error can happen anytime in the cloud due to limited transparency and user control. Hence, it remains paramount to outline the cloud and virtualization process before examining energy efficiency for cloud data centres. The data centre resource management is dealt with and considered a significant critical cloud computing problem. The load balancing approach uses stochastic data modelling for resource optimization in cloud data centres. The VM placement strategy reduces the total energy consumed and undetermined requirements by many cloud servers.

7. REFERENCES:

[1] R. Ashalatha, J. Agarkhed, S. Patil, "Network virtualization system for security in cloud computing", Proceedings of the 11th International Conference on Intelligent Systems and Control, Coimbatore, India, 5-6 January 2017, pp. 346-350.

[2] M. Roohitavaf, R. Entezari-Maleki, A. Movaghar, "Availability Modeling and Evaluation of Cloud Virtual Data Centers", Proceedings of the International Conference on Parallel and Distributed Systems, Seoul, Korea, 15-18 December 2013, pp. 675-680.

[3] R. Ghosh, K. S. Trivedi, V. K. Naik, D. S. Kim, D "End-to-end performability analysis for infrastructure-as-a-service cloud: An interacting stochastic models approach", Proceedings of the IEEE 16th Pacific Rim International Symposium on Dependable Computing, 2010 pp. 125-132.

[4] X. Chang, R. Xia, J. K. Muppala, K. S. Trivedi, J. Liu, "Effective modeling approach for iaaS data center performance analysis under heterogeneous workload", IEEE Transactions on Cloud Computing, Vol. 6, No. 4, 2016, pp. 991-1003.

[5] B. Wei, C. Lin, X. Kong, "Dependability modeling and analysis for the virtual data center of cloud computing", Proceedings of the IEEE 13th International Conference on High Performance Computing and Communications, 2011, pp. 784-789.

[6] W. Zhang et al. "Energy-optimal mobile cloud computing under stochastic wireless channel", IEEE Transactions on Wireless Communications, Vol. 12, No. 9, 2013, pp. 4569-4581.

[7] P. Zhang, C. Lin, K. Meng, Y. Chen, "A Comprehensive Optimization for Performance, Energy Efficiency, and Maintenance in Cloud Datacenters", Proceedings of the Trustcom/BigDataSE/I SPA, Tianjin, China, 23-26 August 2016, pp. 1264-1271.

[8] Y. Xia, M. Zhou, X. Luo, S. Pang, Q. Zhu, "A stochastic approach to analysis of energy-aware DVS-enabled cloud datacenters", IEEE Transactions on Systems, Man, and Cybernetics: Systems, Vol. 45, No. 1, 2015, pp. 73-83.

[9] Y. Xia, Y. Han, M. Zhou, J. Li, "A stochastic model for performance and energy consumption analysis of rejuvenation and migration-enabled cloud", Proceedings of the International Conference on Advanced Mechatronic Systems, 2014, pp. 139-144.

[10] F. Aït-Salaht, H. Castel-Taleb, "Stochastic bounding models for performance analysis of clouds", Proceedings of the IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, 2015 pp. 603-610.

[11] M. Anastasopoulos, A. Tzanakaki, D. Simeonidou, "Stochastic energy efficient cloud service provisioning deploying renewable energy sources", IEEE Journal on Selected Areas in Communications, Vol. 34, No. 12, 2016, pp. 3927-3940.

[12] R. Ghosh et al. "Stochastic model driven capacity planning for an infrastructure-as-a-service cloud", IEEE Transactions on Services Computing, Vol. 7, No. 4, 2014, pp. 667-680.

[13] Y. Xia et al. "Stochastic modeling and quality evaluation of infrastructure-as-a-service clouds", IEEE Transactions on Automation Science and Engineering, Vol. 12, No. 1, 2015, pp. 162-170.

[14] S. El Kafhali, K. Salah, "Stochastic modelling and analysis of cloud computing data center", Proceedings of the 20th Conference on Innovations in Clouds, Internet and Networks, 2017, pp. 122-126.

[15] S. T. Maguluri, R. Srikant, L. Ying, "Stochastic models of load balancing and scheduling in cloud

computing clusters", Proceedings of the IEEE INFOCOM Conference, 2012, pp. 702-710.

[16] Y. Chen et al. "Stochastic workload scheduling for uncoordinated datacenter clouds with multiple QoS constraints", IEEE Transactions on Cloud Computing, Vol. 8, No. 4, 2016, pp. 1284-1295.

[17] V. A. Memos, K. E. Psannis, Y. Ishibashi, B. Kim, B. B. Gupta, "An efficient algorithm for media-based surveillance system (EAMSuS) in IoT smart city framework", Future Generation Computer Systems, Vol. 83, 2017, pp. 619-628.

[18] B. Gupta, D. P. Agrawal, S. Yamaguchi, "Handbook of research on modern cryptographic solutions for computer and cyber security", IGI Global, 2016.

[19] A. Ouammou, M. Hanini, S. El Kafhali, A. B. Tahar, "Energy Consumption and Cost Analysis for Data Centers with Workload Control", Proceedings of the International Conference on Innovations in Bio-Inspired Computing and Applications, 2017, pp. 92-101.

[20] Y. Gao, L. C. Canon, F. Vivien, Y. Robert, "Scheduling stochastic tasks on heterogeneous cloud platforms under budget and deadline constraints", Proceedings of the IEEE International Conference on Cluster Computing, Albuquerque, NM, USA, 23-26 September 2019.

[21] A. Ghassemi, P. Goudarzi, M. R. Mirsarraf, T. A. Gulliver, "A Stochastic Approach to Energy Cost Minimization in Smart-Grid-Enabled Data Center Network", Journal of Computer Networks and Communications, 2019.

[22] Y. Liu et al. "Adaptive Evaluation of Virtual Machine Placement and Migration Scheduling Algorithms Using Stochastic Petri Nets", IEEE Access, Vol. 7, 2019, pp. 79810-79824.

[23] D. Bruneo, "A stochastic model to investigate data center performance and QoS in IaaS cloud computing systems", IEEE Transactions on Parallel and Distributed Systems, Vol. 25, No. 3, 2014, pp. 560-569.

[24] R. Ashalatha, J. Agarkhed, "Dynamic load balancing methods for resource optimization in cloud computing environment", Proceedings of the Annual IEEE India Conference, 2015, pp. 1-6.

[25] D. Shen et al. "Stochastic modeling of dynamic right-sizing for energy-efficiency in cloud data centers", Future Generation Computer Systems, Vol. 48, 2015, pp. 82-95.

[26] Y. Yagawa, A. Sutoh, E. Malamura, T. Murata, "Modeling and Performance Evaluation of Cloud on-Ramp by utilizing a Stochastic Petri-net", Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics, 2016, pp. 995-1000.

[27] A. Uchechukwu, K. Li, K. Li, "Scalable Analytic Models for Performance Efficiency in the Cloud", Proceedings of the IEEE/ACM 7th International Conference on Utility and Cloud Computing, 2014, pp. 998-1003.

[28] B. Silva, P. Maciel, A. Zimmermann, "Performability models for designing disaster tolerant infrastructure-as-a-service cloud computing systems", Proceedings of the 8th International Conference for Internet Technology and Secured Transactions, 2013, pp. 647-652.

[29] Y. Tian, C. Lin, Z. Chen, J. Wan, X. Peng, "Performance evaluation and dynamic optimization of speed scaling on web servers in cloud computing", Tsinghua Science and Technology, Vol. 18, No. 3, 2013, pp. 298-307.

[30] J. Wang, S. Shen, "Risk and energy consumption tradeoffs in cloud computing service via stochastic optimization models", Proceedings of the IEEE/ACM Fifth International Conference on Utility and Cloud Computing, pp. 239-246.

[31] M. Ranjbari, M., & J. A. Torkestani, J. A. "A learning automata-based algorithm for energy and SLA efficient consolidation of virtual machines in cloud data centers", Journal of Parallel and Distributed Computing, Vol. 113, 2018, pp. 55-62.

[32] I. Narayanan, D. Wang, A. Sivasubramaniam, H. K. Fathy, "A Stochastic Optimal Control Approach for Exploring Tradeoffs between Cost Savings and Battery Aging in Datacenter Demand Response", IEEE Transactions on Control Systems Technology, Vol. 26, No. 1, 2018, pp. 360-367.

[33] T. Li, B. B. Gupta, R. Metere, "Socially-conforming cooperative computation in cloud networks", Journal of Parallel and Distributed Computing, Vol. 117, 2018, pp. 274-280.

[34] Y. Pan et al. "A Novel Approach to Scheduling Workflows Upon Cloud Resources with Fluctuating Performance", Mobile Networks and Applications, 2020, pp. 1-11.

[35] J. Zhou et al. "Stochastic Virtual Machine Placement for Cloud Data Centers Under Resource Requirement Variations", IEEE Access, Vol. 7, 2019, pp. 174412-174424.

Scheduling Algorithms: Challenges Towards Smart Manufacturing

Original Scientific Paper

Abebaw Degu Workneh
Euro-Mediterranean University of Fez,
Euromed Research Center, Fes, Morocco
a.deguworkneh@ueuromed.org

Maha Gmira
Euro-Mediterranean University of Fez,
Euromed Research Center, Fes, Morocco
m.gmira@ueuromed.org

Abstract *– Collecting, processing, analyzing, and driving knowledge from large-scale real-time data is now realized with the emergence of Artificial Intelligence (AI) and Deep Learning (DL). The breakthrough of Industry 4.0 lays a foundation for intelligent manufacturing. However, implementation challenges of scheduling algorithms in the context of smart manufacturing are not yet comprehensively studied. The purpose of this study is to show the scheduling No.s that need to be considered in the smart manufacturing paradigm. To attain this objective, the literature review is conducted in five stages using publish or perish tools from different sources such as Scopus, Pubmed, Crossref, and Google Scholar. As a result, the first contribution of this study is a critical analysis of existing production scheduling algorithms' characteristics and limitations from the viewpoint of smart manufacturing. The other contribution is to suggest the best strategies for selecting scheduling algorithms in a real-world scenario.*

Keywords*: Scheduling Algorithm, Smart Manufacturing, Production Scheduling, Industry 4.0*

1. INTRODUCTION

In smart manufacturing industries, huge amounts of data are generated from heterogeneous sources such as sensors, Radio Frequency Identification (RFID), and networked machines [1], [2]. Moreover, inherently stochastic processes exist in industrial processes [3]. The advancement of Industry 4.0 and industrial intelligence leads to increased complexity, dynamics, and uncertainty on the shop floor [4]. This behavior paves the way for production scheduling challenges.

Scheduling algorithms should consider competing requirements to achieve a high-quality solution while remaining computationally efficient. Existing industrial scheduling solutions, such as heuristic algorithms, are efficient but difficult to implement in complex situations [5].

Heuristics, meta-heuristics, and mathematical programming are prominent tools to solve scheduling problems. However, as the complexity and scale of the problem increase, the solution would be unstable or might lead to unacceptable computing overhead [3], [6], [7]. It requires plenty of time to find a new solution [8] or needs manual configuration efforts during changes because of its model-based implementation and static nature [9]. Moreover, it also lacks the adaptability to a stochastic environment and needs a complex design process [10]. For example, as mentioned in [11], genetic algorithm has shown poor local search and slow convergence.

Despite the unavailability of scheduling algorithm challenges review in the smart manufacturing environment, there are an increasing number of review articles about smart manufacturing scheduling.

The purpose of this paper, unlike the previous review articles, is to emphasize the challenges of using different scheduling algorithms in the production environment, to introduce current scheduling strategies and their characteristics from the viewpoint of complex manufacturing and dynamically changing environment in the context of smart manufacturing, and to show the possible future research directions from different perspectives. The paper consists introduction to scheduling algorithms in production scheduling, a review methodology, a literature review, and a discussion on the properties and challenges of current scheduling solutions followed by a conclusion and future work.

2. REVIEW METHODOLOGY

The review is conducted based on the following five criteria: a) semantic areas of the article search; b) repositories used; c) document types; d) subject areas; e) lan-

guage of the article (English only). The four semantic fields on which the article search was based are: a) field 1: "shop floor scheduling problem"; b) field 2: "smart manufacturing"; c) field 3: "Scheduling technology & tools"; d) field 4: "Scheduling algorithm". Terms for each semantics are selected based on their relevance after individual search and all terms yield a different result in each repository because of their difference in their query system. Four repositories are considered: a) Scopus; b) Crossref; c) PubMed; d) Google Scholar.

The search results were initially obtained using publish or perish tool on Scopus, and the other three repositories are used to complement the search results. Publishers for lots of searched articles are: a) Elsevier; b) IEEE; c) Springer; d) SAGE publications; e) Multidisciplinary Digital Publishing Institute (MDPI); f) Hindawi; g) Wiley Online Library; h) IOP Publishing. The review is conducted based on the following research questions:

a) Which algorithm does the industrial environment need?
b) What has been done so far in the production scheduling field that can contribute to smart manufacturing?
c) What still needs to be done for the practical implementation of scheduling solutions in the smart manufacturing industry?

3. LITERATURE REVIEW

The searched articles, as depicted in Fig. 1., are reviewed based on thematic and content analysis.

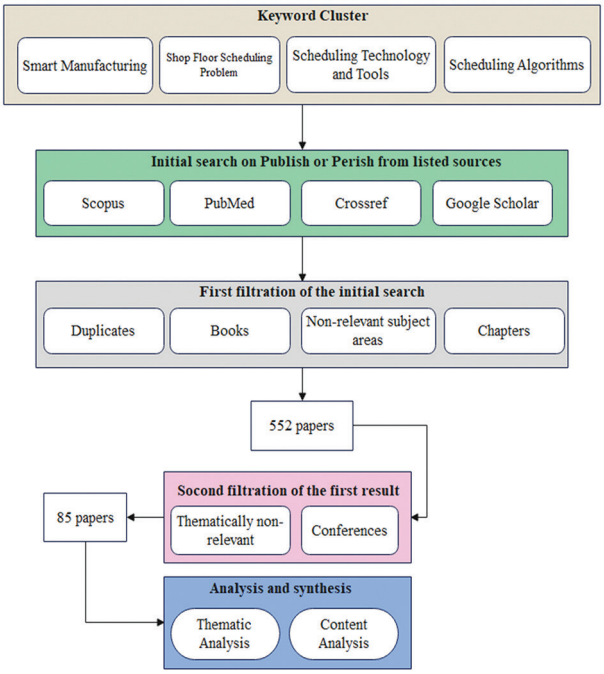


Fig. 1. Searches, collection, analysis, and synthesis methodology

Thematic Analysis: Based on the search term used in the reviewed articles, four thematic areas are identified using ATLAS.ti 9. These themes are smart manufacturing, shop floor scheduling problems, scheduling algorithms, and scheduling technology and tools.

