I. R. IRAN

ISSN: 1728-1431
e-ISSN: 1735-9244

International Journal of Engineering

Journal Homepage: www.ije.ir

MERC

TRANSACTIONS A: BASICS

Volume 39, Number 10, October 2026

Materials and Energy Research Center

In the Name of God

INTERNATIONAL JOURNAL OF ENGINEERING
Transactions A: Basics

DISCLAIMER

The publication of papers in International Journal of Engineering does not imply that the editorial board, reviewers or publisher accept, approve or endorse the data and conclusions of authors.

International **J**ournal of **E**ngineering *Transactions A: Basics* (ISSN 1728-1431) (EISSN 1735-9244)
International **J**ournal of **E**ngineering *Transactions B: Applications* (ISSN 1728-144X) (EISSN 1735-9244)
International **J**ournal of **E**ngineering *Transactions C: Aspects* (ISSN 2423-7167) (EISSN 1735-9244)
Web Sites: www.ije.ir & www.ijeir.info E-mails: ije.editor8@gmail.com, Tel: (+9821) 88771578, Fax: (+9821) 88773352
Materials and Energy Research Center (MERC)

ISSN: 1728-1431, e-ISSN: 1735-9244

International Journal of Engineering, Volume 39, Number 10, October 2026

CONTENTS

Transactions A: Basics

ISSN: 1728-1431, e-ISSN: 1735-9244

International Journal of Engineering, Volume 39, Number 10, October 2026

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2361-2370

International Journal of Engineering

Journal Homepage: www.ije.ir

Microstructural and Mechanical Characteristics of AlCoCrFeNi-YSZ High Entropy Alloy Matrix Composite

M. Askarzadeh Ardestani[a], M. Farvizi*[b], A. Alizadeh Samiyan[a], M. Ghanbariha[a], T. Ebadzadeh[a], M. M. Kalantarian[a]

[a] *Department of Ceramic, Materials and Energy Research Center, Karaj, Iran*
[b] *Department of Materials Engineering, University of Tabriz, Tabriz, Iran*

PAPER INFO

Paper history:
Received 02 September 2025
Received in revised form 05 October 2025
Accepted 25 October 2025

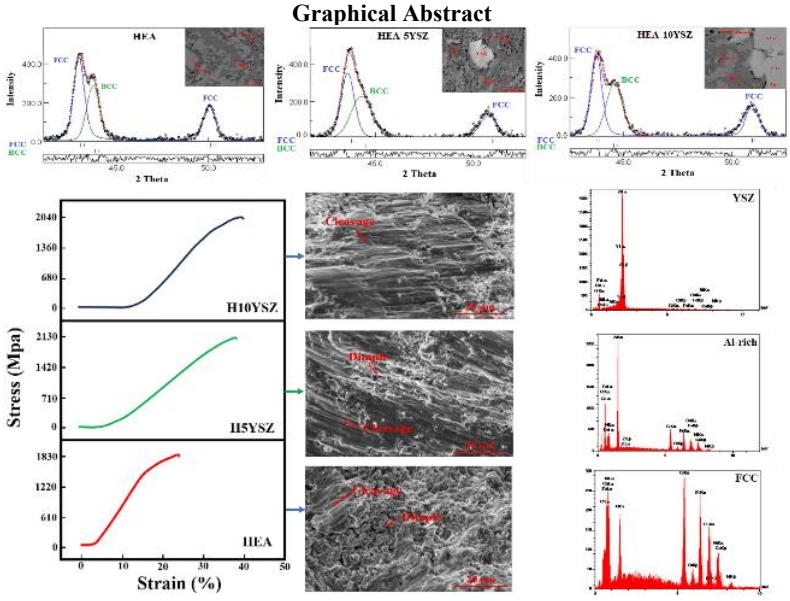
Keywords:
High Entropy Alloy
Ceramic Particulates
Composite Materials
Mechanical Properties

ABSTRACT

In this study, high-entropy alloy matrix composites (HEAMCs) consisting of an AlCoCrFeNi matrix reinforced with zirconia-based particles were fabricated using mechanical alloying (MA) followed by spark plasma sintering (SPS). Microstructural analysis revealed the presence of both FCC and BCC phases after 30 hours of mechanical alloying and subsequent sintering at 1000 °C. Rietveld refinement analysis indicated that the weight fraction of the FCC phase remained nearly constant in both the HEA and HEA–YSZ composites. However, the addition of YSZ particles decreased the BCC phase content and restricted grain growth during sintering. Compressive test results showed that the specimen containing 5 wt.% YSZ exhibited the highest yield strength (1760 MPa), attributed to its higher BCC phase fraction (40%). In contrast, the specimen with 10 wt.% YSZ demonstrated reduced yield strength due to a lower BCC phase content (36%). These findings indicate that AlCoCrFeNi composites reinforced with YSZ possess higher yield strength than those reinforced with ZrO_2 when comparable reinforcement levels are used. This improvement is attributed to the higher BCC phase content and the superior intrinsic properties of YSZ. Fracture surface analysis confirmed the coexistence of both ductile and brittle fracture modes in the tested specimens.

doi: *10.5829/ije.2026.39.10a.01*

Graphical Abstract



*Corresponding Author Email: *mmfarvizi@yahoo.com* (M. Farvizi)

Please cite this article as: Askarzadeh Ardestani M, Farvizi M, Alizadeh Samiyan A, Ghanbariha M, Ebadzadeh T, Kalantarian MM. Microstructural and Mechanical Characteristics of AlCoCrFeNi-YSZ High Entropy Alloy Matrix Composite. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2361-70.

1. INTRODUCTION

As science and technology advance in material engineering, the alloying industry has undergone major changes. A notable advancement in materials science is the identification of new types of alloys known as high entropy alloys (HEAs) in 2004 (1-4). These alloys can be produced by mixing a minimum of five elements, ensuring their atomic percentages are equal to or nearly the same as one another (5-8). This has led to a high entropy effect in these alloys (4). The constituent elements' types and quantities are the primary factors influencing the microstructure and characteristics of high entropy alloys (6). HEAs can possess interesting properties, including high resistance to softening at elevated temperatures, low diffusion kinetics, high fracture and creep resistance, and good compressive strength at room temperature (7). HEAs mostly have a high melting point and maintain strength at extremely high temperatures (9).

The AlCoCrFeNi system has been extensively studied among various HEAs because of its remarkable characteristics (10-18). AlCoCrFeNi compounds can be used as the matrix of metal matrix composites (MMCs) because of their high specific strength compared to nickel- and titanium-based alloys and polymers (7). AlxCoCrFeNi HEAs are relatively lightweight and exhibit high compressive strength and excellent structural flexibility, particularly when compared to other HEAs such as CoCrFeNiMn and refractory HEAs (19-26).

Recent studies have indicated that optimizing material composition, processing parameters, and microstructural design can significantly enhance the tribological behavior, wear resistance, and energy absorption capacity of advanced composite systems used in dynamic and high-friction environments (27-30). Incorporating hard ceramic particles into MMCs is an effective method to enhance mechanical properties such as wear resistance, hardness, compressive strength, and toughness compared to the base metal (31).

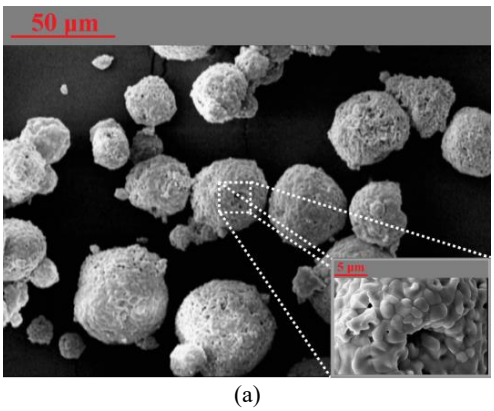
Zirconia is an interesting material for various applications due to its notable characteristics, including a high elastic modulus, outstanding hardness, and elevated melting point (32, 33). Earlier studies have demonstrated that incorporating ZrO_2 into alloys such as Al, Cu, and Mg can enhance both the mechanical and tribological characteristics of these materials (34-36). However, at 1100°C, a phase transformation occurs between the monoclinic and tetragonal phases of zirconia, accompanied by a considerable volume change that can negatively affect the properties of this phase (37). Therefore, different oxides such as Y_2O_3 can be added to ZrO_2 to restrict this transformation. The addition of yttrium oxide stabilizes the tetragonal phase of zirconia at high temperatures and prevents transformation to the stable monoclinic phase at room temperature (37). Previous investigations have demonstrated that YSZ possesses superior mechanical properties compared with ZrO_2. Specifically, YSZ exhibits higher elastic modulus, hardness, and compressive strength than zirconia. Moreover, its higher coefficient of thermal expansion (CTE) relative to monoclinic zirconia helps to minimize thermal mismatch stresses during the sintering process (38).

There are only limited studies on HEA matrix composites reinforced with ceramic particles. In research by Ghanbariha et al. (39), AlCoCrFeNi reinforced with monoclinic zirconia particles was produced by the MA-SPS technique, and the effect of these particles on the microstructural aspects of AlCoCrFeNi-ZrO_2 composites was studied. According to their results, all specimens after the SPS process had two main phases, including FCC and BCC. It was also found that zirconia particles had a positive impact on the wear resistance of HEA-ZrO_2 composites (40).

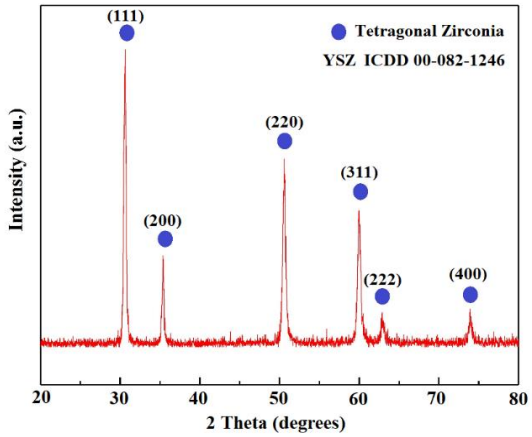
Due to the superior properties of YSZ compared to ZrO_2 in metal matrix composites, this study focuses on producing AlCoCrFeNi-YSZ composites using the MA-SPS method. The microstructural characteristics and mechanical properties of these composites are systematically investigated and compared with AlCoCrFeNi-ZrO_2 composites. The primary objective of this research is to evaluate the effectiveness of YSZ reinforcement in enhancing the mechanical performance of AlCoCrFeNi-based HEA composites and to provide insights for the development of high-performance HEA matrix composites.

2. MATERIALS AND METHODS

Figure 1 presents the XRD and SEM images of the initial YSZ powder. The specimens were fabricated through mechanical alloying combined with the spark plasma sintering method. The details of this procedure were discussed by Ghanbariha et al. (39) and depicted in



(a)



(b)

Figure 1. (a) SEM image, and (b) XRD pattern of YSZ powder

Figure 2. The identification of the phases in the raw powders and the sintered specimens was performed using X-ray diffraction (XRD - CuK_α radiation: at 20 kV, λ = 0.154 nm, and 30 mA) (D-500, Siemens, Germany). Each diffraction pattern was recorded with a step size of 0.2. For the analysis of microstructure, field emission scanning electron microscopy (FESEM, TESCAN MIRA3, Czech Republic) was used in conjunction with energy dispersive spectroscopy (EDS). Material analysis utilizing diffraction (MAUD) software was employed to acquire crystallographic data on the relevant phases, along with quantitative analysis of the specimens through the Rietveld refinement method. During all Rietveld refinement procedures, the R and sigma values were consistently below 10 and 2, respectively.

Compression tests were performed according to E3 standard with a 1 mm/min shear rate at room temperature. For this purpose, cylindrical specimens were prepared with a thickness and diameter of 5.2 mm by an electrical discharge machine (EDM). Every experiment was conducted three times to confirm the reliability of the results.

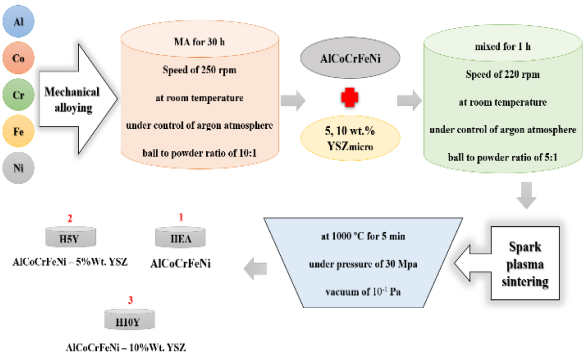


Figure 2. Details of MA and SPS process

3. RESULT AND DISCUSSION

3. 1. Synthesis of HEA Powder Figure 3 displays the SEM micrograph of the HEA powder produced using MA. This image shows that the spherical powders are formed by the cold-welding process during MA. The difference in the inherent hardness of the initial powders activated two main mechanisms — fracture and cold welding (brittle–soft and soft–soft) — that contributed to quasi-spherical particle formation and uniform powder mixing. During the process, due to the increase in temperature, diffusion of atoms within the crystal lattice of other atoms and self-diffusion are activated. The evolved BCC phase results from crushing due to work hardening of the powders and an increase in dislocation density, which enhances diffusion pathways. The particle size reduction and simultaneous diffusion continue until the alloy system appears as a single powder.

3. 2. SPS of AlCoCrFeNi-YSZ (micro) HEA To gain insights into how YSZ reinforcement influences the phase components and the proportion of phases within HEA matrix composites, Rietveld refinement analysis was employed. The regions extracted from Rietveld refinement analysis are shown in Figure 4(a-c). Figure 4 illustrates scans for the (111) plane of the FCC phase as well as the (110) plane of the BCC phase for the HEA, H5Y, and H10Y specimens. The outcomes from this analysis are summarized in Table 1. Prior to SPS, the HEA powder had a crystallite size of approximately 5 nm, while following the SPS process, the crystallite size for the BCC phase of the HEA specimen increased to 80 nm and for the FCC phase to about 38 nm. Notably, when 10 wt.% YSZ was incorporated, the crystallite sizes of the FCC and BCC phases became approximately 29 nm and 79 nm, respectively, indicating that these particles hinder grain growth in the HEA matrix. In general, secondary particles in the matrix, whether formed in situ

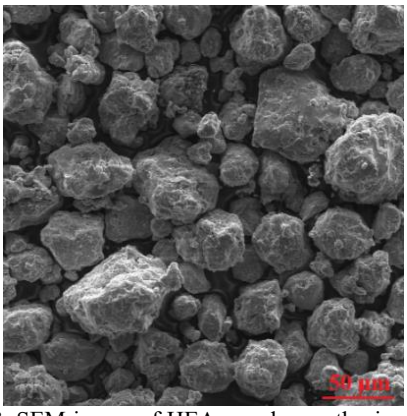
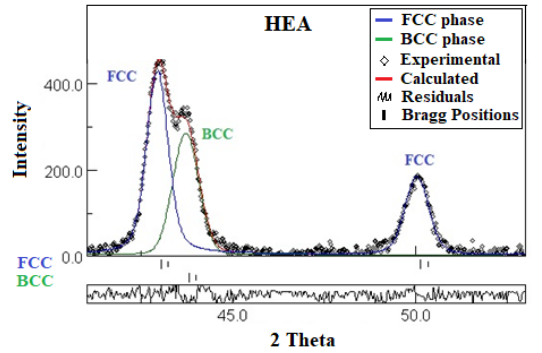
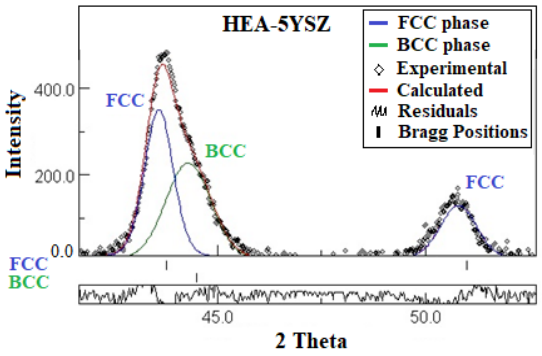


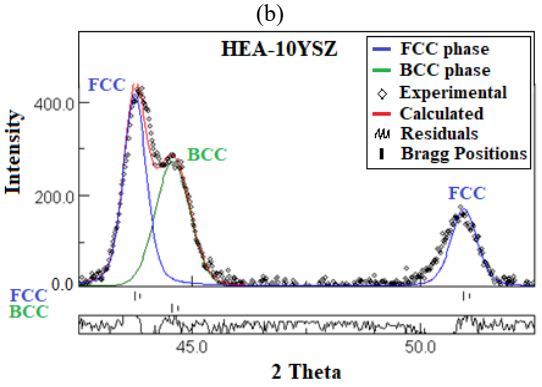
Figure 3. SEM image of HEA powder synthesized by MA after 30 hours



(a)



(b)



(c)

Figure 4. (a, b, c) Result Rietveld refinement with the detailed scans for FCC and BCC phases of all specimens

or added externally, prevent grain growth through pinning (41). In the AlCoCrFeNi-ZrO_2 specimen, incorporating 10 wt.% zirconia resulted in crystallite sizes of the FCC and BCC phases to reach 21 nm and 78 nm, respectively (39). This suggests ZrO_2 particles are more effective at inhibiting grain growth than YSZ particles. The relatively higher diffusion coefficient of YSZ particles, due to oxygen vacancies, facilitates grain growth. Consequently, ZrO_2 exhibits a stronger capacity to prevent grain growth. Additionally, zirconia particles, owing to their irregular shape, provide a larger surface area for nucleation compared to spherical YSZ particles.

