Diagnosing the Defects of the Gears of the Gearbox Using Vibration Analysis and Validating it by the Graphs of the CMM Machine

Mohammadesmaeil Akbari[1], Alireza Mangouri[2], Sajjad Atazadeh[3]
[1]Department of Electrical Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran
[2]Department of Mechanical Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran
[3]Department of Mechanical Engineering. University of Tabriz, Tabriz, Iran
Email: Makbari@iau.ac.ir, a.mangouri@iau.ac.ir,sajad.atazadeh@gmail.com

Received: 5 September2024 Accepted: 19 January 2024 Published :15 February 2025

Abstract

This study focuses on diagnosing gear defects in the S5-380 transmission using advanced vibration analysis and validating findings through the HÖFLER Gear Testing Machine. The research highlights key faults such as backlash, eccentricity, and shaft deflection identified via vibration measurements and confirmed using detailed gear profile and lead inspections. The vibration analysis employed V-SAM Vibrometer and V-SAM-Web Soft to detect critical parameters, including Gear Mesh Frequency (GMF) and Gear Hunting Tooth Frequency (GHTF). Results demonstrated the robustness of vibration analysis as an efficient, cost-effective alternative to Coordinate Measuring Machines (CMM) for end-of-line quality control. Furthermore, integrating vibration analysis with CMM-based validation provided precise fault identification, enabling targeted maintenance and enhanced gearbox performance. The findings underscore the potential of transitioning to AI-driven diagnostic systems in Industry 5.0 for real-time, scalable, and intelligent gearbox monitoring. This approach offers significant implications for improving durability, quality, and operational efficiency in gear manufacturing.

Keywords: Gear Fault Diagnosis, Vibration Analysis, Artificial Neural Network (ANN), Machine Learning (ML), Industry 5.0, Signal Processing

1. Introduction

In heavy-duty vehicles, the gearbox plays a critical and intricate role in the drive train system. Also referred to as the transmission, this component is responsible for transferring engine power to the wheels and managing the gear ratios to adapt to varying driving conditions. Heavy-duty vehicle gearboxes are designed with greater complexity and durability compared to those in lighter vehicles, due to the demands of handling substantial loads and maintaining optimal performance across a wide range of speeds. In the R&D division of Charkheshgar Company, a semi-heavy gearbox was engineered and manufactured to enhance and fortify gearbox performance. To address and resolve issues observed in an existing foreign model, a similar gearbox available for analysis was subjected to detailed vibration condition monitoring. This process allowed for the identification of several critical issues, including:

- **Backlash**: Excessive clearance between meshing gears, leading to loss of precision and increased wear.
- **Eccentricity**: Misalignment or out-of-round conditions in gears, affecting smooth operation and gear engagement.
- **Shaft Deflection**: Inadequate or improper shaft alignment under load, impacting the overall stability and performance of the gearbox.

The insights gained from this vibration analysis were instrumental in refining the design and functionality of the gearbox. The identified problems were addressed in the optimization process, leading to improvements in gear accuracy, load handling, and overall operational efficiency. This approach not only resolved issues present in the foreign model but also provided valuable data to enhance future gearbox designs and applications. At T.P.T. Co.'s R&D department, a proprietary software solution called V-SAM-Web Soft (under web Software) has been specifically developed for diagnosing gear and bearing faults in transmission using vibration analysis. This software is capable of identifying gear meshing frequencies (GMF), gear hunting tooth frequencies (GHTF), Gear assembly phase frequencies (GAPF) or fractional GMF, and also defects such as backlash, gear wear, excessive load on tooth, eccentric gears, misaligned gears, and cracked or broken tooth. In this study, data was collected using the V-SAM Vibrometer from the S5-380 transmission, which was subjected to load testing. Following the fault diagnosis via vibration analysis, the transmission was disassembled, and the gears were sent to the CMM unit for detailed inspection. Subsequently, the Höfler machine was employed to generate the relevant graphs, which validated and confirmed the faults identified through the vibration analysis.

2. Fault Diagnosis of Transmissions and Gears

Fault diagnosis of transmissions and gears is a critical process in the maintenance and optimization of mechanical systems, requiring specialized knowledge and advanced techniques. Below is a detailed, technical overview of the various methods used in transmissions and gear fault diagnosis:

1. Vibration analysis
2. Oil Analysis
3. Thermal Analysis
4. Ultrasonic Testing
5. Modal Analysis
6. Acoustic Emission Testing & etc.

In this study, we conducted a comprehensive analysis of the transmission using advanced vibration analysis techniques. Following the initial fault detection, we employed Coordinate Measuring Machine (CMM-HÖFLER Gear Testing Machine) technology to validate the identified anomalies. This validation was performed through precise measurement and evaluation of the gear profile and lead graphs, ensuring the accuracy and reliability of the fault diagnosis process. The integration of vibration analysis with CMM-based verification provided a robust methodology for identifying and confirming gear defects with high precision.

2.1. *Gear Vibration Theory*

When two or more gears are in mesh, the frequencies such as GMF, GHTF & GAPF, generated depend upon gear speed, number of teeth and common factors. Below we will give a detailed definition of each:

2.1.1. Gear Mesh frequency (GMF)

GMF is a crucial concept in the field of vibration analysis, particularly in the context of rotating machinery and gear systems. It refers to the frequency at which the gear teeth mesh, or engage, during operation. The Gear Mesh Frequency is calculated using the following formula:

$$\mathrm{GMF} = \mathrm{NTG} * \mathrm{RSG} \quad (1)$$

Where:

NTG : Number of Teeth on the Gear is the total number of teeth on the gear being analyzed.

RSG : **Rotational Speed of the Gear** is the rotational speed of the gear in revolutions per second (Hz).

Importance in Vibration Analysis:

- **Condition Monitoring**: GMF is a key indicator in vibration analysis for diagnosing gear-related faults. It helps in detecting issues like misalignment, wear, and tooth damage.
- **Fault Detection**: Any deviation or sideband frequencies around the GMF can indicate potential faults or defects in the gear system, such as gear tooth damage, misalignment, or lubrication problems.
- **Harmonics and Sidebands**: Typically, GMF will have harmonics, and sidebands may appear around the GMF frequency due to modulation effects, which are critical in diagnosing specific types of gear faults.

2.1.2. Gear Hunting Tooth Frequency (GHTF):

GHTF is another important concept in gear vibration analysis, particularly in systems where two or more gears mesh together. It is related to the unique periodic interaction of a specific tooth on one gear with a specific tooth on another gear. This interaction occurs over a longer cycle than the regular gear mesh frequency (GMF).

2.1.3. Understanding Hunting Tooth Frequency (HTF)

- **Hunting Tooth Cycle**: In gear systems, a hunting tooth cycle is the period it takes for a particular tooth on one gear to mesh again with the same tooth on another gear. This cycle is significant because it represents a unique combination of teeth that meet only after a certain number of revolutions.
- **Application**: GHTF is most relevant in gears with non-integral gear ratios, where the same tooth pair does not mesh repeatedly with each revolution. Over multiple cycles, different teeth interact, which can be used to diagnose specific wear patterns or faults that develop over a longer period.

The Hunting Tooth Frequency can be calculated using the following formula:

$$GHTF = \frac{GMF}{GCD} \quad (2)$$

GHTF=GMF/Greatest Common Divisor (GCD) of the number of teeth on both gears

Where:

GMF is defined as (1)

GCD : Greatest Common Divisor is the Greatest Common Divisor of the number of teeth on the driving and driven gears.

Importance of HTF in Gear Analysis:

- **Wear and Tear Detection**: GHTF can help in identifying wear patterns that develop over time, as the same pair of teeth only meet after a certain number of revolutions. This can reveal issues like pitting, spalling, or other damage that might not be evident in the regular GMF spectrum.

- **Long-Term Monitoring**: GHTF provides insight into the long-term interaction between gear teeth, which is crucial for predictive maintenance and avoiding unexpected gear failures.

GHTF is particularly useful for monitoring complex gear systems were identifying unique, long-cycle interactions can help prevent significant damage and extend the life of the machinery.

2.1.4. Gear assembles phase frequencies (GAPF) or Fractional GMF

GAPF or Fractional gear mesh frequency refers to a frequency that occurs in gear systems and is a fraction of the standard gear mesh frequency. When the number of teeth on each meshing gear has a common factor greater than 1, and one of the gears is eccentric, every Nth tooth (where N is the common factor) on the well-aligned gear can be imprinted or worn by the eccentric gear. This imprinting or wearing causes the Nth cycle of the gear mesh frequency to have a higher amplitude than the other cycles.

2.2. GEAR PROBLEMS AND CAUSES

Vibration analysis is a powerful diagnostic tool used to detect and analyze various types of gear failures in mechanical systems. By monitoring and interpreting vibration signals, maintenance professionals can identify potential issues early and prevent catastrophic failures. Below is a detailed examination of common gear failures and how vibration analysis can be used to detect them:

- Eccentric gears
  Eccentric gears can take many forms. For explanation purposes, eccentric gears are divided into four broad categories:
  1. Meshing gears that have a common factor and one gear is eccentric
  2. Gears that do not have a common factor and one or both gears is eccentric
  3. Gears that are out-of-round or have several high places
  4. Gears installed on a bent shaft
- Loose and worn gears
- Misaligned gears
- Backlash Issues and Oscillating Gears
  Backlash issues or oscillating gears can induce a significant second harmonic at the gear mesh frequency. An elevated amplitude at twice the gear mesh frequency is a strong indicator of backlash or oscillating gears. To accurately diagnose this condition, further analysis in the time domain is essential. It is important to note that the second harmonic appears at the peak of the signal, which implies it is 180° out of phase with the fundamental frequency. This phase discrepancy is characteristic of a backlash-related issue. Conversely, if the second harmonic were due to misalignment, both signals would be in phase, and the second harmonic would manifest at the trough of the signal. The observed out-of-phase condition in this instance

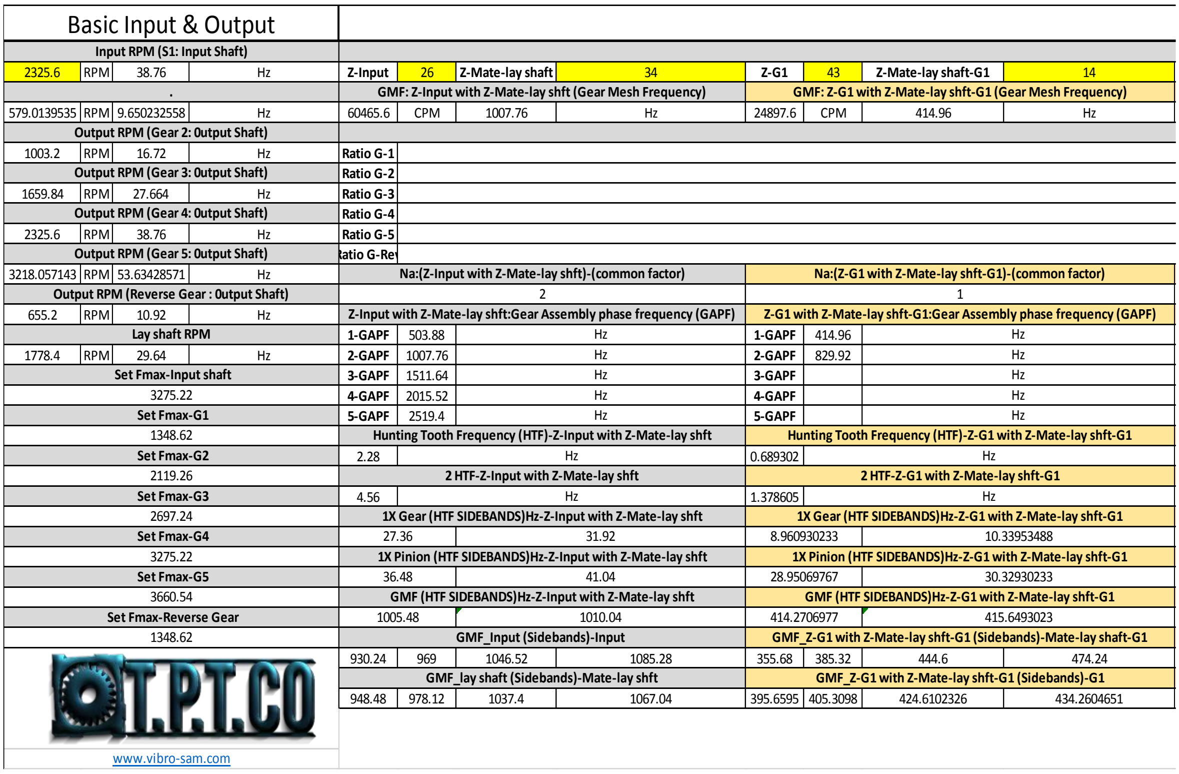
suggests that the gears are experiencing back-and-forth motion. Consequently, inspections for gear looseness or associated bearing issues may not reveal any obvious abnormalities. The underlying causes of backlash and gear oscillation are multifaceted. Some potential sources include:

- **Lightly Loaded Gears:** Gears operating under light loads are prone to oscillation.
- **Inconsistent Loading:** Irregular loads in equipment such as agitators or digesters can lead to gear oscillation.
- **Draw Issues in Paper Machines:** Tension problems in paper machines can exacerbate both of the aforementioned conditions.
- **Electric Drive Power Source Issues:** Malfunctions in the power source for certain electric drives may contribute to these problems.
- **Excessive Backlash:** Excessive gear backlash may also be a contributing factor.
- **Broken, cracked, or chipped teeth**

2.3. All needed Information for frequencies calculation and vibrometer setup

According to the documents received from the Charkheshgar Co., all frequencies related to: Gear Mesh Frequency, Fractional Gear Mesh Frequency, Hunting Tooth Frequency and Gear Life Expectancy are calculated in V-SAM-GearboxFcal software according to the table 2. It is worth noting that software inputs are displayed in yellow according to the table 1

Table 1. V-SAM-GearboxFcal software



Basic Input & Output											
Input RPM (S1: Input Shaft)											
2325.6	RPM	38.76	Hz	Z-Input	26	Z-Mate-lay shaft	34	Z-G1	43	Z-Mate-lay shaft-G1	14
.				GMF: Z-Input with Z-Mate-lay shft (Gear Mesh Frequency)				GMF: Z-G1 with Z-Mate-lay shft-G1 (Gear Mesh Frequency)			
579.0139535	RPM	9.650232558	Hz	60465.6	CPM	1007.76	Hz	24897.6	CPM	414.96	Hz
Output RPM (Gear 2: 0utput Shaft)											
1003.2	RPM	16.72	Hz	Ratio G-1							
Output RPM (Gear 3: 0utput Shaft)				Ratio G-2							
1659.84	RPM	27.664	Hz	Ratio G-3							
Output RPM (Gear 4: 0utput Shaft)				Ratio G-4							
2325.6	RPM	38.76	Hz	Ratio G-5							
Output RPM (Gear 5: 0utput Shaft)				Ratio G-Rev							
3218.057143	RPM	53.63428571	Hz	Na:(Z-Input with Z-Mate-lay shft)-(common factor)				Na:(Z-G1 with Z-Mate-lay shft-G1)-(common factor)			
Output RPM (Reverse Gear : 0utput Shaft)				2				1			
655.2	RPM	10.92	Hz	Z-Input with Z-Mate-lay shft:Gear Assembly phase frequency (GAPF)				Z-G1 with Z-Mate-lay shft-G1:Gear Assembly phase frequency (GAPF)			
Lay shaft RPM				1-GAPF	503.88	Hz		1-GAPF	414.96	Hz	
1778.4	RPM	29.64	Hz	2-GAPF	1007.76	Hz		2-GAPF	829.92	Hz	
Set Fmax-Input shaft				3-GAPF	1511.64	Hz		3-GAPF		Hz	
3275.22				4-GAPF	2015.52	Hz		4-GAPF		Hz	
Set Fmax-G1				5-GAPF	2519.4	Hz		5-GAPF		Hz	
1348.62				Hunting Tooth Frequency (HTF)-Z-Input with Z-Mate-lay shft				Hunting Tooth Frequency (HTF)-Z-G1 with Z-Mate-lay shft-G1			
Set Fmax-G2				2.28	Hz			0.689302	Hz		
2119.26				2 HTF-Z-Input with Z-Mate-lay shft				2 HTF-Z-G1 with Z-Mate-lay shft-G1			
Set Fmax-G3				4.56	Hz			1.378605	Hz		
2697.24				1X Gear (HTF SIDEBANDS)Hz-Z-Input with Z-Mate-lay shft				1X Gear (HTF SIDEBANDS)Hz-Z-G1 with Z-Mate-lay shft-G1			
Set Fmax-G4				27.36		31.92		8.960930233		10.33953488	
3275.22				1X Pinion (HTF SIDEBANDS)Hz-Z-Input with Z-Mate-lay shft				1X Pinion (HTF SIDEBANDS)Hz-Z-G1 with Z-Mate-lay shft-G1			
Set Fmax-G5				36.48		41.04		28.95069767		30.32930233	
3660.54				GMF (HTF SIDEBANDS)Hz-Z-Input with Z-Mate-lay shft				GMF (HTF SIDEBANDS)Hz-Z-G1 with Z-Mate-lay shft-G1			
Set Fmax-Reverse Gear				1005.48		1010.04		414.2706977		415.6493023	
1348.62				GMF_Input (Sidebands)-Input				GMF_Z-G1 with Z-Mate-lay shft-G1 (Sidebands)-Mate-lay shaft-G1			
T.P.T.CO				930.24	969	1046.52	1085.28	355.68	385.32	444.6	474.24
				GMF_lay shaft (Sidebands)-Mate-lay shft				GMF_Z-G1 with Z-Mate-lay shft-G1 (Sidebands)-G1			
				948.48	978.12	1037.4	1067.04	395.6595	405.3098	424.6102326	434.2604651
www.vibro-sam.com											

Based on drawing, the gearbox in question has 5 gears, and all mentioned frequencies have been calculated for all gear engagement conditions. However, this case study focuses only on Gear 1. Considering the gear mesh frequency for Gear 1 is 414.96 Hz, and taking into account a multiplier of 3.25 times the gear mesh frequency, which is 1348.62 Hz, the vibration analyzer device, V-SAM, has been setup as follow:

Instrument: V-SAM 2 Channel, F_{max}: 2500 Hz, No. of line: 6400, Ave. No: 4, Unit: Acceleration (m/s^2), Sensor type: CTC AC-102-1A, Window Type: Hanning, Ave. Type: FFT Linear, FFT Type: Liner.

Based on Figure 2, data acquisition was conducted horizontally under load, with the corresponding acceleration spectrum graphs shown in Figure 3 and the time signal presented in Figure 4. (According to V-SAM-Web Soft)

Based on the spectrum analysis and the strong presence of the second harmonic of the first gear pair's meshing frequency, along with prominent left and right sidebands at the lay-shaft speed frequency, and considering the details provided in Section **Backlash Issues and Oscillating Gears**, a potential backlash issue in the gear is anticipated. For a more detailed examination, the spectrum presented in Figure 1 was magnified, and all meshing frequencies and sidebands are precisely identified in Figure 3. Upon magnification, a sideband with a frequency of 2.03 Hz (the hunting tooth frequency of the gearbox input gear pair) was also observed.

For a more detailed examination, the time signal in velocity units was analyzed as shown in Figure 4, and it matched exactly with the information presented.

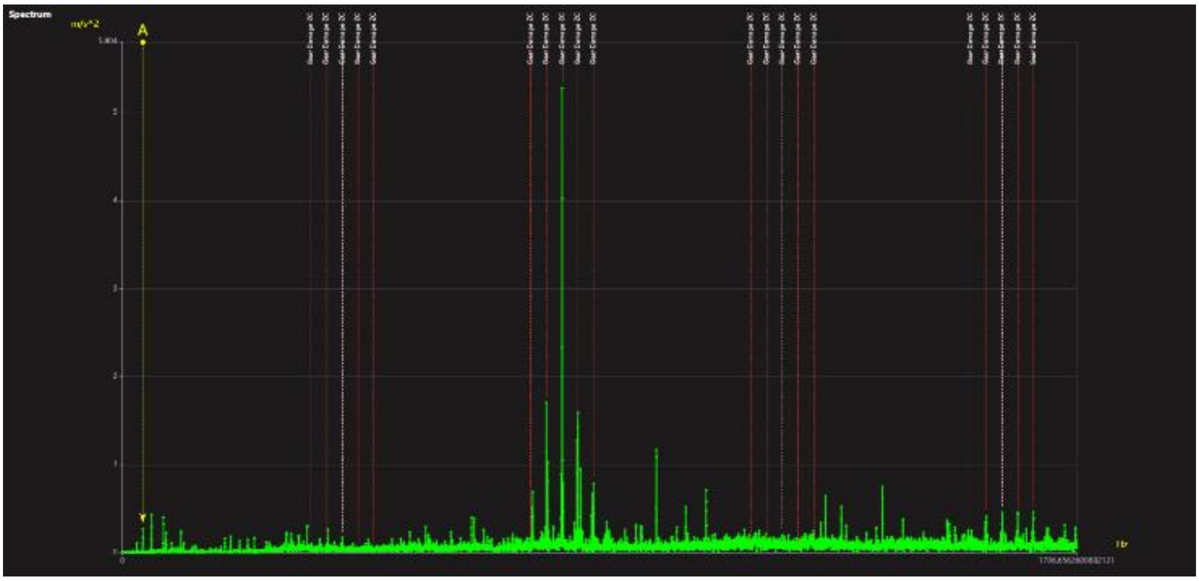


Fig.1. Horizontal -Acceleration Spectrum (1st Gear Mesh Frequency & harmonics)

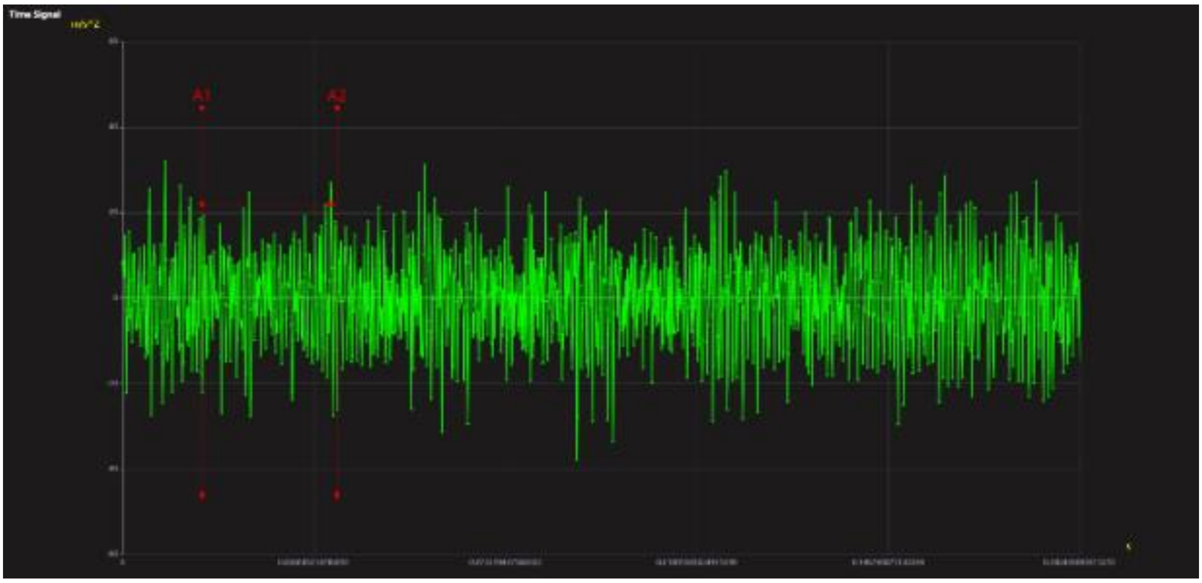
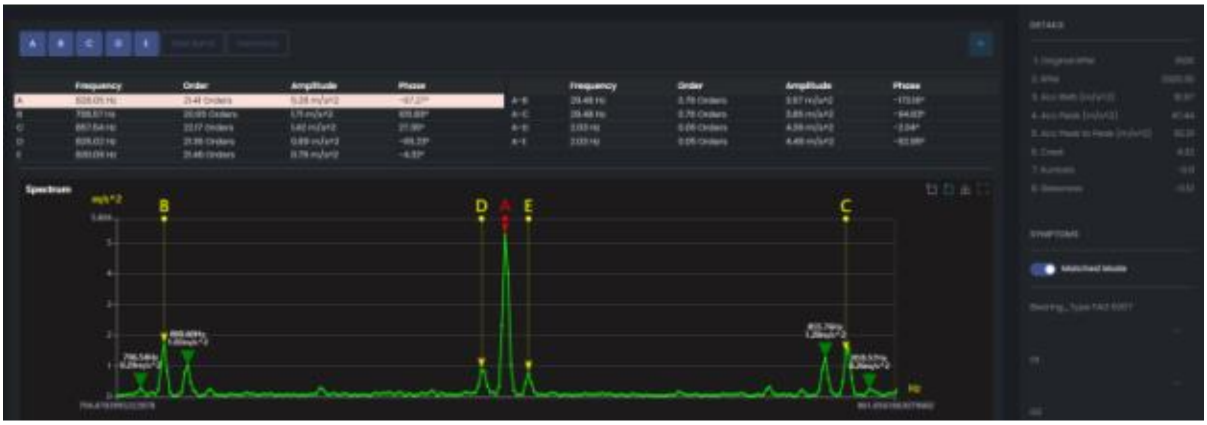


Fig. 2. Horizontal -Acceleration time signal



	Frequency	Order	Amplitude	Phase
A	828.05 Hz	21.41 Orders	5.28 m/s^2	-97.27°
B	798.57 Hz	20.65 Orders	1.71 m/s^2	101.89°
C	857.54 Hz	22.17 Orders	1.42 m/s^2	27.95°
D	826.02 Hz	21.36 Orders	0.89 m/s^2	-88.23°
E	830.08 Hz	21.46 Orders	0.79 m/s^2	-6.53°

	Frequency	Order	Amplitude	Phase
A-B	29.48 Hz	0.76 Orders	3.57 m/s^2	-172.18°
A-C	29.48 Hz	0.76 Orders	3.85 m/s^2	-94.03°
A-D	2.03 Hz	0.05 Orders	4.38 m/s^2	-2.04°
A-E	2.03 Hz	0.05 Orders	4.48 m/s^2	-82.98°

Fig.3. Magnified Horizontal -Acceleration Spectrum including lay-shaft speed and hunting tooth frequencies sidebands around 2nd harmonic of Gear Mesh Frequency

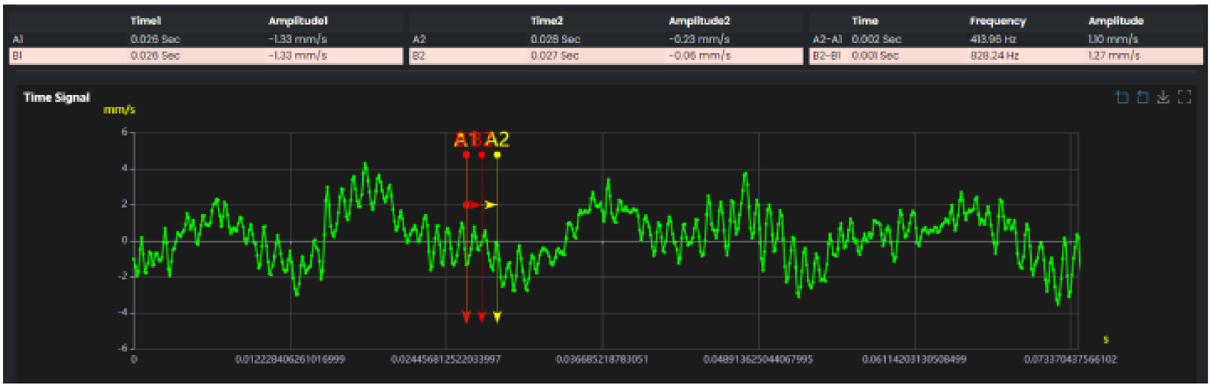
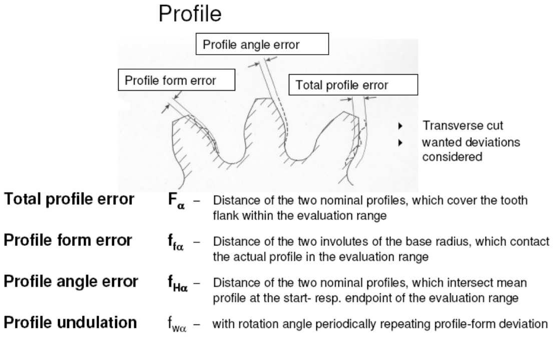

	Time1	Amplitude1		Time2	Amplitude2		Time	Frequency	Amplitude
A1	0.026 Sec	-1.33 mm/s	A2	0.028 Sec	-0.23 mm/s	A2-A1	0.002 Sec	413.96 Hz	1.10 mm/s
B1	0.026 Sec	-1.33 mm/s	B2	0.027 Sec	-0.06 mm/s	B2-B1	0.001 Sec	828.24 Hz	1.27 mm/s

Fig.4. Time Domain Signal of One- and Two-Times Gear mesh.