Content Analysis: The content analysis is performed using the following phases: a) grouping of search articles based on the conceptual scheme of the research, b) the focus, c) experimental evaluation techniques used, and d) contributions and shortcomings of reviewed articles.

2.1. SMART MANUFACTURING

The term Smart manufacturing originated in the USA [12] and has no commonly accepted definition. Based on the study in [12], [13], smart manufacturing is a manufacturing operation that manages manufacturing processes with networked data. Likewise, the study in [14] defines the concept as a creation of manufacturing intelligence throughout all parts of the operation. It is a new manufacturing prototype in which manufacturing devices are entirely linked by wireless connections, supervised by sensors, and managed with cutting-edge computational intelligence [15].

The key technologies in smart manufacturing involve IoT, CPS, cloud computing, machine learning, big data, and mobile internet [14], [16], [17].

These technologies are realized through connected sensors, data interoperability, multi-scale dynamic modeling and simulation, smart digitization, and customizable and multi-level network security [18]. Materials, data, production processes and tools, resource sharing and connectivity, predictive engineering, and sustainability are considered the fundamental components [19]. The main idea behind this paradigm is to accumulate and evaluate massive amounts of manufacturing data to drive knowledge and rules [20].

Smart manufacturing involves the deployment of large amounts of sensors and IoTs, which requires the handling of big manufacturing data[15]. Big data is a key component in transforming today's manufacturing into a smart manufacturing paradigm. It helps companies to be competitive using data-driven strategies [21] and satisfy the needs of the manufacturing industry [14]. Deep learning, with its feature learning and large modeling capabilities, is an advanced analytics method for smart manufacturing. Based on the study in [21], smart manufacturing is divided into four modules i.e., "manufacturing module, data driver module, real-time monitoring module, and problem processing module". In the manufacturing module indicated in Fig. 2., the inputs are raw materials, and the outputs are finished goods.

Smart manufacturing has a different definition from different perspectives. For example, from the engineering point of view [22], smart manufacturing is characterized by the application of advanced intelligent systems that enables rapid production of manufacturing products, dynamic response to demand, and real-time optimization of the production and supply chain networks. In other words, the connected manufacturing

resources take raw materials as input and produce a finished product for a customer. on the other hand, from (IoT & CPS) as well as interconnection perspective [23], smart manufacturing is defined as the collection of all stages of manufacturing data using sensors and different communication technologies to increase production rate and reduce errors and production waste. However, from the viewpoint of predictive analysis and decision making [24], smart manufacturing is the optimization of planning and control of manufacturing activities such as fault diagnosis, risk assessment, resource utilization, predictive supply, and manufacturing. Based on the aforementioned definition, this study focused on the scheduling approaches in interconnection and decision-making perspectives.

2.2. SHOP FLOOR SCHEDULING PROBLEMS

Scheduling is the process of assigning machines to a set of available jobs to optimize objective functions such as earliness or tardiness of jobs, job completion, and processing times [25]. By its nature, scheduling needs details about tasks to be executed and available resources with a set of constraints [17].

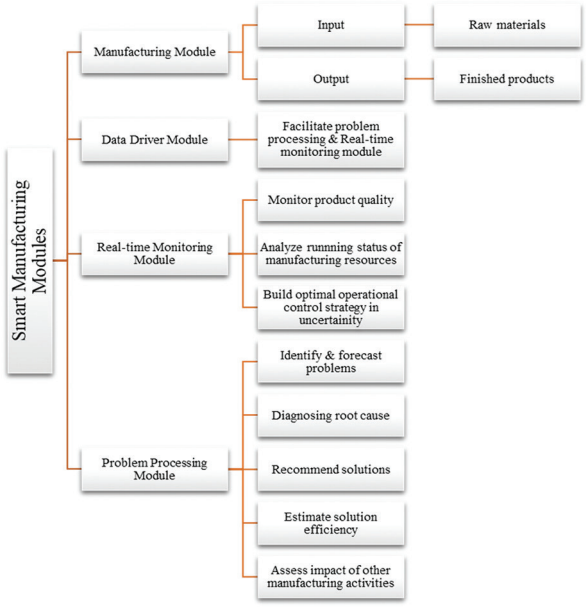


Fig. 2. Smart Manufacturing Modules

On the other hand, re-scheduling or pre-reaction scheduling is a way of scheduling again during the occurrence of new events [26]. It has two stages i.e., the pre-scheduling stage (generating scheduling for actual production) and the re-scheduling stage (re-configure the initial schedule to accommodate dynamic events). Robust scheduling forecasts potential future events based on the existing manufacturing state and generates a pre-schedule that includes different dynamic events. However, its success will be less effective if the dynamic conditions do not happen. Unlike the previous scheduling, the online scheduling is a real-time scheduling technique, that doesn't prepare a schedule in advance and is mainly used after dynamic events happen. This scheduling strategy is mostly used in actual production

In the reactive scheduling approach, decisions at the control and scheduling level don't consider dynamic events. As new events happen in production, a continuous recalculation process will take place, which results to be computationally expensive [27]. Conversely, the preventive scheduling approach considers dynamic events and aims at finding robust scheduling solutions with and without the presence of disruptions. Scheduling accuracy and performance are greatly influenced by the presence of uncertainty [28], [29].

Scheduling can be defined as classical scheduling and dynamic scheduling. In classical scheduling, all machines and constraints such as due date, processing time, etc. are available for scheduling [29]. The production state usually changes with time. However, the predefined states in this scheduling approach cannot address all production states. Moreover, parameters are usually found by statistics or calculations [30].

Investigation of dynamic scheduling is introduced by Jackson in the 1950s [26]. The critical factors in a dynamic manufacturing environment are external disturbances such as failure and production scheduling plan [31]. Despite its advantage over classical scheduling, uncertainties in dynamic scheduling are from direct assumptions, rather than actual production data [30]. As a result, it is not sufficient to provide interactive feedback with the real production. Most of the classical research focused on classical scheduling, in which all system state is known in advance and do not consider changing events [32], [33]. However, in a real production environment, unexpected events may occur at any time. For example, machine breakdown may happen at any time. So, it is also necessary to consider a maintenance plan along with real-time scheduling [34].

Maintenance activities are usually non-separable with production scheduling [32], [33], [35]. Based on [34], there are two main groups of maintenance: corrective maintenance (which involves repair during unexpected machine breakdown) and preventive maintenance (a condition where a planned schedule is executed before machine breakdown happens). On the other hand, the study in [14] classified the industrial maintenance strategies into four: reactive (perform maintenance during complete machine failure); corrective (identify and solve failures when it happens before total machine failure); preventive (performing regular maintenance to prevent partial or complete failure); and predictive (anticipate failures before it happens and guess the remaining useful life of the machine).

2.3. JOB SHOP SCHEDULING

Since the 1960s, Job Shop Scheduling Problems (JSSP) have been considered NP-hard problems [36], [37]. In JSSP, the number of jobs to be scheduled can

be processed at a pre-determined set of machines [38]. Machines can also be re-visited by the job more than once. The Re-entrant Job Shop Scheduling Problem (RJSSP) is more complicated than JSSP [39].

To address scheduling problems, many algorithms have been used on JSSP machine environments such as Hybrid Genetic Algorithm (HGA) [40], multi-agent system [41], variable neighborhood search algorithm [42], hybrid particle swarm optimization [39], genetic algorithm [43], and ant colony optimization [44], [45].

2.4. FLEXIBLE JOB SHOP SCHEDULING

Flexible Job Shop Scheduling Problem (FJSSP) is an extension of JSSP, where each operation can be operated beyond a single machine, and each machine is capable of performing distinct tasks [46]. As a result, it is more complicated than traditional JSSP [47]. It is also known as Integrated Process Planning and Scheduling (IPPS) [48]. Because of its applicability in different industrial applications, FJSSP has received wide attention from researchers. Based on the study in [48], flexibility is classified into three: operation flexibility (a situation where a process can be processed by different alternative machines); process flexibility (a condition where it is possible to finish the product by combining different operations); and sequence flexibility (when operations processing sequence is variable for the product). A flexible job shop can respond quickly to market changes and customer demands [49]. In FJSSP, jobs may re-enter or visit the center more than once before completing the process [38]. This feature is widely known in printed circuit board and semiconductor industries.

Some of the previous research's objective functions and used algorithms are illustrated in Table 1.

Table 1. Previous works on FJSSP

References	Objective	Algorithms
[47]	Makespan and energy	Simulated Annealing (SA) and Artificial Immune Algorithm
[49]	Energy	Ant Colony Optimization (ACO)
[50]	Mean tardiness	Greedy randomized adaptive search
[51]	makespan and energy	simulated annealing
[52]	Makespan and energy	Backtracking search algorithm
[53]	Makespan and energy	Genetic algorithm (GA)
[54]	Makespan	Quantum algorithm
[38]	Makespan	Approximation algorithm
[55]	Makespan and due date	Iterated greedy constructive heuristic
[56]	Makespan, machine workload	Clustering search metaheuristics
[57]	Makespan	Jaya algorithm, Monte Carlo
[58]	Makespan	Genetic algorithm
[37]	Makespan & setup time	Non-Dominated Searching Genetic Algorithm (NSGA-II)
[59]	Makespan	Constraint Programming (CP)
[60]	Makespan	Genetic algorithm

2.5. FLOW SHOP SCHEDULING PROBLEMS

The flow shop is composed of multi-stages, and each stage comprises only one machine [61]. In Flow Shop Scheduling Problem (FSSP), machines are assumed to be available during the entire planning setting [33]. It deals with the sequencing of jobs that enters a specified number of machines usually in the same order. Some of the previous studies on the FSSP machine environment are presented in Table 2.

Table 2. Existing studies on FSSP

References	Objective	Algorithms
[62]	Makespan and cost	Mixed Integer Programming (MIP)
[32]	Makespan and cost	NSGA-II and PSO
[35]	Makespan, earliness, and tardiness	Genetic algorithm and Harmony search
[63]	Makespan	MIP
[64]	Makespan and job flow-time	Dispatching rules
[65]	Makespan and total tardiness	Fruit fly optimization algorithm
[66]	Cost	MIP
[67]	Total flow time	MIP
[68]	Makespan	MIP
[61]	Makespan and total tardiness	Evolutionary algorithm
[69]	Makespan	Simulated annealing
[70]	Makespan	Genetic algorithm & tabu search
[25]	Makespan	NSGA-II

2.6. SCHEDULING ALGORITHMS

The most commonly used scheduling algorithms in previous research are meta-heuristics, exact methods, reinforcement learning, deep reinforcement learning, and multi-agent deep reinforcement learning.

2.6.1. Heuristics/ Meta-heuristics

Metaheuristics algorithm usually adopts optimization and approximation methods [44]. An optimization method is used to find solutions in mathematical computation. However, its application is limited in real-time because it takes too much time to find an optimum solution. It also requires mathematically sophisticated uses so that it is computationally intractable [17]. Con-

trarily, the approximation method is used when it is difficult to apply the optimization method. The approximated optimal solution is found within a specific time for a calculation. The approximation algorithm runs in linear time. As a result, it is computationally efficient [38]. For example, Evolutionary Algorithm (EA) is one of the algorithms that is used to find an approximate solution [52], [71].

Industrial environment scheduling operation needs an efficient algorithm. MIP is an effective approach for solving small-scale instances [62]. This approach is used to find an optimal solution based on the designed constraints. The main weakness of this method is that it tries to solve comprehensive problems by breaking them down into different sub-problems and then using the result of one sub-problem as input to the next sub-problem [72]. As a result, it could be difficult to find the solution in case of different conflicting constraints. For small-scale problems, centralized approaches such as MIP or CPLEX are well suited [73].

In principle, all metaheuristics can be applied to the Flexible Job Shop Problem (FJSP). Due to its fewer parameters, Particle Swarm Optimization (PSO) is much simpler and easier to maintain [46]. Although its convergence speed is fast, PSO will converge to the local optimum and will not be able to jump out with a maximum iteration rate [39]. PSO is known for convenient variable neighborhood search and flexible coding methods for solving some combinatorial optimization problems. Likewise, The Variable Neighborhood Search (VNS) algorithm is a metaheuristic optimization approach for solving combinatorial problems. It finds a solution's neighborhood until a better solution than the existing one is found, and moves to another [74].

2.6.2. Multi-Agent Systems (MAS)

The classical MAS method uses only a single dispatching rule and doesn't consider the impact of environmental changes in selecting dispatching rules [75]. This behavior in turn leads to poor scheduling performance. From the viewpoint of scheduling results, Artificial Intelligence (AI) algorithms perform better than MAS [49].

Multi-Agent System (MAS) is an agent-based system in which distributed agents make their own decisions using available information to ensure the whole system runs smoothly [76]. Another type of MAS approach is one in which agents negotiate while distributed agents make scheduling or production planning decisions. To mention a few, dynamic scheduling algorithm for allocating tasks on MAS with ring structure bidding method and negotiation method [77]; scheduling of distributed machines with negotiation and bidding protocol [78], and agent negotiation protocol to cope with the dynamic manufacturing environment [79].

However, in negotiation and agreement protocol, negotiation between agents is performed through a predetermined rule-based mechanism [80]. As a result, adaptation to the environment remains a challenge.

The combination of decentralized production systems and Industry 4.0 complicates production scheduling optimization. In comparison to the centralized production control system, the decentralized production control system has low complexity, improved scalability, and real-time capability. Implementation of MAS on these problems simplifies the solutions. Despite its solution efficiency, multi-agent systems in this environment tend to show local optimization [81]. To address these challenges, cooperative multi-agents are necessary.

2.6.3. Reinforcement Learning

Reinforcement learning (RL) is concerned with learning from experiences. It describes how agents learn the best policy to achieve the desired objectives by observing an environment, performing possible actions, and obtaining a reward as a result. The agents' goal is to maximize cumulative reward [82].

No algorithm is adaptive enough to address all the wide area of manufacturing problems. Algorithms in previous studies need high computational efforts and failed in the real manufacturing industry where there are dynamic events and uncertainties [83].