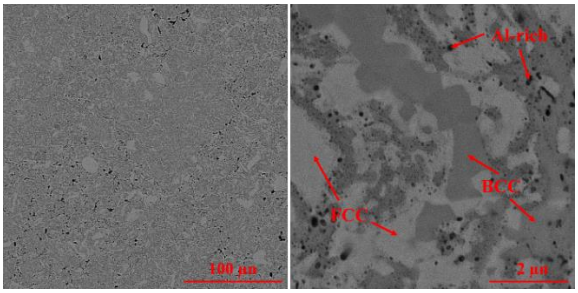
The data shown in Table 1 indicates that the weight percentage of the FCC phase in HEA and HEA composites remains relatively stable as the YSZ content increases during the SPS process; however, the BCC phase content diminishes in the composites. The reduction in BCC phase with increasing reinforcement particles can be attributed to the formation of interfaces and dislocations, which cause active elements such as aluminum to be extracted from the BCC phase—known for its higher aluminum content—and relocate to the interface between HEA and YSZ particles, resulting in the formation of an Al-rich phase. This explanation aligns with the findings presented in Table 1 and Figures 6 and 7.

3. 3. Microstructural Evaluation Figure 5 shows the FESEM results with EDS analysis of AlCoCrFeNi after the SPS process. It is noted that the HEA consists of four primary phases: BCC (dark grey), FCC (bright), Cr-rich (gray), and Al-rich (black) phases. According to the EDS data shown in Table 2, the BCC phase exhibits a high weight percentage of Al and Ni, while displaying lower weight percentages of Cr, Fe, and Co. Conversely, the FCC phase has a greater concentration of Cr, Fe, and Co, with a reduced presence of Al and Ni. Furthermore, the microstructure of this specimen displays small black dots, identified as Al-rich phases.

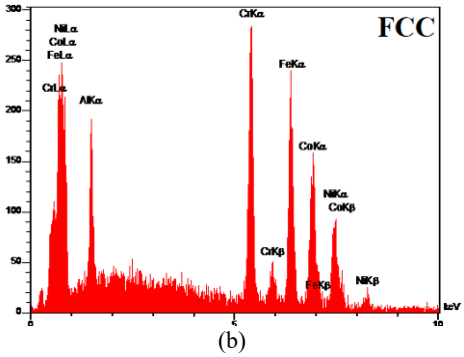
Figure 6 shows FESEM images of AlCoCrFeNi-5wt.% YSZ following the sintering process. In this specimen, in addition to the YSZ reinforcement particles that are visible as white within the HEA matrix, three other phases (FCC, BCC, and Al-rich) have been recognized. The FCC phase is characterized by a high concentration of Fe, Cr, and Co, whereas the BCC phase is predominantly composed of Al and Ni. According to

TABLE 1. Structural data of HEA-YSZ (micro) composites obtained from Rietveld refinement

YSZ (wt.%)	**0** (39)		**5**		**10**	
Crystal system	FCC	BCC	FCC	BCC	FCC	BCC
Space group	Fm-3m	Im-3m	Fm-3m	Im-3m	Fm-3m	Im-3m
Unit cell	0.36	0.28	0.35	0.28	0.35	0.28
Crystallite size (nm)	38.21	80.04	27.17	79.03	29.89	79.84
Phase fraction (wt. %)	55.85	44.15	54.80	40.20	54.00	36.00

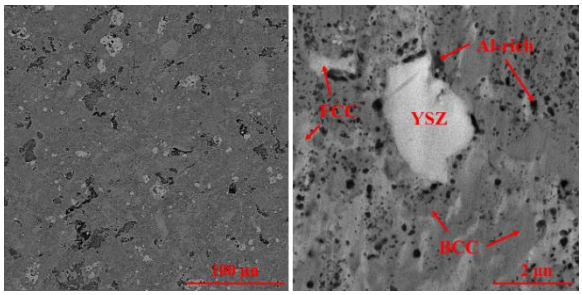


(a)

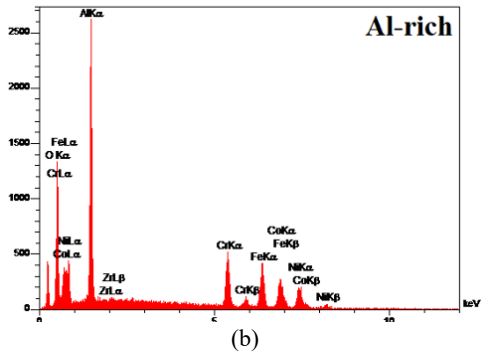


(b)

Figure 5. a) SEM image of the SPS-processed AlCoCrFeNi, and b) EDS analysis



(a)



(b)

Figure 6. a) SEM image of the SPS-processed AlCoCrFeNi-5 wt. % YSZ, and b) EDS analysis

prior research (39), a comparison of HEA-5 wt.% ZrO_2 and HEA-5 wt.% YSZ showed that the amounts of Al and Co in the FCC phase of HEA-5 wt.% YSZ were significantly decreased to 3.72 and 23.69 wt.%, respectively, while the concentrations of these elements in the HEA-5 wt.% ZrO_2 specimen stayed relatively stable. Furthermore, there was a reduction in the weight percentage of Al in the BCC phase from 19.81 to 15.48 wt.%, whereas the data in Table 2 indicates that in the HEA-5 wt.% ZrO_2 specimen, this value increased to 20.77 wt.%. It is widely recognized that YSZ has a higher CTE than ZrO_2 and is more in line with the CTE of metals and alloys. As a result, the difference in thermal expansion coefficients between the matrix and reinforcement in the HEA-YSZ composite is less pronounced than that in the HEA-ZrO_2 composite. In the HEA-ZrO_2 composite specimen, it was observed that by adding zirconia reinforcement particles, the interface was created between the matrix and reinforcement, and the diffusion system was activated. As a result, active elements such as aluminum moved to the interface and an accumulation of Al-rich phase was created at the interface between HEA and ZrO_2 particles. However, in HEA-YSZ composite, due to the smaller difference in coefficient of expansion, the diffusion system is less activated and the rate of Al leaving the BCC phase and accumulating at the interface between HEA phase and YSZ particles will be less. According to these results, HEA-YSZ composite has more BCC phase and less Al-rich phase than HEA-ZrO_2 composite.

Figure 7 illustrates the FESEM micrographs of AlCoCrFeNi-10 wt.% YSZ after undergoing the SPS process. Similar to the earlier specimen, as indicated in Table 2, the predominant phase in the microstructure is the FCC phase, which is rich in Fe, Cr, and Co. Furthermore, the specimen also displayed the existence of BCC, Cr-rich, and Al-rich phases. The formation of the Al-rich phase is likely due to its higher diffusion coefficient. It is proposed that the relatively elevated diffusion coefficient of aluminum, in comparison to the other elements in the HEA system, may promote the development of this Al-rich phase (42). As it was revealed in previous study (35), in FCC phase of HEA-10 wt.% ZrO_2 specimen the amount of Co has decreased from 33.07 to 25.23 wt.%, in comparison with FCC phase of HEA-10 wt.% YSZ. In addition, in BCC phase related to HEA-10 wt.% YSZ, the amount of Ni and Al compared with HEA-10 wt.% ZrO_2 has decreased from 22.07 and 13.28 to 32.68 and 19.95 wt.%, respectively.

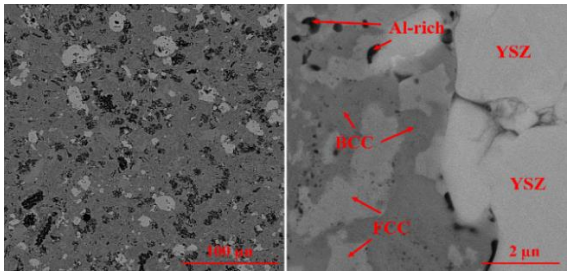
3. 4. Compressive Behavior To understand the influence of YSZ and monoclinic ZrO_2 particle addition on the compressive behavior of AlCoCrFeNi HEA matrix composites, compression tests were performed on the HEA-x wt.% YSZ and HEA-x wt.% ZrO_2 specimens. The related stress-strain curves are shown in Figures 8(a) and 8(b), respectively. The quantitative data obtained from these tests, including ε_p, σ_y, and σ_{max}, are

summarized in Table 3. The results of the compression tests indicate that the addition of ceramic particulates enhances the strength of the HEA matrix. This improvement is more pronounced in specimens with 5 wt.% reinforcement. According to Table 1, while the FCC phase fraction remains relatively unchanged, the addition of reinforcements leads to a reduction in the BCC phase content. There is a trade-off between ceramic reinforcement and BCC content. In the H5Y specimen, the yield strength increased to 1760 MPa. In contrast, the H10Y specimen exhibited a decrease in yield strength, attributed to a lower percentage of the BCC phase (36 wt.%). Similarly, the H5Z specimen, with 39% wt. BCC phase, achieved a yield strength of 1650 MPa, while the H10Z specimen, with a reduced BCC phase percentage (31 wt.%), exhibited a lower yield strength of 1550 MPa.

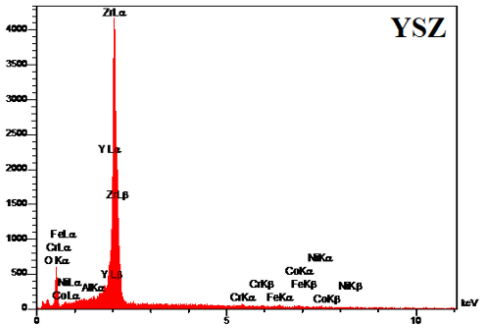
The characteristics of HEAs can differ significantly from those of their constituent elements. The structure type plays a crucial role in influencing the strength and hardness of HEAs. Generally, HEAs with a BCC structure demonstrate high yield strength but limited plasticity, whereas those with an FCC structure tend to have lower yield strength but greater ductility (5, 12, 13). Therefore, high yield strength (σ_y) is associated with the presence of the BCC structure, while ductility is attributed to the FCC structure (43).

Among the different kinds of engineering ceramics, zirconia and YSZ display notable mechanical characteristics like elevated hardness and elastic modulus, as well as significant wear resistance, which contributes to their effective use as reinforcing agents in various MMCs like NiTi (44), Cu (45), Ni-P (46), and Mg (47) composites. Also, inherently YSZ shows better physical and mechanical properties compared to monoclinic zirconia (45) which yields better mechanical properties of HEA-YSZ composites.

Zhou et al. (41) reported a yield strength of 1390 MPa for the AlCoCrFeNi alloy, which is lower than the yield strength (σ_y) reported in this study. In polycrystalline materials, the yield strength is influenced by multiple strengthening mechanisms, including grain boundary, precipitation, dislocation, thermal mismatch, and solid solution strengthening (47). In HEA matrix composites, the enhancement in strength is primarily due to the second phase strengthening resulting from the incorporation of monoclinic and YSZ particles, alongside thermal mismatch strengthening, which increases the density of dislocations and makes their movement more difficult, thereby reinforcing the matrix. Additional contributors to the improved compressive strength and yield strength may be linked to the characteristics of the employed synthesis method (MA+SPS) that can bolster material strength by refining the grain structure and introducing a significant number of defects and strains into the microstructure (48, 49).



(a)

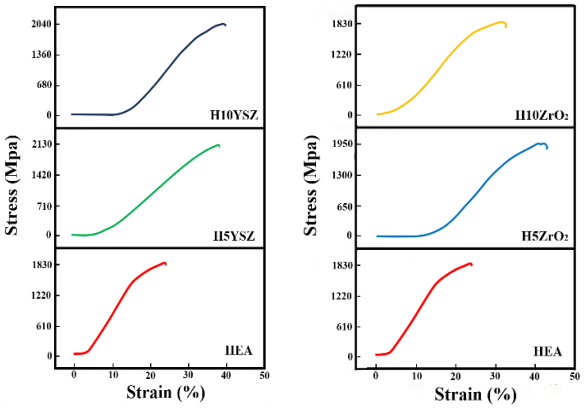


(b)

Figure 7. a) SEM image of the SPS-processed AlCoCrFeNi-10 wt. % YSZ, and, b) EDS analysis

TABLE 2. Results of EDS analysis taken from HEA, H5Y, and H10Y in different phases

	HEA		HEA-5YSZ			HEA-10YSZ		
	FCC	**BCC**	**FCC**	**BCC**	**YSZ**	**FCC**	**BCC**	**YSZ**
Al	4.57	19.81	3.85	16.13	---	4.58	20.12	---
Co	26.56	21.48	24.53	21.59	---	25.29	22.69	---
Cr	23.43	11.58	24.42	11.78	---	25.51	8.14	---
Fe	26.76	17.03	29.49	16.95	---	28.50	16.04	---
Ni	18.68	30.11	17.71	33.55	---	16.11	33.01	---
O	---	---	---	---	27.18	---	---	29.46
Zr	---	---	---	----	67.74	---	---	60.81
Y	---	---	---	---	5.09	---	---	9.73



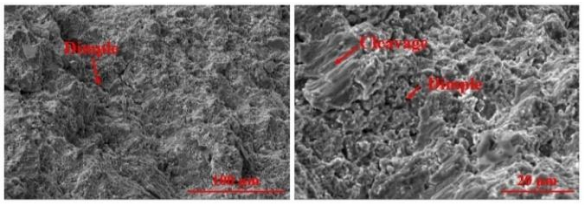
(a) (b)

Figure 8. Compressive engineering stress-strain curves of a) HEA-x wt.% YSZ and, b) ZrO_2 (X=0, 5, and 10)

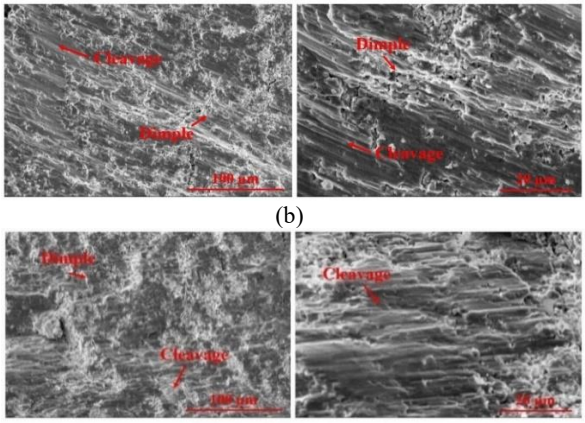
TABLE 3. Mechanical properties of AlCoCrFeNi + x wt.% YSZ and ZrO_2 (x =0, 5, and 10)

Specimens	ε_p (%)	σ_y (MPa)	σ_{max} (MPa)
HEA	10	1520	1830
HEA-5YSZ	9	1760	2130
HEA-5ZrO_2	8	1650	1830
HEA-10YSZ	8	1680	2040
HEA-10ZrO_2	7.5	1550	1950

The SEM fractography inspection of the fracture surfaces of the HEA specimen is presented in Figure 9(a). The fracture surface of the unreinforced HEA specimen exhibits dimple features, indicating that this sample undergoes both elastic and plastic deformation during the compression tests. Therefore, the fracture mechanism identified for this AlCoCrFeNi high entropy alloy is characterized by cleavage fracture and slip separation. The fracture surface of the HEA-5 wt.% YSZ specimens is illustrated in Figure 9(b). It is observed that the dimples on the fracture surface of this specimen have become shallower which correlates with a reduction in ductility for this sample. The fracture surface images of the HEA-10 wt.% YSZ specimen in Figure 9(c) show that the number of dimples has decreased dramatically, indicating a brittle fracture mode. Still, the features of ductile fracture exist.



(a)

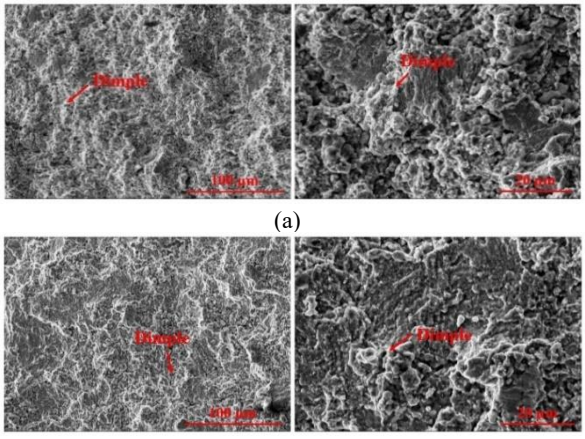


(b)

(c)

Figure 9. Fracture surfaces of a) HEA, b) HEA-5 wt.% YSZ, and c) HEA-10 wt.% YSZ specimens after compression test at room temperature

Figure 10 illustrates the fracture surface of H5Z and H10Z specimens. The figures show that HEA matrix specimens reinforced with ZrO_2 have several dimples on their fracture surface, indicating a localized ductile fracture. The observed cellular fracture surfaces in this figure are similar to those in Figure 9.