2.4. *Gear lead and profile inspection*

There are many tools for increasing durability, quality and decreasing noise and vibration of gears at the step of designing, production and control. One of them is the gear Lead and Profile control chart according to design and production parameters. There are different procedures of gear control chart drawing in the standard documents such as ISO, AGMA and DIN. According to the standard, the control parameters for the lead and profile of the gear are as shown in figure 5 and 6:



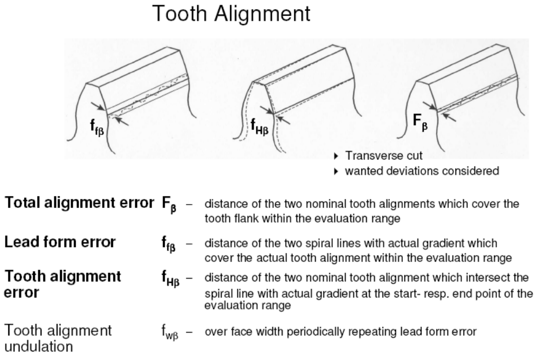
Total profile error F_{α} – Distance of the two nominal profiles, which cover the tooth flank within the evaluation range

Profile form error $f_{f\alpha}$ – Distance of the two involutes of the base radius, which contact the actual profile in the evaluation range

Profile angle error $f_{H\alpha}$ – Distance of the two nominal profiles, which intersect mean profile at the start- resp. endpoint of the evaluation range

Profile undulation $f_{w\alpha}$ – with rotation angle periodically repeating profile-form deviation

Fig.5. Gear profile inspection parameters



Total alignment error F_{β} – distance of the two nominal tooth alignments which cover the tooth flank within the evaluation range

Lead form error $f_{f\beta}$ – distance of the two spiral lines with actual gradient which cover the actual tooth alignment within the evaluation range

Tooth alignment error $f_{H\beta}$ – distance of the two nominal tooth alignment which intersect the spiral line with actual gradient at the start- resp. end point of the evaluation range

Tooth alignment undulation $f_{w\beta}$ – over face width periodically repeating lead form error

Fig.6. Gear Lead inspection parameters.

To evaluate the data obtained from the vibrations and according to gear inspection parameters that mentioned we decided to disassemble the gearbox and to conduct a more detailed examination of the first gear pair, it was sent to the CMM unit using the Höfler machine. Based on the graphs obtained from the Höfler machine, as shown in Figures 7, 8, 9, and 10, the issue related to backlash was confirmed and validated.

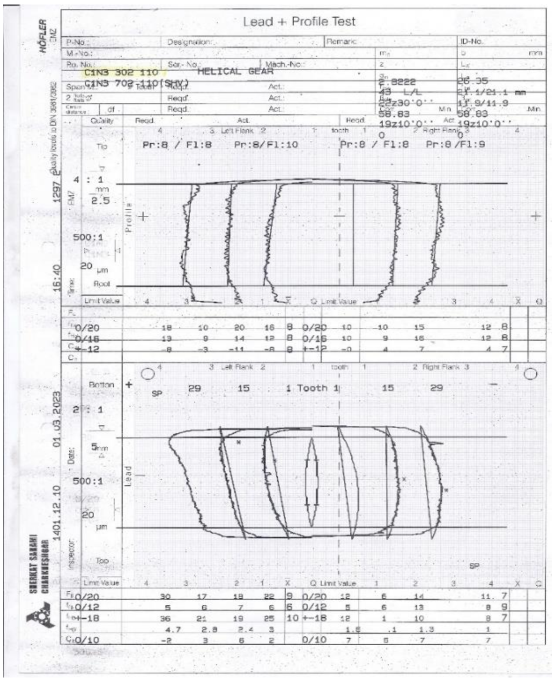


Fig.7. Lead and Profile of Gear 1 according to DIN 3961

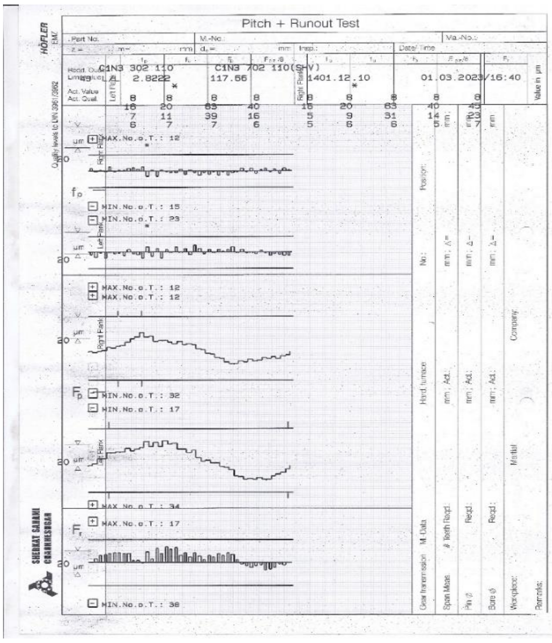


Fig.8. Pitch and Run out of Gear 1 according to DIN 3961

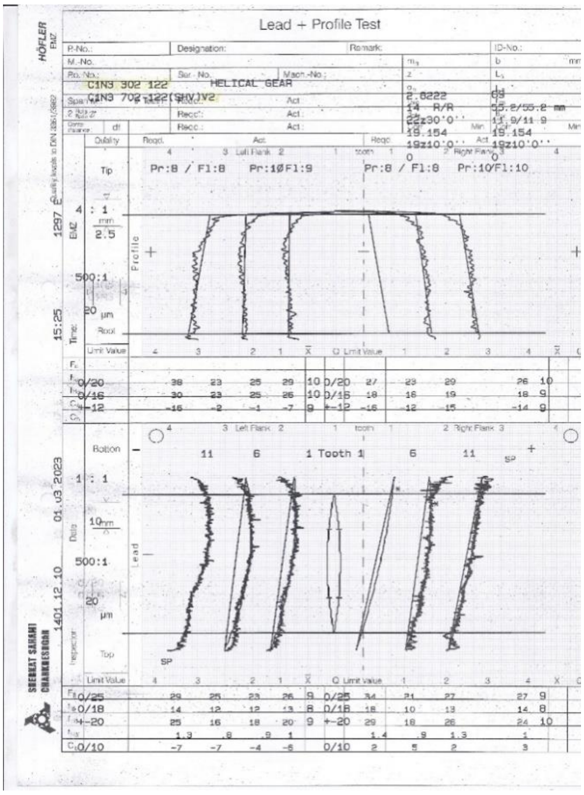


Fig.9. Lead and Profile of Mate Gear 1 according to DIN 3961

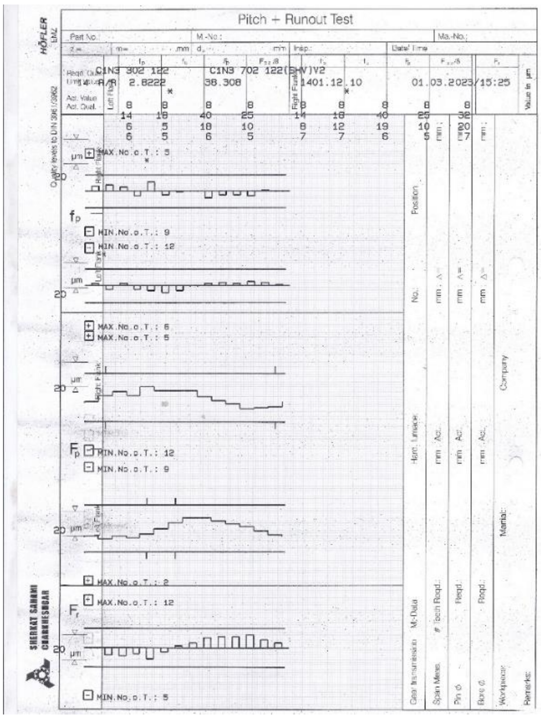


Fig.10. Pitch and Run out of Mate Gear 1 according to DIN 3961

Based on the graphs obtained from the Höfler machine, all parameters related to the surface quality of the gear (lead and profile) were found to be outside the standard range. This issue leads to improper meshing of the gear pair on the pitch circle, confirming and validating the backlash problem that was identified through vibration analysis.

3. Discussion and Conclusion

Given that both vibration measurement and measurement using CMM (Coordinate Measuring Machine) are methods for diagnosing gearbox issues, this paper validated the results of vibration analysis with the graphs obtained from the CMM, and both methods clearly indicated the presence of backlash. Considering that it is not feasible to measure all the gears produced in gearbox production lines using CMM, vibration analysis can be a suitable alternative for end-of-line (quality control and testing) diagnostics. Vibration measurement can test all gears under load with less time and cost.

Currently, gearbox quality is typically assessed at the end of the production line by operators using auditory methods. By using vibration analysis, not only can faults be identified, but the exact source of the fault can also be determined. This allows for precise repairs in the maintenance unit without additional disassembly time.

Given the advancements and the transition to Industry 5.0, there is also the potential to implement AI-based monitoring and diagnostics programs at the end of gearbox production lines.

References

[1] Allianz Versicherungs-AG. *Handbook of Loss Prevention*. Springer-Verlag, 1978.

[2] Blunt, D. M., and B. D. Forrester. “Health Monitoring of Blackhawk and Seahawk Main Transmissions Using Vibration Analysis.” *Proceedings of the Sixth Australian Aeronautical Conference/Second Pacific International Conference on Aerospace Science and Technology*, Melbourne, Australia, 1995.

[3] Braun, S. “The Extraction of Periodic Waveforms by Time Domain Averaging.” *Acustica*, vol. 32, 1975, pp. [insert page numbers].

[4] Braun, S., editor. *Mechanical Signature Analysis: Theory and Applications*. Academic Press Inc., 1986.

[5] Cempel, C., and W. J. Staszewski. “Signal Demodulation Techniques in Vibro-Acoustical Diagnostics of Machinery.” *Machine Dynamics Problems*, vol. 5, 1992, pp. 161-173.

[6] Choy, F. K., D. H. Mugler, and J. Zhou. “Damage Identification of a Gear Transmission Using Vibration Signatures.” *Journal of Mechanical Design*, vol. 125, 2003, pp. 394-403.

[7] Decker, H. J. “Gear Crack Detection Using Tooth Analysis at NASA Research Center.” National Aeronautics and Space Administration and US Army Research Laboratory, NASA/TM-2002-211491, ARL-TR-2681, 2002.

[8] Ebersbach, S., Z. Peng, and N. J. Kessissoglou. “The Investigation of the Condition and Faults of a Spur Gearbox Using Vibration and Wear Debris Analysis Techniques.” *International Conference on Wear of Materials*, vol. 260, 2005, pp. 16-24.

[9] Hu, T., B. C. Lu, and G. J. Chen. “A Rotary Machinery Fault Diagnosis Approach Based on Rough Set Theory.” *3rd World Congress on Intelligent Control and Automation*, Hefei, China, 2000, pp. 589-685.

[10] Lai, W., J. Xuan, T. Shi, and S. Yang. “Research of Vigner-Ville Time Frequency and Application in Detecting Gear Pinion Fault.” *Journal of Vibration Engineering*, vol. 16, 2003, pp. 247-250.

[11] Lin, J., and M. J. Zuo. “Gear Box Fault Diagnosis Using Adaptive Wavelet Filter.” *Mechanical Systems and Signal Processing*, vol. 17, no. 6, 2003, pp. 1259-1269.

[12] Martin, H. R. “Statistical Moment Analysis as a Means of Surface Damage Detection.” *Proceedings of the Seventh International Modal Analysis Conference*, Society for Experimental Mechanics, Schenectady, NY, 1989, pp. 1016-1021.

[13] McFadden, P. D. “Interpolation Techniques for the Time Domain Averaging of Vibration Data with Application to Helicopter Gearbox Monitoring.” *Aero Propulsion Technical Memorandum 437*, Department of Defense, Aeronautical Research Laboratory, September 1986.

[14] Zakrajsek, J. J. “A Review of Transmission Diagnostics Research at NASA Lewis Research Center.” Technical Report NASA TM-106746, ARL-TR-599, NASA and the US Army Research Laboratory, December 1994.

[15] Stewart, R. M. “Some Useful Analysis Techniques for Gearbox Diagnostics.” Technical Report MHM/R/10/77, Machine Health Monitoring Group, Institute of Sound and Vibration Research, University of Southampton, July 1977.

[16] McFadden, P. D. “A Technique for Calculating the Time Domain Averages of the Vibration of the Individual Planet Gears and Sun Gear in an Epicyclical Gearbox.” *Journal of Sound and Vibration*, vol. 144, 1991, pp. 163-172.

[17] Rafiee, J., F. Arvani, A. Harifi, and M. H. Sadeghi. “Intelligent Condition Monitoring of a Gearbox Using Artificial Neural Network.” *Mechanical Systems and Signal Processing*, vol. 21, 2007, pp. [insert page numbers].

A New Asymmetric Cascaded Multilevel Inverter with Reduced Switches and dc Sources Count

Roya Naderi
Department of Electrical Engineering, Heris Branch, Islamic Azad University, Heris, Iran
Email: r_naderi@iau.ac.ir

Received: 18 November 2024 Accepted: 31 December 2024 Published:15 February 2025

Abstract

In this manuscript, a basic unit topology of cascaded multilevel H-Bridge inverter along with the extended topology is proposed in which H-Bridge is utilized to generate positive and negative voltage outputs. A dc source magnitude calculation method is proposed using which the equations for calculating number of IGBTs, isolated dc sources, gate driver circuits and output voltage levels are derived. The proposed topology is compared with recent similar topologies in different aspects using the mentioned equations. The comparison results reveal that the proposed topology has superiority over similar topologies and requires lower number of power electronic switches, gate driver circuits and isolated dc voltage sources which will decrease complexity and cost of the inverter design. Finally, the accurate performance of proposed inverter will be verified through software simulation in PSCAD/EMTDC platform.

Keywords: Cascaded multilevel H-Bridge, symmetric series multilevel inverter, asymmetric series multilevel inverter

1. Introduction

Recent improvements in developing the multilevel inverters (MLIs) and their applications in medium and high voltage, and the advantages they have over traditional two-level inverters, has made MLIs a suitable alternative for replacing these inverters [1,2]. Generating output voltage with lower total harmonic distortion (THD), having very low common voltage, lower switching frequency, lower electromagnetic interference (EMI), lower switching losses, higher efficiency and…, has made MLIs a popular choice for industrial applications [3]. Having overmentioned advantages over traditional two-level inverters, has made multilevel inverters a suitable option for applications such as high voltage direct current (HVDC), hybrid electrical vehicles (HEVs), Flexible AC Transmission System (FACTS), motor drives, grid-connected photovoltaic systems and uninterruptible power supply (UPS) systems [4-7]. The most common multilevel inverter topologies are, Neutral Point Clamped MLIs (NPC MLIs) [8], Flying Capacitor multilevel inverters (FC MLIs) [9] and Cascaded H-Bridge MLIs (CHB MLIs) [10].

To improve the quality of output voltage in multilevel inverters, number of output voltage levels could be increased to minimize the output voltage THD which will create a possible drawback for MLIs. Number of power electronic components will increase in order to generate higher numbers of output voltage levels, which increases the volume, cost and complexity of the inverters. To deal with this issue, efforts in proposing topologies with reduced number of components has been an ongoing research area for the recent years which has led to introduction of different topologies in this regard [11-15].

Topologies introduced in [11,12], are using symmetrical dc magnitude calculation method with equal dc sources, which in comparison to asymmetric dc magnitude calculation, that uses dc sources with different magnitudes, requires lower volume and is less costly for inverter prototyping, but at the same time it has lower number of output voltage levels. Asymmetric topologies presented in [13-15], use bi-directional common emitter switches, which includes two IGBTs, and two parallel diodes and requires only one gate driver circuit. Topology [15], is composed of two series half-bridges, each of which is connected to an isolated dc source. Although the topology requires more power electronic switches, and gate driver circuits, it guarantees lower voltage stress on the switches. Higher number of IGBTs issue in topology [12], could be notably solved by using uni-directional switches, which have and IGBT and a parallel diode, but it still requires high number of isolated dc sources.

In this manuscript, initially the basic unit of proposed inverter which is capable of generating only positive output voltages, is presented. Then to increase the output voltage levels, the basic unit is extended to form the proposed inverter topology, and to generate the negative voltage outputs, an H-Bridge is added to the output of the inverter. Furthermore, the required equations are derived and to increase number of output voltage levels, an algorithm is proposed to calculate magnitude of dc voltage sources. To verify the performance of the proposed inverter and to study the merits and drawbacks of the proposed inverter over previously introduced topologies, it will be compared with most recent MLI topologies. Finally, the accurate performance of proposed MLI, will be verified through simulation results of a 31-level MLI in PSCAD/EMTDC software platform.

2. Proposed Topology

A 16-level multilevel inverter has been proposed in this manuscript, which consists of 4 uni-directional power electronic switches, 5 bi-directional switches, 2 dc voltage sources in the right side and 2 dc voltage source in the left side. The correct switching pattern for this topology, should prevent short circuiting the voltage sources, so the switch pairs $S_{R,1}$ and $S_{R,2}$, $S_{L,3}$ and $S_{L,2}$, K_1, K_2 and $S_{L,1}$should work in a complementary way, and should not be turned at the same time. Table I presents the output voltage for different states of switches for the topology shown in Fig. 1, in which 1 and 0 represent ON and OFF states of switches respectively.

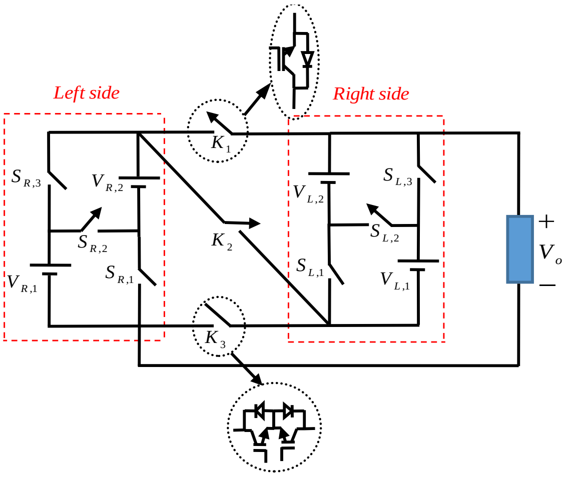


Fig. 1. The basic unit for proposed topology

Table I. Switching states for basic unit of proposed topology

$S_{R,1}$	$S_{R,2}$	$S_{R,3}$	$S_{L,1}$	$S_{L,2}$	$S_{L,3}$	K_1	K_2	K_3	V_o
0	0	0	0	0	0	1	1	1	**0**
1	0	0	0	0	0	1	0	0	$(V_{R,2})V_{dc}$
0	0	1	0	0	0	1	0	0	$(V_{R,1})V_{dc}$
0	1	0	1	0	0	1	0	0	$(V_{R,1} + V_{R,2})V_{dc}$
0	0	0	1	0	0	0	0	1	$(V_{L,2})V_{dc}$
1	0	0	1	0	0	0	1	0	$(V_{R,2} + V_{L,2})V_{dc}$
0	0	1	1	0	0	0	1	0	$(V_{S,2} + V_{L,2})V_{dc}$
0	1	0	0	0	0	0	1	0	$(V_{R,1} + V_{R,2} + V_{L,2})V_{dc}$
0	0	0	0	0	1	0	0	1	$(V_{L,1})V_{dc}$
1	0	0	0	0	1	0	1	0	$(V_{R,2} + V_{L,1})V_{dc}$
0	0	1	0	0	1	0	1	0	$(V_{R,1} + V_{L,1})V_{dc}$
0	1	0	0	0	1	0	1	0	$(V_{R,1} + V_{R,2} + V_{L,1})V_{dc}$
0	0	0	0	1	0	0	0	1	$(V_{L,1} + V_{L,2})V_{dc}$
1	0	0	0	1	0	0	1	0	$(V_{R,2} + V_{L,1} + V_{L,2})V_{dc}$
0	0	1	0	1	0	0	1	0	$(V_{R,1} + V_{L,1} + V_{L,2})V_{dc}$
0	1	0	0	1	0	0	1	0	$(V_{R,1} + V_{R,2} + V_{L,1} + V_{L,2})V_{dc}$

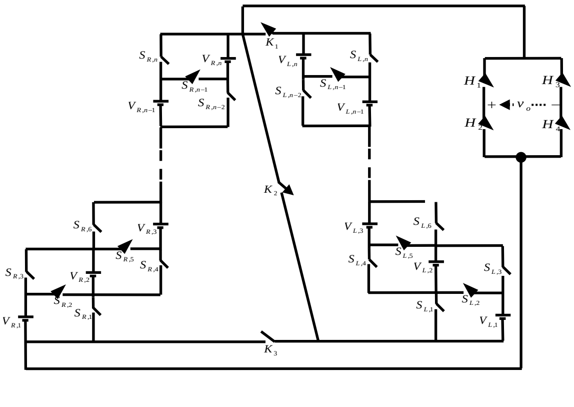


Fig. 2. The proposed Multilevel inverter topology

Error! Reference source not found. The topology shown in **Error! Reference source not found.**, is the extended version of the basic topology shown in Fig. 1, which uses an H-Bridge to generate negative and positive voltage levels in the output.

Number of IGBTs used in the proposed multilevel inverter shown in **Error! Reference source not found.**, (N_{IGBT}), number of isolated dc voltage sources (N_{source}), number of gate driver circuits $(N_{variety})$ could be calculated based on below equations,

$$N_{IGBT} = 10n - 2 \tag{1}$$

$$N_{source} = 2n \tag{2}$$

$$N_{\text{driver}} = 6n + 1 \tag{3}$$

$$N_{variety} = 2n \tag{4}$$

In which, n is the number of dc voltage sources in each side of the inverter. Based on Table, the maximum output voltage and number of generated output voltage levels can also be calculated as;

$$V_{o,max} = (V_{R,1} + V_{R2} + \cdots + V_{Rn} + V_{L1} + V_{L2} + \cdots + V_{L,n})V_{dc} \tag{5}$$

$$N_{step} = 2(V_{R,1} + V_{R,2} + \cdots + V_{R,n} + V_{L,1} + V_{L,2} + \cdots + V_{L,n}) + 1 \tag{6}$$

3. Algorithm for Calculation of dc Source Magnitude

Proper algorithm for calculation of dc source magnitude is of importance in generating different output voltage levels and in avoiding generation of duplicate voltage levels. Therefore, using dc sources with optimum values will increase number of output voltage levels using the same number of basic units.

Based on the proposed algorithm for calculation of dc voltage source magnitude, to avoid generating duplicate output voltage levels for generation of the output voltage levels as presented in Table, magnitude of dc voltage source in each basic unit is calculated by;

$$V_{R,i} = 2^{i-1}V_{dc} \qquad i = 1,2,3,\dots,n \tag{7}$$

$$V_{L,i} = 2^n(2^{i-1})V_{dc} \quad i = 1,2,3,\ldots,n \tag{8}$$

$$V_{o,max} = (2^{2n} - 1)V_{dc} \tag{9}$$

$$N_{step} = 2^{2n+1} - 1 \tag{10}$$

4. Comparison of the Proposed Topology with Common Topologies

To verify the superiority of the proposed topology, a comprehensive comparison on different aspects is performed between this topology and other recently proposed MLI topologies. The first comparison is performed on number of IGBTs, used in different topologies, which based on **Error! Reference source not found.**-a is less than other topologies proposed in [16-26] for $N_{step} \geq 25$. Comparison of number of gate driver circuits is performed in **Error! Reference source not found.**-b, which shows the primacy of proposed topology over other topologies in [16-26] in using lower number of gate driver circuits for $N_{step} \geq 40$. **Error! Reference source not found.**-c reveals that ratio of number of dc voltage sources over number of output voltage levels is equal between proposed topology and the topology in [18] which is lower than all other recent MLI topologies.