Smart manufacturing scheduling differs from job shop scheduling in several ways, including a large number of tasks and services, as well as the dynamic states and uncertainties. Scheduling is a critical process for manufacturing industries to maximize profits while lowering costs. Specifically, in a dynamic and complex manufacturing environment, poor scheduling results in higher costs, longer production times, and higher tardiness [84]. Thereby, to comply with the complexities of a manufacturing site and improve its effectiveness, scheduling must be transformed and enhanced for sustainability and intelligence.

JSSP has been thoroughly researched over the last several decades, and numerous techniques for solving classical JSSP have been developed. Nevertheless, in real manufacturing industries, the environment is mostly dynamic, such as new job arrivals and machine failure [85]. Dynamic systems begin with the jobs that arrive first and are assumed to follow a probabilistic rule [86].

Task scheduling methods are divided into two: precise and approximate scheduling methods [87]. The precise methods search the entire search space for the global optimum solution. consequently, they are computationally complex and are inefficient at solving complex scheduling problems. Conversely, approximate methods have lower complexity and get the appropriate solution faster, while having greater advantages in solving complex scheduling problems. However, approximate methods cannot ensure an optimum

solution to the scheduling. A scheduling algorithm's main objective is to use a small number of machines to process a specified number of jobs while optimizing an objective [88].

The high dynamics, difficulty, and unpredictability of the JSSP environment continue to pose significant challenges [4]. Most JSSP methods are implemented as centralized algorithms with complete knowledge of the manufacturing process [88]. In contrast, one of the visions of Industry 4.0 is decentralized, self-learning, self-organizing, and self-optimizing production control [89].

The use of RL in JSSP has huge benefits. First, it is more adaptable than classical priority dispatching rule heuristics. Furthermore, developing such heuristics is tiresome because they require a great deal of expertise in a scheduling instance to be efficient [8]. RL, unlike traditional COP methods like Linear Programming (LP) or Constraint Programming (CP), can model dynamic uncertainties.

The existing research summary on RL-based DJSSP is presented in Table 3.

Table 3. Studies on RL-based DJSSP

References	Machine environment	Objective functions	Algorithms	Uncertainties
[8]	JSSP	Makespan	Actor-critic	Job order and processing time
[90]	JSSP	Robustness to processing time	DQN	Random processing time (RPT)
[91]	JSSP	Lead time	DQN	Machine Failure (MF)
[92]	FMS	Makespan	PNC & DQN	No
[93]	JSSP	Makespan	DQN	Random Job Arrival (RJA)
[85]	FJSSP	Total tardiness	DQN	New job insertion
[94]	FJSSP	Makespan	Q-learning	RPT
[74]	JSSP	Mean flow time	Q-learning	RJA & MF
[87]	FJSSP	Makespan	DQN	RJA

In dynamic JSSP multi-agent configuration, a Markov property which is considered a precondition for convergence will fail because of the independent updating policy by each agent [93]. However, integrating the whole JSSP into a single agent helps to avoid multi-agent interference with each other and convergence to local optimum. As a result, it has the advantage of stability.

2.6.4. Q-learning

Q-learning is characterized as an off-policy method and with its early convergence behavior [82]. In Q-learning, there are different No.s: the learning process could result in a local optimum solution or it could take longer to succeed and generalization problem [86]. Similarly, the presence of a large number of environmental states limits the accuracy of the applied RL approach [95]. SARSA and Q-learning are model-free Temporal Difference (TD) algorithms. In SARSA, the action is chosen at random with a probability, while in Q-learning, the action is the one that increases the value. That means, Q-learning greedily learns state-action value without looking at the policy [96]. If the environment is entirely observable, the DP approach could be used to infer optimum policy. However, usually, it is unknown, and no precise understanding of the environment exists. Under these scenarios, RL finds the optimum strategy using an iterative process [82].

One of the main challenges of Q-learning in scheduling is its limitation on continuous state space. In the practical industrial environment, where there is a continuous state feature, the total number of states is potentially infinite, establishing a massive Q table is unrealistic [8].

2.6.5. Deep Reinforcement Learning

Deep Reinforcement Learning (DRL) has been recently applied from Traveling Salesman Problem (TSP) in a graph optimization to Satisfiability problem [97]. DRL solutions to scheduling problems, on the other hand, are more recent and limited. DRL has the following features which are suitable for intelligent scheduling: (1) Ability to communicate with its surroundings and utilization of feedback data to optimize its strategy. (2) DRL, like different machine learning algorithms, requires intensive offline training; however, becomes efficient while executed. (3) The synchronization of DNN parameters takes advantage of the scheduling policy consistency between the simulation model and a real factory [92]. Solving dynamic scheduling problems requires the environment to satisfy the MDP requirement.

In DRL, the neural network is used to pick a candidate action. The main advantage of DRL is the ability to demonstrate the complex model in a comparatively simple manner than RL [84]. Furthermore, the agent learns the optimal strategy by trial and error, and this strategy helps the agent to decide in a dynamic environment.

2.6.6. Deep Q Network

When DNN is utilized to approximate the Q-value, it is known as a Deep Q Network (DQN). The problem of RL is its inability to converge because of the correlation between the expected value and Q-value [84]. DQN uses experience replay memory, which stores encountered data, to choose the data at random during learning to eliminate the correlation. The target network's weight is also iteratively updated for optimal convergence of the anticipated Q-value. The only distinction between DQN and Q-learning is; that in DQN, the agent's brain is DNN, whereas, in Q-learning, it is Q-table.

In complex job shop settings, breaking down and adjusting global objectives to local Key Performance Indicators (KPIs) is difficult. The DQN agent optimizes globally rather than locally. This implies that manually breaking down production objectives is not essential [89]. Regardless of its benefits, DQN has also drawbacks. First, training is time-consuming. Second, due to the black-box nature of the neural network, it is difficult to anticipate how DQN agents will behave in uncertain situations. It is also incapable to deal with continuous action spaces. Because of the continuous nature of the training process, each agent's policy changes regularly. This prevents straightforward usage of experience replay, which enables DQN to learn stability. Moreover, Deep Policy Gradient (DPG) also suffers from high variance [98]. Experience replay is the agent's huge experience data pool in which the experience, at each step, will be stored [8].

2.6.7. Drawbacks of DRL

There are two types of RL: model-free and model-based. The latter forecast the future state and understand the entire MDP transition model. Conversely, the majority of JSSP states are usually huge, if not infinite, which makes it impossible to understand the entire changing scenario [8].

The major challenge in model-free DRL is the absence of robustness when the environment changes [99]. It has less potential for reacting to huge environmental uncertainty. These constraints can be solved by retraining the model on the different distributions before deployment. Likewise, combining this approach with a model-based strategy enables solving the problem. Furthermore, DRL involves a large sample, which could be obtained by interacting with the simulator, to learn an optimum strategy. This simulation model should be efficient, but usually hard to construct.

In the case of Policy Gradients (PG) methods, there is no guarantee of optimality [99]. The best strategy to train model-free algorithms is by making a deliberate mistake early on and then determining which actions result in the maximum long-term rewards using a series of Monte Carlo simulations of the scheduling environment. Designing a reward is also a big challenge in DRL [8].

2.7. Scheduling Technology and Tools

The emergence of CPS has led to the development of digital twin technology. Digital twins are a digital copy of the physical machine and are modeled based on different dimensions such as geometric, physical, behavioral, rule, and data modeling.

2.7.1. Digital Twin

Digital Twin (DT) was first introduced by prof. Michael Grieves in 2003 [100]. The main idea of DT is the realization of interoperability and interconnection between virtual and physical elements of the shop floor [30]. There is no commonly agreed definition of a digital twin. However, the general definition of DT is a simulation model of a real-world system that is linked to a physical twin [101]. This linkage aids in the collection of actual data for simulation, and forward responses to the physical environment to fine-tune the behavior of the actual component [102].

DT can be used in a variety of settings, including production and manufacturing processes [31], and in all stages of product lifecycle [102], digital product development, process planning, lean manufacturing, construction of smart cities, energy, and mining solutions [103]. Nevertheless, it is not yet extensively implemented in the production stage [102]. The main advantage of using DT includes a reflection of the real-time working process and direction for the subsequent operational process of the physical model. Apart from simulation, DT is used to showcase unknown problems by predictions [103]. DT enables cyber-physical integration and real-time management between physical objects and digital representation [20].

DT is composed of four levels i.e., geometry, physics, behavior, and rules [104], and it helps not only to show the dynamic and geometric features but also to define the physical attributes and rules [102]. Using DT in production has also some challenges. To monitor composite twin data and extract insights it represents, an effective technique is required [31]. In addition to this, it is time-consuming and costly and requires experts in different areas, for the construction of complete and detailed DT [29]. Accurate and highly efficient communication between physical and digital spaces is needed [101]. Moreover, security No.s are also a critical component that needs to be considered before applying it to a larger scale.

In previous research, the digital twin has been used to assist with a scheduling problems. Machine failure detection and performance evaluation [27], [29], analysis of transportation and production processing stages [60], process simulation and production scheduling [17], and production scheduling for defense weapon systems [83] are among the studies. The reason why the simulation package becomes better than the stochastic Petri-net package is, because of its convenience, timely, and easier to operate nature [102].

Based on the analysis, existing manufacturing paradigms have the following limitations. Interconnection between physical machines and virtual models, the interconnection between the virtual model and physical production, generation of accurate data by converging the data from virtual and physical spaces, a realization of intelligent production simulation and optimization [30]; and lack of consideration of actual transportation condition in shop floors [60] are among the challenges. Most of the existing studies on digital twins focus on individual machines [83] and remain a challenge on how to construct and when to apply them on the shop floor [102].

2.7.2. Petri Net

A Petri net (place/transition net) is a directed connected graph that represents a finite set of arcs and used as a tool for process transitions. Topologically structured graphs or nets, which can represent regulations and connections are more capable of modeling production processes than standard tensors [92]. It is a popular method of process modeling not for searching for optimal scheduling. For example, in heuristic strategies, Petri nets design the manufacturing process, while heuristic rules focus on resolving scheduling conflicts [95]. The main drawback of PN-based scheduling is state space explosion.

4. DISCUSSION

In the era of smart manufacturing, a vast amount of data is being generated from different smart products and resources, which always provide feedback about their status to the system. Despite the extraction of the enormous amount of data, machine interoperability between shop floor environments is still a challenge.

The future of IoT objects will be standardized towards everything-as-a-service, which will bring better interoperability, re-usability, lower complexity, and higher scalability options. However, it will also incur high costs, have a lack of standards, lack of knowledge, and other limitations. The research findings based on the expert ideas in [105] show that service-oriented architecture will be the core component of smart manufacturing. So, this will help to solve the interoperability problems.

The most challenges found in the study are shop floor environment challenges related to CPS and handling of large amounts of information in adaptive manufacturing, machine pro-activeness (suggesting changes by themselves) and scheduling, decentralized and flexible decision making, human-robot collaboration, and constant evolution of new technologies.

Moreover, the choice of algorithm for the industrial environment is still vague. Usually, academic research algorithms' performance is evaluated with existing algorithms on the same setting, parameters, and constraints. This strategy will not help to implement the solution in the real environment. Algorithmic scheduling solutions should be evaluated not only with the existing algorithm but also with the Key Performance Indicators (KPI) of the particular factory.

However, if the solution has to deal with the real industrial environment, then adaptive scheduling such as RL and DRL can be the best fit. Dynamic Programming (DP) operates in fully observable MDP. In other words, DP can only be applied in environments with fully known transition probability. But in the real world, it is difficult to anticipate the entire environment and it is also computationally expensive. Similarly, Monte Carlo (MC) methods cannot be applied in an expensive critical industrial environment. The backup or update is performed at the terminal state. To update the value function, this approach waits for something to happen. In this case, if the machine is broken down, or if it explodes, it is difficult to reverse the initial working state.

Multi-agent Deep Reinforcement Learning (MADRL) scheduling algorithms are used to deal with dynamic uncertainty and a huge environment. However, the social dilemma is the main challenge to implement the solution. In another word, if each agent is competing with each other in a multi-agent environment, then they will waste resources. So, to make them synchronized and achieve a common goal, an appropriate reward function is needed. In MADRL, crafting a reward function is the most difficult task.

5. CONCLUSION AND FUTURE WORK

Scheduling tasks requires a comprehensive accounting of jobs and resources which are available with possible limitations in their use [17]. Scheduling problems are not only NP-hard but also computationally difficult combinatorial problems.

The common bottlenecks in dynamic scheduling include prediction of machine availability, disruption detection, and performance evaluation [29], [31]. Dynamic events and uncertainties are the main cause of scheduling performance deterioration and production disruption. The widely used approach of disturbance detection is, setting predefined constraints as a benchmark to evaluate the change between actual production and the anticipated plan. However, manufacturing states always change with time so the predefined benchmarks cannot correctly visualize currently anticipated production states. The other limitation of existing dynamic scheduling research is that dynamic events are considered from direct assumptions or derived by statistics rather than actual production data. As a result, it fails to provide interactive feedback and is limited in solving real-time problems.

Smart manufacturing system usually fails to achieve the desired objective because of non-reasonable design [106]. Incorporating AI techniques with a digital twin-based design approach can be a solution to such problems. In the majority of existing scheduling solutions, the machine states are modeled as a binary state i.e., up or down. However, it could also be interesting to consider the rate of machine performance degradation and the time to go to an intermediate state before its failure.

States in a job shop environment are infinite. As a result, applying model-based RL methods that know the entire MDP transition model is not recommended. An infinite number of states makes it difficult to understand the entire transition situation. Moreover, the challenging issues in model-based RL scheduling is the exhaustive computation of Q values. When the number of machines and jobs is more than twenty, the agent will find it hard

to find an optimum policy and difficult to converge to the global optimum. Because the value has to be computed for every possible state. However, improving the policy directly using the policy-based approach leads to convergence and an optimum policy.

6. REFERENCES:

[1] N. Stricker, A. Kuhnle, R. Sturm, S. Friess, "Reinforcement learning for adaptive order dispatching in the semiconductor industry", CIRP Annals, Vol. 67, No. 1, 2018, pp. 511-514.

[2] Y.-R. Shiue, K.-C. Lee, C.-T. Su, "Real-time scheduling for a smart factory using a reinforcement learning approach", Computers & Industrial Engineering, Vol. 125, 2018, pp. 604-614.