(a)

(b)

Figure 10. Fracture surfaces a) HEA-5 wt.% ZrO_2 and, b) HEA-10 wt.% ZrO_2

4. CONCLUSION

This research involved the preparation of AlCoCrFeNi + x wt.% YSZ and ZrO_2 (with x = 0, 5, and 10) specimens using mechanical alloying and SPS methods. Following the mechanical alloying of the elemental powders for 30 hours, a single-phase HEA powder with a BCC crystal structure was achieved. The sintering process of the HEA powder resulted in the formation of both BCC and FCC

phases, which can be attributed to the rapid heating rate associated with the SPS process and the limited time available for the complete transformation from BCC to FCC. By the increment of ceramic particles percentage, the phase fraction of FCC remained almost constant but the amount of BCC phase decreased. Due to better CTE compatibility, HEA-YSZ composites consist of a lower amount of Al-rich phase and more BCC phase compared to HEA-ZrO_2 composites. The results of compression tests showed that the addition of ceramic particulates enhanced the strength of the HEA matrix. This improvement is more pronounced in the specimens with 5 wt.% reinforcement. While σ_y in the unreinforced specimen is 1550 MPa, in the H5Y specimen, the yield strength increased to 1760 MPa. In contrast, the H10Y specimen exhibited a drop in yield strength attributed to a lower percentage of the BCC phase (36 wt.%) present in this sample. Similarly, the H5Z specimen, with 39% wt. BCC phase, achieved a yield strength of 1650 MPa, while the H10Z specimen had a reduced percentage of BCC phase (31 wt.%) and lower yield strength (1550 MPa). The observed behavior can be correlated to the phase structure of the specimens. The analysis of the fracture surfaces confirmed the coexistence of both brittle and ductile fracture modes after the compression tests.

Acknowledgements

The authors acknowledge Materials and Energy Research Center (MERC) for providing necessary facilities to conduct present research.

Funding

The authors declare that no financial support was received for this study.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

This article has no conflict with studies by other authors in this field.

Data Availability

All data generated or analyzed during this study are included in this published article.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT to improve the clarity of language and to correct grammar. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Yeh JW, Chen SK, Lin SJ, Gan JY, Chin TS, Shun TT, et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. Advanced engineering materials. 2004;6(5):299-303. https://doi.org/10.1002/adem.200300567
2. Cantor B, Chang IT, Knight P, Vincent A. Microstructural development in equiatomic multicomponent alloys. Materials Science and Engineering: A. 2004;375:213-8. https://doi.org/10.1016/j.msea.2003.10.257
3. Chen J, Zhou X, Wang W, Liu B, Lv Y, Yang W, et al. A review on fundamental of high entropy alloys with promising high–temperature properties. Journal of Alloys and Compounds. 2018;760:15-30. https://doi.org/10.1016/j.jallcom.2018.05.067
4. Ye Y, Wang Q, Lu J, Liu C, Yang Y. High-entropy alloy: challenges and prospects. Materials Today. 2016;19(6):349-62. https://doi.org/10.1016/j.mattod.2015.11.026
5. Wang F, Zhang Y, Chen G, Davies H. Cooling rate and size effect on the microstructure and mechanical properties of AlCoCrFeNi high entropy alloy. 2009. https://doi.org/10.1115/1.3120387
6. Zhang Y, Zuo TT, Tang Z, Gao MC, Dahmen KA, Liaw PK, et al. Microstructures and properties of high-entropy alloys. Progress in materials science. 2014;61:1-93. https://doi.org/10.1016/j.pmatsci.2013.10.001
7. Chikumba S, Rao VV, editors. High entropy alloys: development and applications. Proceedings of the 7th International Conference on Latest Trends in Engineering & Technology (ICLTET'2015); 2015.
8. Miracle DB, Senkov ON. A critical review of high entropy alloys and related concepts. Acta materialia. 2017;122:448-511. https://doi.org/10.1016/j.actamat.2016.08.081
9. Ren B, Liu Z, Li D, Shi L, Cai B, Wang M. Effect of elemental interaction on microstructure of CuCrFeNiMn high entropy alloy system. Journal of Alloys and Compounds. 2010;493(1-2):148-53. https://doi.org/10.1016/j.jallcom.2009.12.183
10. Ji W, Fu Z, Wang W, Wang H, Zhang J, Wang Y, et al. Mechanical alloying synthesis and spark plasma sintering consolidation of CoCrFeNiAl high-entropy alloy. Journal of Alloys and Compounds. 2014;589:61-6. https://doi.org/10.1016/j.jallcom.2013.11.146
11. Zhang A, Han J, Meng J, Su B, Li P. Rapid preparation of AlCoCrFeNi high entropy alloy by spark plasma sintering from elemental powder mixture. Materials Letters. 2016;181:82-5. https://doi.org/10.1016/j.matlet.2016.06.014
12. Mohanty S, Maity T, Mukhopadhyay S, Sarkar S, Gurao N, Bhowmick S, et al. Powder metallurgical processing of equiatomic AlCoCrFeNi high entropy alloy: Microstructure and mechanical properties. Materials Science and Engineering: A. 2017;679:299-313. https://doi.org/10.1016/j.msea.2016.09.062
13. Ma Y, Jiang B, Li C, Wang Q, Dong C, Liaw PK, et al. The BCC/B2 morphologies in Al x NiCoFeCr high-entropy alloys. Metals. 2017;7(2):57. https://doi.org/10.3390/met7020057

14. Wang W-R, Wang W-L, Yeh J-W. Phases, microstructure and mechanical properties of AlxCoCrFeNi high-entropy alloys at elevated temperatures. Journal of Alloys and Compounds. 2014;589:143-52. https://doi.org/10.1016/j.jallcom.2013.11.084
15. Fu Z, Chen W, Wen H, Chen Z, Lavernia EJ. Effects of Co and sintering method on microstructure and mechanical behavior of a high-entropy Al0. 6NiFeCrCo alloy prepared by powder metallurgy. Journal of Alloys and Compounds. 2015;646:175-82. https://doi.org/10.1016/j.jallcom.2015.04.238
16. Chen Z, Chen W, Wu B, Cao X, Liu L, Fu Z. Effects of Co and Ti on microstructure and mechanical behavior of Al0. 75FeNiCrCo high entropy alloy prepared by mechanical alloying and spark plasma sintering. Materials Science and Engineering: A. 2015;648:217-24. https://doi.org/10.1016/j.msea.2015.08.056
17. Manzoni A, Daoud H, Völkl R, Glatzel U, Wanderka N. Phase separation in equiatomic AlCoCrFeNi high-entropy alloy. Ultramicroscopy. 2013;132:212-5. https://doi.org/10.1016/j.ultramic.2012.12.015
18. Wang F, Zhang Y. Effect of Co addition on crystal structure and mechanical properties of Ti0. 5CrFeNiAlCo high entropy alloy. Materials Science and Engineering: A. 2008;496(1-2):214-6. https://doi.org/10.1016/j.msea.2008.05.020
19. Lin C-M, Tsai H-L. Evolution of microstructure, hardness, and corrosion properties of high-entropy Al0. 5CoCrFeNi alloy. Intermetallics. 2011;19(3):288-94. https://doi.org/10.1016/j.intermet.2010.10.008
20. Ghanbariha M, Ketabchi M, Farvizi M. Investigation of alumina reinforcement effects on microstructure, hardness, and tribological behavior of AlCoCrFeNi high entropy alloy. Intermetallics. 2025;185:108915. https://doi.org/10.1016/j.intermet.2025.108915
21. Uporov S, Bykov V, Pryanichnikov S, Shubin A, Uporova N. Effect of synthesis route on structure and properties of AlCoCrFeNi high-entropy alloy. Intermetallics. 2017;83:1-8. https://doi.org/10.1016/j.intermet.2016.12.003
22. Han Z, Liang S, Yang J, Wei R, Zhang C. A superior combination of strength-ductility in CoCrFeNiMn high-entropy alloy induced by asymmetric rolling and subsequent annealing treatment. Materials Characterization. 2018;145:619-26. https://doi.org/10.1016/j.matchar.2018.09.029
23. Shahmir H, Nili-Ahmadabadi M, Shafiee A, Langdon TG. Effect of a minor titanium addition on the superplastic properties of a CoCrFeNiMn high-entropy alloy processed by high-pressure torsion. Materials Science and Engineering: A. 2018;718:468-76. https://doi.org/10.1016/j.msea.2018.02.002
24. Eleti R, Raju V, Veerasham M, Reddy S, Bhattacharjee P. Influence of strain on the formation of cold-rolling and grain growth textures of an equiatomic HfZrTiTaNb refractory high entropy alloy. Materials Characterization. 2018;136:286-92. https://doi.org/10.1016/j.matchar.2017.12.034
25. Han Z, Luan H, Liu X, Chen N, Li X, Shao Y, et al. Microstructures and mechanical properties of TixNbMoTaW refractory high-entropy alloys. Materials Science and Engineering: A. 2018;712:380-5. https://doi.org/10.1016/j.msea.2017.12.004
26. Senkov O, Wilks G, Miracle D, Chuang C, Liaw P. Refractory high-entropy alloys. Intermetallics. 2010;18(9):1758-65. https://doi.org/10.1016/j.intermet.2010.05.014
27. Rahmani M, Petrudi AM. Experimental and numerical optimization study of shock wave damping in aluminum panel sandwich. Fracture and Structural Integrity. 2021;15(55):88-109. https://doi.org/10.3221/IGF-ESIS.55.07
28. Saindane U, Soni S, Menghani J. Dry sliding behavior of carbon-based brake pad materials. International Journal of Engineering Transactions B: Applications. 2021;34(11):2517-24. https://doi.org/10.5829/ije.2021.34.11b.14
29. Saindane U, Soni S, Menghani J. Friction and wear performance of brake Pad and optimization of manufacturing parameters using grey relational analysis. International Journal of Engineering, Transactions C: Aspects. 2022;35:552-9. https://doi.org/10.5829/ije.2022.35.03c.07
30. Saindane U, Soni S, Menghani J. Performance evaluation of brake pads developed using two different manufacturing methods. Materialwissenschaft und Werkstofftechnik. 2023;54(2):186-95. https://doi.org/10.1002/mawe.202100407
31. Jha P, Gupta P, Kumar D, Parkash O. Synthesis and characterization of Fe–ZrO2 metal matrix composites. Journal of Composite Materials. 2014;48(17):2107-15. https://doi.org/10.1177/0021998313494915
32. Ding J, Zhao N, Shi C, Du X, Li J. In situ formation of Cu–ZrO2 composites by chemical routes. Journal of alloys and compounds. 2006;425(1-2):390-4. ttps://doi.org/10.1016/j.jallcom.2006.01.058
33. Yu H, Hou Z, Guo X, Chen Y, Li J, Luo L, et al. Finite element analysis on flexural strength of Al2O3-ZrO2 composite ceramics with different proportions. Materials Science and Engineering: A. 2018;738:213-8. https://doi.org/10.1016/j.msea.2018.05.075
34. Hemanth J. Development and property evaluation of aluminum alloy reinforced with nano-ZrO2 metal matrix composites (NMMCs). Materials Science and Engineering: A. 2009;507(1-2):110-3. https://doi.org/10.1016/j.msea.2008.11.039
35. Grigor'eva T, Letsko A, Talako T, Tsybulya S, Vorsina I, Barinova A, et al. Producing Cu/ZrO2 composites by combining mechanical activation and self-propagating high-temperature synthesis. Combustion, Explosion, and Shock Waves. 2011;47(2):174-8. https://doi.org/10.1134/S0010508211020055
36. Chang C, Wang Y, Pei H, Lee C, Du X, Huang J. Microstructure and mechanical properties of nano-ZrO2 and nano-SiO2 particulate reinforced AZ31-Mg based composites fabricated by friction stir processing. Key Engineering Materials. 2007;351:114-9. https://doi.org/10.4028/www.scientific.net/KEM.351.114
37. Shackelford JF, Doremus RH. Ceramic and glass materials. JF Shackelford, RH Doremus. 2008:28-329.
38. Srivastava M, Srinivasan A, Grips VW. Influence of zirconia incorporation on the mechanical and chemical properties of Ni-Co alloys. American Journal of material science. 2011;1(2):113-22.
39. Ghanbariha M, Farvizi M, Ebadzadeh T. Microstructural development in nanostructured AlCoCrFeNi–ZrO2 high-entropy alloy composite prepared with mechanical alloying and spark plasma sintering methods. Materials Research Express. 2019;6(12):1265b5. https://doi.org/10.1088/2053-1591/ab5fc9
40. Ghanbariha M, Farvizi M, Ebadzadeh T, Samiyan AA. Effect of ZrO2 particles on the nanomechanical properties and wear behavior of AlCoCrFeNi–ZrO2 high entropy alloy composites. Wear. 2021;484:204032. https://doi.org/10.1016/j.wear.2021.204032
41. Zhou P, Xiao D, Yuan T. Microstructure, mechanical and corrosion properties of AlCoCrFeNi high-entropy alloy prepared by spark plasma sintering. 金属学报英文版. 2020;33(7):937-46. https://doi.org/10.1007/s40195-019-00962-8
42. Farvizi M, Ebadzadeh T, Vaezi M, Yoon E, Kim Y, Kim H, et al. Microstructural characterization of HIP consolidated NiTi–nano Al2O3 composites. Journal of Alloys and Compounds. 2014;606:21-6. https://doi.org/10.1016/j.jallcom.2014.03.184
43. George EP, Raabe D, Ritchie RO. High-entropy alloys. Nature reviews materials. 2019;4(8):515-34. https://doi.org/10.1038/s41578-019-0121-4
44. Farvizi M, Bahamirian M, Faraji A, Kim H. Influence of ceramic reinforcement characteristics on the microstructure and wear

behavior of NiTi matrix composites. Materialia. 2023;28:101726. https://doi.org/10.1016/j.mtla.2023.101726

45. Mirazimi J, Abachi P, Purazrang K. Spark plasma sintering of ultrafine YSZ reinforced Cu matrix functionally graded composite. Acta Metallurgica Sinica (English Letters). 2016;29(12):1169-76. https://doi.org/10.1007/s40195-016-0512-0
46. Baba NB, Davidson A, Muneer T. YSZ-reinforced Ni-P deposit: An effective condition for high particle incorporation and porosity level. Advanced Materials Research. 2011;214:412-7. https://doi.org/10.4028/www.scientific.net/AMR.214.412
47. Wang YS, Tan MJ, Chua BW, Bayraktar E. YSZ-Reinforced Mg-Based amorphous composites: processing, characterisation & corrosion. Advanced Materials Research. 2014;939:122-9. https://doi.org/10.4028/www.scientific.net/AMR.939.122
48. Yim D, Sathiyamoorthi P, Hong S-J, Kim HS. Fabrication and mechanical properties of TiC reinforced CoCrFeMnNi high-entropy alloy composite by water atomization and spark plasma sintering. Journal of Alloys and Compounds. 2019;781:389-96. https://doi.org/10.1016/j.jallcom.2018.12.119
49. Jiao W, Li T, Chang X, Lu Y, Yin G, Cao Z, et al. A novel Co-free Al0. 75CrFeNi eutectic high entropy alloy with superior mechanical properties. Journal of Alloys and Compounds. 2022;902:163814. https://doi.org/10.1016/j.jallcom.2022.163814

Persian Abstract

چکیده

در این مطالعه، کامپوزیت‌های آلیاژ با آنتروپی بالا (HEA) متشکل از زمینه AlCoCrFeNi و تقویت‌کننده‌های مبتنی بر زیرکونیا، با استفاده از روش‌های آلیاژی مکانیکی (MA) و تف جوشی پلاسمای جرقه‌ای (SPS) ایجاد شدند. تجزیه و تحلیل ریزساختاری، وجود هر دو فاز FCC و BCC را پس از ۳۰ ساعت آلیاژسازی مکانیکی و سپس SPS در دمای ۱۰۰۰ درجه سانتیگراد نشان داد. همچنین، پالایش ریتولد نشان داد که درصد وزنی فاز FCC در هر دو کامپوزیت HEA و HEA-YSZ تقریباً یکسان بوده و افزودن ذرات YSZ منجر به کاهش فاز BCC گردید و در عین حال رشد دانه فازهای HEA را در طول زمان تفجوشی محدود نمود. نتایج آزمایش فشار نشان داد که در نمونه حاوی ۵ درصد وزنی YSZ به دلیل توزیع مناسب و درصد بالاتر فاز BCC (٤٠٪)، مقادیر استحکام تسلیم به ۱۷٦۰ مگاپاسکال افزایش یافت، در حالی که در نمونه با ۱۰ درصد وزنی YSZ، استحکام تسلیم به دلیل درصد کمتر فاز BCC (۳٦٪) کاهش یافت. یافته‌های این تحقیق نشان می‌دهد که کامپوزیت‌های AlCoCrFeNi تقویت‌شده با YSZ نسبت به کامپوزیت هایی که با ذرات ZrO_2 تقویت شده‌اند، در ترکیب مشابه استحکام تسلیم بیشتری نشان می‌دهند که به نسبت بالاتر فاز BCC در کامپوزیت‌های تقویت‌شده با YSZ و خواص ذاتی بهتر این تقویت کننده نسبت داده می‌شود. بررسی مکانیسم‌های شکست، نشانگر ترکیب شکست نرم و ترد در نمونه‌ها است.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2371-2382