Table2. Switching states for the proposed topology

$S_{R,1}$	$S_{R,}$	$S_{R,3}$	⋯	$S_{R,n-1}$	$S_{R,n}$	$S_{L,1}$	$S_{L,2}$	$S_{L,3}$	⋯	$S_{L,n-1}$	$S_{L,n}$	K_1	K_2	K_3	H_1	H_2	H_3	H_4	V_o
0	0	0	⋯	0	0	0	0	0	⋯	0	0	1	1	1	1	0	0	1	**0**
1	0	0		0	0	0	0	0		0	0	1	0	0	1	0	0	1	$(\boldsymbol{V_{R,2}})\boldsymbol{V_{dc}}$
0	0	1		0	0	0	0	0		0	0	1	0	0	1	0	0	1	$(\boldsymbol{V_{R,1}})\boldsymbol{V_{dc}}$
									⋯										⋮
0	0	1	⋯	0	0	0	1	0	⋯	0	0	0	1	0	1	0	0	1	$(\boldsymbol{V_{R,1}+V_{L,1}+V_{L,2}})\boldsymbol{V_{dc}}$
0	1	0		0	0	0	1	0		0	0	0	1	0	1	0	0	1	$(\boldsymbol{V_{R,1}+V_{R,2}+V_{L,1}+V_{L,2}})$ $\boldsymbol{V_{dc}}$
									⋯										⋮
0	1	0	⋯	1	0	0	1	0	⋯	1	0	0	1	0	1	0	0	1	$((\boldsymbol{V_{R,1}+V_{R,2}+\cdots+V_{R,n}}+$ $\boldsymbol{V_{L,1}+V_{L,2}+\cdots+V_{L,n}})\boldsymbol{V_{dc}}$
1	0	0		0	0	0	0	0		0	0	1	0	0	0	1	1	0	$-(\boldsymbol{V_{R,2}})\boldsymbol{V_{dc}}$
0	0	1		0	0	0	0	0		0	0	1	0	0	0	1	1	0	$-(\boldsymbol{V_{R,1}})\boldsymbol{V_{dc}}$
									⋯										⋮
0	0	1	⋯	0	0	0	1	0	⋯	0	0	0	1	0	0	1	1	0	$-(\boldsymbol{V_{R,1}+V_{L,1}+V_{L,2}})\boldsymbol{V_{dc}}$
0	1	0		0	0	0	1	0		0	0	0	1	0	0	1	1	0	$-(\boldsymbol{V_{R,1}+V_{R,2}+V_{L,1}}$ $\boldsymbol{+V_{L,2}})\boldsymbol{V_{dc}}$
									⋯										⋮
0	1	0	⋯	1	0	0	1	0	⋯	1	0	0	1	0	0	1	1	0	$-(\boldsymbol{V_{R,1}+V_{R,2}+\cdots+V_{R,n}}$ $+$ $\boldsymbol{V_{L,1}+V_{L,2}+\cdots+V_{L,n}})\boldsymbol{V_{dc}}$

Table 3. Quantitative comparison of proposed topology and the recent MLI topologies to generate 75 levels of output voltage

REF →	16	17	18	19	20	21	22	23	24	25	26	PROPOSED
N_{level}	75	75	75	75	75	75	75	75	75	75	75	**75**
N_{IGBT}	76	47	22	78	23	78	66	115	42	34	40	**20**
N_{driver}	76	47	22	78	24	66	66	115	42	34	24	**20**
N_{source}	1	19	6	37	1	37	37	37	18	12	19	**6**
Capacitors	-	-	-	-	12	-	-	-	-	-	-	-
Diodes	-	-	-	-	1	-	-	-	16	-	-	-

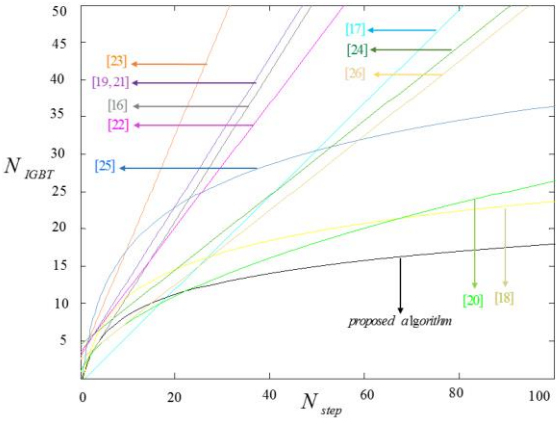


Fig. 3a. Plot of variations of N_{IGBT}, to N_{step}

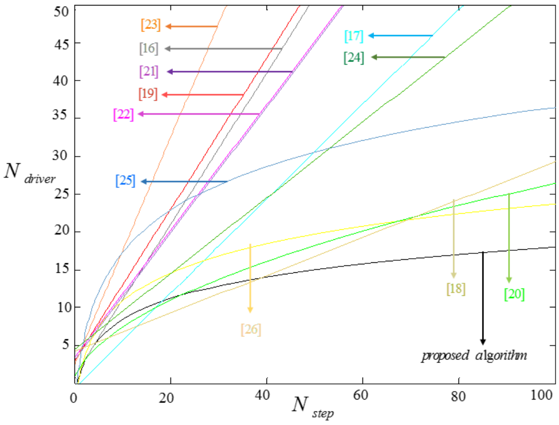


Fig. 3b. Plot of variations of N_{IGBT}, to N_{step}

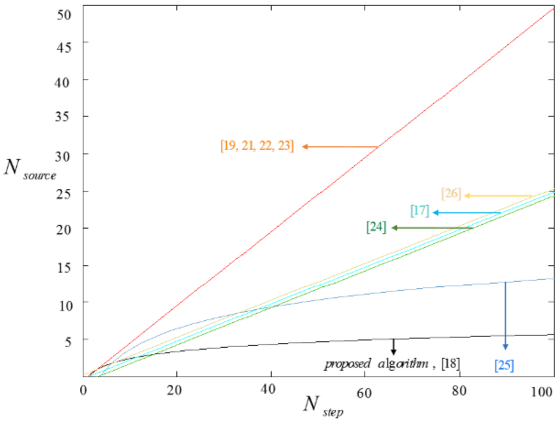


Fig.3c. Plot of variations of N_{source} to N_{step}

Table verifies the over mentioned comparisons by a quantitative comparison on number of power electronic components for inverters at least with 75 levels, between different MLI topologies. Based on this table, the proposed topology uses 20 IGBTs, 20 gate driver circuits and 6 dc voltage sources to generate 75 levels of voltage in the output. However, topologies presented in [16-26] are using 40, 34, 42, 115, 66, 78, 22, 47, and 76 IGBTs 24, 78, 22, 47, 76, 24, 34, 42, 115, 66 and 66 gate driver circuits and 19,12,18,37, 37, 37, 1, 37, 6, 19 and 1 dc voltage source respectively, to generate the same number of output voltage levels in the output.

5. Simulation Results

To verify the performance of the proposed topology, a 31-level cascaded multilevel inverter with output voltage of 150 V and frequency of 50 Hz, is modelled in PSCAD/EMTDC platform with an R-L load of 50 ohms, 90 mH. For the simulation of topology shown in Fig. , based on the proposed dc source magnitude calculation algorithm, values of dc sources $V_{R,1}, V_{R,2}, V_{L,1}, V_{L,2}$ are 10, 20, 40 and 80 volts respectively. In this topology for generating 31 levels of voltage, 4 dc voltage sources, 18 IGBTs and 13 gate driver circuits have been used.

There are several different modulation methods that could be used for generating the switching patterns of MLIs, out of which to keep the switching frequencies lower, the fundamental frequency switching method is used to generate the switching patterns of the MLI switches. In this simulation all the power switches are assumed to be ideal, and output voltage and current waveforms for the proposed inverter is shown in Fig. . Fig. -9 show the standing voltage on OFF switches, maximum of which is 20, 30, 20, 80, 120, 80, 120, 55, 40, 150, 150, 150, 150, volts which leads to a total standing voltage of 1165 the proposed 31-level multilevel inverter. It is worth mentioning that the standing voltage for $S_{R,1}$ has both positive and negative values which proves $S_{R,1}$ to be a bi-directional switch while standing voltage for $S_{R,2}$ only takes positive and zero values which shows that $S_{R,2}$ is a uni-directional switch. THD for output current and voltage is 0.15% and 1.15% respectively. Based on Fig. -b, the output current waveform is almost sinusoidal and free of high order harmonics that has a lower THD value compared to output voltage waveform, which is a result of R-L load acting as a low pass filter.

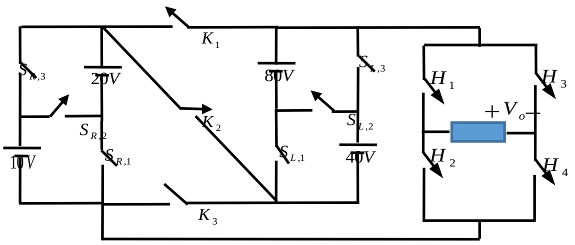


Fig. 4. proposed 31-level MLI

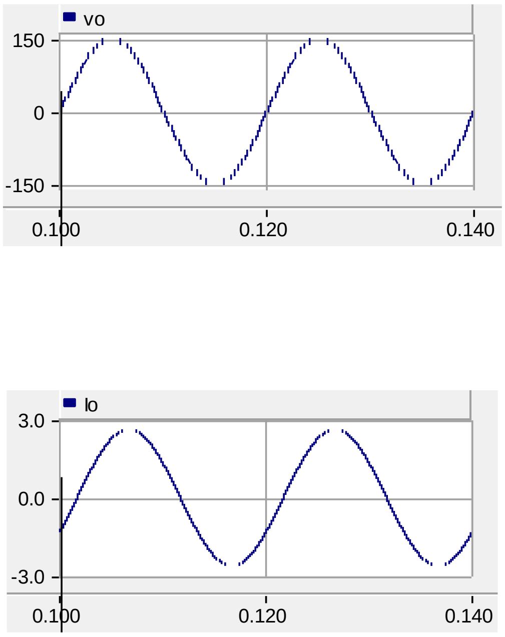


Fig. 5. a) Output Voltage Waveform, b) Output Current Waveform

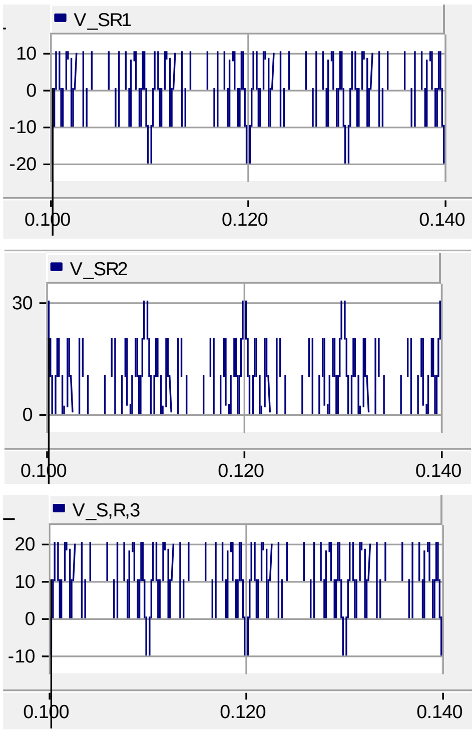


Fig. 6. Standing voltage of different switches in the 31-level MLI

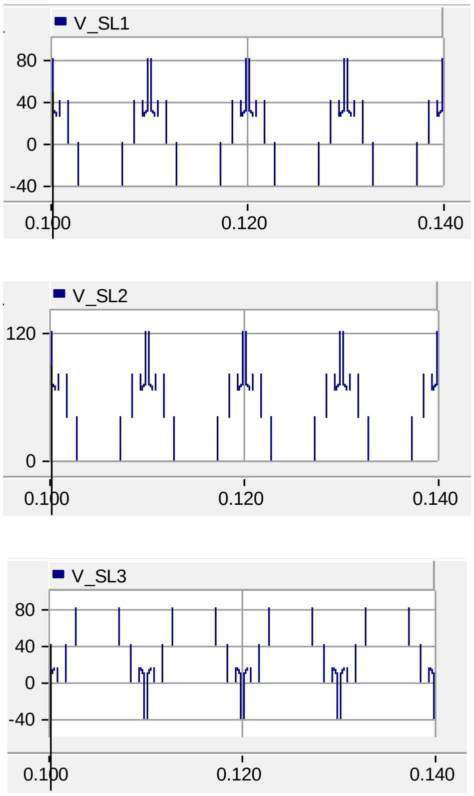


Fig. 7. Standing voltage of different switches in the 31-level MLI

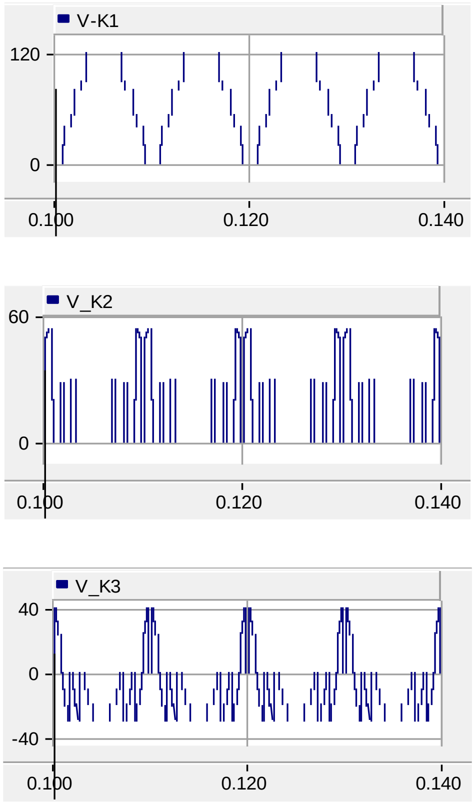


Fig. 8. Standing voltage of different switches in the 31-level MLI

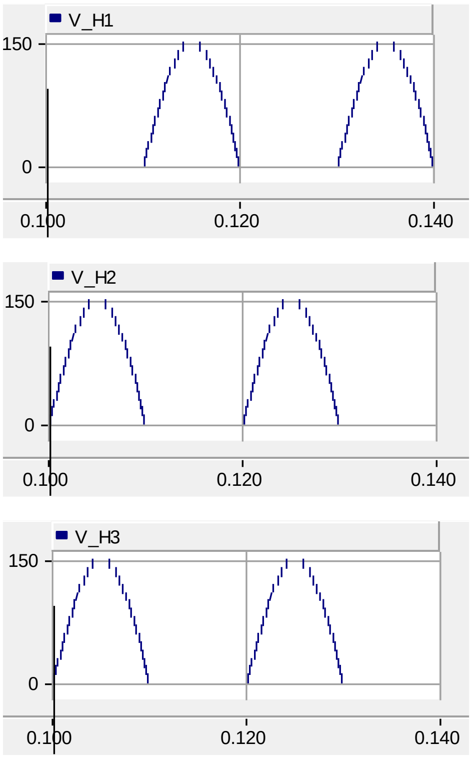


Fig. 9. Standing voltage of different switches in the 31-level MLI

6. Conclusion

In this manuscript, a new basic unit is proposed for a multilevel inverter with reduced number of components. The proposed basic unit, is able to generate 16 levels of output voltage levels using 14 IGBTs and 4 dc sources. To increase the number of output voltage levels, cascaded connection of basic units is presented. To define the merits and drawbacks, the proposed MLI topology is compared with several recently proposed MLIs. The comparisons reveal that, number of IGBTs, gate driver circuits, and dc sources is less for the proposed topology compared to other topologies for generation of the same number of voltage levels in the output, which leads to decreased volume, cost, control complexity and higher efficiency. As an example, to generate 75 levels of output voltage, proposed topology uses 20 IGBTs, 20 gate driver circuits and 6 dc sources whereas the topology in [21], uses 78 IGBTs, 66 gate driver circuits and 37 voltage sources. Finally, to verify the performance of the proposed topology, a 31-level MLI of proposed topology is modelled in PSCAD/EMTDC platform and the results has been reported in the paper.

References

[1] J. Rodríguez, J.S. Lai, and F.Z. Peng, "*Multilevel inverters: A survey of topologies, controls, and applications,*" *IEEE Trans. Ind. Electron*, vol. 49, no. 4, pp. 724–738, Aug. 2002.

[2] H. Abu-Rub, J. Holtz, J. Rodriguez, and G. Baoming, "*Medium-voltage multilevel converters state of the art, challenges, and requirements in industrial applications,*" *IEEE Trans. Ind. Electron*, vol. 57, no. 8, pp. 2581–2596, Aug. 2010.

[3] Y. Naderi, S.H. Hosseini, S.G. Zadeh, B. Mohammadi-Ivatloo, M. Savaghebi, and J.M. Guerrero, "*An optimized direct control method applied to multilevel inverter for microgrid power quality enhancement,*" *International Journal of Electrical Power & Energy Systems,*

vol. 107, no. 0142-0615, pp. 496-506, May. 2019.

[4] G.P. Adam, I.A, Abdelsalam, K.H, Ahmed, and B.W, Williams, "*Hybrid multilevel converter with cascaded H- bridge cells for HVDC applications: operating principle and scalability*," *IEEE Trans Power Electron*, vol. 30, no. 1, pp. 65-77, Jan 2014.

[5] R. Picas, J. Zaragoza, G. Konstantinou, and G.J. Capella, "*Study and comparison of discontinuous modulation for modular multilevel converters in motor drive applications*," *IEEE Trans Power Electron*, vol. 66, no. 3, pp. 2376-2386, Jun. 2018.

[6] M. Calais and V. Agelidis, "*Multilevel converters for single-phase grid connected photovoltaic systems—An overview*," *in Proc. IEEE Int Symp Ind. Electron*, vol. 1, no. 98TH8357, pp. 224–229. Jul. 1998

[7] M. Alsolami, K.A. Potty, and J. Wang, "*A gallium-nitride-device-based switched capacitor multiport multilevel converter for UPS applications*," *IEEE Trans Power Electron*, vol. 32, no. 9, pp. 6853–6862, Sep. 2017.

[8] A. Nabae, I. Takahashi, and H. Akagi, "*A new neutral-point-clamped PWM inverter*," *IEEE Trans Ind Appl*, no. 5, pp. 518–523, Sep. 1981.

[9] T. A. Meynard, H. Foch, P. Thomas, J. Courault, R. Jakob, and M. Nahrstaedt, "*Multicell converters: Basic concepts and industry applications*," *IEEE Trans Ind. Electron*., vol. 49, no. 5, pp. 955–964, Oct 2002.

[10] P. W. Hammond, "*A new approach to enhance power quality for medium voltage AC drives*," *IEEE Trans Ind Appl*, vol. 33, no. 1, pp. 202–208, Jan./Feb 1997.

[11] K. Ramani, M.A.J. Sathik, and S. Sivakumar, "*A new symmetric multilevel inverter topology using single and double source sub-multilevel inverters*," *Journal of Power Electronics*, vol. 15, no.1, pp.96-105. Jan 2015

[12] W. Choi, and F. Kang,"*H-bridge based multilevel inverter using PWM switching function*," *Proc. of in TELEC*, pp. 1-5, Oct 2009.

[13] R. Naderi, E. Babaei, M. Sabahi, and A. Daghigh, "*Optimization and implementation of a new topology for cascaded multilevel inverters with reduced number of semiconductor devices*," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, pp.1-19, Apr 2021.

[14] R. Naderi, E. Babaei, M. Sabahi, and A. Daghigh, "*Optimization of a new extended cascaded multilevel inverter topology to reduce DC voltage sources and power electronic components*," *International Journal of Industrial Electronics Control and Optimization*, vol. 4, no. 4, pp. 465-474, Nov 2021.

[15] G. Waltrich, I. Barbi, "*Three-phase cascaded multilevel inverter using power cells with two inverter legs in series*," *IEEE Trans Ind Appl*, vol. 57, no. 8, pp. 2605-2612, Aug. 2010.

[16] M.J. Uddin, and M.d. Shahidul. Islam, "*Implementation of cascaded multilevel inverter with reduced number of components*," *In 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pp. 669-672. Jan 2021.

[17] M.D.Siddique, S. Mekhilef, N.M. Shah, A. Sarwar, and M.A. Memon, "*A new single- phase cascaded multilevel inverter topology with reduced number of switches and voltage stress*," *International Transactions on Electrical Energy Systems*, vol. 30, no. 2, pp. e12191, Feb 2020.

[18] K.A. Shilpa, M. Revathi, and K. R. Sudha, "*Design and implementation of thirty one level multilevel inverter*," *In Smart and Intelligent Systems*, pp. 547-555, Springer 2022.

[19] M. Revathi, K.A. Shilpa, and K.R. Sudha, "*A novel asymmetric multilevel inverter with low THD*," *In Smart and Intelligent Systems*, pp. 115-132. Springer 2022.

[20] T. Roy, and P.K. Sadhu, "*A step-up multilevel inverter topology using novel switched capacitor converters with reduced components*," *IEEE Trans. Ind. Electron*, vol. 68, no. 1, pp. 236-247, Jan 2020.

[21] M. Farhadi Kangarlu, and E. Babaei, "*A generalized cascaded multilevel inverter using series connection of submultilevel inverters*," *IEEE Trans Power Electron*, vol. 28, no. 2, pp. 625-636, Feb. 2013.

[22] Y. Naderi, M.R. Jannati-Oskuee, M. Karimi, S. Najafi-Ravadanegh, and S.H. Hosseini1, "*A developed asymmetric multilevel inverter with lower number of components*," *AUT Journal of Electrical Engineering*, vol. 50, no. 2, pp. 197-206, Dec 2018.

[23] M. Toupchi Khosroshahi, "*Crisscross cascade multilevel inverter with reduction in number of components*," *IET Power Electron*, vol. 7, no. 12, pp. 2914-2924, Dec. 2014.

[24] Y. Hinago, and H. Koizumi, "*A single-phase multilevel inverter using switched series/parallel dc voltage sources*," *IEEE Trans. Ind. Electron*, vol. 57, no. 8, pp. 2643-2650, Aug. 2010.

[25] S. Saeidabadi, A.A. Gandomi, S.H. Hosseini, M. Sabahi, and Y.A. Gandomi, "*New improved three-phase hybrid multilevel inverter with reduced number of components*," *IET Power Electron*, vol. 10, no. 12, pp. 1403-1412, Oct 2017.

[26] Oskuee, Mohammad Reza Jannati, et al. "A new multilevel voltage source inverter configuration with minimum number of circuit elements." Journal of Central South University 24.4 (2017): 912-920.

Detection of Bone Fracture Area Using Convolutional Neural Network

mohammad fatehi[1], Mohammad Reza baghaei[2], kiumars sharafi[3], mohammad moradi[4], mahdi saneie[5]
[1]Department of Biomedical Engineering, Kazerun Branch, Islamic Azad University, Kazerun, Iran
[2]Department of Biomedical Shahid Beheshti, Tehran Branch, Tehran, Iran
[3]Department of Sports Management, Kerman Branch, Farhangian University, Kerman, Iran
[4]Department of Electrical Engineering, chamran Branch, chamran University, Kerman, Iran
[5]Department of Sports Physiology, Research sciences Branch, Islamic Azad University, Tehran, Iran
Email: mh.fatehi@iau.ac.ir (corressponding Author)

Received: 04 December 2023 Accepted: 22 August 2024 Published:15 February 2025

Abstract

Diagnosis of the fracture site is done using CT-Scan images and based on the doctor's visual diagnosis. This work is very time-consuming and depends on the doctor and his expertise. Systemic methods can help doctors and specialists and can detect the fracture area and the fracture surface. In fractures, only the location of the fracture is determined, but if we want to diagnose the area, high expertise and experience is needed, or in some cases, MRI images are needed. Convolutional neural networks are very powerful in diagnosing diseases and medical complications and can diagnose them correctly. The high accuracy and ability of convolutional neural networks has made this method popular among researchers, and its use is becoming more widespread every day. In this method, fracture location and fracture depth were determined using convolutional neural network. In this work, first the fracture site and then the fracture area are determined. In this study, the location of hip fracture was detected with complete accuracy and the fracture area was obtained with 99.68 accuracy and 99.82% sensitivity. The obtained results indicate that the proposed method is a suitable method for fracture detection.

Keywords: bone fracture, convolutional neural network, aria detection, Pelvis fracture.

1. Introduction

Bone is a living tissue. Bone has blood vessels and living cells. Being alive is what allows it to grow and heal itself. Throughout life, the body's bone tissue is constantly absorbed by the corrosive cells in it (called osteoclasts) and at the same time new bone is replaced by osteoblast cells. In this way, the tissue of all the bones of the body is constantly being renewed [1]. In childhood and youth, bone formation is more than absorption, but in old age, bone is absorbed more than bone is formed, but even in the most severe cases of osteoporosis, the body does not stop building new bone and only its amount decreases. The highest amount of bone density is around the age of twenty [2].

Diagnosis of bone fracture is done by a specialist doctor using MRI and CT_Scan images. In the United States, 2.5 million fractures will occur in 2021. Pelvic fractures are the most serious fractures and the mortality rate is more than 20% per year [3]. More than 50% of patients with hip fracture are not able to return to their previous standing position and approximately 10% of them will need long-term care facilities. Three quarters of hip fractures occur in women. After the age of 50, the risk of hip fracture during life is 20% higher for women [4]. Therefore, accurate and correct and timely diagnosis using computer methods is very important. The faster, less expensive and more accurate this method is, the more desirable it is. The purpose of this research is to accurately detect the location of the fracture and the fracture area using the convolutional neural network.

2. Related works

Fractures usually occur due to impact, and the more severe the impact, the greater the fracture. Hip fracture is one of the severe and slow healing fractures. Hip fracture occurs in two age groups. In young patients and elderly people, which is 70% more in elderly people than in young people, because the quality of their bones is disturbed [5]. Due to demographic changes, this second group is growing rapidly. While extensive research has been conducted on the nature, epidemiology, and treatment of hip fractures, older people are always at greater risk of hip fracture. CT-scan is a common method in fracture diagnosis. Pickhardt and colleagues have described a CT scan-based method for bone quality analysis. They were able to distinguish fracture and non-fracture with 90% accuracy [6].

Also, Chowdhury and his colleagues introduced a method to diagnose hip fracture based on the cutting diagram theory and morphological examination of the hip bone. They diagnosed the fractures by examining the gap in the hip bone. They reached 98.91% accuracy in diagnosis [7]. Also, in [8], using X-ray images based on deep convolutional neural networks (DCNN), healthy and broken bones have been identified. In this research, fracture was detected with 94.32% accuracy. Also, in another research presented in [9], based on convolutional neural networks, they proposed a method for fracture detection using X-ray images. The detection accuracy in this proposal was not higher than 91.2%. In a research based on convolutional neural network, a method for diagnosing femoral head fracture based on VGG architecture was performed with 95.5% accuracy, which is more than the accuracy of conventional orthopedic surgeons (92.2%).[10] .

Also, Sato and his colleagues in [11] presented a new method for hip fracture detection using CT images based on convolutional neural network. The experimental results of 300 images provided an accuracy of 96.1%, which was a high result at that time. Similar methods for fractures of other bones of the body have also been performed, and in [12], a method for detecting vertebral fractures (VCFs) using deep networks was presented, which reached the highest accuracy of 89.1%. In all common methods, they focus on the fracture and achieved the highest accuracy of 98%, but our goal of this research is to achieve reliable accuracy and also to detect the fracture area, so that doctors can be helped in this matter.

3. Materials and Methods

3-1 Database

The dataset was used from the Kaggle database site. which is available on the site under the name "ChestPelvisCSpineScans". This database contains 876 images in two groups consisting of 404 normal images and 472 hip fracture images. Figure 1 shows two sample images.