[3] T. J. Ikonen, K. Heljanko, I. Harjunkoski, "Reinforcement learning of adaptive online rescheduling timing and computing time allocation", Computers & Chemical Engineering, Vol. 141, 2020., p. 106994

[4] H. Hu, X. Jia, Q. He, S. Fu, K. Liu, "Deep reinforcement learning based AGVs real-time scheduling with mixed rule for flexible shop floor in industry 4.0", Computers & Industrial Engineering, Vol. 149, 2020, p. 106749.

[5] L. Li et al. "Bilevel Learning Model Towards Industrial Scheduling", arXiv Prepr. arXiv2008.04130, 2020.

[6] Y. Du, T. Wang, B. Xin, L. Wang, Y. Chen, L. Xing, "A data-driven parallel scheduling approach for multiple agile earth observation satellites", IEEE Transactions on Evolutionary Computation, Vol. 24, No. 4, 2019, pp. 679-693.

[7] H. Lu, X. Zhang, S. Yang, "A learning-based iterative method for solving vehicle routing problems", Proceedings of the International Conference on Learning Representations, 2019.

[8] C. L. Liu, C. C. Chang, C. J. Tseng, "Actor-critic deep reinforcement learning for solving job shop scheduling problems", IEEE Access, Vol. 8, 2020, pp. 71752-71762.

[9] A. Kuhnle, L. Schäfer, N. Stricker, G. Lanza, "Design, implementation and evaluation of reinforcement learning for an adaptive order dispatching in job shop manufacturing systems", Procedia CIRP, Vol. 81, 2019, pp. 234-239.

[10] Y. Liu, L. Zhang, L. Wang, Y. Xiao, X. Xu, M. Wang, "A framework for scheduling in cloud manufacturing with deep reinforcement learning", Proceedings of the IEEE 17th International Conference on Industrial Informatics, 2019, Vol. 1, pp. 1775-1780.

[11] H. Zhu, M. Li, Y. Tang, Y. Sun, "A deep-reinforcement-learning-based optimization approach for real-time scheduling in cloud manufacturing", IEEE Access, Vol. 8, pp. 9987-9997, 2020.

[12] S. Mittal, M. A. Khan, D. Romero, T. Wuest, "Smart manufacturing: Characteristics, technologies and enabling factors", Proceedings of the Institution of Mechanical Engineers, Part B, Vol. 233, No. 5, pp. 1342-1361, 2019.

[13] J. Davis et al. "Smart manufacturing", Annual Review of Chemical and Biomolecular Engineering, Vol. 6, 2015, pp. 141-160.

[14] P. O'Donovan, K. Leahy, K. Bruton, D. T. J. O'Sullivan, "An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities", Journal of Big Data, Vol. 2, No. 1, 2015, pp. 1-26.

[15] J. Wang, Y. Ma, L. Zhang, R. X. Gao, D. Wu, "Deep learning for smart manufacturing: Methods and applications", Journal of Manufacturing Systems, Vol. 48, 2018, pp. 144-156.

[16] F. Tao and Q. Qi, "New IT driven service-oriented smart manufacturing: framework and characteristics", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 49, No. 1, 2017, pp. 81-91.

[17] A. Koulouris, N. Misailidis, D. Petrides, "Applications of process and digital twin models for production simulation and scheduling in the manufacturing of food ingredients and products", Food and Bioproducts Processing, Vol. 126, 2021, pp. 317-333.

[18] H. S. Kang et al. "Smart manufacturing: Past research, present findings, future directions", International Journal of Precision Engineering and Manufacturing-Green Technology, Vol. 3, No. 1, 2016, pp. 111-128.

[19] A. Kusiak, "Smart manufacturing",International Journal of Production Research, Vol. 56, No. 1-2, 2018, pp. 508-517.

[20] Q. Qi, F. Tao, "Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison", IEEE Access, Vol. 6, 2018, pp. 3585-3593.

[21] F. Tao, Q. Qi, A. Liu, A. Kusiak, "Data-driven smart manufacturing", Journal of Manufacturing Systems, Vol. 48, 2018, pp. 157-169.

[22] S. Manufacturing and others, "Implementing 21st Century Smart Manufacturing", Smart Manufacturing Leadership Coalition, 2011.

[23] P. Zheng et al. "Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, future perspectives", Frontiers in Mechanical Engineering, Vol. 13, No. 2, 2018, pp. 137-150.

[24] J. Lee, E. Lapira, B. Bagheri, H. Kao, "Recent advances and trends in predictive manufacturing systems in big data environment", Manufacturing Letters, Vol. 1, No. 1, 2013, pp. 38-41.

[25] M. Zandieh, S. M. Sajadi, R. Behnoud, "Integrated production scheduling and maintenance planning in a hybrid flow shop system: a multi-objective approach", International Journal of System Assurance Engineering and Management, Vol. 8, No. 2, 2017, pp. 1630-1642.

[26] C. Zhang, Y. Zhou, K. Peng, X. Li, K. Lian, S. Zhang, "Dynamic flexible job shop scheduling method based on improved gene expression programming", Measurement and Control, Vol. 54, No. 7-8, 2021, pp. 1136-1146.

[27] E. Negri, V. Pandhare, L. Cattaneo, J. Singh, M. Macchi, J. Lee, "Field-synchronized Digital Twin framework for production scheduling with uncertainty", Journal of Intelligent Manufacturing, Vol. 32, No. 4, 2021, pp. 1207-1228.

[28] Y. Wang, Z. Wu, "Model construction of planning and scheduling system based on digital twin", The International Journal of Advanced Manufacturing Technology, Vol. 109, No. 7-8, 2020, pp. 2189-2203.

[29] M. Zhang, F. Tao, A. Y. C. Nee, "Digital twin enhanced dynamic job-shop scheduling", Journal of Manufacturing Systems, Vol. 58, 2021, pp. 146-156.

[30] H. Zhang, G. Zhang, Q. Yan, "Digital twin-driven cyber-physical production system towards smart shop-floor", Journal of Ambient Intelligence and Humanized Computing, Vol. 10, No. 11, 2019, pp. 4439-4453.

[31] Z. Liu, W. Chen, C. Zhang, C. Yang, Q. Cheng, "Intelligent scheduling of a feature-process-machine tool supernetwork based on digital twin workshop", Journal of Manufacturing Systems, Vol. 58, 2021, pp. 157-167.

[32] R. Boufellouh, F. Belkaid, "Bi-objective optimization algorithms for joint production and maintenance scheduling under a global resource constraint: Application to the permutation flow shop problem", Computers & Operations Research, Vol. 122, 2020, p. 104943.

[33] A. Branda, D. Castellano, G. Guizzi, V. Popolo, "Metaheuristics for the flow shop scheduling problem with maintenance activities integrated", Computers & Industrial Engineering, Vol. 151, 2021., p. 106989

[34] H. Ye, X. Wang, K. Liu, "Adaptive Preventive Maintenance for Flow Shop Scheduling With Resumable Processing", IEEE Transactions on Automation Science and Engineering, Vol. 18, No. 1, 2021., pp. 106-113

[35] W. Cui, Z. Lu, C. Li, X. Han, "A proactive approach to solve integrated production scheduling and maintenance planning problem in flow shops", Computers & Industrial Engineering, Vol. 115, 2018, pp. 342-353.

[36] C. Anand Deva Durai, M. Azath, J. S. C. Jeniffer, "Integrated Search Method for Flexible Job Shop Scheduling Problem Using HHS--ALNS Algorithm", SN Computer Science, Vol. 1, No. 2, 2020, pp. 1-6.

[37] X. Wu, J. Peng, X. Xiao, S. Wu, "An effective approach for the dual-resource flexible job shop scheduling problem considering loading and unloading", Journal of Intelligent Manufacturing, Vol. 32, No. 3, 2021, pp. 707-728.

[38] W. Kubiak, Y. Feng, G. Li, S. P. Sethi, C. Sriskandarajah, "Efficient algorithms for flexible job shop scheduling with parallel machines", Naval Research Logistics, Vol. 67, No. 4, 2020, pp. 272-288.

[39] R. Li, H. Ma, "Integrating preventive maintenance planning and production scheduling under reen-

trant job shop", Mathematical Problems in Engineering, Vol. 2017, 2017.

[40] M. Ghaleb, H. Zolfagharinia, S. Taghipour, "Real-time production scheduling in the Industry-4.0 context: Addressing uncertainties in job arrivals and machine breakdowns", Computers & Operations Research, Vol. 123, 2020, p. 105031.

[41] O. J. Shukla, G. Soni, R. Kumar, A. Sujil, "An agent-based architecture for production scheduling in dynamic job-shop manufacturing system", Automatisierungstechnik, Vol. 66, No. 6, 2018, pp. 492-502.

[42] M. Abderrahim, A. Bekrar, D. Trentesaux, N. Aissani, K. Bouamrane, "Bi-local search based variable neighborhood search for job-shop scheduling problem with transport constraints", Optimization Letters, 2020, pp. 1-26.

[43] B. Grosch, T. Kohne, M. Weigold, "Multi-objective hybrid genetic algorithm for energy adaptive production scheduling in job shops", Procedia CIRP, Vol. 98, 2021, pp. 294-299.

[44] W.-R. Jong, H.-T. Chen, Y.-H. Lin, Y.-W. Chen, T.-C. Li, "The multi-layered job-shop automatic scheduling system of mould manufacturing for Industry 3.5", Computers & Industrial Engineering, Vol. 149, 2020, p. 106797.

[45] E. K. A. Pakpahan, S. Kristina, A. Setiawan, "Proposed algorithm to improve job shop production scheduling using ant colony optimization method", in IOP Conference Series: Materials Science and Engineering, Vol. 277, No. 1, 2017, p. 12050.

[46] R. Zarrouk, I. E. Bennour, A. Jemai, "A two-level particle swarm optimization algorithm for the flexible job shop scheduling problem", Swarm Intelligence, Vol. 13, No. 2, 2019, pp. 145-168.

[47] X. Chen, J. Li, Y. Han, H. Sang, "Improved artificial immune algorithm for the flexible job shop problem with transportation time", Measurement and Control, Vol. 53, No. 9-10, 2020, pp. 2111-2128.

[48] S. Zhang, S. Wang, "Flexible assembly job-shop scheduling with sequence-dependent setup times and part sharing in a dynamic environment: Constraint programming model, mixed-integer programming model, dispatching rules", IEEE Transactions on Engineering Management, Vol. 65, No. 3, 2018, pp. 487-504.

[49] S. Tian, T. Wang, L. Zhang, X. Wu, "An energy-efficient scheduling approach for flexible job shop problem in an Internet of manufacturing things environment", IEEE Access, Vol. 7, 2019, pp. 62695-62704.

[50] A. Baykasouglu, F. S. Madenouglu, A. Hamzadayi, A. Baykasoğlu, F. S. Madenoğlu, A. Hamzaday, "Greedy randomized adaptive search for dynamic flexible job-shop scheduling", Journal of Manufacturing Systems, Vol. 56, 2020, pp. 425-451.

[51] X. Wu, X. Shen, C. Li, "The flexible job-shop scheduling problem considering deterioration effect and energy consumption simultaneously", Computers & Industrial Engineering, Vol. 135, 2019, pp. 1004-1024.

[52] R. H. Caldeira, A. Gnanavelbabu, T. Vaidyanathan, "An effective backtracking search algorithm for multi-objective flexible job shop scheduling considering new job arrivals and energy consumption", Computers & Industrial Engineering, Vol. 149, 2020, p. 106863.

[53] M. Dai, D. Tang, A. Giret, M. A. Salido, "Multi-objective optimization for energy-efficient flexible job shop scheduling problem with transportation constraints", Robotics and Computer-Integrated Manufacturing, Vol. 59, 2019, pp. 143-157.

[54] B. Denkena, F. Schinkel, J. Pirnay, S. Wilmsmeier, "Quantum algorithms for process parallel flexible job shop scheduling", CIRP Journal of Manufacturing Science and Technology, Vol. 33, 2021, pp. 100-114.

[55] J. L. Andrade-Pineda, D. Canca, P. L. Gonzalez-R, M. Calle, "Scheduling a dual-resource flexible job shop with makespan and due date-related criteria", Annals of Operations Research, Vol. 291, No. 1, 2020, pp. 5-35.

[56] D. C. Bissoli, N. Zufferey, A. R. S. Amaral, "Lexicographic optimization-based clustering search metaheuristic for the multiobjective flexible job shop scheduling problem", International Transactions in Operational Research, Vol. 28, No. 5, 2021, pp. 2733-2758.

[57] R. H. Caldeira, A. Gnanavelbabu, "A simheuristic approach for the flexible job shop scheduling problem with stochastic processing times", Simulation, Vol. 97, No. 3, 2021, pp. 215-236.

[58] S. L. Aquinaldo, N. R. Cucuk, others, "Optimization in job shop scheduling problem using Genetic Algorithm (study case in furniture industry)", IOP Conference Series: Materials Science and Engineering, Vol. 1072, No. 1, 2021, p. 12019.

[59] J. M. Novas, "Production scheduling and lot streaming at flexible job-shops environments using constraint programming", Computers & Industrial Engineering, Vol. 136, 2019, pp. 252-264.

[60] J. Yan, Z. Liu, C. Zhang, T. Zhang, Y. Zhang, C. Yang, "Research on flexible job shop scheduling under finite transportation conditions for digital twin workshop", Robotics and Computer-Integrated Manufacturing, Vol. 72, No. June, 2021, p. 102198.

[61] Y. Fu, M. Zhou, X. Guo, L. Qi, "Scheduling dual-objective stochastic hybrid flow shop with deteriorating jobs via bi-population evolutionary algorithm", IEEE Transactions on Systems, Man, and Cybernetics, Vol. 50, No. 12, 2019, pp. 5037-5048.

[62] Z. Zhang, Q. Tang, M. Chica, "Maintenance costs and makespan minimization for assembly permutation flow shop scheduling by considering preventive and corrective maintenance", Journal of Manufacturing Systems, Vol. 59, 2021, pp. 549-564.

[63] S. Aghighi, S. T. A. Niaki, E. Mehdizadeh, A. A. Najafi, "Open-shop production scheduling with reverse flows", Computers & Industrial Engineering, Vol. 153, 2021, p. 107077.

[64] T.-S. Yu, M. Pinedo, "Flow shops with reentry: Reversibility properties and makespan optimal schedules", European Journal of Operational Research, Vol. 282, No. 2, 2020, pp. 478-490.

[65] Y. Fu, M. C. Zhou, X. Guo, L. Qi, "Stochastic multi-objective integrated disassembly-reprocessing-reassembly scheduling via fruit fly optimization algorithm", Journal of Cleaner Production, Vol. 278, 2021, p. 123364.