International Journal of Engineering

Journal Homepage: www.ije.ir

A Comprehensive Approach to Design of a Temperature Control System in Production Process of Aluminum Alloy Rolled Billets

Y. V. Bazhin, H. D. Tran*

Empress Catherine II Saint Petersburg Mining University, Saint Petersburg, Russia

PAPER INFO

Paper history:
Received 11 September 2025
Received in revised form 26 September 2025
Accepted 13 December 2025

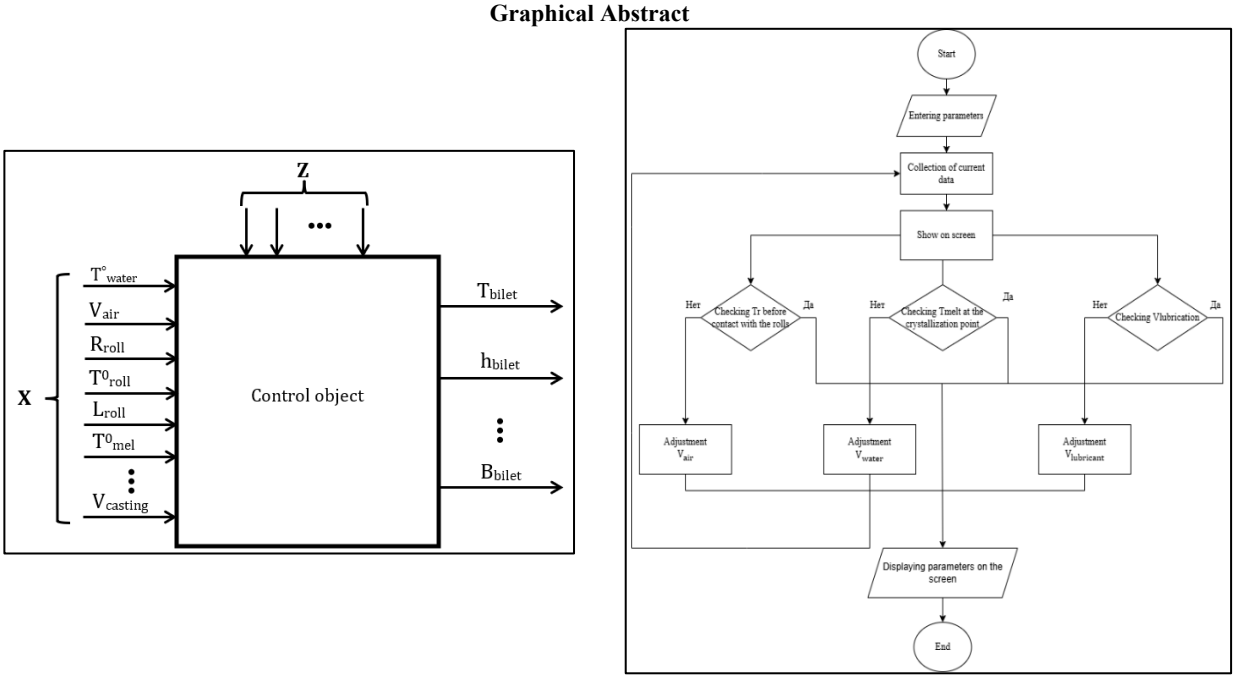
Keywords:
Aluminum Billets
Automatic Control
Continuous Casting and Rolling
Model Predictive Control
Thermal Regime

ABSTRACT

The present work discusses the creation and validation of an automatic control system designed to manage the thermal conditions during the continuous casting and rolling of aluminum billets. At the foundation of this study lies a mathematical model describing heat transfer phenomena in the billet, with particular emphasis on the influence of phase transitions and the release of latent heat during solidification. To ensure the accuracy of the model, experimental temperature data were collected from several points along the crystallizer rolls, providing a reliable basis for comparison. The simulations showed strong agreement with experimental observations, with deviations of less than five percent, which confirmed the robustness of the proposed formulation. Building on these findings, a model predictive control algorithm was incorporated into dedicated software, enabling real-time stabilization of casting and rolling conditions. The results demonstrated clear benefits, including enhanced product quality, improved process stability, and noticeable reductions in energy consumption.

doi: *10.5829/ije.2026.39.10a.02*

Graphical Abstract



*Corresponding Author Email: hieumd1998hp@gmail.com (H. D. Tran)

Please cite this article as: Bazhin YV, Tran HD, A Comprehensive Approach to Design of a Temperature Control System in Production Process of Aluminum Alloy Rolled Billets. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2371-82.

1. INTRODUCTION

In recent decades, considerable research efforts have been directed toward improving casting and rolling processes for aluminum and other non-ferrous metals. A particular focus has been placed on studies addressing thermal regimes and their impact on the quality of the final product. Numerous investigations have demonstrated that temperature parameters play a decisive role in both the development of microstructure and the formation of defects caused by thermal stresses (1, 2). Huang et al. (3) illustrated the evolution of aluminum alloy microstructures under different casting techniques, enabling a direct comparison between conventional and semi-continuous processes. In their review, Amirkhanlou and Ji (4) discussed the prospects of developing lightweight and high-strength aluminum alloys, but their analysis concentrated primarily on metallurgical aspects without addressing the issues of temperature field control.

In the studies conducted by Ilyushin and Boronko (5), as well as Zaitseva and Semykina (6), particular emphasis is placed on the necessity of introducing intelligent control systems into the aluminum industry. Nevertheless, although these contributions underline the significance of such systems, they remain largely conceptual in nature and do not provide a detailed examination of thermal management under strip casting and rolling conditions.

Over the past years, the theoretical basis for modeling thermal processes during metal solidification and deformation has advanced considerably. Brodova et al. (7) investigated phase–structural transformations under extreme conditions, while Churyumov and Pozdnyakov (8) focused on modeling the evolution of microstructure during hot plastic deformation. These studies established an important foundation for further analysis, yet they remain distant from addressing the practical challenges of real-time temperature control in industrial settings. In addition, Delouei et al. (9) and Avramenko et al. (10) proposed comprehensive mathematical models of unsteady heat transfer, which, however, have not yet been integrated into applied control systems.

A significant contribution to the advancement of continuous casting theory was made in the review by Thomas (11), which systematized existing models and numerical approaches for simulating heat transfer and deformation processes. Nevertheless, the author emphasized that most of these models were developed for the steelmaking industry and therefore require adaptation for aluminum alloys. In a related investigation, Wang et al. (12) explored the rolling behavior and interfacial phenomena in Cu–Al composite billets, highlighting the significance of understanding thermal distributions in multicomponent systems. Nevertheless, their study did not consider the application of predictive control techniques.

Recent developments have underscored the growing role of artificial intelligence and machine learning in enhancing temperature regulation processes. For example, Song et al. (13) introduced a multimodal deep neural network designed to control temperature during the continuous casting of steel, achieving high predictive accuracy and demonstrating the promise of intelligent computational approaches in this domain. Simultaneously, increasing attention has focused on the combined process of continuous casting followed by rolling, known as Continuous Casting Direct Rolling (CCDR). The effects of different heat treatment strategies on the properties of 6056 aluminum alloy were examined by Zhao et al. (14) during the direct integration of casting and cold rolling. Jiang et al. (15) further explored the microstructural characteristics and corrosion resistance of aluminum sheets produced via CCDR with ultrasonic melt treatment, highlighting the essential role of precise temperature management in attaining uniform structural properties. Collectively, these findings emphasize that effective thermal control is a key determinant of the material's ultimate performance.

An important direction in technological advancement is strip casting, which represents a refined form of continuous casting and rolling. However, the application of such methods to aluminum strip casting has not yet been realized. The production process based on strip casting technology offers significant advantages over conventional manufacturing techniques (16, 17). This approach can be regarded as an improved variation of the continuous casting method. A comparative overview of selected aspects of the two production methods is presented in Table 1.

Comparable conclusions are reported by Hu et al. (18), who provided a review of the application of soft computing techniques in rolling process control. They emphasized that, although significant progress has been achieved in the steel and non-ferrous sectors, research on aluminum remains notably limited. In a related study Sotnik (19) proposed an automated process control architecture for steelmaking, demonstrating the feasibility of integrating real-time models into closed-loop control systems. Nevertheless, the direct transfer of this approach to aluminum alloys is constrained by their distinct thermophysical properties and the specific characteristics of phase transformations.

Although significant scientific progress has been achieved, the development of specialized temperature control systems for strip casting of aluminum remains an unresolved challenge. Existing publications predominantly address either metallurgical aspects and microstructural analysis, mathematical models of heat transfer, or broader issues related to digitalization and

TABLE 1. Comparative technical parameters of continuous casting and rolling and conventional methods

Criteria	Continuous casting and rolling method	Conventional method
Freezing rate (°C/s)	1 - 10	50 - 100
Rolling speed (m/min)	0,05-6	30 - 60
Energy consumption (kWh/ton)	500 - 800	200 - 400
Rolling pressure (MPa)	20 - 100	5 - 50
Surface roughness (Ra, μm)	0,5 - 3,0	1,5 - 5,0

intelligent systems. However, studies that integrate mathematical modeling of thermal processes with predictive control algorithms specifically for aluminum strip casting have not yet been undertaken.

Thus, several critical gaps can be identified:

– a lack of studies addressing the application of model predictive control (MPC) in the aluminum industry, in contrast to the extensively investigated steelmaking processes;

– limited research on thermal regulation specifically within the "melt–roll" zone during continuous strip casting;

– insufficient integration of numerical heat-transfer models and modern control algorithms into unified real-time systems.

An urgent task, therefore, is the development of a mathematical model and a control algorithm capable of capturing the specific features of heat transfer in the melt–roll contact zone and providing reliable temperature predictions over a given time horizon. The present study addresses this gap by proposing a comprehensive framework that combines numerical simulation of heat transfer with phase transformations and latent heat of solidification, alongside the design and software implementation of a predictive control algorithm. The novelty of this work lies in the fact that model predictive control methods are applied for the first time to regulate the thermal regime under the conditions of continuous strip casting of aluminum billets, whereas this aspect has been virtually absent in the existing literature.

2. MATHEMATICAL MODELING OF THERMAL PROCESS

To construct the model, the primary step is to analyze the fundamental physical phenomena occurring during the process. This is an unsteady heat-transfer problem in a variable material medium, where such parameters as specific heat capacity and thermal conductivity depend on temperature and phase state (solid/liquid) (9, 20). Furthermore, the presence of latent heat during solidification introduces nonlinearity into the problem.

To simplify the analysis while maintaining adequate accuracy, the following assumptions are made (21, 22):

– predominantly one-dimensional heat transfer (normal to the contact surface): when the thickness of the cast strip or billet is small relative to its other dimensions and the principal cooling mechanisms act through the contacting rolls or an externally applied cooling spray, through-thickness temperature gradients dominate the thermal response. Under these conditions, one–dimensional formulations capture the principal transient conduction and phase-change behaviour while substantially reducing numerical complexity. This reduction is especially advantageous for parametric studies and real-time control applications where many forward model evaluations are required.

– macroscopic homogeneity with temperature-dependent bulk properties: at the length scales relevant to thermal modelling of casting and rolling (millimeters to centimeters), microstructural details such as grain morphology or particle dispersions can be represented through effective, temperature-dependent thermophysical properties (density, specific heat, thermal conductivity). This homogenization preserves the macroscopic energy balance and latent-heat effects while avoiding the intractable computational cost of resolving microstructural heterogeneity explicitly;

– negligible radiative heat exchange compared to conduction and forced convection: in typical roll-contact casting environments the dominant heat-sink terms are conduction into the rolls and convective cooling provided by water sprays or cooling jackets. Radiative losses become significant only at substantially higher surface temperatures or in the absence of convective cooling. Omitting radiation from the boundary conditions simplifies the model and introduces only minor error for the process conditions examined; radiation terms may be reinstated if future operating points warrant it;

– geometric evolution and interface motion treated via fixed-grid formulations (enthalpy or effective heat-capacity approaches) rather than continuous remeshing of a moving boundary: moving solid–liquid interfaces in metal solidification can be handled accurately and efficiently by enthalpy-based or effective-specific-heat formulations on a fixed computational mesh. These approaches account for latent heat release and interface progression without the overhead of dynamic mesh adaptation, making them well suited for transient simulations coupled with control algorithms.

The general heat conduction equation with a heat source term is expressed as follows (23, 24):

$$\rho , \left(\frac{Di}{d\tau}\right) = div(\lambda , grad(t)) + q_v \quad (1)$$

where:

ρ – metal density, kg/m³;

t – temperature, °C;

λ – thermal conductivity, W/(m·°C);
Di/dτ– total derivative of enthalpy with respect to time;
q_v – volumetric heat source, W/m³ (including latent heat of solidification). In the two-phase region (between the liquidus and solidus temperatures), average thermophysical properties are applied.

The thermal processes associated with solidification are described by following relations (25, 26):

$$\psi = \frac{V_{cr}}{V_0} \tag{2}$$

$$q_v = q_{cr}, \rho, C, \left(\frac{D\psi}{d\tau}\right) \tag{3}$$

$$C_{eff}(t) = \begin{cases} C_l & t > t_l \\ C(t) - q_{cr}, \left(\frac{D\psi}{d\tau}\right) & t_s < t < t_l \\ C_s = f(t) & t < t_s \end{cases} \tag{4}$$

where:
C_{eff} – effective heat capacity;
ψ – solid fraction ($0 \le \psi \le 1$);
V_s – volume of the solid phase;
V_{cr} – total volume of the crystallizing melt;
q_{cr} – latent heat of crystallization, J/kg;
$D\psi/d_\tau$ – solidification rate of the alloy;
t_s – solidus temperature, °C;
C_l – specific heat capacity of the liquid metal, J/(kg·°C);
C_s – specific heat capacity of the solid metal, J/(kg·°C.

Effective Properties in the Two-Phase Zone (7, 8):

$$C_{eff}(t) = \frac{(C_l + C_s)}{2} + \frac{q_{cr}}{t_l - t_s} \tag{5}$$

$$\lambda_{eff} = \lambda_l, \psi + \lambda_s, (1 - \psi) \tag{6}$$

$$\lambda_s = \varepsilon, \lambda_l(t) \tag{7}$$

$$\rho_{eff}(t) = \rho_l, \psi + \rho_s, (1 - \psi) \tag{8}$$

where:
λ_l , λ_s – thermal conductivity of liquid and solid metal, W/(m·°C);
ε – enhancement coefficient of liquid-phase thermal conductivity due to convective flow;
ρ_l , ρ_s – density of liquid and solid metal, kg/m³.

Generalized unsteady-state heat conduction equation:

$$\rho_{eff}(t), C_{eff}(t), \frac{dt}{d\tau} = div(\lambda_{eff}(t), grad(t)) \tag{9}$$

Heat transfer at the roll surface (27, 28):

$$-k\frac{dT}{dn} = h_w(T - T_w) \tag{10}$$

$$\frac{dT}{dn} = 0 \quad 0 \le \varphi \le 2\pi \tag{11}$$

$$-kr\frac{dT}{dr} = h_{cont}(T - T_x) + h_{ro} \tag{12}$$

$$(T - T_x)\ \frac{\pi}{2} \le \phi \le (\pi - \phi_1) \tag{13}$$

$$-k\frac{dT}{dn} = h_{cont}(T_K - T)\ \ (\phi_1 \le \phi \le \frac{\pi}{2}) \tag{14}$$

where:
hcont – heat transfer coefficient in the contact zone with the molten metal in the sump W/(m²·K);
Tw– roll surface temperature, °C;
n– normal vector to the contact surface;
hw– heat transfer coefficient from rolls, W/(m²·K);
r– roll diameter, m;
φ– roll inclination angle relative to the horizontal, degrees;
hro– Heat transfer coefficient during roll rotation.