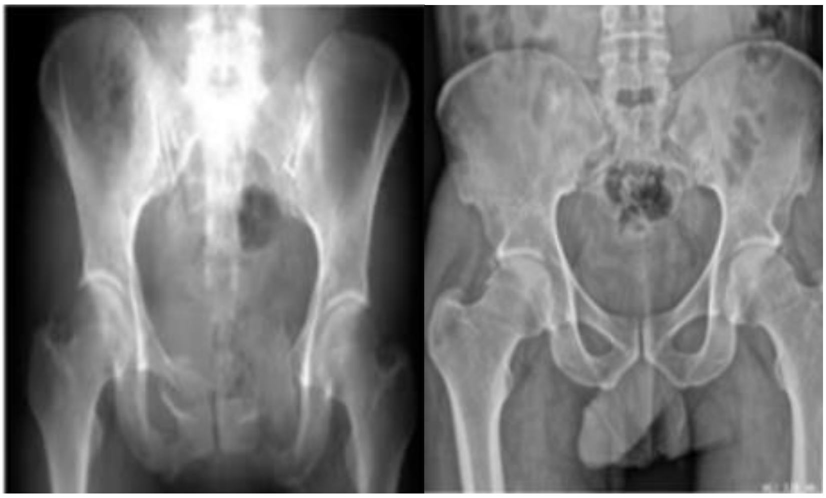


Fig.1. Images of a healthy pelvis and a fractured pelvis

3.2 Preprocessing

At this stage, the images are first converted to the standard size to enter the convolutional neural network so that they can enter the next stage. In the first part of the processing, two groups were separated. For this stage, two groups of broken bones and healthy bones were separated.

In this part, the two-dimensional convolutional neural network with the common and widely used AlexNet architecture was used, which has been used in many studies and has had favorable results. Figure 2 shows the 2D convolutional neural network classifier.

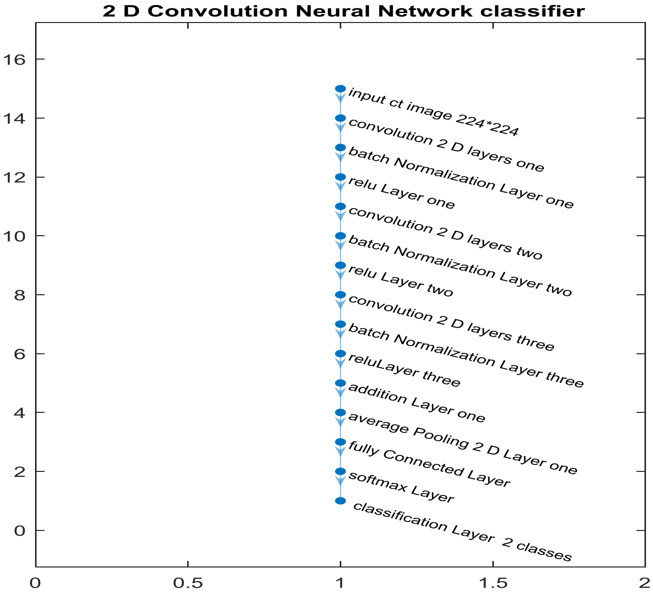

2 D Convolution Neural Network classifier
16
14
12
10
8
6
4
2
0
0
0.5
1
1.5
2
input ct image 224*224
convolution 2 D layers one
batch Normalization Layer one
relu Layer one
convolution 2 D layers two
batch Normalization Layer two
relu Layer two
convolution 2 D layers three
batch Normalization Layer three
reluLayer three
addition Layer one
average Pooling 2 D Layer one
fully Connected Layer
softmax Layer
classification Layer 2 classes

Fig.2. Two-dimensional convolutional neural network classifier

In the next step, the fracture site is diagnosed in the unhealthy group. In this part, the fracture location is detected using the RCNN method. Figure 3 shows the structure of Region convolutional neural network.

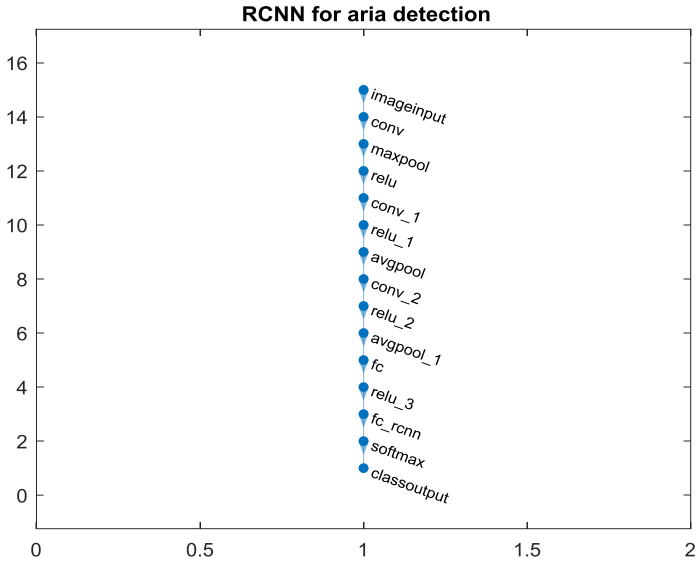


Fig. 3. Region convolutional neural network structure

4. Results

All steps have been done using Asus Cori7-8700 laptop with 16GB RAM by MATLAB 2017b software. The results are calculated based on the confusion matrix. Figure 4 shows the confusion matrix of one of the situations for distinguishing healthy and broken bones.

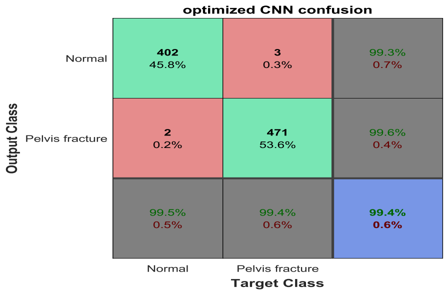


Fig. 4. Confusion matrix of one of the states to distinguish healthy and broken bone.

The accuracy and sensitivity results for the detection of healthy and broken bones have been calculated using 10-fold cross-validation. Figure 5 shows the accuracy results for fracture detection with 10-fold validation.

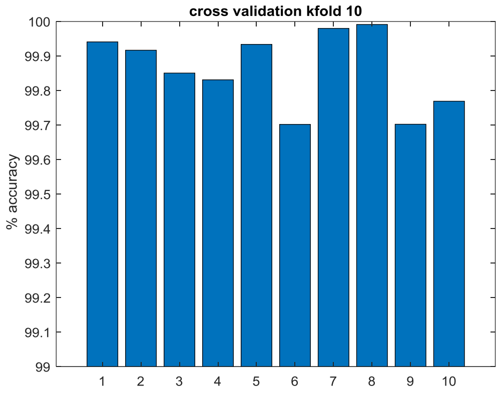


Fig.5. Accuracy results for fracture detection with 10 validations.

Figure 6 shows the sensitivity results for fracture detection with 10-fold validation.

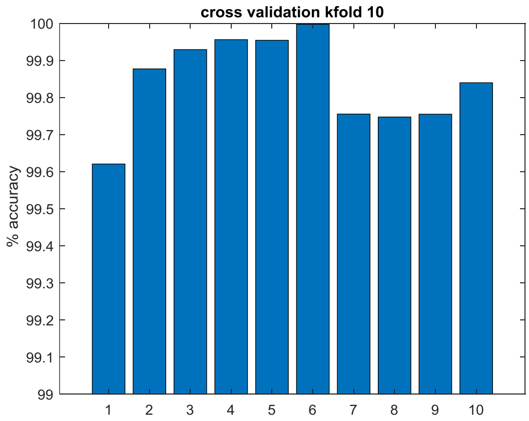


Fig.6. Sensitivity results for fracture detection with 10-fold validation

Finally, the location of the fracture was obtained using RCNN .Figure 7 shows the results of calculating the broken area using RCNN.

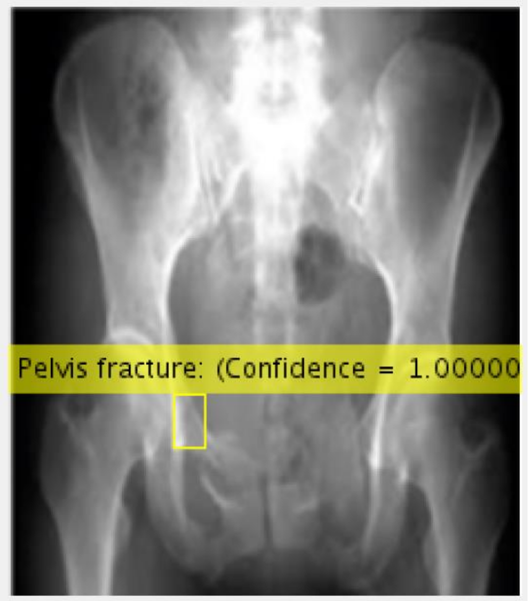


Fig.7. Results of calculating the broken area using RCNN

5. Discussion

Fracture of the femur has many complications due to its location, and if the fracture is not diagnosed, it will cause many problems. In all the common methods examined, healthy and broken bones are detected, but in the conducted research, fractures are detected and then the fracture area is obtained using the RCNN method. In this article, the location of hip fracture was detected with complete accuracy and the fracture area was obtained with 99.68 accuracy and 99.82% sensitivity with 10-fold validation.

References

[1] Mohamed A. Kassem, Soaad M. Naguib, Hanaa M. Hamza, Mostafa M. Fouda, Mohamed K. Saleh, Khalid M. Hosny, "Explainable Transfer Learning-Based Deep Learning Model for Pelvis Fracture Detection", International Journal of Intelligent Systems, vol. 2023, Article ID 3281998, 10 pages, 2023.

[2] S. Atasever, N. Azginoglu, D. S. Terzi, and R. Terzi, “A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning,” Clinical Imaging, vol. 94, pp. 18–41, 2023.

[3] K. M. Hosny and M. A. Kassem, “Refined residual deep convolutional network for skin lesion classification,” Journal of Digital Imaging, vol. 35, no. 2, pp. 258–280, 2022.

[4] M. A. Kassem, K. M. Hosny, R. Damaševičius, and M. M. Eltoukhy, “Machine learning and deep learning methods for skin lesion classification and diagnosis: a systematic review,” Diagnostics, vol. 11, no. 8, p. 1390, 2021.

[5] Y. Ma and Y. Luo, “Bone fracture detection through the two-stage system of crack-sensitive convolutional neural network,” Informatics in Medicine Unlocked, vol. 22, Article ID 100452, 2021.

[6] C. Rainey, J. McConnell, C. Hughes, R. Bond, and S. McFadden, “Artificial intelligence for diagnosis of fractures on plain radiographs: a scoping review of current literature,” Intelligence-Based Medicine, vol. 5, Article ID 100033, 2021.

[7] Chowdhury, A.S., Burns, J., Sen, B., Mukherjee, A., Yao, J. & Summers, R.M. Detection of pelvic fractures using graph cuts and curvatures. In 18th IEEE International Conference on Image Proceedings. 1573–1576.

[8] K. He, Z. Xiangyu, R. Shaoqing, and S. Jian, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on computer vision and pattern recognition, pp. 770–778, Las Vegas, NV, USA, June 2016.

[9] F. Chollet, “Xception: deep learning with depthwise separable convolutions,” in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1800–1807, Piscataway, NJ, USA, July 2017.

[10] Fathi, M., taghizadeh, M., moradi, M., shojaat, G. Diagnosis of Covid-19 using optimized convolutional neural network. journal of Artificial Intelligence in Electrical Engineering, 2023; 11(44): 25-32.

[11] Fatehi, M., khajooee, M., adlband, N., moradi, M. Detection of healthy and unhealthy ECG signal using optimized convolutional neural network. *journal of Artificial Intelligence in Electrical Engineering*, 2022; 11(43): 61-69.

[12] Moradi, M., Fatehi, M., Masoumi, H., Taghizadeh, M. Deep neural network method for classification of sleep stages using spectrogram of signal based on transfer learning with different domain data. Scientia Iranica, 2022; 29(4): 1898-1903.

[13] Kaggle, “ Chest, pelvic and C-spine fractures,” 2022,https://www.kaggle.com/datasets/pardonndlovu/chestpelviscspinescans.

Service Placement in A Fog Computing Environment with Knowledge of Service Quality

Babak Anari , Yousef Abofathi, Mohammad Masdari
[1] Department of Computer Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran
[2,3]Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran
Email: anari322@gmail.com

Received: 03 January 2024 Accepted: 22 August 2024 Published:15 February 2025

Abstract

Fog computing has been proposed to meet the growing demands of Internet of Things users with low latency and high bandwidth. Fog computing has extended cloud computing services to the edge of the network. In this research work, the methods of deploying services in the fog computing environment have been investigated with a focus on service quality. Considering challenges such as resource constraints, heterogeneous environments, and dynamic network conditions, a new framework for deploying services aware of multifaceted aspects of service quality, including response time, availability, reliability, and latency, is proposed. In this article, a new method called DLA-SPSQ is used for optimal placement of clustered services based on service quality criteria. The proposed algorithm is a combination of FCS and DLA-FMP algorithms. The fuzzy FCS algorithm clusters IOT user requests based on the quality-of-service criteria. The quality of clusters is validated based on evaluation criteria. A single objective cost function is used to evaluate the loop delay of modules/services. The results show the improvement of the proposed method in clustering services compared to the case of not clustering them.

Keywords: fog computing, distributed learning automata, quality of service, fuzzy clustering of services, placement of services.

1. Introduction

Cisco announced that more than 75 billion devices will be connected to the Internet by 2025, generating massive amounts of data[1]. Processing large amounts of data with limited resources in real-time is practically impossible. Cloud computing has been used to solve the problem of data processing. Cloud computing also faces increased latency, bandwidth limitations, privacy, and security issues. Fog computing was used to overcome the problems of cloud applications. Fog computing has extended cloud computing services to the network's edge, processing, analysing, and storing data at locations close to users (IoT). Using fog computing results in reducing the amount of data sent to the cloud, reducing delay and calculation costs, and increasing scalability. With all its benefits, fog computing has challenges such as optimal placement of modules/services, resource management, network interference and latency, data management, adaptation to changing environments, security issues, and privacy protection. Optimum placement of services in the fog environment due to the heterogeneous nature and limited capacity of most fog nodes (limited resources), environment dynamics, creation and removal of resources in the fog network, moving people, changing infrastructure and application information over time (e.g., workload change) and geographic distribution of fog devices on an extensive infrastructure is a complex issue. Due to the mentioned reasons, the problem of placing services/modules of Internet of Things applications can be considered an NP-Hard combinatorial optimization problem[2, 3].

In this research work, optimal placement is done on services selected in fuzzy clustering. Fuzzy clustering is performed on them based on service quality criteria. The selected service quality criteria are response time, availability, reliability, and delay. The working method is that the services are initially clustered based on the service quality criteria with the proposed DLA-SPSQ algorithm after removing outliers from the loaded data set. Clustering evaluation criteria have been used to validate the quality of clusters. According to the need of the problem and the acceptable values for the evaluation criteria, clustering with three clusters has been chosen in this research. The XB and average evaluation criteria in each cluster have been used to determine the priority among the three clusters for implementation. After clustering the data set, the services of each cluster are optimally placed in a distributed learning automata system according to the DLA-FMP algorithm. Most of the presented methods for solving the problem of placement of services/modules are based on evolutionary algorithms and without clustering the requests of Internet of Things users, which are unsuitable for online applications due to the time-consuming search for the solution space.

A local and global search framework is designed to locate the modules in the fog topology to implement the DLA-FMP algorithm. To map fog topology to distributed learning automata, a proposed framework has been modelled in the paper x. The proposed algorithm can be used in dynamic environments, and local and global search can be performed simultaneously in the fog environment. The proposed method can be applied to any topology with a high convergence speed. With this method, it is possible to make maximum use of the available capacity of fog nodes in the lower layers and close to the Internet of Things layer, and it is possible to achieve the minimum delay and reduce the execution time of tuples and network consumption

The main contribution of this article is as follows:

- Designing a framework for fuzzy clustering of services based on service quality criteria.
- Providing a new approach for optimal service placement according to service quality criteria.
- Ability to implement the proposed method for homogeneous and heterogeneous topology.
- Validation of the proposed algorithm regarding service execution delay criteria, tuple executiondelay, and network usage reduction.

This paper is further organized as follows: Section 2 provides a summary of related studies, and Section 3 provides concepts related to fuzzy clustering of services, module/service placement problems, learning automata, distributed learning automata, and the evaluated criteria given in the optimal placement of services. The proposed method is explained in Section 4. In section 5, the simulation results and their analysis are given. Finally, Section 6 includes conclusions and future directions.

2. Related Works

This section summarizes several studies related to fuzzy clustering strategies of services and their optimal placement in the fog environment. These studies are motivated by fuzzy system clustering and optimization goals such as minimizing delay or improving service quality. The work related to the fuzzy clustering of services has been done in this article. Most of the studies about optimization have been done in the first article of this research, and here, they are listed in a categorized manner to complete the scope of the proposed algorithm.

A) Several articles of the work done about fuzzy clustering of systems are summarized below:

In[4], the authors address some common challenges in the Internet of Things, such as managing large volumes of data and the need for real-time processing, with the possibility of more efficient and accurate data clustering. This is particularly important for IoT applications where data classification is fundamental to decision-making processes, such as smart cities, healthcare monitoring systems, and industrial automation.

In[5], the researchers propose an innovative algorithmic solution that concurrently optimizes for latency, bandwidth, throughput, cost, and energy consumption to make deployment decisions. They incorporate a Pareto optimization technique, which allows stakeholders to understand the trade-offs and make informed choices about their deployment strategies.

In [6], the authors present an innovative mechanism for embedding Internet of Things (IoT) services in a fog computing environment that prioritizes Quality of Service (quality of service). Acknowledging the importance of meeting quality of service criteria such as latency, throughput, and reliability, the authors have proposed a new service placement strategy using an open-source development model.

In [7],a set of developed algorithms for managing the deployment process is described. These algorithms consider the different QoS requirements of different IoT applications and balance these requirements against the current capabilities and load on cloud and fog resources. Key to this process is the dynamic assessment of network conditions, user demands, and service importance to ensure that QoS objectives are consistently met.

In[8], various algorithms for service placement and resource management, which are designed to maintain quality of service standards, have been discussed. These algorithms focus on minimizing latency, maximizing bandwidth, and ensuring reliability and fault tolerance in IoT services.

The model proposed in[9]considers different quality of service requirements in the context of IoT, such as ensuring low latency, high throughput, and fixed network stability. It features a new architectural solution that uses fog nodes—gateways, routers, or other edge devices—to distribute the computing load between the cloud and end devices. For this, it uses an adaptive mechanism that can dynamically evaluate the service demands of IoT applications in real-time and align them with the capabilities of fog nodes. For example, an intelligent traffic system requires immediate data processing for effective traffic control, which this model prioritizes with efficient use of edge computing resources.

The authors in [10]propose a new approach that combines fuzzy logic with meta-heuristic algorithms for resource provisioning. Fuzzy logic is applied to provide a more flexible decision-making process that can handle the imprecision and uncertainty inherent in cloud environments. It adjusts resource provisioning more subtly than binary logic, which is either too aggressive or too conservative. In this article, fuzzy clustering is used to classify the demand for input resources into different fuzzy categories, and a meta-heuristic algorithm optimizes the allocation of resources in each cluster.

In[11], fuzzy clustering can be used to improve data analysis, resource allocation, and load balancing by assigning workloads to the most appropriate computing layers (cloud canter or edge devices in the fog) and managing inherent uncertainty and variability. The performance and capacity of these resources have been investigated.

In [12], a fuzzy approach for deploying IoT applications in cloud computing environments better aligns with real-world deployment scenarios' dynamic and complex nature and outperforms rigid, rule-based systems.

B)Here, based on the optimization strategies used, a classification of advanced studies is presented, which are (i) mathematical optimization, (ii) heuristic techniques, (iii) meta-heuristics, (iv) machine learning, and (v) Other materials and methods.

Mathematical optimization techniques

Mathematical optimization is finding the best value for an objective function in a permissible set, calculated as the maximum or minimum value based on some criteria. In this method, unlike the complex problem of module placement, more minor problems can be solved because examining the entire solution space in complex problems requires high execution time. Different mathematical optimization models that have been used in fog computing can be called integer linear programming[13], mixed integer linear programming[14], and non-linear integer programming methods [15, 16] which are done as discrete and continuous optimization. Mathematicians have long used this method to solve optimization problems in all sciences, including computers and engineering.

Heuristic techniques

The module placement problem is computationally very complex due to the dynamic nature of the fog infrastructure, and analyzing the entire solution space can be more practical. In this case, heuristic techniques are often used to arrive at the answer. Heuristic techniques use information from previous experiences with similar problems to solve problems. Heuristic techniques include rules that make it easy to implement a practical solution to complex problems but have no guarantee of performance[3, 14].

Metaheuristic techniques

Today, more meta-heuristic methods are used to solve complex and challenging problems. These methods are inspired by nature and are high-level techniques for modeling and optimization. Meta-heuristic methods are optimized by selecting random solutions and acting as a black box. Effective and efficient exploration of the search space avoids local optimization due to the stochastic nature of the method. It improves better solutions in a reasonable amount of time through an iterative search process[17-19].

Machine learning techniques

In several cases, machine learning techniques have been used to solve the problem of placing modules. This method uses data to improve performance among a set of tasks. Based on sample data (training data), it makes a model for decision-making and necessary predictions without explicit planning. Machine learning algorithms have been used in various cases, such as speech recognition, health care[20], machine vision, etc. Developing conventional algorithms to perform the required tasks is impossible and difficult.

Other techniques

Several other strategies have also been reviewed in the literature. A new method called multi-fog placement (MFP) has been introduced [21]to place resources in the Internet of Things systems in the fog environment. In this paper, the authors have used multi-region fog architecture, including several fog nodes, to reduce delay and energy consumption. Comprehensive studies on service placement algorithms have been done[22-26].

3. Preliminares

This section describes the basic elements underlying the challenge of fuzzy clustering of services/modules and their optimal placement, the principles of machine learning, and the framework of distributed learning machines.

3-1- Fuzzy clustering of services

The concept of fuzzy clustering of Internet of Things (IoT) users' service quality criteria using Fuzzy C-Means (FCM) in a fog network environment entail creating groups (clusters) of IoT service quality experiences and preferences that are not sharply defined but overlap. This approach recognizes the subjective and varied nature of user experiences in IoT applications and the importance of latency-sensitive and context-aware processing provided by fog networking. Here is a summary of how FCM applies to IoT quality of service metrics in a fog network deployment:

- **Quality of Service Data Collection:**

This involves collecting various quality of service metrics from IoT devices spread throughout the cloud network. These metrics can include response time, availability, reliability, latency, and others, which may vary in importance depending on the specific IoT context and application.

- **Application of FCM for Clustering:**

Deploy the FCM algorithm to analyze the service quality data. Since FCM allows for fuzzy membership, each IoT user or device can belong to multiple clusters that represent different service quality profiles or experiences.

- ✓ **Initialization:** Choose the number of clusters and initialize the cluster centres.
- ✓ **Membership Assignment:** Compute the membership degree of each IoT user's service quality data for each cluster.
- ✓ **Centroid Update:** Update the cluster centroids based on the calculated membership degrees and the service quality data.
- ✓ **Iterative Process:** Iterate the assignment and update steps until the centroids stabilize within a small tolerance limit.

➢ **Interpretation of Clusters:**

Assess the resulting clusters to understand different categories or levels of service quality experienced by users across the fog network. It can be insightful to identify clusters that reflect high satisfaction, moderate satisfaction, and low satisfaction, for example.

➢ **Action based on Cluster Analysis:**

Utilize the insights from cluster analysis to optimize resource distribution, improve service delivery, and forecast future demands or the need for infrastructure adjustments in the fog network.By using FCM, the fog network can reliably interpret the nuanced, user-reported experiences of service quality, accommodating the inherently imprecise and overlapping evaluation that different IoT users might have. Furthermore, the localized data analytics capability of fog computing allows for real-time or near-real-time clustering and analysis, whichis crucial for promptquality of service adjustments and enhancements in IoT systems.

3-2- Module Placement ProblemDefinition

Suppose an array of IoT applications for processing as$App\{App_1, \dots, App_k, \dots, App_n\}$ defined.App_k representing a collection of m modules, as $\{M_{(k,1)}, \dots, M_{(k,t)}, \dots, M_{(k,m)}\}$. These modules are mandated to execute within the virtualized spheres of either the cloud or the fog on demand from a user. The optimal placement of each module unto an appropriate fog node is contingent upon a specific cost metric. Modules are characterized by the Equation (1).

$$M_{k,t} = \{M_{k,t}^{CPU}, M_{k,t}^{RAM}, M_{k,t}^{Disk}, M_{k,t}^{BW}, M_{k,t}^{DL}, M_{k,t}^{size}\} \quad (1)$$

In Equation (1), $M_{k,t}^{CPU}$, $M_{k,t}^{RAM}$, $M_{k,t}^{Disk}$, $M_{k,t}^{BW}$, $M_{k,t}^{DL}$, and $M_{k,t}^{size}$ correspond to the module's requisites for processing capability, RAM, main memory, network traffic to serve the module, execution deadline (in milliseconds), and the volume of module instructions (in MIPS),respectively. The process of assigning modules,$M_{k,t}$ where t belongs *to t ε{1,…, |*App_k*| }* from applicationsApp_k, *k* ∈*{1,…, n}* to fog nodes $Fog_{(i,j)}$, *iε{0,…, L}, jε{0,…, n*ᵢ*}* is performed such that the best possible solution can be reached with the lowest value of the cost function without violating the service level agreement (SLA) and disrupting the QoS. It is known as the service/module placement problem.

3-3- Learning Automata

The method of automata-based learning involves selecting the most suitable action from a range of possible actions. An automaton selects an action from its limited action set to use in a random environment. This environment then the chosen action and provides feedback to the automaton. With the feedback, the automaton refines its action-selection mechanism, specifically its probability distribution of actions. This process — from action selection by the automaton to environmental application and assessment to the adjustment of the action probability distribution —iterates until the automaton achieves a predefined goal state. An automaton that adapts to a random and unknown environment and enhances its performance is called a learning automaton. This concept was introduced by [27, 28]. Learning automata have found utility across diverse areas, such as computer networks[29],fuzzy logic systems[30], image analysis[31], chaos theory[32], structure identification in Bayesian Networks[33, 34], advanced reinforcement learning[35], and the dissection and interpretation of speech [36]. Figure 1 illustrates the interaction between a learning automaton and its stochastic environmental context.