[66] J. Seif, M. Dehghanimohammadabadi, A. J. Yu, "Integrated preventive maintenance and flow shop scheduling under uncertainty", Flexible Services and Manufacturing Journal, Vol. 32, No. 4, 2020, pp. 852-887.

[67] P. P. Suryadhini, S. Sukoyo, S. Suprayogi, A. H. Halim, "A batch scheduling model for a three-stage flow shop with job and batch processors considering a sampling inspection to minimize expected total actual flow time", Journal of Industrial Engineering and Management, Vol. 14, No. 3, 2021, pp. 520-537.

[68] J. Bautista-Valhondo, R. Alfaro-Pozo, "Mixed integer linear programming models for Flow Shop Scheduling with a demand plan of job types", Cent. European Journal of Operational Research, Vol. 28, No. 1, 2020, pp. 5-23.

[69] J. Krishnaraj and S. Thiagarajan, "A two-phase simulated annealing algorithm to minimise the completion time variance of jobs in flowshops", International Journal of Process Management and Benchmarking, Vol. 10, No. 2, 2020, pp. 261-281.

[70] W. Li, D. Han, L. Gao, X. Li, Y. Li, "Integrated production and transportation scheduling method in hybrid flow shop", Chinese Journal of Mechanical Engineering, Vol. 35, No. 1, 2022, pp. 1-20.

[71] L. Sun, L. Lin, M. Gen, H. Li, "A hybrid cooperative coevolution algorithm for fuzzy flexible job shop scheduling", IEEE Transactions on Fuzzy Systems, Vol. 27, No. 5, 2019, pp. 1008-1022.

[72] A. Al-Shayea, E. Fararah, E. A. Nasr, H. A. Mahmoud, "Model for Integrating Production Scheduling and Maintenance Planning of Flow Shop Production System", IEEE Access, Vol. 8, 2020, pp. 208826-208835.

[73] W. Bouazza, Y. Sallez, B. Beldjilali, "A distributed approach solving partially flexible job-shop scheduling problem with a Q-learning effect", IFAC-Papers OnLine, Vol. 50, No. 1, 2017, pp. 15890-15895.

[74] J. Shahrabi, M. A. Adibi, M. Mahootchi, "A reinforcement learning approach to parameter estimation in dynamic job shop scheduling", Computers & Industrial Engineering, Vol. 110, 2017, pp. 75-82.

[75] H. Zhu, M. Chen, Z. Zhang, D. Tang, "An adaptive real-time scheduling method for flexible job shop scheduling problem with combined processing

constraint", IEEE Access, Vol. 7, 2019, pp. 125113-125121.

[76] S. Karnouskos, P. Leitao, "Key contributing factors to the acceptance of agents in industrial environments", IEEE Transactions on Industrial Informatics, Vol. 13, No. 2, 2016, pp. 696-703.

[77] M. Owliya, M. Saadat, R. Anane, M. Goharian, "A new agents-based model for dynamic job allocation in manufacturing shopfloors", IEEE Systems Journal, Vol. 6, No. 2, 2012, pp. 353-361.

[78] C.-J. Huang, L.-M. Liao, "A multi-agent-based negotiation approach for parallel machine scheduling with multi-objectives in an electro-etching process", International Journal of Production Research, Vol. 50, No. 20, 2012, pp. 5719-5733.

[79] S. Wang, J. Wan, D. Zhang, D. Li, C. Zhang, "Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination", Computer Networks, Vol. 101, 2016, pp. 158-168.

[80] Y. G. Kim, S. Lee, J. Son, H. Bae, B. Do Chung, "Multi-agent system and reinforcement learning approach for distributed intelligence in a flexible smart manufacturing system", Journal of Manufacturing Systems, Vol. 57, 2020, pp. 440-450.

[81] L. Monostori, J. Váncza, S. R. T. Kumara, "Agent-based systems for manufacturing", CIRP Annals, Vol. 55, No. 2, 2006., pp. 697-720

[82] M. H. Moghadam, S. M. Babamir, "Makespan reduction for dynamic workloads in cluster-based data grids using reinforcement-learning based scheduling", Vol. 24, 2018, pp. 402-412.

[83] H. Latif, B. Starly, "A simulation algorithm of a digital twin for manual assembly process", Procedia Manufacturing, Vol. 48, No. 2019, 2020, pp. 932-939.

[84] S. Lee, Y. Cho, Y. H. Lee, "Injection Mold Production Sustainable Scheduling Using Deep Reinforcement Learning", Sustainability, Vol. 12, No. 20, 2020, p. 8718.

[85] S. Luo, "Dynamic scheduling for flexible job shop with new job insertions by deep reinforcement learning", Applied Soft Computing, Vol. 91, 2020, p. 106208.

[86] M. E. Aydin, E. Öztemel, "Dynamic job-shop scheduling using reinforcement learning agents", Robotics and autonomous systems, Vol. 33, No. 2-3, 2000, pp. 169-178.

[87] L. Zhou, L. Zhang, B. K. P. Horn, "Deep reinforcement learning-based dynamic scheduling in smart manufacturing", Procedia CIRP, Vol. 93, 2020, pp. 383-388.

[88] M. S. A. Hameed, A. Schwung, "Reinforcement learning on job shop scheduling problems using graph networks", arXiv Prepr. arXiv2009.03836, 2020.

[89] B. Waschneck et al. "Deep reinforcement learning for semiconductor production scheduling", Institute of Electrical and Electronics Engineers Inc., 2018, pp. 301-306.

[90] D. Shi, W. Fan, Y. Xiao, T. Lin, C. Xing, "Intelligent scheduling of discrete automated production line via deep reinforcement learning", International Journal of Production Research, Vol. 58, No. 11, 2020, pp. 3362-3380.

[91] T. Altenmüller, T. Stüker, B. Waschneck, A. Kuhnle, G. Lanza, "Reinforcement learning for an intelligent and autonomous production control of complex job-shops under time constraints", Production Engineering, Vol. 14, 2020, pp. 319-328.

[92] L. Hu, Z. Liu, W. Hu, Y. Wang, J. Tan, F. Wu, "Petri-net-based dynamic scheduling of flexible manufacturing system via deep reinforcement learning with graph convolutional network", Journal of Manufacturing Systems, Vol. 55, 2020, pp. 1-14.

[93] B. Luo, S. Wang, B. Yang, L. Yi, "An improved deep reinforcement learning approach for the dynamic job shop scheduling problem with random job arrivals", Journal of Physics: Conference Series, Vol. 1848, No. 1, 2021, p. 12029.

[94] I.-B. Park, J. Huh, J. Kim, J. Park, "A reinforcement learning approach to robust scheduling of semiconductor manufacturing facilities", IEEE Transactions on Automation Science and Engineering, Vol. 17, No. 3, 2019, pp. 1420-1431.

[95] M. Drakaki, P. Tzionas, "Manufacturing Scheduling Using Colored Petri Nets and Reinforcement Learning", Applied Sciences, Vol. 7, No. 2, 2017, p. 136.

[96] N. Chauhan, N. Choudhary, K. George, A comparison of reinforcement learning based approaches to appliance scheduling. Institute of Electrical and Electronics Engineers Inc., 2016, pp. 253-258.

[97] P. Tassel, M. Gebser, K. Schekotihin, "A reinforcement learning environment for job-shop scheduling", arXiv Prepr. arXiv2104.03760, 2021.

[98] S. Li, Y. Wu, X. Cui, H. Dong, F. Fang, S. Russell, "Robust multi-agent reinforcement learning via minimax deep deterministic policy gradient", Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, No. 01, 2019, pp. 4213-4220.

[99] C. D. Hubbs, C. Li, N. V. Sahinidis, I. E. Grossmann, J. M. Wassick, "A deep reinforcement learning approach for chemical production scheduling", Computers & Chemical Engineering, Vol. 141, 2020, p. 106982.

[100] Y. Wang, Z. Wu, "Digital twin-based production scheduling system for heavy truck frame shop", Proceedings of the Institution of Mechanical Engineers, Part C, Vol. 236, No. 4, 2022, pp. 1931-1942.

[101] A. Villalonga et al. "A decision-making framework for dynamic scheduling of cyber-physical production systems based on digital twins", Annual Reviews in Control, 2021.

[102] C. Zhuang, T. Miao, J. Liu, H. Xiong, "The connotation of digital twin, the construction and application method of shop-floor digital twin", Robotics and Computer-Integrated Manufacturing, Vol. 68, 2021, p. 102075.

[103] H. Yu, S. Han, D. Yang, Z. Wang, W. Feng, "Job Shop Scheduling Based on Digital Twin Technology: A Survey and an Intelligent Platform", Complexity, Vol. 2021, 2021.

[104] F. Tao, M. Zhang, "Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing", IEEE Access, Vol. 5, 2017, pp. 20418-20427.

[105] M. Moghaddam, M. N. Cadavid, C. R. Kenley, A. V Deshmukh, "Reference architectures for smart manufacturing: A critical review", Journal of Manufacturing Systems, Vol. 49, 2018, pp. 215-225.

[106] Q. Liu et al. "Digital twin-based designing of the configuration, motion, control, optimization model of a flow-type smart manufacturing system", Journal of Manufacturing Systems, Vol. 58, 2021, pp. 52-64.

Genetic algorithm for the design and optimization of a shell and tube heat exchanger from a performance point of view

Original Scientific Paper

Mohammed Bakr
Director of the Department of Instrumentation and Control Laboratory, Delta Factory for Fertilizers and Chemical Industries,
Computers and Control Systems engineering Department, Faculty of Engineering, University Mansoura, Egypt, Mbakr.q@gmail.com

Ahmed A. Hegazi
Mechanical Power Engineering Department, Faculty of Engineering, Mansoura University,
El-Mansoura 35516, Egypt

Amira Y. Haikal
Computers and Control Systems Engineering Department, Faculty of Engineering,
Mansoura University, Mansoura 35516, Egypt,
amirayh@mans.edu.eg

Mostafa A. Elhosseini
Computers and Control Systems Engineering Department, Faculty of Engineering,
Mansoura University, Mansoura 35516, Egypt,
melhosseini@mans.edu.eg
College of Computer Science and Engineering,
Taibah University, Yanbu 46421, Saudi Arabia,
melhosseini@ieee.org

Abstract *– A new approach to optimize the design of a shell and tube heat exchanger (STHX) is developed via a genetic algorithm (GA) to get the optimal configuration from a performance point of view. The objective is to develop and test a model for optimizing the early design stage of the STHX and solve the design problem quickly. GA is implemented to maximize heat transfer rate while minimizing pressure drop. GA is applied to oil cooler type OKG 33/244, and the results are compared with the original data of the STHX. The simulation outcomes reveal that the STHX's operating performance has been improved, indicating that GA can be successfully employed for the design optimization of STHX from a performance standpoint. A maximum increase in the effectiveness achieves 57% using GA, while the achieved minimum increase is 47%. Furthermore, the average effectiveness of the heat exchanger is 55%, and the number of transfer units (NTU) has improved from 0.475319 to 1.825664 by using GA.*

Keywords*: Genetic algorithm, Optimization, Overall heat transfer coefficient, Shell and tube heat exchanger*

1. INTRODUCTION

The heat exchanger is a thermal medium that transfers heat between two or more fluids at different temperatures [1-3]. Heat exchangers are widely utilized in industrial applications such as chemical processing systems, waste heat recovery units, power plants, food processing systems, air conditioning systems, refrigeration, heating, and automobile radiators [2].

According to specific heat exchange requirements, various types of heat exchanger equipment such as casing and tube, bare tube, finned tube, spiral, plate, frame, and plate coil are used [4].

Among these heat exchangers, STHX is the most commonly used type due to its easy maintenance, application versatility, and resistance to high temperature and pressure [5-7]. This type comprises several round tubes mounted inside a cylindrical shell and has five major components [8]: the shell, tube bundle, front head, rear head, and the baffles [8,9]. The fluid enters and exits the tube side through the rear and front headers. Baffles support the tubes by increasing the turbulence of the shell fluid and directing the fluid flow to the tubes (approximately transversely), increasing the heat transfer intensity. Heat exchange occurs when one fluid flows outside the tubes, and the other fluid flows through the tubes [8].

Several geometric parameters determine the STHX performance [9,10], including the flow rate ratio between the tube and shell sides, the heat transfer coefficient on the shell and tube sides, the type and spacing of baffles, pressure drop, fouling, and turbulence. There are three common types of STHX as follows; STHX with segmental baffles (STHX-SG), STHX with continuous helical baffles (STHX-CH), STHX possessing

staggered baffles (STHX-ST). STHX-SG is most common and widely used because of its ease of installation and low cost. STHX-SG, on the other hand, offers high heat transfer performance due to its crossflow on the shell side. A type of STHX known as STHX-CH produces shell-side helical flow. The last one, STHX-ST used both continuous helical baffles and segmental; it has the convenience of segmental baffles in terms of fabrication and installation and the helical flow generated by helical baffles. Shell inner diameter, outer tube diameter, baffle spacing, baffle cut, and baffle orientation angle are all design parameters that substantially impact the overall performance of this heat exchanger [5].

The design of STHXs that meets a specified set of design constraints and provides the optimum heat duty includes many geometrical and operative variables [11].

An optimum heat exchanger configuration has been extensively applied with artificial intelligence (AI) methodology, particularly AI-based on metaheuristics. For the cost-effective design of STHX, GAs have been adopted as an optimization method to improve the design [9].

Many standards for STHX aimed to help designers, engineers, and users work more efficiently. Many producers and consumers widely use tubular exchanger manufacture association (TEMA) standards, covering manufacturing tolerances, thermal relationships, performance data, installation, maintenance and operation, vibration standards, mechanical standards, and recommended good practices [12].

Optimization of STHX has been conducted with metaheuristics and deterministic methods [13]. High-dimensional problems cannot be solved with an exact optimization algorithm. It is impossible to conduct a comprehensive search with the size of the problem because the search space grows exponentially with size. The population-based optimization algorithms can be used to find near-optimal solutions to difficult optimization problems. Metaheuristic algorithms are optimization methods based on a stochastic approach that can produce solutions with good and reliable approximations in a reasonable amount of time [16]. These approaches are one of the most complex computational intelligence models that greatly approximate optimization problems [13,17].