Similarity criteria and heat transfer coefficients:

$$Nu = \frac{(f/8)RePr}{1+12\ 7(f/8)^{1/2}(Pr^{2/3}-1)\frac{h_w D_w}{k_w}} \tag{15}$$

$$f = (1\ 821\ log_{10}\ Re - 1\ 64)^{-2} \tag{16}$$

$$Re = \rho_w U D_h / \mu_w \tag{17}$$

$$\mathrm{Pr} = \mu_w C_w / k_w \tag{18}$$

where:
D_w – characteristic dimension, m;
k_w – thermal conductivity coefficient of the melt, W/(m·K);
Pr – Prandtl number;
f – hydraulic friction coefficient;
U – characteristic velocity, m/s;
D_h – hydraulic diameter, m;
C_w – specific heat capacity of the medium, J/(kg·K);
μ_w – dynamic viscosity of the medium, kg/(m·s);
ρ_w – density of the medium, kg/m³;
Re – Reynolds number.

3. EXPERIMENTAL METHODOLOGY

3. 1. Main Equipment used in the Experiment Included crystallizer rolls with internal cooling channels;

– melt feeding system (trough and nozzle);
– forced-air cooling unit;
– water-cooling system comprising a pump, flowmeter, pressure gauges, and thermometers at the inlet and outlet;
– chromel–alumel thermocouples (type K) with a diameter of 1.0–1.5 mm, together with mounting holders and protective tubes (asbestos or ceramic);
– contact thermocouples for roll surface temperature measurements;
– KSP-4 potentiometer;

– data acquisition system with thermocouple inputs and digital input/output channels;
– thermocouple amplifiers/adapters with a cold-junction block;
– mechanical fasteners for thermocouples, as well as thermally conductive adhesives/pastes;
– instruments for measuring sample geometry.

The measurement points and the locations of the thermocouples are presented in Figure 1.
T0–T4: Positions of the installed thermocouples.
Point D: Initial contact between the molten metal and the roll surface.
Point I: Solidification zone where the metal has fully transformed from liquid to solid.
D, I, B1, B2, B3: Temperature measurement locations on the roll surface.

3. 2. Measurement of Strip Temperature T0 (before the nozzle): The thermocouple was positioned along the axial line within the melt supply channel, 10–20 mm from the nozzle exit. To protect the sensor from direct flow, a ceramic sleeve with a diameter of 3–4 mm was used. The thermocouple was secured in a holder with a screw clamp, preventing displacement due to flow fluctuations.

T1 (center of the strip): Thermocouples for melt temperature measurement were enclosed in protective ceramic or asbestos sleeves and fixed in a holder above the casting cavity. The junctions were located 1–2 mm from the metal surface. At the start of casting, the sensor tip was immersed in the liquid metal; after a few seconds, solidification occurred, anchoring the sleeve and sensor within the billet. The external portion of the wires remained outside the metal, guided along the holder from the heated zone and connected to the measurement system.

T2 and T3 (across the strip width): These were installed symmetrically relative to the axis of the feeder nozzle, at a distance of 45–50 mm from the center. T4 (edge of the strip): Positioned 1–2 mm from the edge to capture temperature gradients near the boundary. Its design was similar to that of T1. All thermocouples (T0–T4) were installed before the melt was introduced.

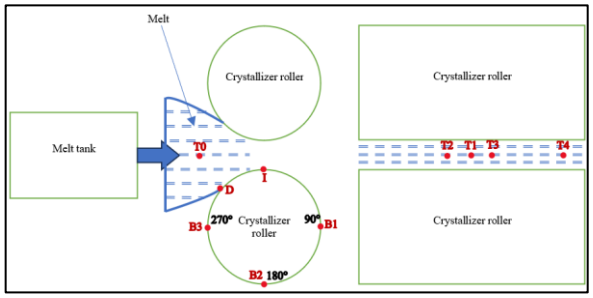


Figure 1. Measurement points and thermocouple locations

Subsequently, the electrical circuit and cold-junction compensation were verified. Measurement uncertainty was ±5 °C.

3. 3. Measurement of Roll-crystallizer Temperature A hole with a depth of 2 mm and a diameter of 3 mm was drilled into the lateral wall of the roll, located 2 mm from the roll surface. A contact sensor was inserted into this hole and fixed using heat-resistant adhesive and a mica spacer. Temperature values were recorded at points in contact with the melt, within the crystallization zone, and after rotating the roll by 90°, 180°, and 270° relative to the crystallization point.

4. ANALYSIS AND VERIFICATION OF THE THERMAL MODEL

Validation of the mathematical model of the thermal regime was performed using the finite difference method (29, 30). This approach relies on discretization of the computational domain, dividing space and time into grid nodes, and replacing derivatives in the governing equations with approximate finite-difference expressions. Instead of solving a complex system of differential equations, the problem is reduced to a set of algebraic equations that can be readily computed (31).

The finite-difference scheme is the central step in modeling transient heat transfer. In the one-dimensional, non-stationary problem considered here, the computational domain is divided into uniform spatial and temporal nodes. The derivatives of the governing differential equation are approximated by algebraic difference formulas for stepwise calculation.

In each grid point, the temperature at the next time step is computed from the current temperature at that point and the temperatures of two neighboring nodes. Explicit schemes were initially applied because of their simplicity and ease of programming. In such a scheme, the update formula for the temperature at node i at time $t + \Delta t$ is given as the current temperature plus the thermal diffusivity term, proportional to the difference between neighboring temperatures, along with the effect of any internal heat source (32, 33).

The thermal diffusivity coefficient α characterizes the rate of heat propagation in the material and is calculated as the ratio of thermal conductivity to the product of density and specific heat capacity. For computational stability—particularly in explicit schemes—the Courant–Friedrichs–Lewy condition must be satisfied, requiring the ratio of the time step to the square of the spatial step to be less than or equal to 0.5. Violation of this condition can lead to oscillations or non-convergence of the numerical solution.

When higher accuracy is required, or when a larger grid step is desirable to reduce computational cost, implicit schemes or the Crank–Nicolson method (a combination of explicit and implicit formulations) can be employed. These methods require solving a system of linear equations at each time step but offer the advantage of unconditional stability (34, 35).

In the numerical implementation, the heat transfer problem was addressed using the finite difference method, which involves constructing a grid of nodal temperature values over the domain of interest. The spatial domain was discretized with a step size of Δx, while the temporal evolution was discretized with a step size of Δt. An explicit scheme was employed to approximate the heat conduction equation:

$$T_i^{n+1} = T_i^n + \frac{\alpha \Delta t}{\Delta x^2}(T_i^n - 2T_i^n + T_{i-1}^n) + \frac{\Delta t, q_v}{\rho, C_{eff}} \quad (19)$$

where:

T_i^{n+1} – temperature at node i at time step n;

$\alpha = \frac{\lambda}{\rho, C_{eff}}$ – thermal conductivity coefficient.

The model was developed in Python, with NumPy handling array and matrix computations, SciPy used to solve the governing equation systems, Matplotlib employed for result visualization, and Tkinter providing the user-interface framework. In the modeling process, parameters such as melt temperature, casting speed, and cooling water flow rate were set according to experimental conditions, while geometric parameters such as roll length, roll radius, and roll gap were fixed. Based on these inputs, the model calculated the temperature distribution along the surface of the crystallizer rolls. Simulation results are presented in Figures 2 and 3.

The roll surface temperature was determined at the initial contact point with the melt (point D), where it reached the maximum range of 440–475 °C. At the crystallization point (point I), the temperature decreased

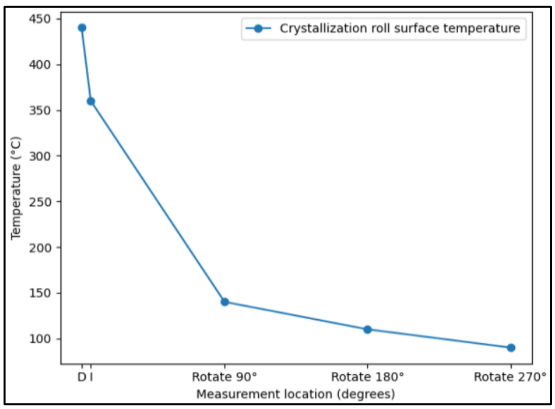


Figure 2. Surface temperature of the crystallizer rolls under lower-bound conditions

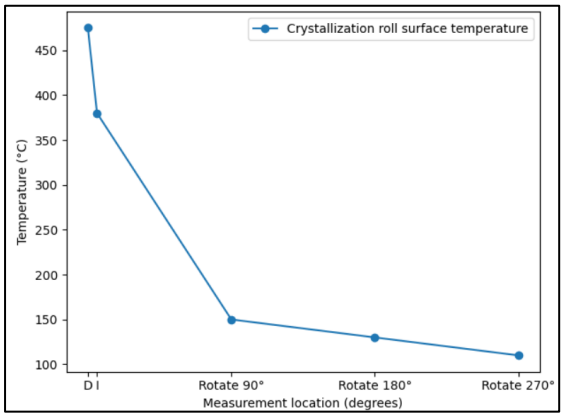


Figure 3. Surface temperature of the crystallizer rolls under upper-bound conditions

to 360–380 °C. As the rolls continued rotation through 90°, 180°, and 270°, the temperature gradually dropped to approximately 145 °C, 120 °C, and 100 °C, respectively. The temperature of the rolled billet was recorded within the range of 445–460 °C.

Table 2 summarizes the comparison of simulated temperatures with experimental values obtained using contact thermocouples at selected points on the roll surface.

The comparison clearly demonstrates that the applied mathematical model provides high accuracy, with a maximum deviation not exceeding 5%, which is acceptable for technical analysis. Such models can be effectively employed in the development of digital control systems as well as for predictive adjustment of process parameters.

5. DESIGN OF THE CONTROL ALGORITHM

The continuous casting and rolling process is characterized by significant transport delay, which complicates control tasks. According to control theory (36, 37), the stability and efficiency of a control system are largely determined by the ratio of delay time (τ) to the time constant of the process (T).

This ratio serves as a criterion for selecting the regulator type and its corresponding control law. When $\tau/T > 0.5$, conventional schemes such as proportional (P), integral (I), proportional–integral (PI), proportional–derivative (PD), or proportional–integral–derivative (PID) often fail to deliver the required system performance (38, 39). In such cases, it is necessary either to reduce the τ/T ratio or to implement more advanced control strategies that incorporate process models. A promising solution is predictive control, where the control action is generated not only on the basis of the current system state but also with regard to the predicted

TABLE 2. Comparison of modeling and experimental results

	Parameter	Experiment	Simulation	ΔT °C	ΔT %
Input data	Melt temperature, °C	680-690			
	Casting speed, mm/s	10-12			
	Cooling water flow rate, L/s	10-15			
	Initial cooling water temperature, °C	25			
	Ambient temperature, °C	25			
	Roll length, mm	1200		-	-
	Roll diameter, mm	800			
	Roll gap, mm	7,6			
	Melt feed rate, kg/s	5-6			
	Initial roll surface temperature, °C	70-90			
	Air cooling rate, m³/s	2			
Output data	Roll temperature at crystallization point, °C	370-400	360-380	15	3.89
	Roll temperature at initial melt contact, °C	450-500	440-475	17.5	3.68
	Roll temperature after 90° rotation, °C	130-175	140-150	7.5	4.92
	Roll temperature after 180° rotation, °C	100-150	110-130	5	4
	Roll temperature after 270° rotation, °C	80-120	90-110	5	5
	Rolled billet temperature, °C	450-480	445-460	12.5	2.69

future behavior of the process. This method offers a significant benefit by incorporating time delays and predicting the future behavior of system parameters, which leads to improved accuracy and stability during actual operation.

For predictive control of the continuous casting and rolling process, the objective function is defined to minimize both the deviation of the actual temperature from the setpoint and the variation of control actions:

$$\min \sum_{i=1}^{N} (T_{act}(t+i) - T_{set}(t+i))^2 + \lambda \sum_{i=1}^{N} \Delta U(t+i)^2 \quad (20)$$

where:

T_{act}— actual temperature;

T_{set}— setpoint temperature, required for compliance with casting specifications, defined by process requirements and entered by the operator via the command module;

$\Delta U(t+i)$— change in the control action (melt feed rate, water or air cooling rate, etc.);

λ— weighting factor balancing tracking accuracy against smoothness of control signals;

N — prediction horizon, typically set to 5–20 steps; here N = 10.

Strict adherence to technological conditions is essential to ensure billet quality. The continuous strip-casting stage consists of three primary phases: melt feeding into the crystallizer rolls, solidification, and lubrication of the cast billet. At the first phase, two key requirements are imposed:

- reduction of melt temperature from 760–800 °C to 680–690 °C;
- melt feed rate onto the rolls maintained at 5–6 kg/s.

In the subsequent phase, the temperature of the semi-solid billet must be decreased from 680–690 °C to 450–480 °C (from roll contact to the solidification point). Finally, continuous lubrication at a rate of 1–2 L/min must be ensured to prevent billet deformation after roll exit.

The control algorithm for the principal parameters of the continuous casting and rolling process is illustrated in Figure 4.

The flowchart uses the following designations:

1. Process start.
2. Input of initial parameters:
   - casting speed V_{cast};
   - water cooling rate V_{water_cool};
   - air cooling rate V_{air_cool};
   - lubricant feed rate $V_{lubricant}$.
3. Collection of real-time sensor data:
   - air and water temperature;
   - melt temperature before roll contact T_r;
   - melt temperature at the solidification point (billet temperature);
   - roll surface temperature;
   - air cooling rate V_{air_cool};
   - lubricant feed rate $V_{lubricant}$;

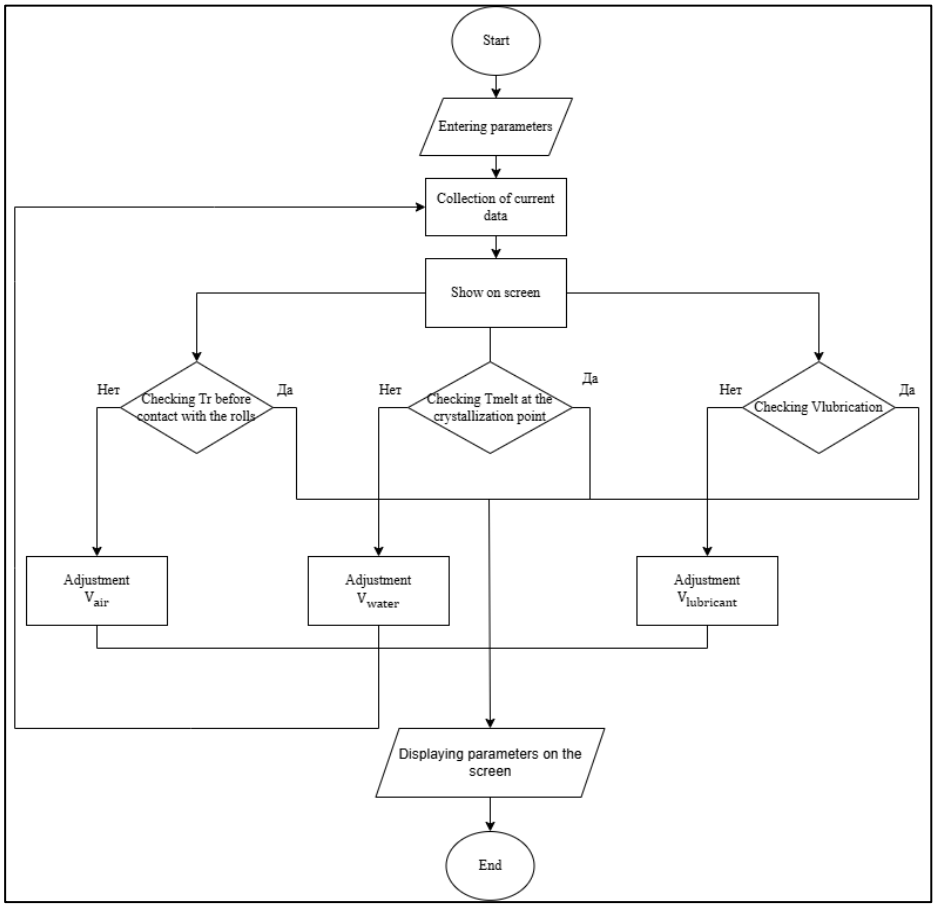


Figure 4. Algorithm for controlling key parameters in the continuous casting and rolling process

4. Real-Time Monitoring and Control of the Strip-Casting Process:
   - the system displays real-time diagrams of the melt temperature before contact with the rolls and at the crystallization point, as well as predicted temperature over the subsequent 10 seconds;
   - the melt temperature before roll contact is maintained within the allowable range of 680–690 °C;
   - at the crystallization point, the melt temperature is controlled within 450–480 °C.
5. Verification of lubricant feed rate.
6. Adjustment of air-cooling rate.
7. Adjustment of water-cooling rate.
8. Adjustment of lubricant feed rate.
9. Process completion and display of final parameters.