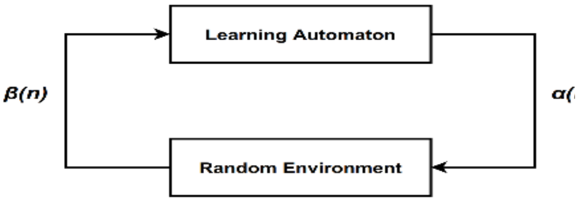


Fig.1. Mutual interaction between learning automata and the random environment

Learning automata are classified into fixed-structure learning automata (FSLA) and variable-structure learning automata (VSLA). Within a fixed-structure automaton, the probabilities associated with selecting actions and the probabilities governing transitions between states are static. Conversely, in a variable-structure automaton, these same probabilities—the action selection likelihoods and state transition frequencies—are dynamic, evolving as time progresses. The variable-structure learning automaton (VSLA) can be conceptualized as LA={α,β,P,T}, where it encapsulates the permitted actions available to the automaton, represents the environmental feedback (reinforcement signals) in reaction to the automaton's selected action, constitutes the set of probabilities for each action (Element Pi indicates the probability of choosing action αi). T denotes the learning algorithm updating the automaton's action probability vector based on the received environmental feedback. In binary environments, the feedback from the environment to the automaton's action is interpreted as either positive/rewarding (β=0) or negative/ penalty (β=1). The most fundamental learning algorithm is linear, which is formalized in Equations (2) and (3). When the automaton's action is rewarded by the environment, the action probability vector gets updated as specified in Equation (2). Conversely, if the environment penalizes the automaton's action, the update is carried out per Equation (3). Within Equations (2) and (3), the variables 'a' and 'b' act as tuning parameters for rewards and penalties, respectively, affecting the increment and decrement rates of the action probabilities.

$$p_j(n+1) = \begin{cases} p_j(n) + a\left(1 - p_j(n)\right), & if\ j = i \\ (1-a)p_j(n), & if\ j \neq i \end{cases} \qquad (2)$$

$$p_j(n+1) = \begin{cases} (1-b)p_j(n), & if\ j = i \\ \dfrac{b}{r-1} + (1-b)p_j(n), & if\ j \neq i \end{cases} \qquad (3)$$

3-4- Distributed learning automata

A distributed learning automata (DLA) is a new model of interconnected automata in which a set of automatons cooperate to solve a specific problem. Choosing an action by an automaton in the network activates the automata corresponding to this action. Only one automaton is active in the network at any time. A DLA with n learner automatons can be defined by a directed graph (A, E) where A= {A_1, A_2,⋯,A_n}is the set of automata and E⊂A×A is the set of edges such that the edge (i,j) corresponds to the j th action from automata Ai to automata A_j. $A_j$will be activated when the learning automata action A_j is selected. The number of actions of the learning automata A_k, $k = 1,2,3,\dots,n$,equals the degree of the output of the node corresponding to the learning automata A_k (Figure 2).

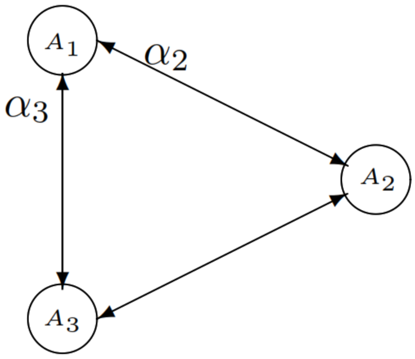


Fig. 2. Distributed learning automata with three automatons

3-5- Application Loop Delay

Each fog node has a waiting queue in which the tuples are placed as soon as they arrive at the nodes. In fog nodes, each tuple is removed from the queue and processed with the default FIFO policy. The application execution loop delay (D) between the user and the deployed fog nodes is calculated using Equation (4).

$$\mathrm{D}=\sum_{\text{for each APP}_j}^{\infty}\left(\mathrm{T}^{s\to f}+\sum_{i=1}^{|\mathrm{APP}_j|-1}\left(\mathrm{PD}_{f\to f'}+\mathrm{T}_{f\to f'}^{\mathrm{Trans}}+\mathrm{T}_f^{\mathrm{Queue}}+\mathrm{T}_f^{\mathrm{process}}\right)+\mathrm{T}^{f\to a}\right) \tag{4}$$

where $T^{s\to f}$ is the transfer time from the sensor to the first fog node, and $T^{f\to a}$ is the transfer time from the last fog node to the actuator; moreover, $\mathrm{T}_{f\to f'}^{\mathrm{Tran}}$, The transmission delay is based on the length of the tuples and the network bandwidth between two end-to-end nodes, which is calculated using equation (5).

$$\mathrm{T}_{f\to f'}^{\mathrm{Trans}}=\frac{\mathrm{TupleNwLength}}{\mathrm{Bandwidth}} \tag{5}$$

In equation (5), Bandwidth and tupleNwLength, respectively, indicate the link's Bandwidth and the size of the tuple used between two end-to-end fog nodes in the network. The waiting time of each fog node is calculated using equation (6).

$$T_f^{Queue}=\sum_{i=1}^{n}T_i^{wait} \tag{6}$$

where T_i^{wait},iε{1,2, 3,…,n}is the waiting time for the tuple i th in the waiting queue. The execution time at each fog node is calculated using equation (7).

$$T_f^{process}=\sum_{i=1}^{n}T_i^{process} \tag{7}$$

In the above equation, $T_i^{process}$ is the execution time of the i th tuple in each fog and iscalculated as Equation (8).

$$T_{\mathrm{f}}^{process} = \frac{TupleCpuLength}{Total\_MIPSoffog} \quad (8)$$

In equation (8), *TupleCpuLength* indicates the processing required to execute a tuple, and *Total_MIPS* of the fog indicates the total MIPS allocated to the processing element node of each fog (PE). The propagation delay is calculated using equation (9) between two adjacent fog nodes.

$$PD_{f\to f'} = \frac{D_{f\to f'}}{PS} \quad (9)$$

The end-to-end distance between two adjacent fog nodes is indicated by $D_{f\to f'}$, PS indicates the speed of light, and its size is equal to$3 * 10^{8^{\mathrm{m/s}}}$.

3-6- Tuple CPU Execution Delay criterion

The execution delay of tuple processing is calculated using Equation (10).

$$Tuple\ Cup\ Execution\ Delay = \quad (10)$$
$$\sum_{for\ All\ Tuple\ Type} AverageCpuTime_{(TupleType)}$$

where *AverageCpuTime* indicates the average execution time of each type of entered tuple.

3-7- Network Usage criterion

IOT input devices transmit the data required for processing in multiple byte units to fog or cloud nodes in higher layers. The amount of data sent and received by the network is called network usage. Network utilization is calculated using equation(11). Network usage depends on the size and overall delay of the request between the source device and the user's desired destination.

*Network usage= Latency*TupleNwSize* (11)

In equation (11), Latency indicates the sending delay of each node in the fog network, and TupleNwSize indicates the file size of each sent tuple.

4. The Proposed DLA-SPSQ Algorithm

The proposed Distributed Learning Automata-Service Placement based Service Quality (DLA-SPSQ) method is performed in two steps, which are:

1. Fuzzy clustering of services
2. Optimum placement of services based on distributed learning automata

Fuzzy clustering of services is done by FCS Algorithm and optimal placement of services is done based on distributed learning automata by DLA-FMP method, each of the methods are explained in order below.

4-1- Fuzzy clustering of services (FCS)

Web service quality of service (QoS) is essential to the overall user experience. This includes response time, availability, reliability, and latency. Due to the wide variation and uncertainty in these parameters, the fuzzy logic approach is often used. Fuzzy logic enables more flexible and realistic modeling of complex systems by handling fuzzy and ambiguous data. Fuzzy-based clustering is a classification method in machine learning where data elements are grouped based on their similarity. In the QoS of web services, fuzzy-based clustering can separate services into different quality levels or groups based on specific QoS characteristics. This allows easy management and prioritization of services.

Fuzzy-based QoS clustering of web services has been widely used and researched. Several models and techniques have been proposed to improve service selection, matching, and composition in web services. Continued development in this area promises more efficient and user-centric web services.

```
AlgorithmFCS
    1.  Load dataset.
    2.  Remove extraneous data from dataset.
    3.  fuzzy clustering.
    4.  Calculate below the Evaluation criteria of clustering for validating cluster quality.
            A)  PC (Partition Coefficient)
            B)  CE (Classification Entropy)
            C)  XB (Xie-Beni Index)
            D)  avgSilhouette
    5.  Determining the priority of clusters for placing services according to the XB values of each cluster and the average value of clustering quality criteria in each cluster.
```

Each of the steps of the FCS algorithm is explained in order below.

- **Load dataset**

In this section, the method of loading datasets in different formats for fuzzy clustering of web services requested by Internet of Things users to run in the fog network by the Load dataset algorithm is explained in two modes.

```
Algorithm Load dataset
Input:your dataset
    1.  Load datasetfrom mat file:
        //Path and file name with extensionmat.
        path='D:\aaaFuzzy110\fuzzy classification\datasetName.mat'
        Data=load(path);
    2.  Load datasetfrom xlsx file:
        /Path and file name with extensionxlsx.
        file Path = 'C:\Users\Home\Desktop\ xlsxFileName.xlsx'
        Data = xlsread(filePath);
```

- **Remove extraneous data from dataset.**

Outliers can significantly bias clustering results because they may be inappropriately attributed to the degree of membership in a cluster. To remove outliers, you should consider preprocessing your data.

While MATLAB does not have a built-in function specifically to detect and remove outliers for direct FCM, it is certainly possible to combine several steps and techniques to detect outliers before performing FCM. Use statistics, including:

1) **Standard deviation method**: If the data is normally distributed, about 68% of the data values fall within one standard deviation of the mean and 95% fall within two standard deviations. Data points that lie beyond a certain threshold may be considered outliers.

2) **Interquartile range (IQR):** The spread of the middle 50% of values. Anything more than 1.5 times the IQR above the third quartile and below the first quartile can be considered an outlier.

3) **Standard Deviation Method:** If data is normally distributed, then around 68% of data values will lie within one standard deviation of the mean, and 95% within two standard deviations.

4) **Boxplot Analysis:** Using boxplots can help visualize potential outliers.

5) **Proximity Based Methods:** Such as DBSCAN, where data points that do not fall within a cluster are considered outliers.

This section explains how to remove outliers from a data set using the IQR method by the remove outlier's algorithm.

```
Algorithm remove outliers
Input: dataset
Output: filteredData
% Compute the lower and upper bounds for each feature
Q1 = quantile (dataset, 0.25);
Q3 = quantile (dataset, 0.75);
IQR = Q3 - Q1;
lowerBound = Q1 - 1.5 * IQR;
upperBound = Q3 + 1.5 * IQR;
% Initialize a logical index vector assuming all values are not outliers
nonOutlierIdx = true (size (dataset, 1), 1);
% Check for outliers in each feature dimension
for i = 1: size (Data.data10, 2)
nonOutlierIdx = nonOutlierIdx& ...
        (dataset (: i) >lowerBound(i)) & ...
        (dataset (: i) <upperBound(i));
end
% Use the nonOutlierIdx to filter the non-outliers
filteredData = dataset (nonOutlierIdx, :);
```

The standard methods of outlier detection might incorrectly remove non-outliers, especially when the data is not normally distributed or when the "outliers" actually represent valuable extremes that are of interest.

FCM is less sensitive to outliers than hard clustering methods because it assigns a degree of belonging to each cluster rather than absolute membership. But the presence of outliers can still affect the centroids and therefore the results of the clustering process.

- **Fuzzy clustering**

The FCM is a soft clustering method that allows a single web service to belong to multiple clusters to varying degrees. This characteristic is particularly beneficial in fog environments where web services exhibit varying quality-of-service (QoS) attributes and may not fit strictly into a single category. For instance, a video streaming service might rank highly in bandwidth but lower in response time, fitting into different clusters for different QoS assessments. By applying FCM, fog nodes can categorize services into clusters based on criteria like latency, bandwidth, reliability, and throughput.

This section explains how to use the FCM (Fuzzy C-Means) function for fuzzy clustering of web services based on service quality criteria using a fuzzy clustering algorithm.

```
Algorithm fuzzy clustering
Input: dataset, number of clusters
Output:clusters, U,centers
X= dataset
[centers, U] = fcm (X,numberofclusters)
maxU=max(U); // U is the membership matrix from fcm
// Identifies data points that are most strongly associated with each cluster.
$index_k$=find (U (k, :)==maxU); □ k=1,2, 3, … ,numberofclusters
//In this step, a subset of the dataset is created.
$cluster_k$= X ($index_k$, :) □ k=1,2, 3,…, numberofclusters
```

- **Calculate Evaluation criteria of clustering for validating cluster quality**

Calculating cluster validity indices is an essential step in evaluating the performance of a clustering algorithm in MATLAB. Cluster validity indices can help assess the quality of cluster formation—for example, how distinct they are, how dense they are, and how well they fit the underlying data distribution.

For fuzzy clustering, such as that performed by the FCM (Fuzzy C-Means Clustering) function in MATLAB, standard validity indices are:

- **PC (Partition Coefficient)**

This method is specific to fuzzy clustering. The PC is calculated for different numbers ofclusters, and a higher value indicates a better clustering structure.It can be calculated using equation 12.

PC = sum (sum (U.^2)) / size (X, 1) (12)

where U is the membership matrix from FCM, and X is dataset.

CE (Classification Entropy)

Evaluates the distribution of membership values across clusters. Lower values are generally better, indicating a more defined partition.It can be calculated using equation 13.

CE = -sum(sum((U.*log(U)))) / size (X, 1) (13)

XB (Xie-Beni Index)

Provides a ratio of compactness and separation of clusters; a lower value of the Xie-Beni index indicates a better partition due to compact and well-separated clusters. FCM algorithm can be run for different numbers of clusters and XB index can be calculated for each partition. The number of clusters with the lowest XB index can be considered as optimal clustering, which can be calculated using the Xie-Beni Index algorithm.

AlgorithmXie-Beni Index

```
Input: dataset,centers, U //U is the membership matrix from FCM
Output:XB
X= dataset
sum_dist = sum(sum(sum((U.^2))).* pdist2(X, centers).^2);
min_dist = inf;
for i = 1:2
   for j = i+1:3
      d = norm (centers (i, :) - centers (j, :), 2);
min_dist = min (min_dist, d);
   end
end
XB = sum_dist / (size(X,1) * min_dist^2);
```

- **Silhouette Coefficient**

Calculating the silhouette coefficient in fuzzy clustering, especially in Fuzzy C-Means (FCM), can be tricky since each data point has a degree of belonging to every cluster, not just a single one. The silhouette coefficient measures how similar a point is to its own cluster compared to other clusters, which is straightforward in hard clustering methods. Still, for fuzzy clustering, one must first determine the degree to which a data point is assigned to its clusters.

Here's how to calculate the silhouette coefficient for a clustered dataset, which the FCM function in MATLAB can calculate.

crispy the Clusters:

Although in FCM, each data point has a membership degree to each cluster, to calculate the silhouette coefficient we need a “hard” assignment. Assign each data point to the cluster for which it has the highest membership value.

Calculate a(i)Within-ClusterDissimilarity:

For each data point i, calculate the average distance from i to all other points in its assigned cluster.

Calculate b(i)Between-Cluster Dissimilarity:

For the same data point i, calculate the average distance from i to all points in the next nearest cluster that it is not assigned to. One way to define the “next nearest” cluster could be finding the cluster for which the average distance to all its points is minimal, excluding the cluster to which i is assigned.

Compute the Silhouette Value:

The silhouette value s(i) for each data point is calculated byequation 14:

$$s(i) = (b_i - a_i) / \max (a_i, b_i) \qquad (14)$$

Here, a(i) represents the average distance from i to all other points in its own cluster, and b(i) represents the smallest average distance from i to points in any other cluster that i is not a part of.

Average Silhouette Coefficient:

To find the overall silhouette coefficient for the dataset, average s(i) over all data points i.

In MATLAB, the silhouette algorithm is designed for hard clustering outputs. It is necessary to rewrite the existing function to implement the silhouette coefficient in fuzzy mode. It is explained here under the Silhouette Coefficient Algorithm.

AlgorithmSilhouette Coefficient

```
 Input: dataset, U //U is the membership matrix
from FCM
 Output:avgSilhouette//Average silhouette value of
the dataset
 X= dataset
 [~, hardClustering] = max (U, [], 1);
 % Peral locate silhouette values array
 s = zeros (size (X, 1), 1);
  for i = 1: size (X, 1)
 % within-cluster distance
 a_i = mean (pdist2(X(i, :), X(hardClustering ==
 hardClustering(i), :)));
       % between-cluster distances excluding the own
cluster
 b_i = inf;
 for k = setdiff(unique(hardClustering),
hardClustering(i))
 distToOtherCluster = mean (pdist2(X (i, :),
 X(hardClustering == k, :)));
 b_i = min (b_i, distToOtherCluster);
 end
 s(i) = (b_i - a_i) / max (a_i, b_i);
 end
  avgSilhouette = mean(s)
```

This simplified algorithm does not handle some cases (such as when a_i is zero), so this must be modified for the specific application and data set.

4-2- DLA-FMP Algorithm

After the data set's clustering, each cluster's services are optimally placed in a distributed learning automata system according to the DLA-FMP algorithm[37]. With the optimal placement performed by the Distributed Learning Automata – FogModule/Service Placemen(DLA-FMP) algorithm, it is possible to significantly reduce the delay in the execution of services in clusters with higher priority and respond to the requested users' services who have paid more money at a suitable time. Provide higher quality services to increase user satis faction. The general steps of the DLA-FMP algorithm are illustrated by the diagram presented in Figure 3.

This section describes the main concepts of the algorithm proposed in article x. The DLA-FMP algorithm's working method is that the required parameters are first defined and then quantified. In the next step, a fog topology is created using the iFogSim tool after the fuzzy clustering service. To perform optimal placement, a distributed learning automata is mapped on the defined fog network so that each of the fog nodes in the fog network corresponds to a learning automaton.

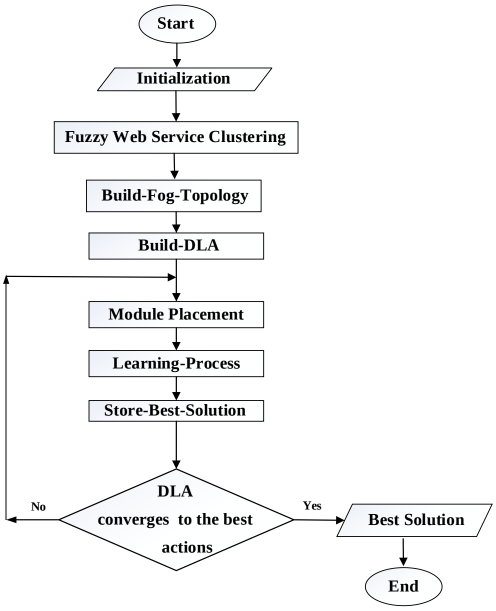


Fig. 3. DLA-FMP algorithm steps diagram

This section describes the main concepts of the algorithm proposed in article x. The DLA-FMP algorithm's working method is that the required parameters are first defined and then quantified. In the next step, a fog topology is created using the iFogSim tool after the fuzzy clustering service. To perform optimal placement, a distributed learning automata is mapped on the defined fog network so that each of the fog nodes in the fog network corresponds to a learning automaton.

The proposed algorithm for solving the module placement problem uses three main phases in each iteration: Phase 1: Module placement, Phase 2: Learning process, and Phase 3: Cost function evaluation phase. These three phases are explained below:

Phase1-Module Placement: In this phase, each application is placed in a module in one of the fog nodes, starting from the edge-level fog nodes. Each edge-level automaton chooses one action from its set of actions to place each module. After selecting the action, the capacity of the fog node corresponding to the selected action is checked using Equation (15).

$$M_{(k,t)}^{RAM} \le Fog_{(i,j)}^{RAM}, M_{(k,t)}^{cpu} \le Fog_{(i,j)}^{cpu} \qquad (15)$$

The desired module is deployed if the above conditions apply to the selected fog node.Otherwise, the learning automaton corresponding to the selected action will be responsible for placing that module. This process continues until the module can be placed in one of the fog nodes or that module is placed in the cloud. The important point in this process is that the cloud can only choose one action. In this phase, learning automata in DLA finds a suitable place for each desired module of each application.

Phase 2 - Learning process: The output of Phase1 indicates that each application module is located in which fog node. In Phase 2, we check whether the action selected by the automaton corresponding to each fog node could minimize the objective function. If that automaton has minimized the objective function, it is inferred that the action chosen by that automaton is suitable. Thus, that action will be rewarded according to Equations (2) and (3); otherwise, it will be penalized. This learning process is repeated until DLA converges to the best actions.

Phase 3 - The cost function evaluation and Store-Best-Solution: The cost function is used to evaluate the rewards or penalties in the learning process. The cost function (D) is considered as a single objective function in the proposed method. Therefore, the goal is to reduce the average delay of IoT user services (D) in all fog nodes during the execution of applications.In this phase, the best solution is saved after learning and evaluating the cost function.

These three main phases continue until the distributed learning automata converges to the best solution.

5. Evaluation and Experimental Analysis

The evaluation, similar to the proposed DLA-SPSQ method, is done in two stages, which are:

5-1- Fuzzy clustering of services

In these stages, the selected data set is clustered by Fuzzy C-Means (FCM) in MATLAB after removing extraneous data. Clustering evaluation criteria have been calculated to validate the quality and optimal number of clusters. High-priority clusters are determined by calculating the average service quality indicators in each cluster. The data set (QWS Dataset) has been used for clustering services[39]. From the service quality criteria, four criteria have been selected according to the purpose of the problem, which are given in Table 1.

Table 1. Selected service quality criteria

Title	Description
Response Time	Time taken to send a request and receive a response
Availability	Number of successful invocations/total invocations
Reliability	Ratio of the number of error messages to total messages
Latency	Time taken for the server to process a given request

To remove outlier data from the data set, the algorithm (remove outliers) mentioned in section 4-1 is used; also, for easy work, the fuzzy k-means method can be used. The calculation results of the PC, CE andavgSilhouetteof clustering evaluation criterion for validating the quality of the clusters are given in Table 2.According to the need of the problem and the acceptable values for the evaluation criteria, clustering with three clusters has been chosen in this research work (row 2 of Table 2).

Table 2. Selected clustering from clustering evaluation criterion

Clusters	PC	CE	avgSilhouette
2 Cluster	0.964617	0.064732	0.872612
3 Cluster	0.92457999	0.1478412	0.7893497
4 Cluster	0.808208	0.347637	0.574982
5 Cluster	0.774274	0.41459	0.522598
6 Cluster	0.736786	0.492458	0.479991
⋮	⋮	⋮	⋮

The calculation results of the XB clustering evaluation criterion for validating the quality of the clusters are given in Table 3.

Table 3. Selected clustering from XB criteria

Clusters	XB1	XB2	XB3	XB4	XB5	XB6
2 Cluster	164	2335				
3 Cluster	2498	928	21345			
4 Cluster	198006	8608	7893	33356		
5 Cluster	107645	11261	31998	12773	316640	
6 Cluster	19119	830225	388270	85181	22014	25598
⋮	⋮	⋮	⋮	⋮	⋮	⋮

In this research work, the priority of the clusters can be determined according to the lowest amount of response time, delay, reliability, XB criterion and the highest amount of availability.

Table 4.Selected cluster from the average value of QoS criteria in each cluster

Evaluation Criteria	Cluster1	Cluster 2	Cluster 3
Response Time	1214.233	237.0746	3269.776
Availability	77.1105	81.5802	77.12698
Reliability	71.67403	69.56739	72.11111
Latency	216.926	32.65847	379.0551
XB	2498	928	21345
Priority	2	1	3

In Table 4, the lowest average result for the evaluation criteria of response time, reliability, delay, XB and the highest average availability are obtained for cluster 2. Therefore, with certainty, the first priority can be assigned to cluster 2. (lighted with yellow color). According to the values obtained for other clusters, clusters 1 and 3 are placed in the second and third priorities, respectively.

5-2- Simulation and evaluation of the proposed method

In these stages, the DLA-FMP method has been used to calculate *AppLoopDelay*, *Tuplecupexecutiondelay*, and *Networkusage* criteria in high-priority clusters of different sizes, and the results of the proposed DLA-FMP method have been evaluated and compared in two modes of clustering and without clustering of services.

5-2-1- Simulation Environment

The iFogSim simulator tool and Java programming language have been used to simulate the proposed method. The specifications of the system used are given in Table 5. The iFogSim simulator uses a tree structure to generate fog topology. Since the proposed fog topology structure is a graph, the iFogSim simulator has been developed to simulate the proposed method so that any graph can be defined and created.

Table 5. Specifications of the simulation system

Name	Description
CPU	Core i7-3720QM CPU @ 2.60GHZ
RAM	32.0GB
Memory	1TB+ 128GB SSD
OS	Window10-64 bit

5-2-2- Parameter Setting

The proposed method considers the reward (a) and penalty (r) parameters of 0.3 and 0.003, respectively. The values assigned to their parameters in related articles have been used in implementing the compared algorithms.

5-2-3- Fog Device Characteristic

In the simulation of the proposed method, a heterogeneous topology with 63 fog nodes, which has 32 edge nodes and is shown as $T_{63}(32)$, is used. The CPU and RAM specifications of fog nodes of topology $T_{63}(32)$ are given in Table 6. In addition, the specifications related to the power consumption of each fog node in the idle state (idlePower), high bandwidth (UpBW) and low bandwidth (downBW) are given in Table 7.

Table 6. Topology with 63 fog nodes $T_{63}(32)$[37]

Level	Fog$_{(\mathbf{i,j})}$$(\boldsymbol{CPU}\ ,\boldsymbol{RAM})$
0	Fog$_{(0,0)}$(44800,40000)
1	Fog$_{(1,0)}$(22400,20000),Fog$_{(1,1)}$(33600,30000)
2	Fog$_{(2,0)}$(18000,16000),Fog$_{(2,1)}$(17000,15000), Fog$_{(2,2)}$(19000,17000),Fog$_{(2,3)}$(20000,18000)
3	Fog$_{(3,0)}$(8400,8000),Fog$_{(3,1)}$(8400,8000), Fog$_{(3,2)}$(5600,6000),Fog$_{3,3)}$(5600,6000), Fog$_{(3,4)}$(16000,14000),Fog$_{(3,5)}$(15000,13000), Fog$_{(3,6)}$(14000,12000),Fog$_{(3,7)}$(12000,10000)
4	Fog$_{(4,00)}$(8400,8000),Fog$_{(4,01)}$(8400,8000), Fog$_{(4,02)}$(18000,16000),Fog$_{(4,03)}$(8600,6000), Fog$_{(4,04)}$(12000,10000),Fog$_{(4,05)}$(13000,8000) , Fog$_{(4,06)}$(16000,14000),Fog$_{(4.07)}$(9400,6000), Fog$_{(4,08)}$(16000,10000),Fog$_{(4,09)}$(5600,6000), Fog$_{(4,10)}$(4200,2000),Fog$_{(4,11)}$(5600,4000), Fog$_{(4,12)}$(14000,8000),Fog$_{(4,13)}$(12000,10000) , Fog$_{(4,14)}$(8000,6000),Fog$_{(4,15)}$(6000,4000),
5	Fog$_{(5,00)}$(4000,2000),Fog$_{(5,01)}$(4000,2000) , ... ,Fog$_{(5,31)}$(4000,2000)

Table 7. Fog nodes parametervalues

Fog node	UpBW, downBW, idlePower
Cloud	100, 10000, 1332
Fog nodes	10000, 10000,83.4333

5-2-4- Application Characteristic

The iFogSim simulator tool was used to create programs and related modules. The general structure of the program used in the iFogSim simulator is shown in Figure 4. It includes three modules: ***Object_Detector***, ***Motion_Detector***, and ***Object_Tracker***.One ***Object_Detector*** module is considered for App_1, and therequired ***Object_Detector*** modules are considered for applications $App_2,\ App_3\ and\ \ App_4$ according to Table 8.