The objective function does not need to be differentiated for metaheuristics. Metaheuristics are more efficient than simple heuristics or calculus-based methods. As a result, they may be used to search over many solutions with less computational effort than traditional calculus-based methods [14,15]. However, algorithms of this sort are often constructed on disordered solving strategies based on random numbers rather than robust and accurate computations and hence may not always reach the global optimal point [16]. Although they have no guarantee of good performance, metaheuristic algorithms have been found to perform acceptably in many use cases [18,19].

Bio-inspired and physics/chemistry-based algorithms are the major divisions. Biogeography-based optimization (BBO), estimation of distribution algorithms, differential evolution (DE), (EDAs), and flower pollination algorithm (FPA) are mentioned as an example of the so-called bio-inspired algorithms. Some other algorithms are swarm intelligence-based, a subcategory of bio-inspired algorithms such as artificial bee colony (ABC), ant colony optimization (ACO), cuckoo search (CS), grey wolf optimizer (GWO), particle swarm optimization (PSO), and whale optimization algorithm (WOA). Simulated annealing (SA), big bang-big crunch (BBBC), and harmony search algorithm (HSA) are examples of physics/chemistry-based algorithms that were inspired by physical or chemical phenomena [13].

GA has successfully obtained optimal designs for STHE in several works, including [20, 21].

Selbas et al. [22] applied GA to optimize the STHX economically by varying the design variables: outer shell diameter, outer tube diameter, baffle cut, baffle spacing, number of tube passes, and tube layout. In addition, they determined the heat transfer area as an objective function using the logarithmic mean temperature difference (LMTD) method. They concluded that the heat transfer area increases as the total cost increases.

Antonio et al. [23] used GA in Toolbox to optimize a heat exchanger; the objective function is based on the heat exchanger's total cost. They compare their results to conventional approaches by reducing the objective function while considering decision variables such as tube diameter, casing diameter, and septum area. Compared with traditional methods, the results showed that the performance of the heat exchanger was improved.

Guo et al. [24] developed a new approach for STHX optimization design using entropy generation minimization and GA. The rate of dimensionless entropy generation was used as the objective function. A variety of design variables were taken into account. They found that the effectiveness of the STHX was significantly increased while pumping power was reduced simultaneously.

Patel et al. [25] investigated the optimization of STHXs from an economic viewpoint using PSO. They compared the optimization results to those obtained by the GA and found that the PSO algorithm outperforms the GA in terms of predictive performance.

Vahdat Azad and Amidpour [26] optimized STHX using a GA to lower the total cost of the heat exchanger. Although GAs, CS, and firefly algorithm (FA) were used by Khosravi et al. [27], they concluded that when GAs were implemented, it was impossible to find designs that met the constraints while FA could come up with good designs.

Dastmalchi et al. [28] investigated the PSO algorithm in a double pipe heat exchanger with finned tubes.

Their findings revealed that as the Reynolds number increased, the optimum height of the fin increased as well.

Saijal and Danish [5] designed the STHX-ST by incorporating helical and segmental baffles features. They investigated the influence of five design parameters through numerical analysis: outer tube diameter, inner shell diameter, septal orientation angle, septum cut-out, and septum spacing on STHX-ST performance. They implemented multi-objective optimization using GA, where the heat transfer rate is maximized while the pressure drop is minimized. They used artificial neural networks (ANNs) to approximate the optimization of the objective function. Using the computational fluid dynamics (CFD) and Taguchi orthogonal test table analysis, the training data for ANNs are generated. Finally, they provided the optimal design parameters for minimum pressure drop and maximum heat transfer rate.

There may be contradictions regarding the efficiency of using GA to optimize STHX because of the confusing relationships between optimizing STHXs economically and optimizing them from a performance standpoint.

The GA enables the design problem to be solved quickly and enables the designer to examine some high-quality alternative solutions, giving the designer more flexibility concerning traditional methods in his final selection [23].

Improving the heat exchange capacity of the heat exchanger used in the industry by improving its effectiveness to increase production capacity is a great challenge. Therefore, this study aims to improve the effectiveness of STHX, which already works for an industrial application, by using GA. We used GA to improve the design optimization of STHX from a performance point of view. MATLAB and the optimization toolbox of MATLAB are used to apply our mathematical model. The proposed algorithm is compared with the STHX data of each run to demonstrate the effectiveness and best points under each run.

The main contributions of this research paper are:

- Formulate a mathematical model for oil cooler type OKG 33/244 STHX
- Applying the GA to an industrial model of STHX (oil cooler type OKG 33/244).
- Improve the effectiveness of the oil cooler type OKG 33/244.
- Deciding the issue of whether or not GA can improve the effectiveness of STHX, from a performance point of view.

Following is the remainder of this paper; a description of the hydraulic thermal design formula of an STHX can be found in section 2, and an overview of the GA can be found in section 3. Next, section 4 describes the results and computational analysis, while the last section (section 5) provides the concluding remarks.

2. DESIGN FORMULATIONS OF A SHELL AND TUBE HEAT EXCHANGER

This section presents the equations used in the current study to calculate the heat transfer coefficients (HTC) of STHX and the objective function of the study.

2.1. SHELL AND TUBE HEAT TRANSFER COEFFICIENT

The convective HTC depends on the flow regime and the fluid velocity. The HTC for the flow in the tubes can be determined using several equations. Regarding the phenomenon of intra-tube flow and according to the pressure drop (PD) calculations and the flow regime in HTCs, the intra-tube flow is divided into transition, relaxation, and developed turbulence. The dimensionless Reynolds number in the mobility factor concept is the criterion for separating these three areas. The fluid acts as a barrier to its movement [29]. Laminar HTC is calculated using the Seider-Tate correlation described in [30,31]. Hausen correlation [32] is applied to transient conditions, whereas Dittus-Boelter correlation [29] is widely used to describe fully developed turbulent (turbulent area) flow conditions in tubes.

The heat transfer surface area, A, for the exchanger is firstly determined according to the following equations (1) to (6) [33]:

$$d_{t,o} = d_{t,i} + S_t \tag{1}$$

Calculate the number of tubes

$$\mathrm{n_t} = \frac{\mathrm{m_c^{\cdot}}}{\rho_c * \frac{\pi}{4} * \mathrm{d_{t,i}^2} * \mathcal{V}_t} \tag{2}$$

Calculate heat transfer surface area

$$\mathrm{A_s} = \pi * \mathrm{d_{t,o}} * \mathrm{l_t} * \mathrm{n_t} * \mathrm{n_p} \tag{3}$$

Calculate the Tube side Reynolds number

$$\mathrm{Re_t} = \frac{\rho_t \mathcal{V}_t \mathrm{d_{t,i}}}{\mu_t} \tag{4}$$

Calculate Darcy friction factor

$$\mathrm{f_t} = (1.82 \log_{10} \mathrm{Re_t} - 1.64)^{-2} \tag{5}$$

Calculate tube side convective coefficient

$$h_t = \begin{cases} \frac{k_t}{d_{t,i}} \left[3.657 + \frac{0.0677 \left(R_{et} P_{rt} \left(\frac{d_{t,i}}{l} \right) \right)^{1.33}}{1 + 0.1 P_{rt} \left(R_{et} \left(\frac{d_{t,i}}{l} \right) \right)^{0.3}} \right]; R_{et} < 2300 \\ \frac{k_t}{d_{t,i}} \left[\frac{\left(\frac{f_t}{8} \right) (R_{et} - 1000) P_{rt}}{1 + 12.7 \left(\frac{f_t}{8} \right)^{\frac{1}{2}} \left(P_{rt}^{\frac{2}{3}} - 1 \right)} \left(1 + \frac{d_{t,i}}{l} \right)^{0.67} \right]; 2300 < R_{et} < 10^3 \\ 0.027 \frac{k_t}{d_{t,i}} {R_{et}}^{0.8} {P_{rt}}^{1/3} \left(\frac{\mu_t}{\mu_{tw}} \right) 0.14 \; ; R_{et} > 10^4 \end{cases} \tag{6}$$

In the above equation (6), the coefficients are calculated; for laminar flow (R_{et}<2300), for transition flow (2300<R_{et}<10^3), for fully developed turbulent flow (R_{et}>10^4).

Where ft and kt are the Darcy friction factor and the tube side thermal conductivity, respectively. All unmentioned symbols are listed in Table (7) in the nomenclature table.

The hydraulic shell diameter D_e is computed as:

$$\mathrm{D_e} = \left(\frac{4 * \mathrm{A}}{\mathrm{P}}\right) \tag{7}$$

$$\mathrm{A} = \frac{\pi}{4}\mathrm{D}^2 - \frac{\pi}{4}\mathrm{d_{t,o}}^2\mathrm{n} \qquad \& \qquad \mathrm{P} = \pi \mathrm{d_{t,o}}\,\mathrm{n} + \pi\,\mathrm{D} \tag{8}$$

The fluid velocity inside the tube, Reynolds number, and Prandtl are calculated from the equations 9-10:

$$\mathrm{v_s} = \frac{\mathrm{m_h}}{\rho_\mathrm{s}\frac{\pi}{4}\mathrm{D}_\mathrm{e}^{\,2}\mathrm{n_t}} \tag{9}$$

$$\mathrm{Re_s} = \frac{\rho_\mathrm{s}\mathcal{V}_s\mathrm{D_e}}{\mu_\mathrm{s}} \tag{10}$$

The shell side heat transfer coefficient h_s is calculated using Kern's formula for STHX-SG [9].

$$h_s = 0.36\frac{k_t}{D_e}Re_s^{0.55}Pr_s^{1/3}\left(\frac{\mu_s}{\mu_{sw}}\right)^{0.14} \tag{11}$$

On both the shell and tube sides, the overall heat transfer coefficient (U) is calculated using the heat transfer coefficients and fouling resistances. Fouling resistances are calculated based on literature data for various fluid types and operating temperatures. The overall heat transfer coefficient is calculated using the formula [11]:

$$U = \frac{1}{\frac{1}{h_s} + R_{fs} + \left(\frac{d_{t,o}}{d_{t,i}}\right)\left(R_{ft} + \frac{1}{h_t}\right)} \tag{12}$$

The minimum and maximum thermal capacity, C_{min} and C_{max}, respectively, are defined as below,

$$C_c = m_c^{\cdot} * C_{p,c} \text{ and } C_h = m_h^{\cdot} * C_{p,h}\,, if \qquad C_c > C_h \tag{13}$$

$$C_{max} = C_c\,,\ C_{min} = C_h \quad or \quad C_{max} = C_h,\ \ C_{min} = C_c \tag{14}$$

2.2. PROBLEM DESIGN

In this study, the objective function for the design optimization issue is the thermal efficiency of the oil cooler type OKG 33/244 STHX (Fig.1) by varying the design variables: tube inside diameter($d_{t,i}$), tubes length (L), shell diameter (D), number of tubes (n_t), effectiveness (ε), and output temperature of hot fluid ($T_{h,o}$).

The effectiveness-number of transfer units (ε-NTU) method and the LMTD method are often used for heat exchanger design and analysis [34]. In heat exchanger analysis, LMTD is straightforward when both the outlet and inlet temperatures of the hot and cold fluids can be determined from the energy balance. In addition, it is great for determining the size of a heat exchanger to achieve the right outlet temperature.

On the other hand, NTU is a direct measure of the surface area of the heat transfer; consequently, the size of the heat exchanger is proportional to the NTU [35].

Fig.1. Oil cooler-Type: OKG 33/244

Accordingly, the ε-NTU approach is chosen in the proposed work. It can estimate outlet temperatures without requiring a numerical iterative solution to the nonlinear equations system.

The heat exchanger's size and heat transfer rate can be measured by the number of thermal units (NTU) using Eq. 15 [36]. All symbols are listed in table [7].

$$NTU = \frac{A_s * U}{C_{min}} \tag{15}$$

The heat capacity ratio C_r is measured according to Eq. 16 [37].

$$C_r = \frac{C_{min}}{C_{max}} \tag{16}$$

The effectiveness ($\mathcal{E}$) is calculated according to Eq. 17 [37].

$$\begin{aligned}\mathcal{E} = 2 \times \Bigg[1 + C_r + \left(1 + C_r^{\,2}\right)^{1/2} \\ \times \left[\frac{1 + exp\,\left(-NTU\left(1 + C_r^{\,2}\right)^{1/2}\right)}{1 - exp\,\left(-NTU\left(1 + C_r^{\,2}\right)^{1/2}\right)}\right]\Bigg]^{-1}\end{aligned} \tag{17}$$

The following equation gives the heat exchange rate between cold and hot currents [13].

$$\begin{aligned}T_{h,o} = T_{h,i} - \ \varepsilon\left(T_{h,i} - T_{c,i}\right) \\ and\ \ T_{c,o} = \varepsilon\left(T_{h,i} - T_{c,i}\right) + \ T_{c,i}\end{aligned} \tag{18}$$

Where C_{min}=C_h

The transfer rate (heat duty) (q_C=q_h) are calculated as follows [13] :

$$\begin{aligned}q_c = \ m_c^{\cdot} * C_{p,c} * \left(T_{c,o} - T_{c,i}\right) \\ \&\ \ q_h = m_h^{\cdot} * C_{p,h} * \left(T_{h,i} - T_{h,o}\right)\end{aligned} \tag{19}$$

$$T_{h,o} = -\frac{q_c}{m_h^{\cdot} * C_{p,h}} + T_{h,i} \tag{20}$$

$$T_{c,o} = \frac{q_h}{m_c^{\cdot} * C_{p,c}} + T_{c,i} \tag{21}$$

3. OPTIMIZATION TECHNIQUE

GAs are parameter search procedures for artificial systems based on the mechanics of natural genetics [38]. The GA methodology is used to solve optimization problems by performing a stochastic search of the solution space using strings of integers representing the optimized parameters, known as chromosomes. For these modeling applications, each integer within these chromosomes is referred to as a gene, and each gene has a decimal value between 0 and 9 [39]. They begin with a population, a collection of solutions (represented by chromosomes). Next, a population's solutions are taken and used to create a new population. This is driven by Darwinian survival of the fittest and a structured random exchange of information using reproduction, crossover, mutation, and permutation operators. First, solutions (parents) are chosen to create new solutions (offspring), which are chosen based on their fitness—the more fit they are, the better their chances of reproducing. This is repeated until certain conditions are met, such as the number of generations or the improvement of the best solution [38].

This paper uses the GA to solve the optimization design problem for the STHX with a single tube pass. The original design data and the original data for shell and tube are given in Table 1 and Table 2, respectively.