6. SOFTWARE IMPLEMENTATION

The continuous casting and rolling process can be represented as a structural control scheme of the managed object, is shown in Figure 5.

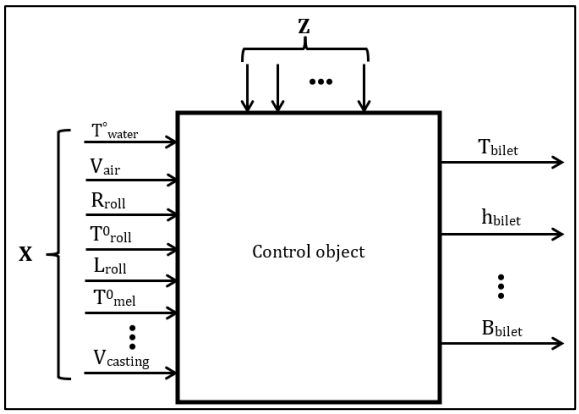


Figure 5. Structural scheme of the continuous casting and rolling process

The input controlled and manipulated parameters X include:
- initial cooling water temperature, T^{0}_{water};
- air-cooling rate in the mixer, V_{air};
- cooling water feed rate to the crystallizer rolls, V_{water};
- casting speed, $V_{casting}$;

− melt temperature before contact with the crystallizer rolls, T^0_{melt};

− initial surface temperature of the crystallizer rolls, T^0_{roll};

− length of the crystallizer rolls, L_{roll};

− melt feed rate, V_{melt};

− radius of the crystallizer rolls, R_{roll}.

Uncontrolled and unregulated input variables Z represent external influences, such as ambient temperature and humidity, voltage stability, and equipment operating consistency.

The outputs correspond to the achieved target parameters, including billet temperature T_{billet}, thickness h_{billet}, width B_{billet}, and surface quality. The structural-functional organization of the continuous casting and rolling process control system is presented in Figure 6.

Drawing on both experimental data and modeling outcomes, a software system was created to regulate the main parameters of the continuous casting and rolling process. Before starting the program, the operator is required to enter the initial process values. The input fields provide suggested default settings that fall within the allowable ranges. An example of the input interface is presented in Figure 7.

When the operator provides inputs that fall outside the recommended limits or are entered in the wrong format, the system issues a warning that calls for correction. Running the equipment with values beyond the allowable ranges can lead to safety risks such as fire, possible equipment malfunction, or unnecessary energy losses. The program will only start once all input parameters are entered correctly. Figure 8 illustrates the input interface in the case of an incorrect entry.

When all input values fall within the permitted range, the system switches to its operating mode. At this stage, the display shows live updates of the selected parameters. Examples of the monitoring interface for the key variables of the strip casting and rolling process, together with real-time graphical outputs, are provided in Figures 9, 10, and 11.

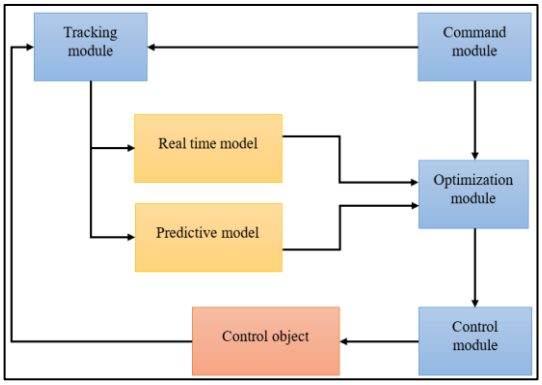


Figure 6. Structural and functional organization of the continuous casting and rolling process control system

Figure 7. Input interface for process parameters

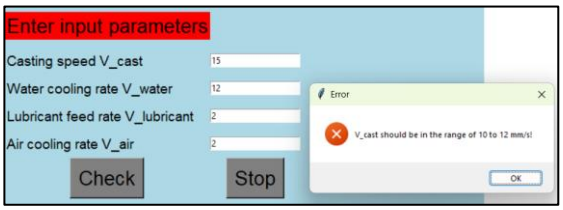


Figure 8. Input interface in the case of an incorrect entry

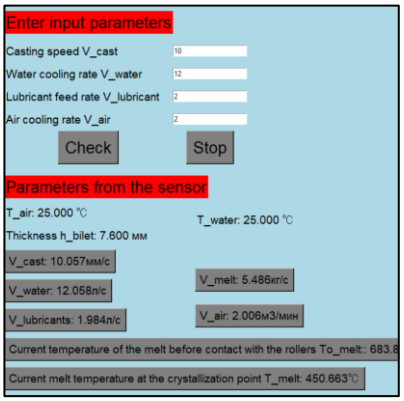


Figure 9. Operational interface of the continuous casting and rolling process control program

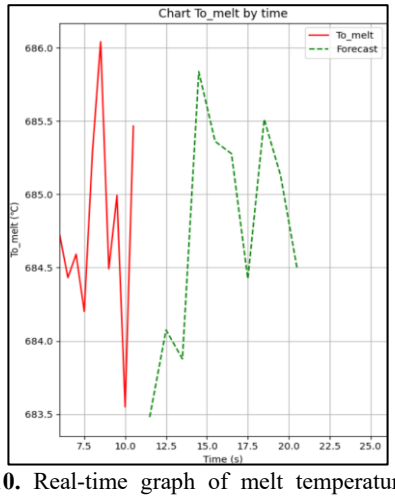


Figure 10. Real-time graph of melt temperature before contact with rolls

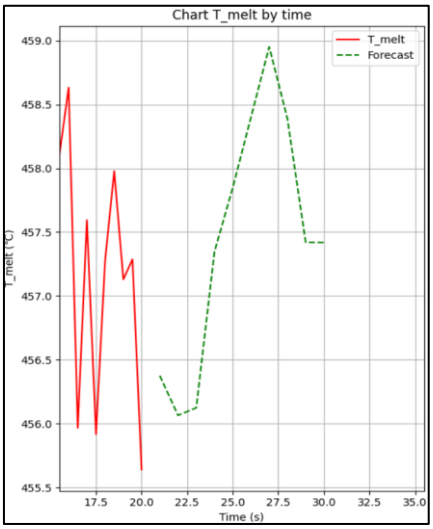


Figure 11. Real-time graph of melt temperature at the crystallization point

The reliability of the system was evaluated using several essential criteria. To begin with, the system blocks initiation whenever input values exceed the allowed range, while simultaneously issuing a warning to the operator. In cases where a component malfunctions, the system does not cease functioning entirely; rather, it switches to a safe operating mode and notifies the operator. Furthermore, even when substantial variations occur during the process, the accuracy of the displayed information remains unaffected, with data being refreshed promptly.

7. CONCLUSION

The conducted study demonstrates that the application of model-predictive control methods in the continuous casting and rolling of aluminum billets represents a promising approach for improving product quality and enhancing the energy efficiency of the technological process. The developed mathematical model and control algorithm confirm the feasibility of integrating numerical simulation of thermal processes with a real-time automatic regulation system. The practical value of these findings lies in the proposed method's ability to mitigate defects caused by thermal instabilities while offering greater adaptability to external disturbances. The developed software also provides a basis for advancing the digital transformation of production lines and incorporating elements of "smart manufacturing" within the aluminum industry.

Nevertheless, the study has certain limitations. The model is largely based on a one-dimensional heat transfer approximation and does not fully consider the effects of complex roll geometries or the heterogeneous microstructure of the alloy. Moreover, the predictive control algorithm is limited by its forecast horizon and would benefit from optimization to improve computational efficiency. Future research should focus on developing multidimensional models of heat and mass transfer, integrating machine learning methods for adaptive control parameter tuning, and creating comprehensive digital twins for production systems.

Acknowledgements

The authors sincerely thank Saint Petersburg Mining University for providing the necessary facilities and academic support to conduct this research. The author(s) also acknowledge the helpful discussions and guidance from colleagues in the related laboratories and departments.

Funding

This research received no external funding and was conducted with institutional support and resources.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

This article has no conflict with studies by other authors in this field.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT (generative AI) exclusively for minor language editing and readability improvement. After using this tool, the author(s) carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

1. Boczkal S, Augustyn B, Hrabia-Wiśnios J, Kapinos D, Lewis G, Bareel P-F, et al. Preparation of High-Quality 6xxx Aluminium Eco Alloys Cast in Billets. Engineering Proceedings. 2023;43(1):23. 10.3390/engproc2023043023

2. Greß T, Glück Nardi V, Mittler T, Schmid S, Buchberger P, Tonn B, et al. Interface formation and characterization of brass/aluminum compounds fabricated through die casting and semi-continuous casting. International Journal of Metalcasting. 2020;14(2):564-79. 10.1007/s40962-019-00387-0
3. Huang SH, Zhu BH, Zhang YA, Liu HW, Guo SL. Microstructure Evolution of AlSn20Cu Alloy Prepared by Conventional Casting and Semi Continuous Casting during Rolling and Annealing. Key Engineering Materials. 2021;904:31-8. 10.4028/www.scientific.net/KEM.904.31
4. Amirkhanlou S, Ji S. Casting lightweight stiff aluminum alloys: a review. Critical reviews in solid state and materials sciences. 2020;45(3):171-86. 10.1080/10408436.2018.1549975
5. Ilyushin YV, Boronko EA. Analysis of Energy Sustainability and Problems of Technological Process of Primary Aluminum Production. Energies. 2025;18(9):2194. 10.3390/en18092194
6. Zaitseva NM, Semykina IY. Intelligent control system for technological complexes of aluminum industry enterprises. Bulletin of the Tomsk Polytechnic University Geo Assets Engineering. 2024;335(11):119-32. 10.18799/24131830/2024/11/4613
7. Brodova IG, Zeldovich, V.I., Khomskaya, I.V. Phase-Structural Transformations and Properties of Non-Ferrous Metals and Alloys under Extreme Conditions. Physics of Metals and Metallography. 2020;121(7)::696–730. 10.31857/S0015323020070025
8. Churyumov AY, Pozdnyakov, A.V. . Modeling the Microstructure Evolution of Metallic Materials during Hot Plastic Deformation and Thermal Treatment. Physics of Metals and Metallography. 2020;121(11):1162-86. 10.31857/S0015323020110030
9. Delouei AA, Emamian A, Ghorbani S, He F. A general unsteady Fourier solution for orthotropic heat transfer in 2D functionally graded cylinders. Mathematical Methods in the Applied Sciences. 2024;47(6):3942-59. 10.1002/mma.9796
10. Avramenko A, Shevchuk I, Dmitrenko N, Kovetska YY, Tyrinov A. Unsteady theory of heat transfer and fluid flow during instantaneous transition to film boiling. International Journal of Thermal Sciences. 2020;153:106345. 10.1016/j.ijthermalsci.2020.106345
11. Thomas BG. Review on modeling and simulation of continuous casting. steel research international. 2018;89(1):1700312. 10.1002/srin.201700312
12. Wang J, Zhao F, Xie G, Hou Y, Wang R, Liu X. Rolling deformation behaviour and interface evaluation of Cu-Al bimetallic composite plates fabricated by horizontal continuous composite casting. Journal of Materials Processing Technology. 2021;298:117296. 10.1016/j.jmatprotec.2021.117296
13. Song GW, Tama BA, Park J, Hwang JY, Bang J, Park SJ, et al. Temperature control optimization in a steel-making continuous casting process using a multimodal deep learning approach. steel research international. 2019;90(12):1900321. 10.1002/srin.201900321
14. Zhao J-R, Hung F-Y, Chen B-J. Effects of heat treatment on a novel continuous casting direct rolling 6056 aluminum alloy: Cold rolling characteristics and tensile fracture properties. Journal of Materials Research and Technology. 2021;11:535-47. 10.1016/j.jmrt.2021.01.037
15. Jiang R, Zhao W, Zhang L, Li X, Guan S. Microstructure and corrosion resistance of commercial purity aluminum sheet manufactured by continuous casting direct rolling after ultrasonic melt pre-treatment. Journal of Materials Research and Technology. 2023;22:1522-32. 10.1016/j.jmrt.2022.12.025
16. Belsky A, Ngyen V, Sheikhi M, Starshaia V. Analysis of specifications of bifacial photovoltaic panels. Renewable and Sustainable Energy Reviews. 2025;224:116092. 10.1016/j.rser.2025.116092
17. Gorlanov ES, Leontev LI. Directions in the technological development of aluminium pots. Записки Горного института. 2024(266 (eng)):246-59.
18. Hu Z, Wei Z, Sun H, Yang J, Wei L. Optimization of Metal Rolling Control Using Soft Computing Approaches: A Review. Archives of Computational Methods in Engineering. 2021;28(2). 10.1007/s11831-019-09380-6
19. SV S. Development of automated control system for continuous casting. Radio Electronics, Computer Science, Control. 2024(2). 10.15588/1607-3274-2024-2-18
20. Matveeva VA, Chukaeva MA, Semenova AI. Iron ore tailings as a raw material for Fe-Al coagulant production. Записки Горного института. 2024(266 (eng)):433-43.
21. Vynnycky M. Applied mathematical modelling of continuous casting processes: A review. Metals. 2018;8(11):928. 10.3390/met8110928
22. Lu J, Liu L, Pan W, Wang W, Dou K. New model for heat transfer of copper roller in twin-roll strip casting to assist experimental and industrial process. International Journal of Thermal Sciences. 2025;213:109826. 10.1016/j.ijthermalsci.2025.109826
23. Huang W, Li J, Liu D. Real-time solution of unsteady inverse heat conduction problem based on parameter-adaptive pid with improved whale optimization algorithm. Energies. 2022;16(1):225. 10.3390/en16010225
24. Imane A, Hamid H, Jawad L, Khalid Z, Abdelaziz O. Numerical solution of unsteady conduction heat transfer in anisotropic cylinders. Journal of Thermal Science and Engineering Applications. 2016;8(3):031013. 10.1115/1.4033467
25. Proell SD, Wall WA, Meier C. On phase change and latent heat models in metal additive manufacturing process simulation. Advanced Modeling and Simulation in Engineering Sciences. 2020;7(1):24. 10.1186/s40323-020-00158-1
26. Dai Z, Wang Z, Zhu J, Chen X, Li Q, Jin Z. Three-dimensional solidification modeling of various materials using the lattice Boltzmann method with an explicit enthalpy equation. Physical Review E. 2024;110(2):025301. 10.1103/PhysRevE.110.025301
27. Lekhov OS. Investigation of the Stress-Strain State of the Roll-Strip System during Wide-Flange Beam Rolling in Universal Beam Mills Part 2. . Proceedings of Higher Education Institutions Ferrous Metallurgy. 2015;57(12):15–9. 10.17073/0368-0797-2014-1215–199
28. Yury I, Martirosyan A. The development of the soderberg electrolyzer electromagnetic field's state monitoring system. Scientific Reports. 2024;14(1):3501. 10.1038/s41598-024-52002-w
29. Barjoee SS, Gendler S. Sustainable illumination: Experimental and simulation analysis of illumination for workers wellbeing in the workplace. Heliyon. 2024;10(24). 10.1016/j.heliyon.2024.e40745
30. Nguyen MP, Ponomarenko T, Nguyen N. Energy transition in vietnam: a strategic analysis and forecast. Sustainability. 2024;16(5):1969. 10.3390/su16051969
31. Mazo A. Fundamentals of theory and methods of calculation of heat transfer [Osnovy teorii i metody rascheta teploperedachi]: textbook. Kazan Kazan University. 2013.
32. Fedorova E, Morgunov V, Pupysheva E. Effect of variation of internal diameter along the length of a rotary kiln on material movement. Non-ferrous Metals. 2024;56(1):28-34. 10.17580/nfm.2024.01.05
33. Simakov AS, Mas'ko, O.N., Nikolaev, M.Yu. . Machine Vision System for Quality Assessment of Quartz Raw Materials. Non-Ferrous Metals. 2025;3:81-8. 10.17580/tsm.2025.03.12

34. Skamyin A, Shklyarskiy Y, Lobko K, Dobush V, Sutikno T, Jopri MH. Impedance analysis of squirrel-cage induction motor at high harmonics condition. Indonesian Journal of Electrical Engineering and Computer Science. 2024;33(1):31-41. 10.11591/ijeecs.v33.i1

35. Shklyarsky YE, Lobko, K.K., Kuznetsova, Yu.N., Vorobyov, M.S. Study of Composite Motor Load in the Presence of Higher Harmonics in Electrical Networks. Energy Proceedings of Higher Education Institutions and Energy Associations of the CIS. 2024;67(4):285-99. 10.21122/1029-7448-2024-67-4-285-299

36. Li G, Wang X, Zhu W. Theoretical and experimental investigation on an integral time-delay feedback control combined with a closed-loop control for an infinitely variable transmission system. Mechanism and Machine Theory. 2021;164:104410. 10.1016/j.mechmachtheory.2021.104410

37. Ai B, Sentis L, Paine N, Han S, Mok A, Fok C-L. Stability and performance analysis of time-delayed actuator control systems. Journal of Dynamic Systems, Measurement, and Control. 2016;138(5):051005. 10.1115/1.4032461

38. George T, Ganesan V. Optimal tuning of PID controller in time delay system: A review on various optimization techniques. Chemical Product and Process Modeling. 2022;17(1):1-28.