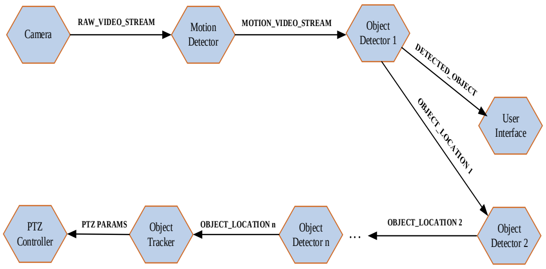


Fig. 4. Modules and edges of the smart monitoring application [38]

Table 8. Specifications of the generated applications

Application Name	Modules	$\boldsymbol{App_k}(\text{"}\ \boldsymbol{M^{CPU}_{(K,1)}}, \boldsymbol{M^{RAM}_{(K,1),}}\ \text{"}, \ldots, \text{"}\ \boldsymbol{M^{CPU}_{(K,m)}}, \boldsymbol{M^{RAM}_{(K,m),}}\ \text{"})$
App_1	3	App_1(“50,10”,”400,10”,”200,10”)
App_2	6	App_2(“50,10”,”400,10”,”200,10”, ”600,10”,”200,10”,”400,10”)
App_3	9	App_3(“50,10”,”400,10”,”200,10”, ”600,10”,”200,10”,”400,10”, ”200,10”,”600,10”,”200,10”)
App_4	12	App_4(“50,10”,”400,10”,”200,10”, ”600,10”,”200,10”,”400,10”, ”200,10”,”600,10”,”200,10”, ”300,10”,”400,10”,”300,10”)

5-3- Experiments

In this section, the results related to the simulation and evaluation of each parameter in both clustering and non-clustering modes of services have been reviewed and compared

5-3-1-Experiment1

In this experiment, the application loop delay is calculated by the DLA-FMP algorithm in the two modes of clustering and non-clustering of services by Equation 4. The simulation results in both cases are given in Figure 5. According to this figure, due to the optimal placement of high-priority cluster services in resources close to the network's edge, they are more efficient than running the service randomly. As a result, the proposed method performs better in reducing the delay of the program loop.

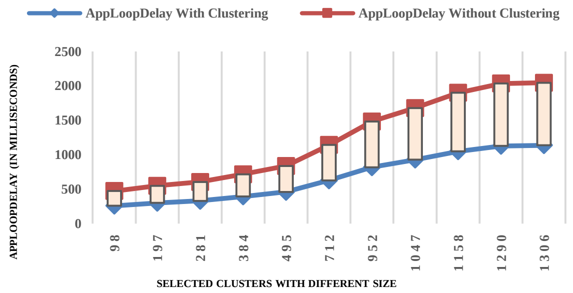


Fig. 5. Comparison of AppLoopDelay in two modes of clustering and non-clustering of services

According to the results obtained in Figure 5 and comparing the results in both cases, the proposed algorithm has improved the App Loop Delay parameter in the service clustering mode compared to the service non-clustering mode. On average, the application loop delay of applications is an 18.5% reduction.

5-3-2-Experiment 2

This experiment aims to investigate the amount of delay caused by the processing of tuples for services in both clustered and non-clustered modes during the simulation process. According to Figure 6, the average processing delay of tuples in the clustering mode of services compared to the non-clustering mode is acceptable by increasing the number of services in the selected cluster and has experienced an improvement of 17.78 percent. Due to the optimal placement of the modules, the increase in the delay in the processing of tuples in the proposed method has been less in the clustering mode than in the non-clustering mode. Therefore, the processing delay of tuples is significantly increased with the increase of the number of modules in non-clustering mode.

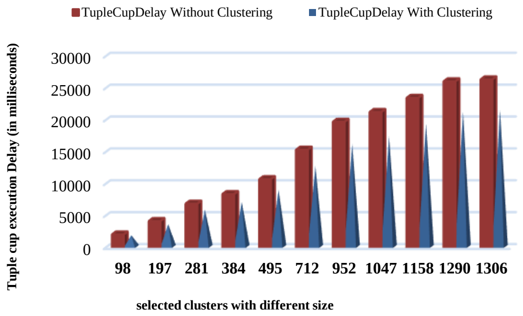


Fig. 6. Comparison of tuple cup execution delay in two modes of clustering and non-clustering of services

5-3-3-Experiment 3

In this test, the amount of network usage is calculated in two modes of clustering and non-clustering services during the simulation process. This parameter depends on the number of links in the fog network, and with the increase in the number of links, the size of the network usage increases. As shown in Figure 7, the use of the network in the service clustering mode performs better than the service non-clustering mode. Due to the optimal placement of the modules, there has been a significant reduction in the delay in creating communication links between the modules/services in the source and destination devices, and the proposed algorithm has experienced a 15.82% reduction.

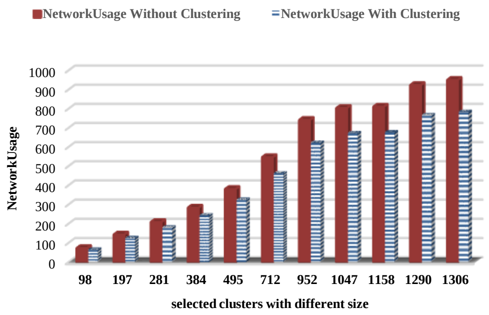


Fig. 7. Comparison of network usage in two modes of clustering and non-clustering of services

5-3-4-Experiment 4

In this section, the p values obtained from the statistical analysis for the evaluated parameters such as *AppLoopDelay*, Tuple cup execution delay, and Network usage are given in Table 9. A value of $p \leq 0.05$ indicates that H0 can be rejected with a confidence interval of 95%. In other words, p-values smaller than 0.05 show that the proposed method is better than service execution in non-clustering mode due to module/services loop delay, tuples processing delay, and network usage rate.

Table 9. Statistical analysis for the proposed algorithm

Metric	Algorithms	P-value	Significance
AppLoopDelay	Non-clustering mode:clusteredmode	P = 0.0028	extremely statistically
Tuple cup execution Delay	Non-clustering mode:clusteredmode	P = 0.0189	very statistically
Network usage	Non-clustering mode:clusteredmode	P = 0.0382	Statistically

6. Conclusion

We can determine the optimal service placement in the fog network based on the clustering results, minimize latency, and improve service responsiveness by considering the proximity between IoT devices and fog nodes. In addition, we can allocate services according to the availability of resources and the capacity of fog nodes and ensure the efficient use of resources.This article proposes a DLA-SPSQ method based on fuzzy clustering of services and distributed learning automata for optimal placement of services/modules in heterogeneous fog nodes. The DLA-SPSQ algorithm combines the FCS algorithm and the DLA-FMP algorithm. Fuzzy clustering of services, the optimal number of clusters, and the determination of high-priority clusters based on service quality criteria have been done using cluster evaluation criteria. The optimal placement of services in two modes, clustered and non-clustered, is based on the cost function provided by the DLA-FMP. This article proposes a method called DLA-SPSQ based on fuzzy clustering of services and distributed learning automata for optimal placement of services/modules in heterogeneous fog nodes. The DLA-SPSQ algorithm combines the FCS algorithm and the DLA-FMP algorithm. The FCS algorithm has done fuzzy clustering of services, the optimal number of clusters, and the determination of high-priority clusters based on service quality criteria. The optimal placement of services in two modes, clustered and non-clustered, is based on the cost function provided by the DLA-FMP algorithm. For the optimal placement of services, the topology of the fog network is modelled by a directed graph and mapped to a DLA. DLA performs the deployment of services from the edge nodes upwards in the fog topology with the cooperation of automatons in an optimal and hierarchical manner. In this search method, due to the use of the maximum capacity of the edge level fog nodes, the cost function (service execution delay), tuple processing execution delay, and network usage have decreased by 18.5%, 17.78%, and 15.82%, respectively. P-Value ≤ 0.05 in the results of Experiment 4 confirms the efficiency and improvement of the proposed algorithm in the state of clustering of services compared to the state of not clustering them.

future works:

Research areas related to the proposed topic that can be done in future works are:

- ✓ Creating multi-objective optimization models for optimal placement of services considering QoS requirements and resource constraints simultaneously.
- ✓ Research the placement of sustainable services considering renewable energy sources in fog nodes
- ✓ Investigating the impact of load balancing on service quality in fog nodes with heterogeneous workloads.
- ✓ Investigating security and privacy in the optimal deployment of QoS-based services.
- ✓ Creating criteria and standards for service quality evaluation in cloud computing environments.
- ✓ Investigate economic models for billing based on the QoS provided, such as premium pricing for higher service levels.
- ✓ Investigating the effect of network conditions on the performance of real-time analytical programs in fog environments.

References

[1] Y. Al Mtawa, A. Haque, and B. Bitar, "The mammoth internet: Are we ready?," *IEEE Access*, vol. 7, pp. 132894-132908, 2019.

[2] P. Maiti, B. Sahoo, A. K. Turuk, A. Kumar, and B. J. Choi, "Internet of Things applications placement to minimize latency in multi-tier fog computing framework," *ICT Express*, vol. 8, no. 2, pp. 166-173, 2022, doi: http://dx.doi.org/10.1016/j.icte.2021.06.004.

[3] T. Huang, W. Lin, C. Xiong, R. Pan, and J. Huang, "An ant colony optimization-based multiobjective service replicas placement strategy for fog computing," *IEEE Transactions on Cybernetics*, vol. 51, no. 11, pp. 5595-5608, 2020, doi: http://dx.doi.org/10.1109/TCYB.2020.2989309.

[4] S. Gorikapudi and H. K. Kondaveeti, "A novel clustering model via optimized fuzzy C- means algorithm and sandpiper optimization with cycle crossover process in IoT," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 23, p. e7776, 2023.

[5] M. Hosseini Shirvani and Y. Ramzanpoor, "Multi-objective QoS-aware optimization for deployment of IoT applications on cloud and fog computing infrastructure," *Neural Computing and Applications*, vol. 35, no. 26, pp. 19581-19626, 2023.

[6] D. Zhao, Q. Zou, and M. Boshkani Zadeh, "A QoS-aware IoT service placement mechanism in fog computing based on open-source development model," *Journal of Grid Computing*, vol. 20, no. 2, p. 12, 2022.

[7] A. Brogi and S. Forti, "QoS-aware deployment of IoT applications through the fog," *IEEE Internet of Things Journal*, vol. 4, no. 5, pp. 1185-1192, 2017.

[8] M. Haghi Kashani, A. M. Rahmani, and N. Jafari Navimipour, "Quality of service- aware approaches in fog computing," *International Journal of Communication Systems*, vol. 33, no. 8, p. e4340, 2020.

[9] H. K. Apat, B. Sahoo, and S. Mohanty, "A Quality of Service (QoS) aware Fog Computing model for intelligent (IoT) applications," in *2021 19th OITS International Conference on Information Technology (OCIT)*, 2021: IEEE, pp. 267-272.

[10] A. N. Al-Masri and M. Nasir, "A novel fuzzy clustering with metaheuristic based resource provisioning technique in cloud environment," *Fusion: Practice and Applications*, vol. 6, no. 1, pp. 08-8-16, 2021.

[11] M. Kupriyanov, I. Holod, and A. Shorov, "Fuzzy Clustering Based on Cloud and Fog Computing," in *2019 XXII International Conference on Soft Computing and Measurements (SCM))*, 2019: IEEE, pp. 1-5.

[12] F. Tavousi, S. Azizi, and A. Ghaderzadeh, "A fuzzy approach for optimal placement of IoT applications in fog-cloud computing," *Cluster Computing*, pp. 1-18, 2022.

[13] A. M. Maia, Y. Ghamri-Doudane, D. Vieira, and M. F. de Castro, "An improved multi-objective genetic algorithm with heuristic initialization for service placement and load distribution in edge computing," *Computer Networks*, vol. 194, p. 108146, 2021, doi: http://dx.doi.org/10.1016/j.comnet.2021.108146.

[14] M. Ghobaei-Arani and A. Shahidinejad, "A cost-efficient IoT service placement approach using whale optimization algorithm in fog computing environment," *Expert Systems with Applications*, vol. 200, p. 117012, 2022, doi: http://dx.doi.org/10.1016/j.eswa.2022.117012.

[15] F. Khosroabadi, F. Fotouhi-Ghazvini, and H. Fotouhi, "Scatter: Service placement in real-time fog-assisted iot networks," *Journal of Sensor and Actuator Networks*, vol. 10, no. 2, p. 26, 2021.

[16] A. Yousefpour, G. Ishigaki, and J. P. Jue, "Fog computing: Towards minimizing delay in the internet of things," in *2017 IEEE international conference on edge computing (EDGE)*, 2017: IEEE, pp. 17-24.

[17] C. Liu, J. Wang, L. Zhou, and A. Rezaeipanah, "Solving the multi-objective problem of IoT service placement in fog computing using cuckoo search algorithm," *Neural Processing Letters*, vol. 54, no. 3, pp. 1823-1854, 2022, doi: http://dx.doi.org/10.1007/s11063-021-10708-2.

[18] B. Natesha and R. M. R. Guddeti, "Adopting elitism-based Genetic Algorithm for minimizing multi-objective problems of IoT service placement in fog computing environment," *Journal of Network and Computer Applications*, vol. 178, p. 102972, 2021, doi: http://dx.doi.org/10.4108/eai.22-2-2022.173492.

[19] M. Salimian, M. Ghobaei-Arani, and A. Shahidinejad, "An evolutionary multi-objective optimization technique to deploy the IoT Services in fog-enabled Networks: an autonomous approach," *Applied Artificial Intelligence*, pp. 1-34, 2022, doi: http://dx.doi.org/10.1080/08839514.2021.2008149.

[20] F. M. Calisto, N. Nunes, and J. C. Nascimento, "Modeling adoption of intelligent agents in medical imaging," *International Journal of Human-Computer Studies*, vol. 168, p. 102922, 2022, doi: http://dx.doi.org/10.2139/ssrn.4116048.

[21] M. Dadashi Gavaber and A. Rajabzadeh, "MFP: an approach to delay and energy-efficient module placement in IoT applications based on multi-fog," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 7, pp. 7965-7981, 2021, doi: https://link.springer.com/article/10.1007/s12652-020-02525-7.

[22] G. Baranwal and D. P. Vidyarthi, "FONS: a fog orchestrator node selection model to improve application placement in fog computing," *The Journal of Supercomputing*, vol. 77, no. 9, pp. 10562-10589, 2021, doi: https://link.springer.com/article/10.1007/s11227-021-03702-x.

[23] M. Masdari, A. B. Sangar, and K. Majidzadeh, "A Hybrid Multi-objective Algorithm for Imbalanced Controller Placement in Software-Defined Networks," *Journal of Network and Systems Management*, vol. 30, no. 3, pp. 1-54, 2022, doi: http://dx.doi.org/10.1007/s10922-022-09650-y.

[24] Z. M. Nayeri, T. Ghafarian, and B. Javadi, "Application placement in Fog computing with AI approach: Taxonomy and a state of the art survey," *Journal of Network and Computer Applications*, vol. 185, p. 103078, 2021, doi: http://dx.doi.org/10.1016/j.jnca.2021.103078.

[25] F. A. Salaht, F. Desprez, and A. Lebre, "An overview of service placement problem in fog and edge computing," *ACM Computing Surveys (CSUR)*, vol. 53, no. 3, pp. 1-35, 2020, doi: http://dx.doi.org/10.1145/3391196.

[26] E. Torabi, M. Ghobaei-Arani, and A. Shahidinejad, "Data replica placement approaches in fog computing: a review," *Cluster Computing*, pp. 1-29, 2022, doi: http://dx.doi.org/10.1007/s10586-022-03575-6.

[27] K. Narendra and M. Thathachar, "Learning Automata: An Introduction Prentice-Hall," *New Jersey*, 1989.

[28] M. A. Thathachar and P. S. Sastry, "Varieties of learning automata: an overview," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 32, no. 6, pp. 711-722, 2002, doi: https://doi.org/10.1109/tsmcb.2002.1049606.

[29] K. S. Narendra and M. A. Thathachar, "On the behavior of a learning automaton in a changing environment with application to telephone traffic routing," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 10, no. 5, pp. 262-269, 1980, doi: https://doi.org/10.1109/TSMC.1980.4308485.

[30] Z. Anari, A. Hatamlou, and B. Anari, "Automatic Finding Trapezoidal Membership Functions in Mining Fuzzy Association Rules Based on Learning Automata," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 7, no. 4, 2022, doi: https://doi.org/10.1142/S0218001421590266.

[31] B. Anari, J. A. Torkestani, and A. M. Rahmani, "Automatic data clustering using continuous action-set learning automata and its application in segmentation of images," *Applied Soft Computing*, vol. 51, pp. 253-265, 2017, doi: https://doi.org/10.1145/3391196.

[32] B. Zarei and M. R. Meybodi, "Improving learning ability of learning automata using chaos theory," *The Journal of Supercomputing*, vol. 77, no. 1, pp. 652-678, 2021, doi: https://doi.org/10.1007/s11227-020-03293-z.

[33] K. Asghari, M. Masdari, F. Soleimanian Gharehchopogh, and R. Saneifard, "A fixed structure learning automata- based optimization algorithm for structure learning of Bayesian networks," *Expert Systems*, vol. 38, no. 7, p. e12734, 2021, doi: https://doi.org/10.1111/exsy.12734.

[34] F. Farahbakhsh, A. Shahidinejad, and M. Ghobaei- Arani, "Multiuser context- aware computation offloading in mobile edge computing based on Bayesian learning automata," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4127, 2021, doi: https://doi.org/10.1002/ett.4127.

[35] F. Jazayeri, A. Shahidinejad, and M. Ghobaei-Arani, "Autonomous computation offloading and auto-scaling the in the mobile fog computing: a deep reinforcement learning-based approach," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 8265-8284, 2021, doi: https://link.springer.com/article/10.1007/s12652-020-02561-3.

[36] N. Kumar, J.-H. Lee, and J. J. Rodrigues, "Intelligent mobile video surveillance system as a Bayesian coalition game in vehicular sensor networks: Learning automata approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 3, pp. 1148-1161, 2014, doi: http://dx.doi.org/10.1109/TITS.2014.2354372.

[37] Y. Abofathi, B. Anari, and M. Masdari, "A learning automata based approach for module placement in fog computing environment," *Expert Systems with Applications*, vol. 237, p. 121607, 2024.

[38] U. Arora and N. Singh, "IoT application modules placement in heterogeneous fog–cloud infrastructure," *International Journal of Information Technology*, vol. 13, no. 5, pp. 1975-1982, 2021, doi: http://dx.doi.org/10.1007/s41870-021-00672-4.

[39] https://qwsdata.github.io/

Optimal Planning of Conductors and Capacitors in a Distribution Network Using a Hybrid Evolutionary Algorithm

Leila Mohammadian
[*]Department of Electrical Engineering, Shabestar Branch, Islamic Azad University, Shabestar, Iran
Email: Lemohammadian@iau.ir

Received: 03 November 2024 Accepted: 31 December 2024 Published:15 February 2025

Abstract

The optimum planning of power distribution networks is one of the most important research fields for electrical engineers. Normally in a distribution system, operational costs are high because of their losses. In this paper, the practical planning of the distribution system includes the selection of optimal conductor size and capacitor placement in the radial distribution network considering the increasing rate of loads. Technical operational constraints are available conductors and capacitors, voltage limit, maximum permissible carrying current of conductors, and maximum reactive power that could be injected, without overvoltage. The objective function includes the cost of power losses, capacitors, and conductors, also the above constraints are added as penalty functions to the objective function. In this paper, the minimization problem is solved using an effective hybrid method of GA and PSO, which is called HGAPSO. By applying the proposed method, the final cost of network planning, losses, and their cost are considerably reduced and the voltage profile of the network has improved to a semi-flat shape. In the minimization process, an efficient algorithm is used to solve the radial distribution power flow problem in complex mode, which makes it easier to get system data iteratively. Simulation results are investigated on a sample radial distribution network. Finally, the effectiveness of the proposed hybrid method is proved by comparing the results with the results obtained from PSO.

Keywords: distribution system, conductor size, capacitor placement, hybrid evolutionary algorithm, loss reduction, optimization.

1. Introduction

Power distribution systems constitute the largest segment of power systems, delivering electrical energy to end-users via substations, distribution feeders, transformers, secondary conductors, and service mains. Typically designed in a radial configuration, these systems operate at lower voltage levels than transmission systems. Common challenges include technical losses and inadequate voltage regulation, which are addressed through planning techniques such as network reconfiguration, conductor selection, capacitor placement, and the inclusion of distributed generation. Many of these systems were established decades ago and have not been updated, resulting in undersized conductors that increase power losses and lead to significant voltage drops, especially at nodes far from the substation. Additionally, expanding the network can be costly due to environmental permit processes and geographic constraints. In this context, optimal conductor selection presents a viable alternative to network upgrades. This approach aims to replace existing conductors with more suitable types to reduce energy losses, enhance current capacity, and improve voltage levels when they drop below acceptable standards[1-4].The authors of [10] introduced a Mixed-Integer Linear Programming Model along with a heuristic approach to determine the Pareto front for the conductor size selection problem. In [11], a Mixed-Integer Nonlinear Programming (MINLP) formulation for optimal conductor selection was demonstrated and solved using the General Algebraic Modeling System (GAMS) and the DICOPT solver. Additionally, the authors of [12] developed an exact nonlinear model, which was addressed using existing MINLP solvers to identify the optimal conductors. Furthermore, a MINLP model specifically for optimal conductor selection in DC radial distribution systems was proposed by the authors of [13].The optimal placement of capacitor banks in distribution systems is essential for minimizing voltage drops and reducing line losses. This essentially involves installing capacitors at strategic locations within the network to compensate for reactive power demands and enhance technical performance. Proper distribution of reactive power within the system not only helps decrease overall power consumption but may also lead to lower electricity costs due to reduced losses in the system.. In [17], a heuristic methodology based on graph search was proposed to optimally size and allocate reactive compensation in distribution networks. Deterministic methods and Genetic Algorithms (GAs) were employed in [18] to address the optimal placement of capacitors. In [19], the authors utilized a GA for reactive compensation in power systems with varying customer load patterns. A GA was also applied in [20] for optimal capacitor placement, where the authors limited the maximum number of operations for switched capacitive banks to account for equipment aging. In [21], Shannon’s Entropy was leveraged for the optimal allocation of capacitors in distribution networks, taking into consideration multiple criteria. The authors in [22] examined unbalanced networks and implemented a micro-GA to tackle the optimization challenges in radial distribution networks. Additionally, in [23], a Particle Swarm Optimization (PSO) approach was proposed for the optimal placement and sizing of capacitors, while factoring in the effects of harmonics in unbalanced networks. This technique was further applied in [24] for optimal capacitor allocation in microgrids and in [26], considering the impact of switchable capacitive banks. In [27], the authors introduced a multiverse optimizer approach that partially modifies conventional loss sensitivity factors. Multi-stage methodologies have also been explored to address the optimal capacitor allocation problem. A two-stage method presented in [28] incorporates a loss sensitivity technique to identify potential locations for capacitor placement. Planning techniques such as optimal conductor selection, capacitor placement, reconfiguration, the installation of new substations, and the integration of distributed generation (DG) are often examined in isolation. However, in distribution systems with heavily loaded feeders and inadequate voltage profiles, relying on a single technique may not suffice to minimize power losses and enhance voltage levels. Thus, combining these techniques could lead to a more effectively planned system. The optimal conductor selection problem has been integrated with optimal capacitor placement in numerous studies [1,5–16,19,20,23,25]. The main contribution of this paper is the simultaneous modeling of optimal conductor sizing along with the optimal placement of capacitors. In general, attention has focused on reducing cost through optimizing the conductor profile, capacitor cost, and in some cases cost of losses. But in these all, increasing rate of load for coming years is not considered. In addition in most articles available in the literature, there is not any special way to solve the power flow problem in the distribution system and they simultaneously have solved the power flow problem and minimized the objective function, but as it will be shown we have developed a software pack which is able to determine radial distribution system parameters just having line data matrix, and technical data for available capacitors and conductors .further more we have used HGAPSO method for solving the optimization problem, which uses both effective GA and PSO‘s benefits simultaneously. The proposed method is tested on a sample radial distribution network, with 5 types of available conductors and 11 different sizes of capacitors, considering 8 years for load growth. The results show that the proposed objective function minimizes the loss of the system by considering all of the constraints and incorporating capacitors and conductors selection.

The proposed method identifies better solutions than those reported in the reviewed literature. The rest of this document is organized as follows: Section 2 describes the proposed model, and Section 3 describes the method to solve the simultaneous optimization problem in the power distribution network that can be carried out either jointly or separately. Section 4 presents the results of the proposed method applied to a test system. Finally, Section 5 presents the research conclusions.

2. Problem Formulation

1.1. Power flow

Load flow is very important and fundamental tool for analysis of any power system and is used in the operational as well as planning stages. Certain application, particularly in distribution automation and optimization of power system, requires repeated load flow solution. In these applications, it is very important to solve the load flow problem as efficiently as possible. The Newton-Raphson and the fast decoupled power flow solution techniques and a host of their derivatives have efficiently solved for “well behaved” power systems. Researchers however have been aware of the shortcomings of these algorithms when they are generally implemented and applied to ill conditioned power systems.

Power flow in a distribution system obeys physical laws such as (Kirchhoff laws and Ohms law) which became part of the constraints in the capacitor placement problem. The distribution system power flow solution is to be used as a subroutine in each iteration. Therefore, it is essential to have a computationally efficient and numerically robust method for solving the distribution system power flow. By radial distribution system, we mean a system which has a single simultaneous path of power flow to the load. We have used a method, that exploits the radial structure of the distribution network and the relationship between the bus powers and branch powers is expressed as a non-singular square matrix known as element incidence matrix.

The power flow equations for a radial distribution system are derived as the relationship between the specified complex bus powers and the bus voltages.