Table 1. Original design data of oil cooler type OKG 33/244 under study

Data of cold fluid		Data of hot fluid	
$T_{c,i}$	30 (C°)	$T_{h,i}$	60 (C°)
$T_{c,o}$	35 (C°)	$T_{h,o}$	45 (C°)
Cp_c	4/86 (kJ/kg k)	Cp_h	2035 (kJ/kg k)
k_t	0.613 (w/m k)	k_c	141 * 10^{-3}
m_c	245.6197 (kg/s)	m_h	169.0715 (kg/s)
ρ_t	1000 (kg/$m^{3)}$	ρ_t	865.8 (kg/m^3)
μ_t	855 * 10^{-6} Pa.s	μ_t	8.36 * 10^{-2} Pa.s
Pr_t	5.83	Pr_s	1205

Table 2. Original data for shell and tube

Data of cold fluid		Data of hot fluid	
n_t	420	D	0.62 (m)
$d_{t,i}$	0.019 (m)	n_p	1
$d_{t,o}$	0.0192 (m)		
S_t	0.0021 (m)		
L	3.050 (m)		
P_t	0.005 (m)		

The range of design variables of the geometric parameters is given in Table 3. The GA employs the Roulette Wheel Selection method. The probability of being chosen increases with increasing fitness. To create the offspring population, uniform crossover and random uniform mutation are used. With a probability of 0.85, the integer-based uniform crossover operator switches each corresponding binary bit between two different parents. After crossover, the mutation operator modifies each binary bit with a 0.01 mutation probability [40].

In MATLAB a first generation of 30 individuals is generated. There are three genes on each chromosome: tube inside diameter, shell diameter, and tube length. These three gens are binary coded. The range of each chromosome is shown in Table 3.

The roulette method is used for selection. According to most scholars, crossover and mutation probabilities are 0.85 and 0.01, respectively [40-41].

Table 3. The range of design variables of the geometric parameters

Data of input			
Variable	Original design	Range	
		From	to
$d_{t,i}$	0.015 (m)	0.011 (m)	0.02 (m)
D	0.62 (m)	0.58 (m)	0.67 (m)
L	3.050 (m)	2.65 (m)	3.55 (m)

3.1 CASE STUDY

The known information of the STHX, six design variables (tube inside diameter ($d_{t,i}$), tubes length L_t, shell diameter (D), the number of tubes (n_t), effectiveness (ε), and the output temperature of the hot fluid ($T_{h,o}$) ˚C) are selected and listed in Table (4).

Table 4. The six chosen variables to improve oil cooler type OKG 33/244 understudy.

Data of input		Data of output	
$d_{t,i}$	tube inside diameter (m)	n_t	number of tubes
L	tubes length (m)	ε	effectiveness
D	shell diameter (m)	$T_{h,o}$	Output temperature of hot fluid (C°)

The flowchart for the proposed implementation steps of GA solving the STHX design problem is shown in Fig. 2 as follows:

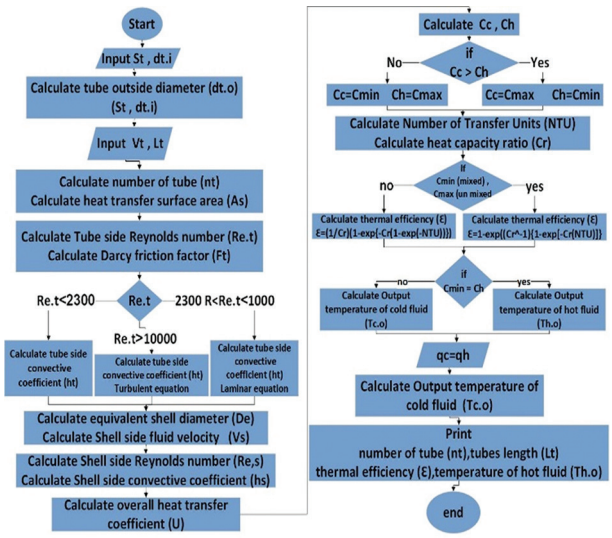


Fig.2. Flowchart for the proposed steps of GA solving design problem of STHX

The proposed implementation steps of GA solving the STHX design problem can be summarized as shown in Algorithm1:

Algorithm1:

Start

Input: m_c, $d_{t,i}$, S_t, Rf_t, n_p, Rf_s.

1. Calculate tube outside diameter $dt,o=dt,i+S_t$
2. Calculate the number of tubes $n_t = \frac{\dot{m_c}}{\rho_c * \frac{\pi}{4} * d_{t,i}^2 * \mathcal{V}_t}$
3. Calculate heat transfer surface area of tubes $A_s = \pi * d_{t,o} * l_t * n_t * n_p$
4. Calculate the Tube side Reynolds number $Re_t = \frac{\rho_t \mathcal{V}_t d_{t,i}}{\mu_t}$
5. Calculate Darcy friction factor $f_t = (1.82 \log_{10} Re_t - 1.64)^{-2}$
6. Calculate tube side convective coefficient (h_t)

If $Re_t < 2300$ *then*

$h_t = \frac{k_t}{d_{t,i}}\left[3.657 + \frac{0.0677\{Re_t Pr_t (d_{t,i}/l)\}^{1.33}}{1+0.1Pr_t\{Re_t(d_{t,i}/l)\}^{0.3}}\right]$

Else if $2300 < Re_t < 1000$

$h_t = \frac{k_t}{d_{t,i}}\left[\frac{(f_t/8)(Re_t - 1000)Pr_t}{1+12.7(f_t/8)^{1/2}\left(Pr_t^{2/3}-1\right)}\left(1+\frac{d_{t,i}}{l}\right)0.67\right]$

Else if $Re_t > 10000$

$h_t = 0.027 \frac{k_t}{d_{t,i}} Re_t^{0.8} Pr_t^{1/3} \left(\frac{\mu_t}{\mu_{tw}}\right)^{0.14}$

End if

7. Calculate equivalent shell diameter $D_e = \left(\frac{4*A}{P}\right)$
8. Calculate Shell side fluid velocity $V_s = \frac{\dot{m_h}}{\rho_s \frac{\pi}{4} D_e^2 n_t}$
9. Calculate Shell side Reynolds number $Re_s = \frac{\rho_s \mathcal{V}_s D_e}{\mu_s}$
10. Calculate Shell side convective coefficient $h_s = 0.36 \frac{k_t}{D_e} Re_s^{0.55} Pr_s^{1/3} \left(\frac{\mu_s}{\mu_{sw}}\right) 0.14$
11. Calculate the overall heat transfer coefficient $U = \frac{1}{\frac{1}{h_s} + R_{fs} + \left(\frac{d_{t,o}}{d_{t,i}}\right)\left(R_{ft} + \frac{1}{h_t}\right)}$
12. Calculate $C_c = \dot{m_c} * C_{p,c}$, $C_h = \dot{m_h} * C_{p,h}$
    $C_{max} = \max(C_c, C_h)$
    $C_{min} = \min(C_c, C_h)$
13. Calculate the Number of Transfer Units $NTU = \frac{A_s * U}{C_{min}}$
14. Calculate heat capacity ratio $C_r = \frac{C_{min}}{C_{max}}$
15. Calculate effectiveness (Ɛ)

If $C_{min}\,(mixed), C_{max}\,(unmixed)$ then

$\varepsilon = 1 - \exp(-C_r^{-1}\{1 - \exp[-C_r(NTU)]\})$

Else $C_{min}\,(unmixed), C_{max}\,(mixed)$

$\varepsilon = \left(\frac{1}{C_r}\right)(1\text{-}\exp\{\text{-}C_r[1 - \exp(-NTU)]\})$

End if

16. Calculate Output temperature $T_{h,o}$ and $T_{c,o}$

If $C_{min} = C_h$ then

$T_{h,o} = T_{h,i} - \varepsilon\,(T_{h,i} - T_{c,i})$

Else

$T_{c,o} = \varepsilon(T_{h,i} - T_{c,i}) + T_{c,i}$

End if

17. Heat duty ($q_c = q_h$)

$q_c = \dot{m_c} * C_{p,c} * (T_{c,o} - T_{c,i})$ and

$q_h = \dot{m_h} * C_{p,h} * (T_{h,i} - T_{h,o})$

$T_{h,o} = -\frac{q_c}{\dot{m_h} * C_{p,h}} + T_{h,i}$ and $T_{c,o} = \frac{q_h}{\dot{m_c} * C_{p,c}} + T_{c,i}$

End

4. RESULTS AND DISCUSSION

MATLAB (version: R2018a(9.4.0.813654)) genetic algorithm toolbox is used to solve the optimization problem described. This section compares the effectiveness of the original oil cooler type OKG 33/244 with the proposed design. The results of the effectiveness of STHX are presented in Figures 3-6, where the relationship between the number of individuals in populations and the effectiveness has been plotted for each generation (1, 10, 20, 30). By applying GA to the presented case study discussed in section 3.1, the heat transfer coefficient should exceed 0.5 (of the original design) for success.

In Figure 3, the effectiveness of generation number 10 has been compared with that of generation number 1. The figure shows that the effectiveness has improved in generation 10 compared to generation 1. However, in the first generation, effectiveness suffers from a wide range (0.31268 to 0.72207). At the same time, the effectiveness in the tenth generation falls between 0.65804 and 0.7432 and mostly between 0.73142 and 0.7432.

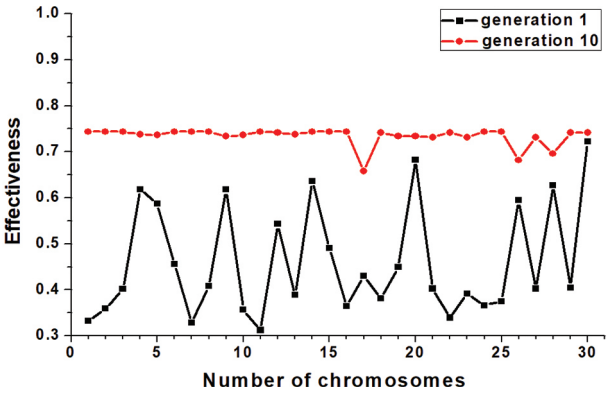


Fig. 3. Number of chromosomes in populations versus effectiveness for generations 1 and 10

Figure. 4 shows the effectiveness of generation 20 compared to the first generation. Figure 4 clearly illustrates that the effectiveness in generation 20 has improved compared to generation 1. Generation 20 has average effectiveness in the range of 0.73293 - 0.76883.

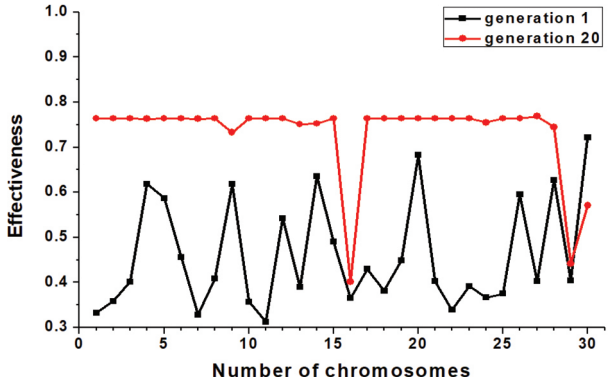


Fig. 4. Number of chromosomes in populations versus effectiveness for generations 1 and 20

The effectiveness of generation number 30 compared to the first generation is shown in Figure 5 as well, Generation 30 has effectiveness in the range of 0.73334 - 0.76923.

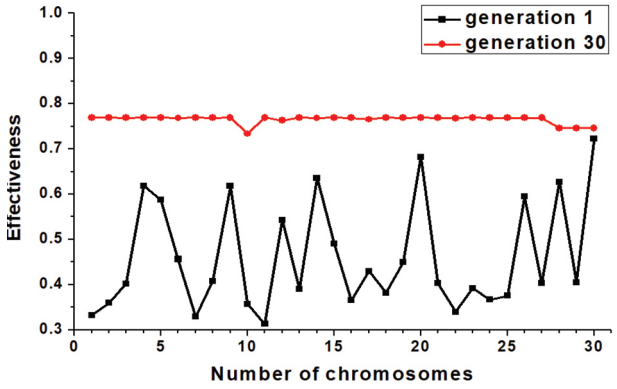


Fig. 5. Number of chromosomes in populations versus effectiveness for generations 1 and 30

Figure 6 proves the proposed claim for effectiveness improvement over generations.

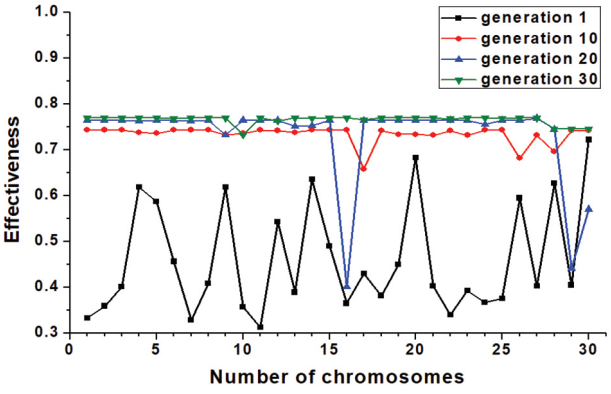


Fig. 6. Number of chromosomes in populations versus effectiveness for generations 1, 10, 20, and 30

Figure. 7 shows the relationship between the best effectiveness value for each generation. The effectiveness significantly improves through generations number 1 to 20. While the effectiveness in generation number 30 shows a slight increase in effectiveness compared to generation number 20.

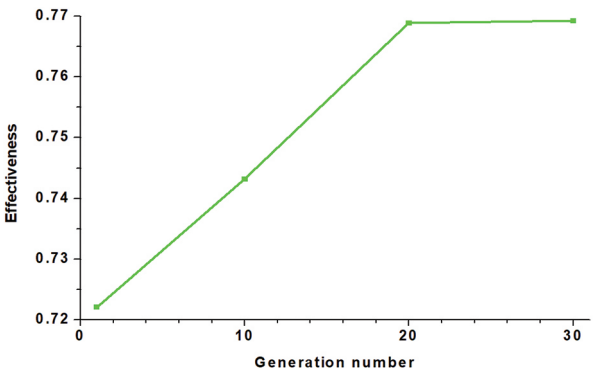


Fig. 7. The best value of effectiveness in each generation

Figure 8 illustrates the best point in each run of 50 runs; the robustness of the proposed method is illustrated due to the obvious convergence of best points in the range of (0.73456 to 0.78458).