39. Zhuo-Yun N, Yi-Min Z, Qing-Guo W, Rui-Juan L, Lei-Jun X. Fractional-order PID controller design for time-delay systems based on modified Bode's ideal transfer function. IEEE Access. 2020;8:103500-10. 10.1109/ACCESS.2020.2996265

Persian Abstract

چکیده

کار حاضر به بحث در مورد ایجاد و اعتبارسنجی یک سیستم کنترل خودکار طراحی شده برای مدیریت شرایط حرارتی در طول ریخته‌گری و نورد مداوم شمش‌های آلومینیومی می‌پردازد. اساس این مطالعه یک مدل ریاضی است که پدیده‌های انتقال حرارت در شمش را توصیف می‌کند، با تأکید ویژه بر تأثیر گذارهای فازی و آزادسازی گرمای نهان در طول انجماد. برای اطمینان از دقت مدل، داده‌های دمای تجربی از چندین نقطه در امتداد غلتک‌های کریستالیزر جمع‌آوری شد و مبنای قابل اعتمادی برای مقایسه فراهم کرد. شبیه‌سازی‌ها با مشاهدات تجربی، با انحراف کمتر از پنج درصد، تطابق قوی نشان دادند که استحکام فرمول پیشنهادی را تأیید می‌کند. بر اساس این یافته‌ها، یک الگوریتم کنترل پیش‌بینی‌کننده مدل در نرم‌افزار اختصاصی گنجانده شد که امکان تثبیت شرایط ریخته‌گری و نورد را در زمان واقعی فراهم می‌کند. نتایج، مزایای آشکاری از جمله افزایش کیفیت محصول، بهبود پایداری فرآیند و کاهش قابل توجه مصرف انرژی را نشان داد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2383-2394

International Journal of Engineering

Journal Homepage: www.ije.ir

Simulation and Analysis of Voltage Distribution and Leakage Current in Porcelain and Glass Insulator Strings Considering of Pollution, String Deviation, and Defective Units Effects

S. J. Ghasemnezhad, M. Mirzaie*, A. Gholami

Faculty of Electrical & Computer Engineering, Babol Noshirvani University of Technology, Babol, Iran

P A P E R I N F O

Paper history:
Received 18 August 2025
Received in revised form 08 October 2025
Accepted 26 October 2025

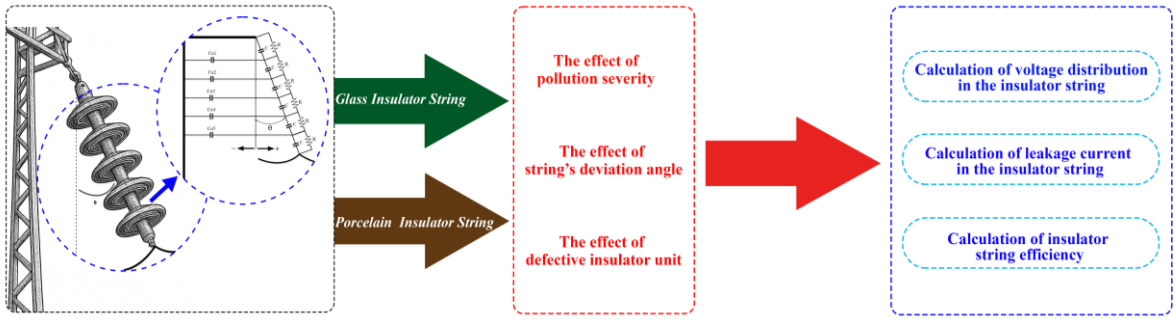
Keywords:
Voltage Distribution
Porcelain and Glass Insulator
Pollution Levels
Insulator String Deviation
Defective Unit

A B S T R A C T

Insulator strings are important and critical components of overhead lines, providing insulation properties and preventing leakage currents. This paper assesses the influence of several factors on the voltage distribution and leakage currents of insulator strings. Non-uniform voltage distribution among the units of an insulator can lead to degradation of the string and reduced reliability of the overhead lines and power network. In this study, a detailed analysis of five-unit porcelain and glass insulator strings under various conditions is conducted using circuit modeling and simulation in the MATLAB environment. The investigated scenarios include different levels of surface pollution, deviations of the insulator string toward or away from the tower, and the presence of a defective unit within the string. The results indicate that increasing pollution levels improves voltage distribution uniformity among the units, thereby enhancing the overall efficiency of the string; however, it also results in an increase in leakage current. Moreover, when the string is deflected outward (away from the tower) compared to inward deviation, a more uniform voltage distribution is achieved along with a notable reduction in leakage current. In the scenario involving a defective unit, the voltage stress on that unit decreases, and its voltage share is redistributed among the remaining healthy units depending on their relative positions. Simulations show that glass insulator strings exhibit a more uniform voltage distribution and higher efficiency compared to porcelain insulator strings under clean conditions and also under low to medium pollution levels.

doi: *10.5829/ije.2026.39.10a.03*

Graphical Abstract



NOMENCLATURE

C	Capacitance of each unit in the porcelain insulator string	DBD	Dry Band Discharges
C_e	Capacitance between the pin of each insulator unit and the tower structure	ESDD	Equivalent Salt Deposit Density
R	Resistance parallel to the capacitance of each insulator unit	NSDD	Non-Soluble Deposit Density

*Corresponding Author Email: *mirzaie@nit.ac.ir* (M. Mirzaie)

Please cite this article as: Ghasemnezhad SJ, Mirzaie M, Gholami A. Simulation and Analysis of Voltage Distribution and Leakage Current in Porcelain and Glass Insulator Strings Considering of Pollution, String Deviation, and Defective Units Effects. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2383-94.

1. INTRODUCTION

With the continuous growth of the population and increasing demand for electrical energy, the importance of reliable and sustainable transmission and distribution of energy is more critical than ever. One of the essential components in transmission systems is the insulator. Insulators play a crucial role in providing electrical insulation between conductors and the metallic structures of transmission towers in overhead lines (1). Insulators are critical components of transmission lines and are highly sensitive. Their proper and stable operation is essential for ensuring the reliability, safety, and efficiency of the power network. Any malfunction or degradation in their performance can result in serious damage, outages, reduced service life, and adverse effects on power grids. They are available in various types depending on the voltage level, environmental conditions, and application. One common type is the insulator string, which is widely used in extra-high voltage distribution and transmission systems. These insulators are installed in suspension or tension configurations on high-voltage towers and are typically made of porcelain or glass (2, 3). Due to their favorable electromechanical properties and high resistance to pollution, porcelain insulators are widely used in power network design and high-voltage transmission lines (4). The number of units in an insulator string depends on factors such as the system's rated voltage, mechanical strength requirements, installation altitude above sea level, severity of over-voltages (lightning and switching), and environmental pollution levels. One of the major challenges in operating this equipment is the phenomenon of non-uniform voltage distribution among insulator strings (5-7). This phenomenon causes excessive electrical stress on certain units, which accelerates their deterioration and reduces the service life of the entire string. Voltage distribution in an insulator string depends on multiple factors, including environmental conditions (such as wind speed and direction, temperature, humidity, and pollution level), the capacitance values of the insulators themselves and between the insulator units and the ground, the capacitance of metallic components adjacent to the tower structure, and the presence or absence of devices such as corona rings (8, 9). Instabilities caused by environmental variations can lead to significant fluctuations in voltage sharing. Therefore, to prevent potential damage from these fluctuations, the precise design and implementation of preventive measures in insulation systems are of paramount importance (10).

Non-uniform voltage distribution increases the intensity of corona and electrical discharges around the insulators (11). Additionally, environmental pollution on some units can intensify surface discharges and leakage currents through the insulators, as well as increase surface temperature (12-15). To accurately examine the voltage and electric field distribution in the insulator string, common numerical methods such as the Finite Element Method (FEM) and other simulation techniques can be employed (16). Due to the inherent non-uniformity of voltage and electric field along insulator strings, analyzing these parameters under various environmental conditions and considering the presence or absence of components such as towers, conductors, and corona rings is crucial. The analysis of the electrical behavior of insulator strings has mainly focused on porcelain insulators because these types exhibit high resistance to pollution. They are also widely used in high-voltage transmission lines due to their suitable electromechanical properties and are considered primary options in power network design (17). In some cases, applying excessive voltage and mechanical force to a unit within the insulator string can cause breakage(18, 19). However, the remaining units may still retain some insulating properties. A broken unit can alter the voltage distribution of the insulator string, as the voltage previously assigned to that unit under intact conditions is redistributed among the other units. This redistribution can increase electrical stress on the remaining units and may lead to their gradual failure (20-23). Voltage distribution along an insulator is inherently non-uniform due to the mutual interactions among the tower, line conductors, and metallic fittings. This non-uniformity can increase electrical stress on specific insulator units, thereby reducing their service life (24). To mitigate this uneven distribution, a corona ring is typically installed at the end of the insulator string (on the energized conductor side). The corona ring modifies the electric field intensity around specific units and improves voltage distribution. For more accurate analysis, a circuit model based on capacitive coupling that accounts for the presence of the corona ring is employed (25, 26).

On the other hand, electrical power in the distribution and transmission networks may be wasted due to losses caused by the malfunctioning of defective insulators. The studies have shown that at least 25% of energy losses in the network can result from the use of defective insulators (27). Another factor affecting the performance of the insulator string is its deviation or angle deviation relative to the hanging and vertical position. The deviation of the insulator string can be caused by asymmetry of the applied forces, improper design of the tower or insulator, unbalanced weight of the connected equipment, environmental conditions such as wind, improper installation or wear of the structure, and the angle of the tower. Strong winds and changes in wind direction can lead to the deviation of the insulator string from its standard position and its transfer to the outside of the tower (28, 29). Deviation of an insulator string can reduce the electrical insulation distance and increase the probability of electric arc. Moreover, this deviation alters

the voltage distribution pattern along the insulator strings and can negatively affect the efficiency and dielectric strength of the insulator strings (30). In addition, since the surface of insulators is in direct contact with the surrounding environment, there is a risk of surface electrical discharge. This phenomenon becomes more pronounced when the insulator surface is exposed to factors such as dust and pollution combined with moisture, so that it provides a conductive surface (31, 32). Under these conditions, leakage current flows along the surface of the insulator, reducing its dielectric strength and ultimately degrading its operational performance (33).

To accurately evaluate and analyze voltage distribution in an insulator string, an equivalent circuit model can be used. This model incorporates electrical parameters such as the self-capacitance of each insulator unit, the capacitance between each unit's pin and the energized conductor, the capacitance between the pin and the metallic tower structure, the mutual capacitance between adjacent units, and other electrostatic coupling effects (34).

In heavily polluted areas, polymer insulators face challenges such as loss of hydrophobicity and increased leakage current. Laboratory studies on polymer insulators have shown that surface contamination of different ages, combined with aging effects, leads to a significant increase in leakage current and dry band discharge (DBD) intensity. Finite element modeling has confirmed these findings, demonstrating that dynamic variations in the conductivity of the contamination layer, influenced by the electric field, play a key role in this process (35). Experimental results also indicate that the simultaneous deposition of soluble and insoluble salts on the insulator surface, especially under wet conditions, increases the conductivity of the contamination layer and the leakage current, reducing the flashover voltage. Variations in parameters such as equivalent salt deposit density (ESDD), non-soluble deposit density (NSDD), and their ratio, along with the pH of the contaminated layer, significantly influence these effects (36).

In this paper, the voltage distribution and also the leakage current in 63 kV porcelain and glass insulator strings are calculated and analyzed. The analyses are conducted considering various factors such as surface resistance amount on the insulators, different deviation angles from the vertical position (toward and way from tower), and the presence of a defective unit within the insulator string. These evaluations are carried out using circuit modeling and simulation in MATLAB software. The study investigates variations in voltage distribution, leakage current through the insulator strings, and the overall efficiency of the string. Furthermore, the impacts of different scenarios, including surface pollutions, deviation angle of the insulator string, and the existence of a defective unit on the performance of the insulator string, are assessed. This study can serve as a foundation for more accurate analysis of insulator string behavior under various environmental conditions and in the presence of several issues. The results of this research can be contributed to improve the design and operation of power transmission lines and help to reduce energy losses and extend the service life of insulators.

2. INSULATOR STRING MODELING

In this study, 63 kV porcelain and also glass insulator strings were examined for the intended objectives. The insulator string was analyzed under the influence of factors such as pollution resistance, insulator deviation, and defective units. Figure 1 illustrates a 5-unit insulator string with a deviation angle. Variations in voltage distribution, leakage current, and the overall efficiency of the insulator string were determined and extracted using the simulation of a circuit-based model in MATLAB. In this model, each unit of the insulator string is represented as a capacitor (C). Surface pollution and also condition of a defective unit are modeled using a resistance parallel to the capacitance of each insulator unit (R). The capacitance of each unit in the insulator string (from C_1 to C_5) is considered to be 30 pF. The capacitance between the pin of each insulator unit and the tower structure (C_e) is 2 pF under vertical position, but it can vary with the deviation angle of the string (37, 38). In the modeling, the capacitive effects between the pins of each insulator unit and the conductor, as well as between the pins of adjacent units, are neglected. In this study, the voltage of the first insulator, which is the closest to the tower cross arm, is defined as V_1, and the insulator numbering increases with distance from the cross arm toward the energized conductor.

Figure 2 shows the electrical model of a typical insulator string. This electrical model analyzes the voltage distribution and leakage current by considering pollution levels, deviation angles, and the presence of defective units at different positions along the porcelain and glass insulator strings.

In this research, four scenarios are defined for the simulation of the 5-unit porcelain and glass insulator strings under conditions of pollution, deviation angle, and the presence of a defective unit. Then, voltage sharing, efficiency, and leakage current are calculated.

3. SIMULATION AND RESULTS

In this section, the results of the carried out simulations for the 5-unit porcelain and glass insulator strings, in transmission line 63kV, considering various conditions such as pollution resistance, deviation angle, and defective units are presented and analyzed in terms of

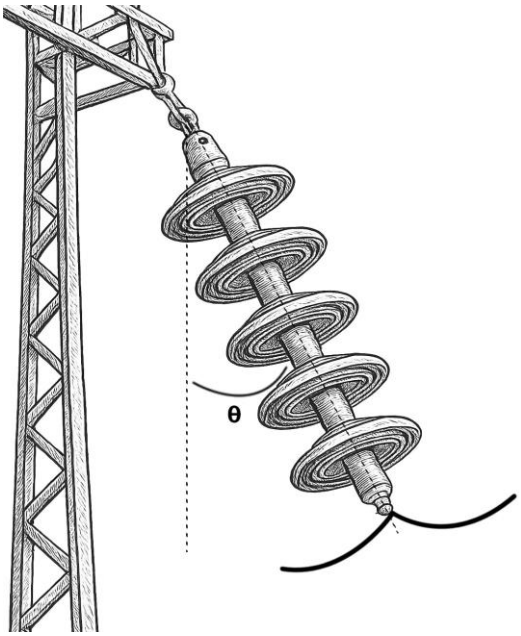


Figure 1. Schematic of the insulator string with deviation angle (θ)

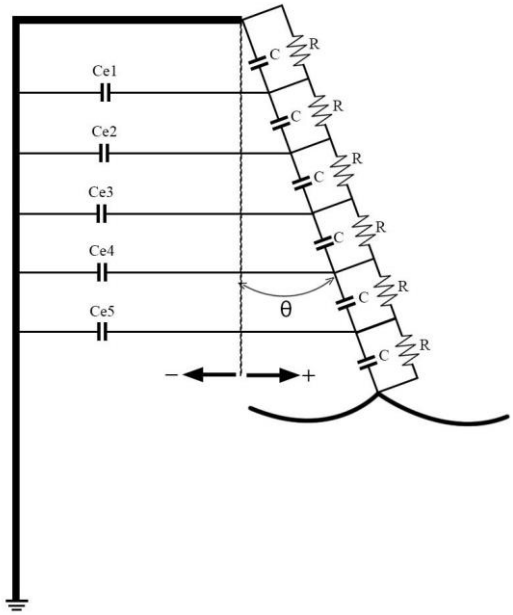


Figure 2. Equivalent circuit of the insulator string considering pollution levels and deviation angle

voltage distribution, leakage current, and overall string efficiency.