Let Sijis the complex power flowing from bus ‘i’ to bus’j’:

$$S_{\mathrm{ij}} = \mathrm{P_{ij}} + \mathrm{jQ_{ij}} = \mathrm{V_i}\,(\mathrm{V_i^*} - \mathrm{V_j^*})\mathrm{Y_{ij}^*} \tag{1}$$

The ‘i’th bus powers are expressed as:

$$\begin{aligned}\mathrm{P_i} + \mathrm{Q_i} &= \sum_{\mathrm{i\epsilon k(i)}} \mathrm{P_{ij}} + \mathrm{Q_{ij}} \\ &= \sum_{\mathrm{i\epsilon k(i)}} \mathrm{V_i}\,(\mathrm{V_i^*} - \mathrm{V_j^*})\mathrm{Y_{ij}^*}\end{aligned} \tag{2}$$

k(i) is the set of nodes connected to node i, and Pi /Qi denotes the real/reactive power at node i. The above complex nonlinear equations are to be solved to determine the bus voltages. The real and imaginary parts of the equations are separated and solved using numerical methods.

1.2.Formulation of used method

The basis for the method is that an N bus radial distribution network has only N-1 lines (elements) and the branch currents (powers) can be expressed in terms of bus currents (powers). For an element ij connected between nodes ‘i’ and ‘j’ the bus current of node j can be expressed as a linear equation.

$$I_j = I_{ij} - \sum I_{jk(j)} \tag{3}$$

k(j) is the set of nodes connected to node j. For the slack bus the power is not specified so it is excluded and the relationship between bus currents and branch currents are derived as a non-singular square matrix.

$$\begin{aligned}&I_{\mathrm{bus}} = \mathrm{kI_{branch}} \\ &I_{bus} = [I_{b2} I_{b3} \cdots I_{bn}]^T\end{aligned} \tag{4}$$

The matrix K is element incidence matrix. It is a non singular square matrix of order N-1.The elemental incidence matrix is constructed in a simple way same like bus incidence matrix. In this matrix K each row is describing the element incidences. The elements are numbered in conventional way i.e. the no of element ‘ij’ is j-1.

1. The diagonal elements of matrix K are one. The variable j is denoting the element number.

K(j, j) =1

2. For each ‘j’th element let m (j) is the set of element numbers connected at its receiving end.

K(j, m(j)) = -1

3. All the remaining elements are zero. It can be observed that all the elements of matrix K below the main diagonal are zero.

$$I_{branch} = k^{-1} I_{bus} \tag{5}$$

The relationship between the branch currents and bus currents can be extended to complex branch powers and bus powers. The sending end power and the receiving end powers are not same due to transmission loss. The transmission loss is included as the difference between the sending end/receiving end powers. The relationship between branch powers and bus powers is established in same way of bus/branch currents. Multiplying both sides by element incidence matrix K:

$$\begin{aligned}&S_{bus} = K[S_{branc\square}^{sending} - TL_{branc\square}] \\ &S_{branc\square} = K^{-1}.S_{bus} \\ &\qquad\qquad + TL_{branc\square}\end{aligned} \tag{6}$$

The power flow equations are complex quadratic equations. These are solved as [2]and the results we can get are voltage magnitude, voltage angle, active and reactive losses, and complex current of branches. The advantage of this method is that it does not require a flat start. The formulation can be extended to unbalanced three-phase networks. The reactive power injections at multiple ends can be effectively calculated to improve the voltage profile.

1.3. Objective Function

In each optimization problem, objective function should be defined. In the proposed approach, objective function can be formulated as following equation; The proposed objective function aims at minimizing the total annual cost due to capacitor placement, conductor selection and power losses considering load growth in period of instrument life, with constraints that include limits on voltage, maximum permissible carrying current of conductors, size of installed capacitors and type of selected conductors, and maximum permissible reactive current to be injected to avoid over voltage.

$$\begin{aligned}\mathrm{J} = &\sum_{\mathrm{l=0}}^{\mathrm{n_S}} \mathrm{L_l} \times \mathrm{C(con_l)} + \sum_{\mathrm{j=1}}^{\mathrm{J}} \mathrm{K_j^c} \times \mathrm{Q_j^c} + \\ &+ \sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{P_{loss,i}} \times \mathrm{C_E} \times \mathrm{P_W^i} \times 8760 \times \mathrm{LSF_i} + \\ &+ \sum_{\mathrm{k=1}}^{\mathrm{m}} \{\max(0, \mathrm{V_{min}} - \mathrm{V_k})^2 \\ &\qquad + \max(0, \mathrm{V_k} - \mathrm{V_{max}})^2\} \\ &\qquad + \\ &+ \sum_{\mathrm{l=1}}^{\mathrm{n_S}} (\mathrm{I_l} - \mathrm{I_{max}^l})^2 + \sum_{\mathrm{j=1}}^{\mathrm{m}} (\mathrm{Q_j^c} - \mathrm{TQlossj})^2\end{aligned} \tag{7}$$

where:

$$P_W = \left(\frac{1 + Intr}{1 + \mathrm{Inf}r}\right) \tag{8}$$

According to the mentioned constraints we should have:

$$I_{max}^l = \begin{cases} I_{\max(ConType_l - 1)} & if \quad I_{\max(ConType_l)} \leq |I_l| \ , ConType_l \neq 1 \\ I_{\max(ConType_l\ )} & else \end{cases} \tag{9}$$

$$V_{min} \leq |V_i| \leq V_{max} \tag{10}$$

$$I_l \leq I_{\max(l)} \tag{11}$$

While computing the cost by defined objective function, load growth is considered as (12):

$$load_i = \begin{cases} load \times (1+r)^i \ i = 1,2,3,\dots,M \\ load \times (1+r)^M i = M+1,\dots,N \end{cases} \tag{12}$$

Where, load i is the load in i'th year, r is the annual growth rate and M is a plan period up to which the feeder can take load growth.

The peak load growth during the planning period is illustrated in figure 1.

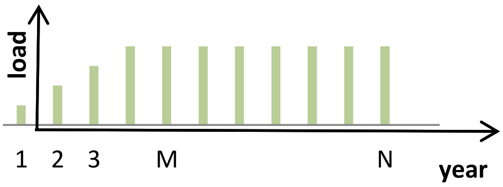


Fig1. Peak load growth during planning period

Variation of load is considered with loss factor parameter in objective function.

The objective function includes six statements which are described as followed:

Statement1: The cost of power losses considering load growth,

Statement2: the cost of the installed capacitors,

Statement3: the cost of the installed conductors,

Statement4: the constraint of voltage limit,

Statement5: the constraint of maximum permissible carrying current of the conductors,

Statement 6: the constraint to avoid over-voltage, or having sufficient capacitor installation.

3. Proposed Computational Algorithm

3.1.HGAPSO algorithm

In GA, a candidate solution for a specific problem is called an individual or a chromosome and consists of a linear list of genes. Each individual represents a point in the search space, and hence a possible solution to the problem.

A population consists of a finite number of individuals. Each individual is decided by an evaluation mechanism to obtain its fitness value. Based on this fitness value and undergoing genetic operators, a new population is generated iteratively with each successive population referred to as a generation. The GAs use three basic operators (reproduction, crossover, and mutation) to manipulate the genetic composition of a population.

The PSO conducts searches using a population of particles corresponding to GA individuals. A population of particles is randomly generated, initially. Each particle represents a potential solution and has a position represented by a position vector$\bar{x}_i$,A swarm of particles moves through the problem space, with the moving velocity of each particle represented by a velocity vector $\bar{v}_i$, At each time step, a function f_i representing a quality measure is calculated by using $x_i$as input. Each particle keeps track of its own best position, which is associated with the best fitness it has achieved so far in a vectorv_i. Further, the best position among all the particles obtained so far in the population is kept track of as p_g. At each time step t, by using the individual best position$p_i(t)$and global best position $p_g(t)$, a new velocity for particle i is updated as follows:

$$\begin{aligned}\bar{v}_i(t+1) &= \chi[(\bar{v}_i(t) + c_1\varphi_1\{\bar{p}_i(t) \\ &\qquad - \bar{x}_i(t)\} \\ &+ c_2\varphi_2\{\bar{p}_g(t) - \bar{x}_i(t)\}]\end{aligned} \tag{13}$$

Where c1 and c2 are positive constants, φ1 and φ2 are uniformly distributed random numbers in [0,1] interval, andχcontrols the magnitude of v. Changing velocity in this way enables the particle i to search around its individual best position, $\mathrm{p_i}$, and global best position, p_g . Based on the updated velocities, each particle changes its position according to the following equation:

$$\bar{x}_i(t+1) = \bar{x}_i(t) + \bar{v}_i(t+1) \tag{14}$$

The computation of PSO is easy and adds only a slight computation load when it is incorporated into GA. The detailed design algorithm of HGAPSO consists of three major operators: enhancement, crossover and mutation [4]. In HGAPSO, GA and PSO both work with the same population. Based on the encoding scheme, Ps individuals forming the population are randomly generated. These individuals may be regarded as chromosomes in terms of GA, or as particles in terms of PSO. Then, new individuals on the next generation are created by enhancement, crossover and mutation operations [4]. For clarity, the flow of these operations is illustrated in figure 2.

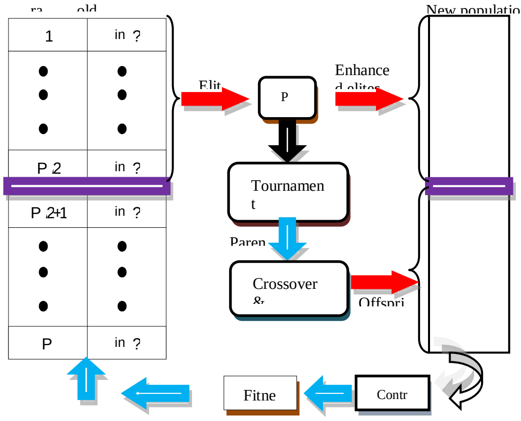


Fig2. Flowchart of the HGAPSO method

Enhancement, crossover, and mutation operators are described as follows:

(a)Enhancement: In each generation, after the fitness values of all the individuals in the same population are calculated, the top-half best-performing ones are marked. These individuals are regarded as elites. Instead of reproducing the elites directly to the next generation as elite GAs do, we first enhance the elites by PSO. By using these enhanced elites as parents, the generated offspring will usually achieve better performance than those bred by original elites. The group constituted by the elites is regarded as a swarm, and each elite corresponds to a particle in it. By performing PSO on the elites, we may increase the search ability. Half of the population in the next generation is occupied by the enhanced individuals, the remainder by crossover operation.

(b)Crossover: To produce well performing individuals, in the crossover operation parents are selected from the enhanced elites only. To select parents for the crossover operation, the tournament-selection scheme is used, in which two enhanced elites are selected at random, and their fitness values are compared to select the elite with better fitness value as one parent. Then the other parent is selected in the same way. Two off springs are created by performing crossover on the selected parents. Two-point crossover operation is used, where two crossover sites are selected randomly within the range of an individual and swapping occurs. These produced offspring occupy half of the population in the next generation.

(c)Mutation: In HGAPSO, mutation occurs in conjunction with the crossover operation. Here, uniform mutation is adopted, that is, the mutated gene is drawn randomly, uniformly from the corresponding search interval. In the following simulations, a constant mutation-probability Pm=0.1 is used.

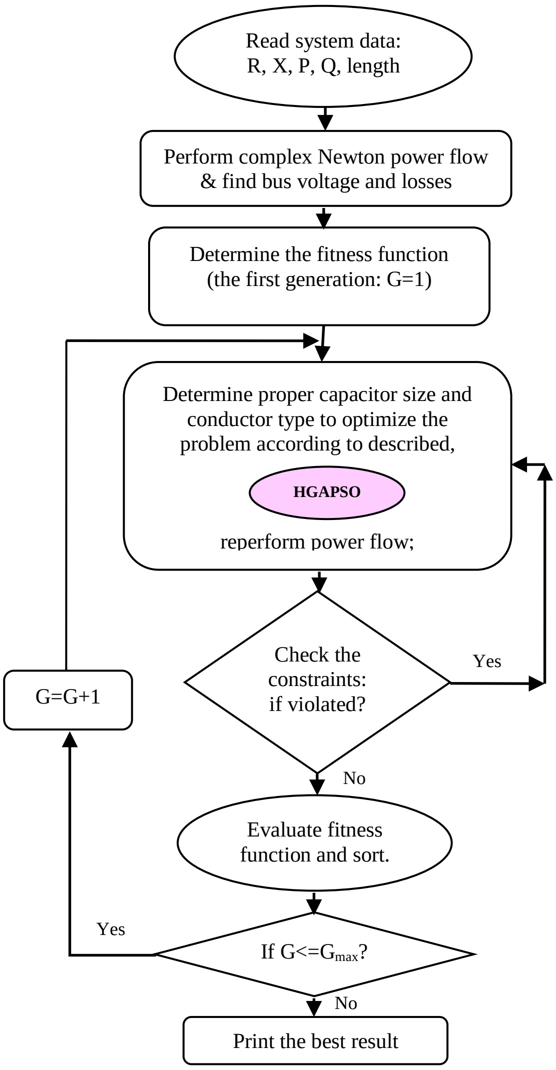

Read system data:
R, X, P, Q, length
Perform complex Newton power flow & find bus voltage and losses
Determine the fitness function (the first generation: G=1)
Determine proper capacitor size and conductor type to optimize the problem according to described,
HGAPSO
reperform power flow;
Check the constraints: if violated?
Yes
No
G=G+1
Evaluate fitness function and sort.
Yes
If G<=Gmax?
No
Print the best result

Fig3. Flowchart of the whole proposed method

4. Computational Test

Based on the proposed algorithm, software was developed using MATLAB for proper conductor and capacitor selection considering load growth in distribution networks. The proposed method was tested on a sample distribution network by use of prepared software to evaluate its effectiveness. The test case is a 20 kV radial distribution network that has 13 nodes and 12 sections. Network and load data of radial test feeder are shown in tables 1 and 2.

The technical and economical data of available conductors and capacitors are given in tables 3 and 4, respectively. Other input data needed for evaluating the objective function are as follows:

N: 20, M: 5, Intr: 17%, Infr: 14%, r: 7%,

Vmax: 1.03 pu, V_{min}: 0.95 pu, LSF: 0.63 for all loads, Cost of energy: 50 ($/kWh)

Table1. Data for the test feeder

Sending end(i)	Receiving end(i)	R(Ω)	X(Ω)	Section length(km)
0	1	0.7822	0.2835	0.835
1	2	0.7822	0.2835	0.415
2	3	0.7822	0.2835	0.215
3	4	0.7822	0.2835	0.420
4	5	0.7822	0.2835	0.215
4	6	0.7822	0.2835	0.580
6	7	0.7822	0.2835	0.455
7	8	0.7822	0.2835	0.350
7	9	0.7822	0.2835	0.350
6	10	0.7822	0.2835	0.300
10	11	0.7822	0.2835	0.320
6	12	0.7822	0.2835	0.415

Table2. Load data of the feeder

Bus no.	P(kw)	Q(Kvar)
1	890	468
2	628	470
3	1112	764
4	636	378
5	474	344
6	1342	1078
7	920	292
8	766	498
9	662	480
10	690	186
11	1292	554
12	1124	480

Table3. Technical and economical data of available capacitors

capacitor type	size(kvar)	price($/kvar)
1	0	0
2	150	0.5
3	300	0.35
4	450	0.253
5	600	0.22
6	750	0.276
7	900	0.183
8	1050	0.228
9	1200	0.170
10	1350	0.207
11	1500	0.201

Table4. Technical and economical data of available conductors

conductor type	R(Ω/km)	X(Ω/km)	price($/km)	Max. current
1	0.7822	0.2835	151	500
2	0.0625	0.279	1155	1000
3	0.0353	0.259	1733	1800
4	0.0745	0.285	1026	900
5	0.0429	0.267	1500	1500

5. Results and Discussion

The simulation results are clearly illustrated in following figures and tables. By solving the optimization problem, the size of the capacitor banks, the types of the conductors and the amplitude of the bus voltages are determined. The total cost, power losses, minimum and maximum of the voltages in each node are obtained. Figures4 shows the system performance. Figure5 compares voltage magnitudes before and optimization. Figure 6 compares voltage angles before optimization with them after performing optimization process:

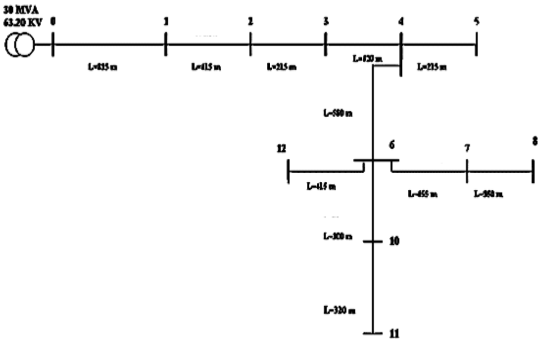


Fig4. The chosen distribution network for implementation of the proposedmethod

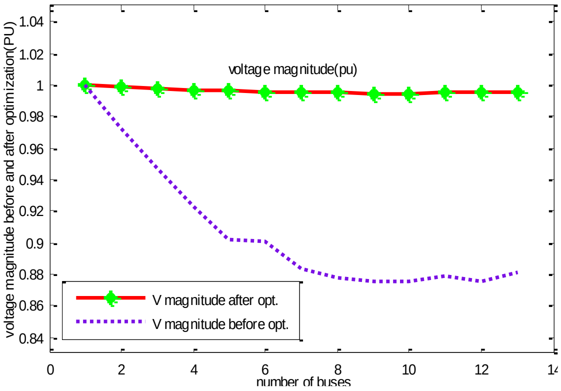


Fig5. Comparison of voltage magnitude before and after performing HGAPSO

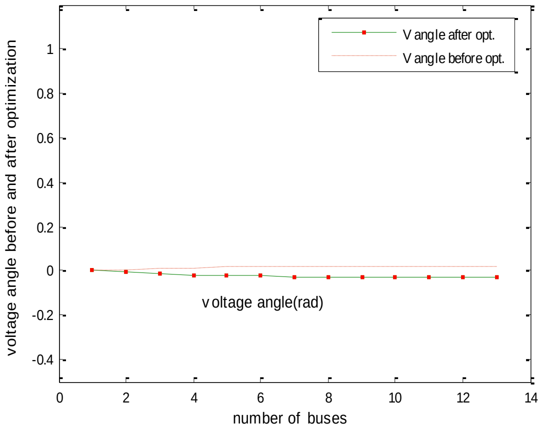


Fig6. Comparison of voltage angle before and after performing HGAPSO

Table5. Comparison of voltage magnitudes before and after performing HGAPSO

Voltage magnitude before optimization	Voltage magnitude after optimization
1.0000	1.0000
0.9722	0.9983
0.9466	0.9971
0.9227	0.9963
0.9017	0.9959
0.9004	0.9956
0.8837	0.9953
0.8774	0.9949
0.8753	0.9946
0.8756	0.9945
0.8787	0.9954
0.8754	0.9948
0.8808	0.9958

Table6. Comparison of power losses before and after performing HGAPSO

	Before optimization	After optimization
TPloss(kw)	1.3219	0.0383
TQloss(kvar)	12.4630	2.0548

Table7. Result of conductor selection and capacitor placement after performing HGAPSO

Sending end(i)	Receiving end(i)	Conductor type	Capacitor type
0	1	3	1
1	2	3	1
2	3	3	1
3	4	5	11
4	5	4	1
4	6	3	1
6	7	3	5
7	8	3	2
7	9	2	2
6	10	3	9
10	11	2	1
6	12	4	11

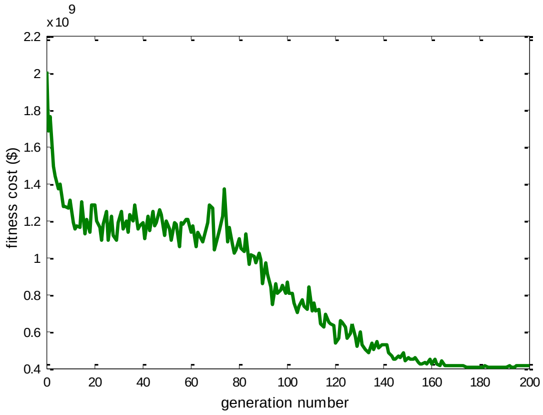


Fig7. Fitness optimized using HGAPSO

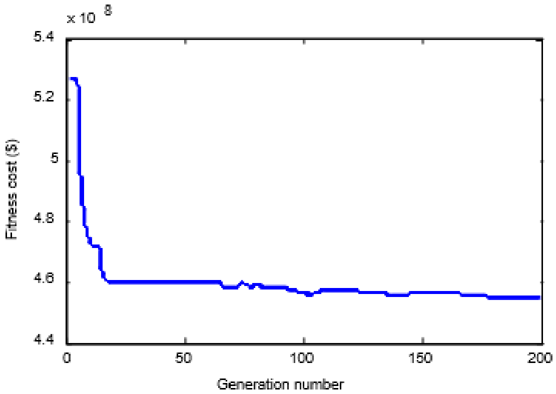


Fig8. Fitness optimized using PSO

Optimization process in HGAPSO, PSO and comparing the two methods are clearly illustrated in figures 8,9. The PSO best cost is 4.52241871*10^8$, and HGAPSO best cost is 3.969713*10^8$. Results obtained from PSO are included in below tables.

Table8. Result of conductor selection and capacitor placement after performing PSO

Sending end(i)	Receiving end(i)	Conductor type	Capacitor type
0	1	3	1
1	2	3	1
2	3	5	6
3	4	3	1
4	5	1	4
4	6	5	11
6	7	3	1
7	8	3	1
7	9	2	10
6	10	3	1
10	11	3	4
6	12	1	1

Table9. Comparison of voltage magnitudes after performing PSO and HGAPSO

Voltage magnitude performing PSO	Voltage magnitude performing HGAPSO
1.0000	1.0000
0.9979	0.9983
0.9963	0.9971
0.9949	0.9963
0.9938	0.9959
0.9929	0.9956
0.9929	0.9953
0.9928	0.9949
0.9924	0.9946
0.9932	0.9945
0.9925	0.9954
0.9924	0.9948
0.9903	0.9958

Table10. Comparison of power losses after performing PSO & HGAPSO

	performing PSO	performing HGAPSO
TPloss(kw)	0.0428	0.0383
TQloss(kvar)	3.2583	2.0548

6. Conclusion

In this paper, using HGAPSO the conductor selection has been incorporated in the conventional optimal capacitor placement, considering load growth in life period of instruments (like capacitors). By solving the optimization problem by the HGAPSO method, the optimal size and place of the capacitors and the conductors are defined. The method has been applied to a sample radial distribution network and the results show the reduction of total loss especially active power loss, in addition to the improvement of voltage profile. According to the results, the bus voltages of the ending buses are in the permissible limits. At last the proposed method is compared with PSO. The obtained results show the efficiency of the hybrid GA and PSO method. The proposed hybrid method has reduced power losses more than PSO, also the cost obtained from HGAPSO is a better answer. In conclusion, the hybrid method performs better than PSO.

References

[1] V. Marvasti, M. Nozari, H. Shariati Dehaghan, J. Azizi,” Combination of Optimal Conductor Selection and Capacitor Placement in Radial Distribution Systems for Maximum Loss Reduction ”,ACM portal,2009.

[2] RM. S. Danaraj, SH.F.Kodad“An algorithm for radial distribution power flow in Complex mode including voltage controlled buses”, Indian Journal of Science and Technology, Vol.1 No.2, 2007.

[3] De Souza, “Micro genetic algorithms and fuzzy logic applied to the optimal placement of capacitor banks in distribution networks”, IEEE Trans. On Power System, vol.19, Issue:2, May 2004, p.942-947.

[4] Chun-Feng Lu, Chia-Feng Juang, “Evolutionary fuzzy control of flexible AC transmission system,” IEE Proc.-Gener. Transm. Distrib, Vol. 152, pp. 441-448, No. 4, July 2005.

[5] H. Falaghi, M. Ramezani, M.-R. Haghifam, K. Roshan milani, ”optimal selection of conductors in radial distribution systems with time varying load” 18th International Conference on Electricity Distribution, Turin, 6-9 June 2005

[6] Ramon. .A. Gallego, and etc, “Optimal capacitor placement in radial distribution”, IEEE Trans. On power systems vol. 16, No.4, November 2001.

[7] Z. Wang, H. Liu, D. C. Yu, X. Wang, H. Song, “A Practical Approach to the Conductor Size Selection in Planning Radial Distribution Systems,” IEEE Transaction on Power delivery, Vol. 15, No. 1, pp. 350-353, 2000.

[8] AJ Rabih ,”Radial Distribution Load Flow Using Conic Programming” , IEEE Transactions on power systems,21,1458-1459.2006.

[9] S.Mandal, A.Pahwa, “Optimal Selection of Conductors for Distribution Feeders,” IEEE Transaction on Power Systems, Vol. 17, Issue 1, pp.192-197, 2002.

[10] R. A. Hooshmand, M.Ataei, “Optimal Capacitor Placement in Actual Configuration and Operational Conditions of Distribution Systems Using RCGA,” Journal of Electrical Engineering, Vol. 58, no. 4, pp. 189–199, 2007.

[11] R.A. Jabr," Optimal Placement of Capacitors in a Radial Network Using Conic and Mixed Integer Linear Programming, " Elsevier Ltd, Electric Power Systems Research, vol.78, pp. 941–948, 2008.

[12] A.Seifi, M.R.Hesamzadeh, "A Hybrid Optimization Approach for Distribution Capacitor Allocation Considering Varying Load Conditions," Elsevier Ltd., Electrical Power and Energy Systems, 2009. (Article in press).

[13] H.M.Khodr, F.G.Olsina, "Maximum Savings Approach for Location and Sizing of Capacitors in Distribution Systems, "Elsevier Ltd., Electric Power Systems Research, vol.78, pp. 1192–1203, 2008.

[14] R. Ranjan, B. Venkatesh, D. Das, "A New Algorithm for Power Distribution System Planning", Electric power system research, vol. 62, pp. 55-65, 2002.

[15] Jorge Angarita Marquez, Mohammad Ahmad A. Al-Ja'Afreh, GeevMokryani, Sohag Kabir, FelicianCampean, Cuong Dao, Sana Riaz, Optimal planning and operation of distribution systems using network reconfiguration and flexibility services, Energy Reports, Volume 9, 2023, Pages 3910-3919.

[16] Jagtap, K.M., Shukla, A. &Baboria, S.A. Optimal planning for distribution networks considering system uncertainties using pseudo-inspired gravitational search algorithm. ElectrEng (2024). https://doi.org/10.1007/s00202-024-02382-z.

[17] Z. Yang, F. Yang, Y. Shen, L. Su, Y. Lei and F. Yan, "Optimal Planning of Distribution Network Considering Photovoltaic Energy Consumption," *2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2)*, Wuhan, China, 2020, pp. 4133-4137, doi: 10.1109/EI250167.2020.9346975.