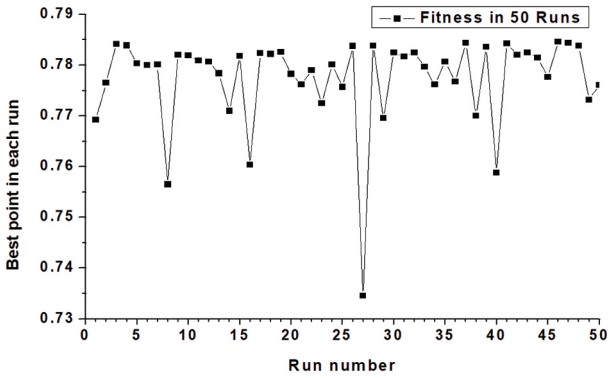


Fig. 8. Best fitness in 50 Runs

The simulation results using GA for design optimization of an oil cooler type OKG 33/244 reveal that the performance of STHX has been improved, indicating

that GA can be successfully employed for solving the design optimization problem of a STHX from a performance point of view. Using GA, a maximum increase in the effectiveness of 57% and a minimum increase in the effectiveness of 47% have been achieved. Furthermore, the average effectiveness of the presented oil cooler type OKG 33/244 has improved by 55%, and NTU improved from 0.475319 to 1.825664 using GA. The details for optimal primitive and effectiveness for maximum and minimum fitness in 50 runs are reported in Tables 5 and 6, respectively.

Table 5. Details of the maximum fitness in 50 runs

Optimal primitive			effectiveness	NTU
$d_{t,i}$	L	D	0.784583	1.825664
0.011	3.549969	0.58		

Table 6. Details of the minimum fitness in 50 runs

Optimal primitive			effectiveness	NTU
$d_{t,i}$	L	D	0.734555	1.535565
0.012062	3.549977	0.58		

5. CONCLUSION

The design optimization of an STHX is developed from a performance point of view by using GA to achieve the optimal configuration. The objective is to develop and test a model (Fig.1) for optimizing the early design stage of the STHX, which is quick. GA is implemented to maximize heat transfer rate while minimizing pressure drop. The GA is applied to the oil cooler type OKG 33/244. By comparing the results obtained using GA with the original data of the STHX, the following conclusions are obtained:

- A maximum increase in the effectiveness of 57% was achieved using GA.
- A minimum increase in the effectiveness of 47% was also achieved. Furthermore, the heat exchanger's average was 55% by using GA.
- NTU improved from 0.475319 to 1.825664
- Finally, the simulation outcomes reveal that the STHX's operating performance has been improved, indicating that GA can be successfully employed for design optimization of a STHX from a performance standpoint.

In the future investigate, the simultaneous approach using GA on a larger scale (heat exchanger network) can be employed to improve overall plant performance.

6. REFERENCES

[1] M. E. H. Assad, M. A. Nazari, "Heat exchangers and nanofluids", Design and performance optimization of renewable energy systems, Academic Press, 2021, pp. 33-42.

[2] C. Balaji, B. Srinivasan, S. Gedupudi, "Heat Transfer Engineering: Fundamentals and Techniques", Academic Press, 2020.

[3] E. Edreis, A. Petrov, "Types of heat exchangers in industry, their advantages and disadvantages, and the study of their parameters", IOP Conference Series: Materials Science and Engineering Vol. 963, No. 1, 2020, p. 012027.

[4] S. M. Hall, "Rules of thumb for chemical engineers", Butterworth-Heinemann, 2012.

[5] K. K. Saijal, T. Danish, "Design optimization of a shell and tube heat exchanger with staggered baffles using neural network and genetic algorithm", Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 2021.

[6] S. Yang, Y. Chen, J. Wu, H. Gu, "Performance simulation on unilateral ladder type helical baffle heat exchanger in half cylindrical space", Energy Conversion and Management, Vol. 150, 2017, pp. 134–147.

[7] A. Abd, K. Ali; Q. Mohammed, S. Zaki, "Performance Analysis of Shell and Tube Heat Exchanger: Parametric Study. Case Studies", Thermal Engineering, 2018

[8] L. Pekar, "Advanced Analytic and Control Techniques for Thermal Systems with Heat Exchangers", Academic Press, 2020.

[9] D. K. Mohanty, "Gravitational search algorithm for economic optimization design of a shell and tube heat exchanger", Applied Thermal Engineering, Vol. 107, 2016, pp. 184-193.

[10] U. Salahuddin, M. Bilal, H. Ejaz, "A review of the advancements made in helical baffles used in shell and tube heat exchangers", International Communications in Heat and Mass Transfer, Vol. 67, 2015, pp. 104-108.

[11] D. K. Mohanty, "Application of firefly algorithm for design optimization of a shell and tube heat exchanger from economic point of view", International Journal of Thermal Sciences, Vol. 102, 2016, pp. 228-238.

[12] R. Setiawan, F. Hrdlička, P. S. Darmanto, V. Fahriani, S. R. Pertiwi, "Thermal Design Optimization of Shell-and-Tube Heat Exchanger Liquid to Liquid

to Minimize cost using Combination Bell-Delaware Method and Genetic Algorithm", Journal of Mechanical Engineering Science and Technology, Vol. 4, No. 1, 2020, pp. 14-27.

[13] O. D. Lara-Montaño, F. I. Gómez-Castro, C. Gutiérrez-Antonio, "Comparison of the performance of different metaheuristic methods for the optimization of shell-and-tube heat exchangers", Computers & Chemical Engineering, 2021, p. 107403.

[14] M. Abd Elaziz, A. H. Elsheikh, D. Oliva, L. Abualigah, S. Lu, A. Ewees, "Advanced metaheuristic techniques for mechanical design problems", Archives of Computational Methods in Engineering, Vol. 29, No. 1, 2022, pp. 695-716.

[15] A. Banerjee, D. Singh, S. Sahana, I. Nath, "Impacts of metaheuristic and swarm intelligence approach in optimization", Cognitive Big Data Intelligence with a Metaheuristic Approach. Academic Press, 2022, pp. 71-99.

[16] O. E. Turgut, "Multi-Agent Metaheuristic Framework for Thermal Design Optimization of a Shell and Tube Evaporator Operated with R134a/Al 2O3 Nanorefrigerant", Arabian Journal for Science and Engineering, Vol. 44, No. 2, 2019, pp. 777-801.

[17] M. Reda, M. Elhosseini, A. Haikal, M. Badawy, "A novel cuckoo search algorithm with adaptive discovery probability based on double Mersenne numbers", Neural Computing and Applications, Vol. 33, No. 23, 2021, pp. 16377-16402.

[18] K. Sörensen, M. Sevaux, F. A. Glover, "History of metaheuristics", Handbook of heuristics. Springer, 2018, pp. 791-808.

[19] A. F. Villaverde, F. Fröhlich, D. Weindl, J. Hasenauer, J. Banga, "Benchmarking optimization methods for parameter estimation in large kinetic models", Bioinformatics, Vol. 35, No. 5, 2019, pp. 830-838.

[20] J. M. Ponce-Ortega, M. Serna-González, L. I. Salcedo-Estrada, A. Jiménez-Gutiérrez, "Minimum-investment design of multiple shell and tube heat exchangers using a MINLP formulation", Chemical Engineering Research and Design, Vol. 84, No. 10, 2006, pp. 905-910.

[21] P. Wildi-Tremblay, L. Gosselin, "Minimizing shell-and-tube heat exchanger cost with genetic algorithms and considering maintenance", International journal of energy research, Vol. 31, No. 9, 2007, 867-885.

[22] R. Selbaş, Ö. Kızılkan, M. Reppich, "A new design approach for shell-and-tube heat exchangers using genetic algorithms from economic point of view", Chemical Engineering and Processing: Process Intensification, Vol. 45, No. 4, 2006, pp. 268-275.

[23] A. C. Caputo, P. M. Pelagagge, P. Salini, "Heat exchanger design based on economic optimisation", Applied Thermal Engineering, Vol. 28, No. 10, 2008, pp. 1151-1159.

[24] J. Guo, L. Cheng, M. Xu, "Optimization design of shell-and-tube heat exchanger by entropy generation minimization and genetic algorithm", Applied Thermal Engineering, Vol. 29, No. 14-15, 2009, pp. 2954-2960.

[25] V. K. Patel, R. V. Rao, "Design optimization of shell-and-tube heat exchanger using particle swarm optimization technique", Applied Thermal Engineering, Vol. 30, No. 11-12, 2010, pp. 1417-1425.

[26] A. V. Azad, M. Amidpour, "Economic optimization of shell and tube heat exchanger based on constructal theory", Energy, Vol. 36, No. 2, 2011, pp. 1087-1096.

[27] R. Khosravi, A. Khosravi, S. Nahavandi, H. Hajabdollahi, "Effectiveness of evolutionary algorithms for optimization of heat exchangers", Energy Conversion and Management, Vol. 89, 2015, pp. 281-288.

[28] M. Dastmalchi, G. A. Sheikhzadeh, A. Arefmanesh, "Optimization of micro-finned tubes in double pipe heat exchangers using particle swarm algorithm", Applied Thermal Engineering, Vol. 119, 2017, pp. 1-9.

[29] A. Farzin, M. Ghazi, A. F. Sotoodeh, M. Nikian, "Economic Optimization of a Shell-and-Tube Heat Exchanger (STHE) based on New Method by Grasshopper Optimization Algorithm (GOA)", Journal of Marine Science University of Imam Khomeini, Vol. 7, No. 2, 2020, pp. 112-124.

[30] A. D. Kraus, J. R. Welty, A. Aziz, "Introduction to thermal and fluid engineering", CRC Press, 2011.

[31] R. W. Serth, T. Lestina, "Process heat transfer: Principles, applications and rules of thumb", Academic press, 2014.

[32] R. K. Shah, D. P. Sekulic, "Fundamentals of heat exchanger design", John Wiley & Sons, 2003.

[33] J. P. Sai, B. N. Rao, "Non-dominated Sorting Genetic Algorithm II and Particle Swarm Optimization for design optimization of Shell and Tube Heat Exchanger", International Communications in Heat and Mass Transfer, Vol. 132, 2022, p. 105896.

[34] Z. Y. Guo, X. Liu, W. Tao, R. Shah, "Effectiveness-thermal resistance method for heat exchanger design and analysis", International Journal of Heat and Mass Transfer, Vol. 53, No. 13-14, 2010, pp. 2877-2884.

[35] S. S. Murshed, M. M. Lopes, "Heat exchangers: design, experiment and simulation", BoD-Books on Demand, 2017.

[36] H. S. Dizaji, S. Jafarmadar, M. Abbasalizadeh, S. Khorasani, "Experiments on air bubbles injection into a vertical shell and coiled tube heat exchanger; exergy and NTU analysis", Energy Conversion and Management, Vol. 103, 2015, pp. 973-980.

[37] M. Amini, M. Bazargan, "Two objective optimization in shell-and-tube heat exchangers using genetic algorithm", Applied Thermal Engineering, Vol. 69, No. 1-2, 2014, pp. 278-285.

[38] J. M. Ponce-Ortega, M. Serna-González, A. Jiménez-Gutiérrez, "Use of genetic algorithms for the optimal design of shell-and-tube heat exchangers", Applied Thermal Engineering, Vol. 29, No. 2-3, 2009, pp. 203-209.

[39] D. J. Murray-Smith, "Experimental modelling: system identification, parameter estimation and model optimisation techniques", Modelling and Simulation of Integrated Systems in Engineering, 2012, pp. 165-214

[40] C. Wang, Z. Cui, H. Yu, K. Chen, J. Wang, "Intelligent optimization design of shell and helically coiled tube heat exchanger based on genetic algorithm", International Journal of Heat and Mass Transfer, Vol. 159, 2020, p. 120140.

[41] T. Baklacioglu, "Modeling the fuel flow-rate of transport aircraft during flight phases using genetic algorithm-optimized neural networks", Aerospace Science and Technology, Vol. 49, 2016, pp. 52-62.

7. APPENDIX

Table 7.

Nomenclature

Abbreviations:

A	surface area (m^2)	n_p	number of tubes passages
A_s	heat transfer surface area of tubes (m^2)	Pr_s	shell side Prandtl number
Cp_c	Heat capacity of cold fluid $(kJ/kg\ k)$	Pr_t	tube side Prandtl number
Cp_h	Heat capacity of hot fluid $(kJ/kg\ k)$	P_t	tube pitch (m)
C_c	Heat capacity of cold fluid $(kJ/kg\ k)$	q_c	heat duty of cold fluid (w)
C_h	Heat capacity of hot fluid $(kJ/kg\ k)$	q_h	heat duty of hot fluid (w)
C_{max}	Maximum of heat capacity $(kJ/kg\ k)$	Re_s	Shell side Reynolds number
C_{min}	minimum of heat capacity $(kJ/kg\ k)$	Re_t	Tube side Reynolds number
C_r	heat capacity ratio	Rf_s	shell side fouling resistance $(m^2\ k/w)$
D	shell diameter (m)	Rf_t	shell side fouling resistance $(m^2\ k/w)$
D_e	equivalent shell diameter (m)	$T_{c,i}$	Input temperature of cold fluid (C^o)
$d_{t,i}$	tube inside diameter (m)	$T_{c,o}$	Output temperature of cold fluid (C^o)
$d_{t,o}$	tube outside diameter (m)	$T_{h,i}$	Input temperature of hot fluid (C^o)
ε	effectiveness	$T_{h,o}$	Output temperature of hot fluid (C^o)
f_t	Darcy friction factor of tube	U	overall heat transfer coefficient $(w/m^2\ k)$
f_s	Darcy friction factor of shell	S_t	thickness of tube (m)
h_t	tube side convective coefficient $(w/m^2\ k)$	v_s	Shell side fluid velocity (m/s)
h_s	Shell side convective coefficient $(w/m^2\ k)$	v_t	tube side fluid velocity (m/s)
k_t	thermal conductivity $(w/m\ k)$	μ_s	shell side dynamic viscosity $(Pa\ s)$
L_t	tubes length (m)	μ_t	Tube side dynamic viscosity $(Pa\ s)$
m_c	Mass flow rate of cold fluid (kg/s)	ρ_s	Shell side fluid density (kg/m^3)
m_h	Mass flow rate of hot fluid (kg/s)	ρ_t	tube side fluid density (kg/m^3)
NT	Number of Transfer Units	π	numerical constant
n_t	number of tube		

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

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

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

_____ _____

Author/Authorized Agent **Date**

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