[18] Seyed Abbas Taher, Mohammad Hasani, Ali Karimian, A novel method for optimal capacitor placement and sizing in distribution systems with nonlinear loads and DG using GA, Communications in Nonlinear Science and Numerical Simulation, Volume 16, Issue 2, 2011, Pages 851-862.

[19] Gallego Pareja LA, López-Lezama JM, Gómez Carmona O. A MILP Model for Optimal Conductor Selection and Capacitor Banks Placement in Primary Distribution Systems. *Energies*. 2023; 16(11):4340. https://doi.org/10.3390/en16114340.

[20] Ali, H.; Ullah, S.; Sami, I.; Ahmad, N.; Khan, F. Economic Loss Minimization of a Distribution Feeder and Selection of Optimum Conductor for Voltage Profile Improvement. In Proceedings of the 2018 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET), Islamabad, Pakistan, 10–12 September 2018; pp. 1–6.

[21] Kumari, M.; Ranjan, R. Economical Selection of Conductor in Radial Distribution System using PSO. J. Inst. Eng. Ser. B 2022, 103, 1105–1114.

[22] Ravindra, K.; Prasad, K.R.K.V. Optimal Conductor selection and Capacitor Placement for Cost minimization in Distribution Systems. Int. J. Adv. Sci. Res. Manag. 2019, 11–15.

[23] Ismael, S.; Abdel Aleem, S.; Abdelaziz, A. Optimal Conductor Selection in Radial Distribution Systems using Whale Optimization Algorithm. J. Eng. Sci. Technol. 2019, 14, 87–107.

[24] Montoya, O.D.; Gil-González, W.; Grisales-Noreña, L.F. On the mathematical modeling for optimal selecting of calibers of conductors in DC radial distribution networks: An MINLP approach. Electr. Power Syst. Res. 2021, 194, 107072.

[25] Gupta, S.; Yadav, V.K.; Singh, M. Optimal Allocation of Capacitors in Radial Distribution Networks Using Shannon's Entropy. IEEE Trans. Power Deliv. 2022, 37, 2245–2255.

[26] Navesi, R.B.; Nazarpour, D.; Ghanizadeh, R.; Alemi, P. Switchable Capacitor Bank Coordination and Dynamic Network Reconfiguration for Improving Operation of Distribution Network Integrated with Renewable Energy Resources. J. Mod. Power Syst. Clean Energy 2022, 10, 637–646.

[27] Mtonga, T.P.M.; Kaberere, K.K.; Irungu, G.K. Optimal Shunt Capacitors' Placement and Sizing in Radial Distribution Systems Using Multiverse Optimizer. IEEE Can. J. Electr. Comput. Eng. 2021, 44, 10–21.

[28] Abul'Wafa, A.R. Optimal capacitor allocation in radial distribution systems for loss reduction: A two stage method. Electr. Power Syst. Res. 2013, 95, 168–174.

Improving The Reliability of the Distribution Network with The Approach of Load Response Resources in The Presence of Distributed Generation Resources

Mohammad Rahimzadeh [1], Hosein Nasir Aghdam*[2]
[1,2]Department of Engineering, Ahar Branch, Islamic Azad University, Ahar, Iran.
Email:hossein.nasiraghdam@iau.ac.ir(Corresponding Author)

Received: 26 September 2023 Accepted: 06 August 2024 Published:15 February 2025

Abstract

Traditionally, distribution networks are radial in topology and passive in nature. Over the past few decades, there has been tremendous interest in non-conventional energy sources, largely driven by dwindling fossil fuel supplies and increasing global warming concerns. The purpose of this research is to manage and improve the reliability of the distribution network with the approach of load response resources in the presence of scattered production resources. In this research, we also considered the effect of storage devices on the congestion and observed that the storage devices act as a load when there is no congestion in the network and draw power from the network and are charged, but in a situation where the congestion in the network is high. They act as a source of power production and are discharged and injected into the power grid, which in turn reduces congestion.

Keywords: reliability, distribution network, load response, distributed generation resources

1. Introduction

Electric energy cannot be widely stored at the level of the power system, so the amount of production capacity available at all times must be equal to or greater than the total load of the system's consumers. At some times during a year, the amount of power consumption of the system will increase drastically, and in this situation, without taking into account the response, the amount of production capacity required to provide power and save these hours will increase. Meanwhile, the cost of installing power plant units is very high and time-consuming. However, with the implementation of load response programs, the amount of consumption during peak hours by consumers who are willing to reduce their consumption will be reduced, and as a result, spending additional costs to create production capacity for a short period of time will prevent excessive load increase every year. Load accountability forms a major part of consumption management programs. Because the nature of these programs is very suitable for adapting to the new power system management structure. Today, these programs are considered as a suitable solution to solve some problems of restructured power systems. Load response can change the form of electric energy consumption in such a way that the peak load of the system is reduced and the consumption is transferred to non-peak hours. The implementation of load response programs can improve the utilization of the power system from an economic point of view, maintaining the reliability and efficiency of the retail and wholesale markets.

The integration of any type of resource into the distribution network makes it an active network. There are many advantages due to DG integration such as reduction of power losses, improvement of voltage profile, stability, reliability, cost saving, etc. [3]. These positive effects of DG resources mainly depend on the location and size of DG resources in the distribution network [4]. But due to the higher penetration rate of DG resources, including technical, commercial and regulatory challenges, there can be some disadvantages [5]. One of the most important topics among them is protection coordination [6]. In [7], the importance of replacing protective devices due to the higher penetration of DG sources is discussed. Under certain conditions, the fault current can exceed the minimum or maximum range [8]. Protection failures due to high penetration of DG can be mainly of two types, i.e., undesired shutdown or failure during fault [9]. With the introduction of more and more DG resources on the distribution side, the network topology is changing from radial to mesh. For a mesh network, traditional analysis methods do not provide accurate results. [10] Figure (1-1) shows a typical active distribution network. The output of renewable energy sources and the load profile are usually stochastic in nature and this uncertainty must be considered in the planning stage. [11] The impact of renewable DG output changes when integrated with the grid is taken care of by pumped storage units, plug-in electric vehicles (PEVs) with vehicle-to-grid (V2G) capability, battery storage devices, diesel generators, etc. [12]. If not properly planned, the increased use of PEVs can lead to the deterioration of system reliability, mainly due to the increased system load condition during vehicle charging. Grid-connected battery swapping stations with V2G capability improve reliability [13].

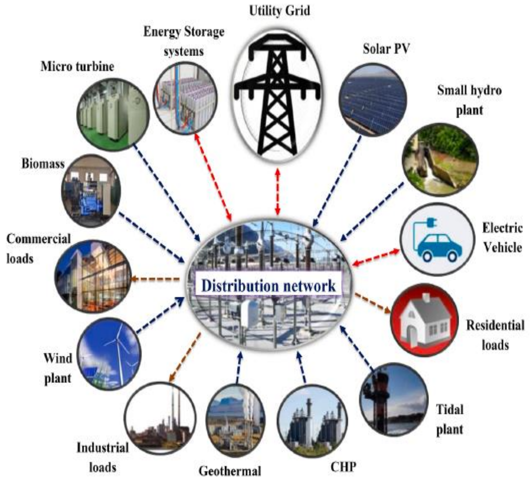


Fig.1. A typical active distribution network.

The positive and negative effects of DG integration in the network are shown in Figure (2).

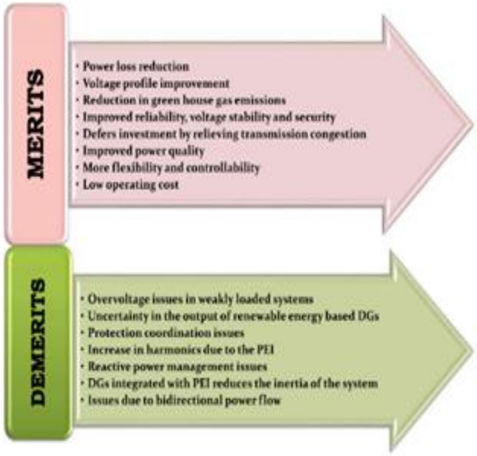

MERITS
• Power loss reduction
• Voltage profile improvement
• Reduction in green house gas emissions
• Improved reliability, voltage stability and security
• Defers investment by relieving transmission congestion
• Improved power quality
• More flexibility and controllability
• Low operating cost
DEMERITS
• Overvoltage issues in weakly loaded systems
• Uncertainty in the output of renewable energy based DGs
• Protection coordination issues
• Increase in harmonics due to the PEI
• Reactive power management issues
• DGs integrated with PEI reduces the inertia of the system
• Issues due to bidirectional power flow

Fig. 2. Effects of DG integration on the grid.

After the recent blackouts in various power grids around the world, access to reliable electricity has become the need of the day.

The complexity of the electric power system is increasing day by day due to the increasing integration of renewable DGs, especially on the distribution grid side. The uncertainty introduced into the system by these renewable DGs can have a detrimental effect on the reliability of the distribution network .[14] Hence reliability assessment and improvement techniques for distribution networks with renewable DG sources have been widely studied in recent years. The implementation of performance-based electricity pricing in some countries has also increased interest in distribution network reliability studies among researchers. In the literature, various reliability assessment methods have been proposed and are broadly classified as analytical and simulation-based approaches [15]. Definition of congestion and its effects in the power network.

Congestion can be defined as the filling of the permitted capacities or the use of the power network outside the permitted operating limits [8]. These limits can be related to the permitted bus voltage limits, generator production limits, transmission lines , the ranges of distribution lines and... In terms of the transmission network, congestion or congestion refers to a megawatt of overload in the network lines that occurs during the operation of the power system in conditions such as peak load or other emergency conditions such as outages of lines and generators [7]. The limitations of the transmission lines are generally complex and some of them can be omitted for the simplicity of the model, but the limitations mentioned below are among the most important limitations that need to be considered for congestion management [9]:

• Thermal limitations: the passage of electric current through the lines causes a part of the power to be lost and the line to heat up. Around a certain temperature, there is a possibility of permanent damage to the line. This heat is generated not only by active power but also by reactive power.

• Limits of the voltage range: the voltage limits define the boundaries of the operation, which limits the power passing through the lines. Considering the voltage limits, both active and reactive power loading of the transmission lines should be considered.

• Stability limits of transmission lines: the difference between the load voltage angle and the generator voltage is called "phase angle". When the phase angle reaches nearly 90 degrees, the power passing through the line decreases and causes instability. This value shows the limit of the physical stability of the lines. In general, angle stability is divided into two categories: small signal stability, which is the ability of the system to remain synchronous when small fluctuations occur; Transient stability, which expresses the system's ability to remain synchronous when a strong transient oscillation occurs.

• Limits of voltage stability: Voltage stability is the ability of the power system to keep the voltage of all buses within the permissible range under normal conditions or after encountering a disturbance. The main reason for voltage instability is the system's inability to supply reactive power [9].

In addition to the four physical limitations stated above, event limitations for line congestion must also be considered. Event analyzes with different simulations deal with potential emergency situations in the network [8].

2. Congestion management methods in the power network

Congestion management refers to all things that are done to prevent the occurrence of congestion or to release the transmission and distribution capacity, which means removing the existing congestion. With a more comprehensive definition, any operation that ensures the balance of load and production conditional on freeing the capacity of the lines is called congestion management. Congestion management of transmission lines, as one of the key tasks of the network operator, is a process that ensures the use of the transmission network within the permitted limits of operation. In general, congestion management can be a systematic solution for production planning and adaptation. and consumption should be considered. In restructured systems, the main goal in congestion management can be considered a set of rules and procedures that guarantee sufficient control over producers and consumers, to maintain an acceptable level of safety and reliability of the power system along with The maximum economic efficiency of the market is at the time of transmission network limitation and can provide sufficient economic signs for the long-term growth of the system. These laws must have characteristics such as being inviolable, transparent and fair, the network operator must be able to solve the problem of congestion management with a competitive approach. in such a way that the maximum possible use of the transmission network is realized and at the same time the largest amount of exchanges is possible in the most economical way possible; That is, in theory, the confrontation of different market forces makes it possible to provide the right to use the condensed line for the user provided that he attaches the highest value to it. [11]

Congestion of system can in some cases lead to the creation of market power for some market participants. This issue can interfere with the creation of a free electric energy market and delay its progress. Since congestion management is very dependent on the market and without market considerations, it does not have the necessary validity, the compilation of this set of rules has been done in different ways in the world, which has led to the presentation of various methods in congestion management based on the market. From a general point of view, existing congestion management methods can be divided into two categories, before congestion under the title of preventive congestion management methods2 and after congestion under the title of corrective congestion management methods. In the following, these methods are briefly explained [10].

The main methods of congestion management in transmission networks

Based on the review of available sources, in general, the main methods of congestion management in corrective congestion management methods can be briefly divided as follows:

1) auction-based methods

2) methods based on pricing

3) Methods based on reuse

4) load shedding and use of load response programs

5) Using tap transformers and phase changers and using tools (FACTS)

6) Using scattered production resources

3. Problem Data

Data related to DGs, ESSs and WTs are given in tables (1) to (6) for 33-base and 85-base IEEE networks. The capacity of the distribution feeder, which is supplied from the upstream network, is assumed for the 33 bus and 85 bus networks, respectively, for active power P_max=1 pu and P_max=1 pu و P_max=1.5 pu and for reactive power Q_max=0.8 pu و Q_max=0.7 pu. For both 33-bus and 85-bus networks, the number of DGs, ESSs, and WTs is 2, 4, and 2 units, respectively.

Table 1.DG data for 33-base network

Encouragement coefficients ($)	Posation	Qmin(Kvar)	Qmax(Kvar)	Pmin(kw)	Pmax(kw)	DG
1/1	6	0	100	0	200	1
1/34	11	0	100	0	200	2
½	16	0	100	0	200	3
1/42	22	0	100	0	200	4

Table 2. ESS data for the 33-base network

Posation	Pdmin (KW)	Pdmax (KW)	Pcmin (KW)	Pcmax (KW)	SOCmin (KWh)	SOCmax (KWh)	ESS
13	0	40	0	40	40	200	1
21	0	20	0	20	20	100	2

Table 3-4. WT data for 33-base network

Encouragement coefficients ($)	Posation	Pmin(kw)	Pmax(kw)	WT
1/1	24	0	450	1
1/2	29	0	650	2

Table .4 Data of DGs for 85-base network

Encouragement coefficients ($)	Posation	Qmin(Kvar)	Qmax(Kvar)	Pmin(kw)	Pmax(kw)	DG
1/1	23	0	130	0	130	1
1/34	47	0	130	0	130	2
½	81	0	130	0	130	3
1/42	53	0	130	0	130	4

Table 5. WT data for 85-base network

Encouragement coefficients ($)	Posation	Pmin(kw)	Pmax(kw)	WT
1/1	24	0	320	1
1/2	32	0	210	2

Simulation results on IEEE 33 bus network in deterministic space

Congestion management modeling in this thesis is implemented on IEEE 33 bus system. Figure (5) shows the single-line diagram of the IEEE 33-bus network. This network has 33 buses, 33 branches and its total load is 3.715 MW and 2.300 MW. The number of DGs used is 4 units, which are located in 〖dg〗_4. are located and their capacity is 200 (kw) each for active power production and 100 (var) for reactive power production. Also, there are 2 storage units in this system, which are ESS1 and ESS2. They are located in busses 21 and 13, respectively, and their capacity is 100 (kw) and 200 (kw), finally, there are 2 wind turbine units named WT1 and WT2 located in busses 24 and 29. Their power output is (kw) 450 and (kw) 650 respectively. For preuniting parameters and variables, V_base=12.66 KV and S_base=1 MVA.

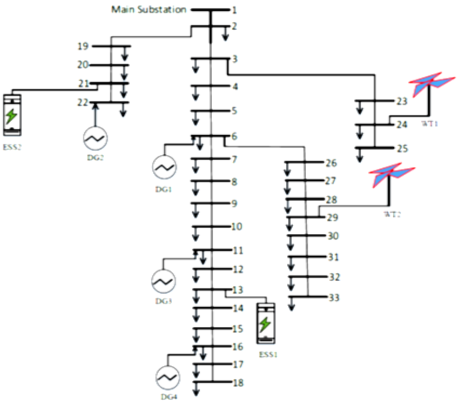


Fig .5. Radial distribution network of 33 bases along with DGs, ESSs and WTs

The profile of the 34-hour consumption load curve and the wind speed pattern are shown as the input data of the simulation in Figure (6).

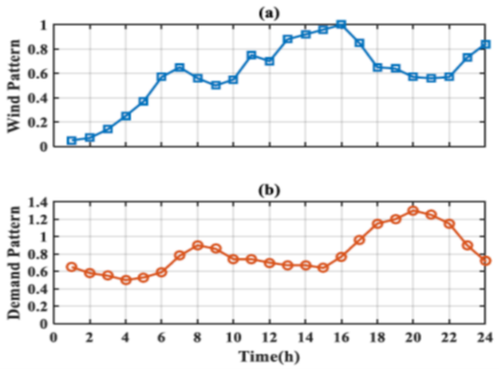


Fig.6. wind speed pattern (a) and load pattern (b) [18]

If we consider the output power of WTs as a parameter and input and the maximum capacity of WT1 = 0.45 pu and WT2 = 0.65 pu, then the output power of WT1 and WT2 according to the proposed wind speed pattern for 24 hours in The decompression time will be as shown in figure (7).

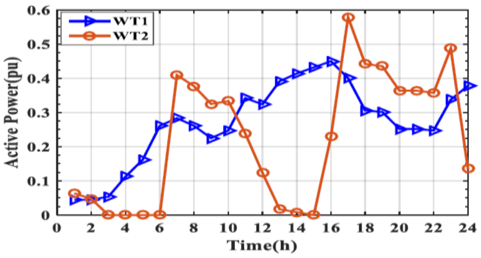


Fig. 7. Production power of wind units

Figure (8) shows the status of clearing network congestion and the time of network congestion for different hours.

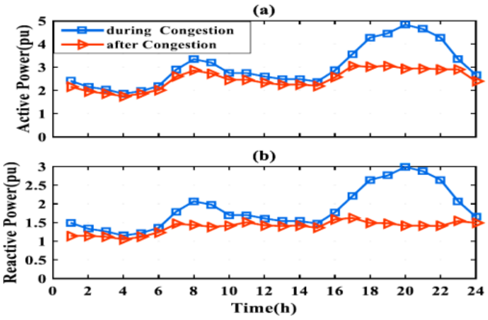


Fig. 8. The total active power consumption of the load in all buses at the time of congestion and the time of decongestion (a) and the total reactive power of the load in all buses at the time of congestion and the time of decongestion (b)

As can be seen in Figure (8), the congestion in the network has been significantly resolved, for example, for hours 18 to 22 and for part (a) where the network congestion is at its peak and the network is under stress and full pressure. In these hours, the amount of active power consumed by loads is mostly in the amount of 3 pryunits, if for the state before the congestion is removed, these values are in the range of 4 to 5 pryunits. In general, the decompression load curve for part (a) mostly controls the loads in the range of 1.5 to 3 priunits, while for part (b) it is more in the range of 1 to 1.5 priunits to create compaction. fix the problem

Figure (9) shows the amount of power produced by DGs to relieve congestion. In this figure, it can be seen that for hours 18 to 22 for part (a) and (b) when the congestion is at its peak value, the production power of DGs is also at its maximum capacity for these hours to solve the related congestion. In fact, the ups and downs in Figure (9) respectively indicate more or less participation of DGs to relieve congestion, for example, for hour 4, when a decrease is observed in the graph for both active and reactive power, it shows that in this hour congestion The grid is low and the DGs have produced less power.

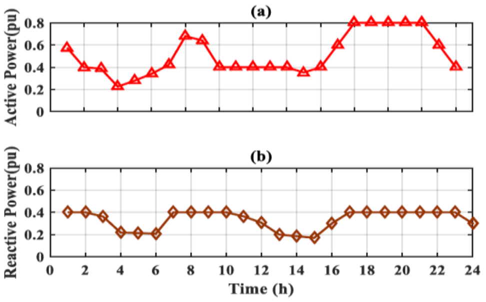


Fig .9. The sum of all the active power generation (DGs) (a) The sum of all the reactive power generation DGs (b) after removing congestion

4. Conclusion

Congestion is defined as a violation of physical limits, security and reliability in the power system or to - use of the power network outside the permitted limits of operation. These limits can be related to bus voltage, generator production limits, transmission and distribution line limits, etc. The problem of congestion in the distribution network is raised as a voltage problem and an overload problem, so that the voltage of the buses must be within the defined limits, usually 10% higher or lower than their nominal voltage, and the loading of the distribution lines must be be very close to the thermal limits of the power system. The effects of congestion in the distribution network make it possible to create additional outputs in the network, cause damage to the electrical equipment in the system, increase the price of electrical energy in some areas, reduce social welfare, frequent blackouts, etc. In order to solve or reduce the problems of voltage drop, overvoltage and overload in the distribution network, the distribution system operator (DSO) must implement congestion management methods in the network to resolve network congestion.

In this article, we presented congestion management in intelligent distribution networks by considering the uncertainty caused by the load and production of DGs in two deterministic and random spaces with the presence of storage devices and wind units for two IEEE 33-bus and 85-bus networks. . The main approach of the thesis is based on congestion management in intelligent distribution networks due to the comprehensive response of loads and increasing the production of DGs. As shown from the results, the proposed model manages congestion in both deterministic and random spaces. In the deterministic and random space, in two cases, taking into account the emission constraint and without considering the emission constraint, we investigated the effect of load response and increasing the power generation of DGs. Decongestion was effective, while without taking into account the pollution constraint, the increase in the power generation of DGs is more than the load response on decongestion. They can produce any amount of power, so to solve the congestion, the load response should be increased, but in the case of not considering the emission limit, the production rate of DGs is higher, which makes DGs produce power in any amount that they have transfer capacity, so in this case More than load response, it has an effect on congestion relief. For times when the congestion in the network is high, as we have seen in the results, all DGs produce power at their maximum value and the responsiveness of loads reaches its maximum value. As mentioned earlier, the type of load response program is the type of incentive-based load response programs and direct load control method (DLC). In this method, we considered the cost coefficients for the loads and DGs, which makes the loads and DGs participate in decongestion according to these cost coefficients, and the distribution network operator makes the corresponding planning to minimize the total cost of decongestion. objective function to do.

The total cost of decongestion, as it was obtained from the results, is higher in the deterministic space than in the random space, which was the reason that in the deterministic space, we force loads and DGs according to predetermined conditions, which may be due to the conditions is not too close to the real one, participate in decongestion, while in the random space, we plan by considering scenarios to more accurately simulate the real conditions, which in turn reduces the total cost of decongestion in the random space compared to the deterministic one. It is possible that pollution also had an effect on the costs, so that in the case of taking pollution into account, the total cost of decongestion and the cost of load response was higher than the case without considering pollution, but the cost of power generation of DGs in the case Considering the pollution restriction was lower than the case without considering the pollution restriction. In this work, we also showed the voltage profile for the state after decompression, as it was seen from the simulations, this curve was smoothed, in fact, in the decompression mode, we see the depression and non-flatness of the voltage curve, but after decompression By increasing power generation in DGs and load response, we flattened the voltage curve in high congestion points, which in turn led to better voltage management. Also, in this thesis, we also considered the effect of storage devices on the congestion and we observed that the storage devices act as a load when there is no congestion in the network and draw power from the network and are charged, but in the case where the congestion is The grid acts as a source of power generation and discharges and injects into the power grid, which in turn reduces congestion. This effect of storage devices is shown in the discussed results, which are discharged during peak hours and It relieves network congestion.

References

[1] Huang, Wujing, et al. "From demand response to integrated demand response: Review and prospect of research and application." Protection and Control of Modern Power Systems 4.1 (2019): 1-13.

[2] McPherson, Madeleine, and Brady Stoll. "Demand response for variable renewable energy integration: A proposed approach and its impacts." Energy 197 (2020): 117205.

[3] McPherson, Madeleine, and Brady Stoll. "Demand response for variable renewable energy integration: A proposed approach and its impacts." Energy 197 (2020): 117205.

[4] Wang, Xiaonan, Ahmet Palazoglu, and Nael H. El-Farra. "Operational optimization and demand response of hybrid renewable energy systems." Applied Energy 143 (2015): 324-335.

[5] Jiang, Wei, et al. "Optimal economic scheduling of microgrids considering renewable energy sources based on energy hub model using demand response and improved water wave optimization algorithm." Journal of Energy Storage 55 (2022): 105311.

[6] Hajiamoosha, Pouria, et al. "Stochastic energy management in a renewable energy-based microgrid considering demand response program." International Journal of Electrical Power & Energy Systems 129 (2021): 106791.

[7] Zeng, Bo, et al. "Optimal demand response resource exploitation for efficient accommodation of renewable energy sources in multi-energy systems considering correlated uncertainties." Journal of Cleaner Production 288 (2021): 125666.

[8] Zeng, Bo, et al. "Optimal demand response resource exploitation for efficient accommodation of renewable energy sources in multi-energy systems considering correlated uncertainties." Journal of Cleaner Production 288 (2021): 125666.

[9] Bogdanov D, Gulagi A, Fasihi M, Breyer C. Full energy sector transition towards 100% renewable energy supply: Integrating power, heat, transport and industry sectors including desalination. Appl Energy 2021;283:116273.

[10] Mehigan L, Deane JP, Gallach´oir B´O, Bertsch V. A review of the role of distributed generation (DG) in future electricity systems. Energy 2018;163:822–36.

[11] Paliwal P, Patidar NP, Nema RK. Planning of grid integrated distributed generators: A review of technology, objectives and techniques. Renew Sustain Energy Rev 2014;40:557–70.

[12] Balamurugan K, Srinivasan D, Reindl T. Impact of distributed generation on power distribution systems. Energy Procedia 2012;25:93–100.

[13] Mahmud N, Zahedi A. Review of control strategies for voltage regulation of the smart distribution network with high penetration of renewable distributed generation. Renew Sustain Energy Rev 2016;64:582–95.

[14] Huda AN, ˇZivanovi´c R. Large-scale integration of distributed generation into distribution networks: study objectives, review of models and computational tools. Renew Sustain Energy Rev 2017;76:974–88.

[15] Meneses CAP, Mantovani JRS. Improving the grid operation and reliability cost of distribution systems with dispersed generation. IEEE Trans Power Syst 2013;28 (3):2485–96.

[16] Hernando-Gil I, Ilie IS, Djokic SZ. Reliability planning of active distribution systems incorporating regulatory requirements and network-reliability equivalents. IET Gener Transm Distrib 2016;10(1):93–106.

[17] Jafari A, Khalili T, Ganjehlou HG, Bidram A. Optimal integration of renewable energy sources, diesel generators, and demand response program from pollution, financial, and reliability viewpoints: a multi-objective approach. J Cleaner Prod 2020;247:119100.

[18] Prajapati VK, Mahajan V. Reliability assessment and congestion management of power system with energy storage system and uncertain renewable resources. Energy 2021;215:119134.