In 63 kV overhead lines, the most common configuration consists of insulator strings with five units. Therefore, the main focus of this study was on five-unit porcelain and glass insulator strings under various conditions. Nevertheless, to evaluate the effect of increasing the number of insulator units, additional simulations were performed for strings with six and seven units. For example, under clean conditions, the efficiency of the five-unit insulator string is 74.1% for porcelain and 84.3% for glass. For the six-unit string, the efficiency is 66.25% for porcelain and 78.3% for glass. Similarly, for the seven-unit string, the efficiency values are 59.16% and 72.32% for porcelain and glass insulators, respectively. In addition, the leakage current for the porcelain insulator was 117.03 μA for the five-unit string, 109.1 μA for the six-unit string, and 105.5 μA for the seven-unit string. For the glass insulator, the corresponding leakage currents were 188.23 μA, 169 μA, and 158.46 μA for the five-, six-, and seven-unit strings, respectively. These results indicate that increasing the number of insulator units from five to seven results in a less uniform voltage distribution along the string and a slight reduction in overall efficiency; however, it also causes a decrease in leakage current, which can enhance the system's overall reliability.

Considering that including all these extended scenarios would substantially lengthen the manuscript and divert focus from the main configuration used in 63 kV systems, only the key outcomes related to the five-unit strings are presented here.

3. 1. The Effect of Pollution Severity (Scenario 1)

Insulator surface pollution is one of the most critical environmental factors, as it can increase leakage current and lead to partial discharges or even flashovers. This factor can alter the voltage distribution along the insulator string and may result in progressive or sudden equipment failure (39, 40). Therefore, in this scenario, the impact of different pollution resistances, from clean to heavy, on the voltage sharing of insulators is determined and calculated. Additionally, the material type of the insulators can also affect their electrical performance. Porcelain and glass insulators are commonly used in transmission lines, each with unique physical and electrical characteristics. Investigating and comparing their performance in voltage distribution and arc behavior under different pollution conditions is important (41, 42).

In this study, the effects of different pollution levels on voltage distribution, efficiency, and leakage current of the insulator strings (porcelain and glass) under different conditions are investigated. In this scenario, according to data in Table 1. different values of surface resistance (R) are assumed and considered (43). These values can

TABLE 1. Range of resistance values for polluted insulator (44)

Pollution Level	Pollution Resistance ® (MΩ)
Clean	$R \geq 100000$
Light and Medium	$100 \leq R \leq 100000$
Heavy	$R \leq 100$

significant effects on the system performance and voltage distribution in the insulator strings. In this research, the surface resistance of the insulator under clean, light pollution, medium pollution, and heavy pollution are assumed to be 100000, 1000, 10, and 0.1 MΩ, respectively.

Based on the simulation results and as shown in Figures 3 and 4, it is observed that when surface pollution increases (or pollution resistance decreases), the voltage variation among the insulator units in the string is negligible and remains nearly constant. However, as pollution resistance decreases, the voltage distribution across the insulator string gradually becomes more uniform. For instance, at a pollution resistance of 0.1 MΩ, the voltage across each unit reaches approximately 7.27 kV in both porcelain and glass insulator strings, indicating a completely uniform voltage distribution. Consequently, although surface pollution may lead to a more uniform voltage distribution, this should not be considered an advantage.

Figure 5 illustrates the variation of leakage current in a 5-unit porcelain and glass insulator string under different surface resistances. For example, in a porcelain insulator, the leakage current increases from 117.03 µA under clean condition (no pollution) to 7270 µA under heavy pollution condition. As the resistance decreases, which corresponds to more severe pollution levels on the insulator surface, the leakage current increases significantly. This trend is observed in both porcelain and glass strings. They are slightly different from each other due to the distinct dielectric properties of the materials. At high resistance values (representing clean or lightly polluted conditions), the leakage current remains low and relatively stable. However, as the resistance drops, the leakage current rises sharply, indicating a higher risk of surface discharges and possible insulation failure. In addition, voltage distribution becomes uniform, and string efficiency potentially improves. It also leads to greater energy losses and significantly increases the risk of thermal degradation and electrical breakdown of the insulators.

This figure highlights the sensitivity of insulator performance to pollution severity and underscores the importance of maintaining adequate surface resistance to ensure safe and reliable operation of high-voltage equipments.

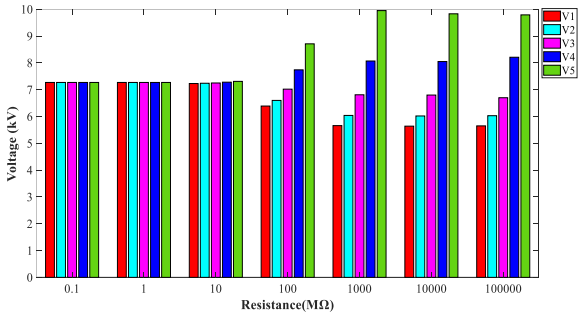


Figure 3. Voltage distribution in 5-unit porcelain insulator strings under different surface resistance values

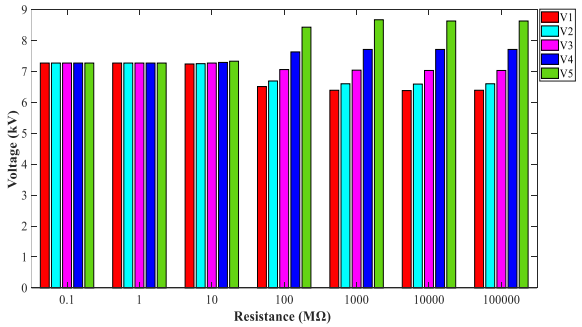


Figure 4. Voltage distribution in 5-unit glass insulator strings under different surface resistance values

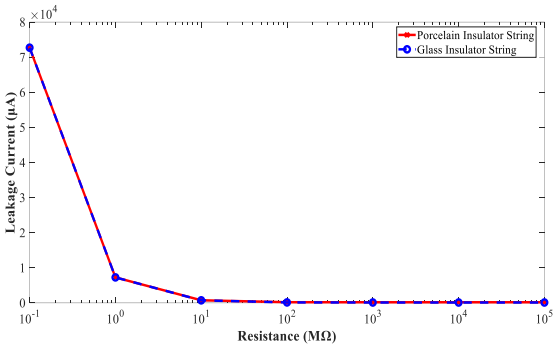


Figure 5. Leakage current in a 5-unit porcelain and glass insulator strings vs. different pollution resistances

3. 2. The Effect of the String's Deviation Angle and Pollution Levels (Scenario 2) In this scenario, the combined effect of the insulator string deviation angle and different pollution levels is investigated. Specific deviation angles considered are −30°, −15°, 0°, +15°, +30°, and +70°, with clean, light, medium, and heavy pollution conditions. These simulations are performed for two porcelain and glass insulator strings. As is well known, due to strong winds or other factors, and in tension towers, the insulator string can deviate from its original position, moving slightly closer to or farther from the tower. Additionally, in suspension strings, significant deviation from the vertical position often occurs (44).

In this scenario, due to changes in the distances between insulator units and the tower, the capacitance between them varies accordingly. In this case, by using computer simulation, the voltage distribution and efficiency of the insulator strings (porcelain and glass) and leakage current are calculated and analyzed by considering different uniform pollutions and deviation

angles. According Table 2, the deviation of the insulator string angle leads to a change in the capacitance values between the pin of each insulator unit and the tower (10). In this analysis, according to Figure 2, deviation angles towards the tower are indicated with a negative sign (-), and deviation angles away from the tower are shown with a positive sign (+). In the circuit-based modeling of insulator strings, the capacitances Ce1 through Ce5 represent the capacitive coupling between the metallic pins of each unit and the tower structure. Specifically, Ce1 denotes the capacitance between the pin of the first unit and the tower, which is typically the closest point to ground. Similarly, Ce2 to Ce5 correspond to the capacitances between the pins of the second through fifth units and the tower. The values of these capacitances depend on various factors such as the spatial position of each unit relative to the tower, the deviation angle of the string, the geometric shape of insulators, and atc. These external capacitances play a significant role in determining the voltage distribution along the insulator string and have a considerable impact on insulator's electrical performance under real operating conditions.

The results of the simulations in Figure 6 show that with increasing the pollution level, the voltage distribution becomes significantly uniform. Also, the orientation of the insulator string relative to the tower plays a significant role in the voltage distribution pattern. According to Figure 6, in clean and light pollution conditions, the voltage distribution variations in both insulator strings are very small, and it can be said that the voltage distribution shows a uniform and stable behavior in these states. As the insulator string deviates farther outward from the tower, the voltage distribution along the string becomes more uniform. However, when the insulator string leans toward the tower, the uniformity decreases. Under heavy pollution condition, the reduction in the surface resistance of the insulator significantly impacts on insulator performance and renders the effect of geometrical and positional parameters largely negligible; such that the voltage across each insulator unit in the string, regardless of the deviation angle, reaches an almost constant value of 7.27 kV, and a completely uniform voltage distribution is

TABLE 2. Capacitance between the pins of each insulator unit and the tower in the porcelain and glass insulator strings under different deviation angles

Deviation angle	Capacitance between insulator pin and tower (pF)				
	Ce1	Ce2	Ce3	Ce4	Ce5
-30	2.15	2.32	2.52	2.77	3.06
-15	2.07	2.15	2.24	2.33	2.44
0	2	2	2	2	2
+15	1.93	1.86	1.8	1.75	1.69
+30	1.87	1.75	1.65	1.56	1.48
+70	1.77	1.58	1.44	1.31	1.21

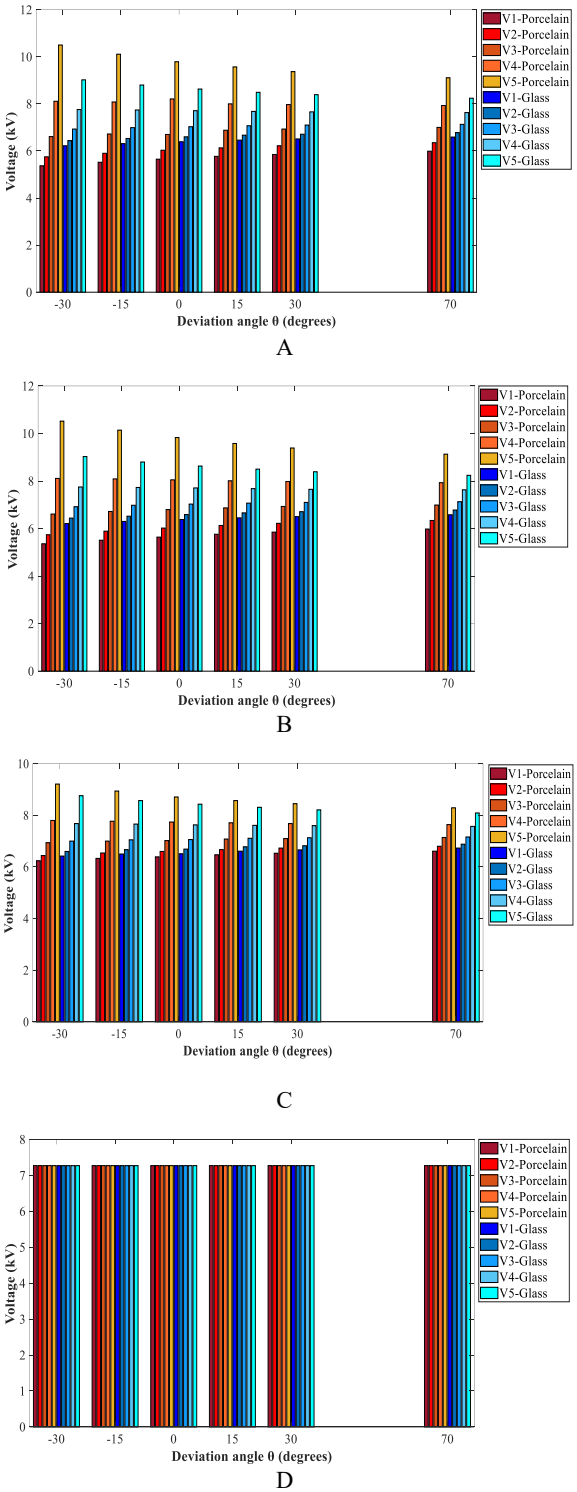


Figure 6. Voltage distribution in the 5-unit porcelain and glass insulator strings considering deviation angle and different pollution levels. A: Clean, B: Light , C: Medium , D: Heavy

observed throughout the string. This shows that at high pollution, the effect of surface resistance is more dominant than other parameters of the insulator. So, special attention should be paid to this point in the design of insulation systems.

Figure 7 illustrates the leakage current flowing through the insulator strings. Under this Scenario, the results have shown that with the increase of surface pollution of insulators (decrease of surface resistance), leakage current increases significantly. Deviation of the insulator string by moving away from the tower (positive angle θ) leads to a reduction in leakage current, whereas deviation toward the tower (negative angle θ) results in an increase in leakage current. This behavior can occur due to changes in the capacitive distance between the insulator and the tower, which leads to a variation in the capacitance Ce. Considering that under heavy pollution conditions, the leakage current at all angles is significantly higher than at other pollution levels and remains constant at 7270 μA, the variations in leakage current in this case have been disregarded in Figure 7.

3. 3. The Effect of a Defective Insulator Unit and Pollution Levels (Scenario 3) In this scenario, the presence of a defective unit in the string—such as an insulator with reduced dielectric strength or minor physical damage—is considered, modeled with a parallel low resistance to the insulator's capacitance.

Variations in voltage distribution, string efficiency, and leakage current are calculated under different conditions, including light, medium, and heavy pollution levels, as well as the presence or absence of defective units located at positions 1, 3, and 5 along the porcelain and glass insulator strings. Each defective unit is modeled in the equivalent circuit using a 1 MΩ resistor in parallel with the capacitive element of that unit. The resistance value of a defective unit is set at 1 MΩ; heavy pollution effects are excluded from this section due to very low surface resistance.

According to the results presented in Figure 8, the voltage distribution across the insulator strings significantly deviates from uniformity in the presence of a defective unit combined with increasing pollution levels. This non-uniformity is more pronounced when the middle unit (unit 3) is defective. It is because the voltage applied to the defective unit drops to a minimum, and its voltage is redistributed among the remaining units. This redistribution mainly affects the units adjacent to the defective one, leading to voltage concentration in those units and thereby increasing the risk of electrical breakdown.

Figure 9 shows that while the defective unit is close to the energized conductor, the leakage current through the insulator strings is maximum amount. The results presented in Figure 9 indicate that under clean and light pollution conditions, the leakage current for both porcelain and glass insulators ranges from 115 to 150 μA and 180 to 240 μA, respectively. With an increase in pollution level to medium, the leakage current ranges from 130 to 180 μA for porcelain insulators and 200 to 260 μA for glass insulators. Furthermore, the presence of a defective unit in the string significantly increases the leakage current; in the most critical condition (defective unit and medium pollution), the leakage current in the glass insulator reaches approximately 260 μA. Comparison between the two types of insulators indicates that glass insulators exhibit higher leakage currents under all conditions, with the difference being more

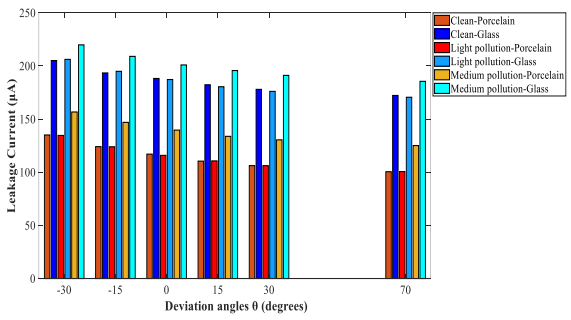


Figure 7. Leakage current in a 5-unit porcelain and glass insulator strings under different pollution levels and deviation angles

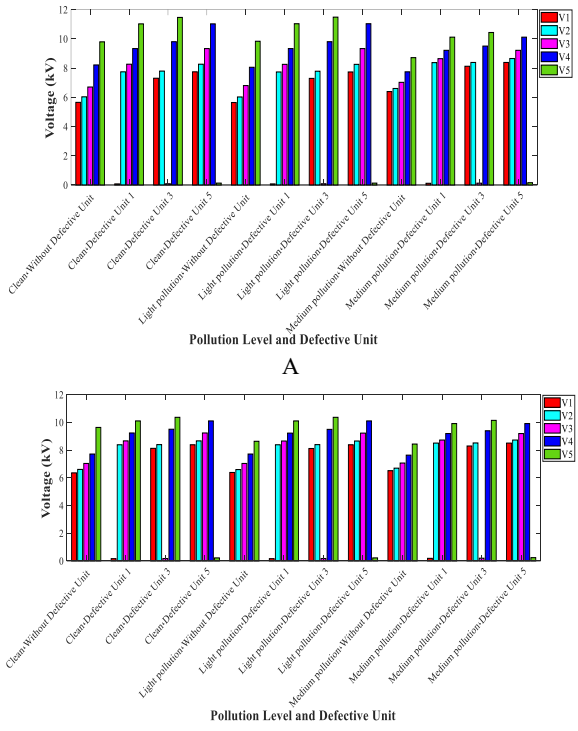


Figure 8. Voltage distribution in a five-unit insulator string under the c effects of a defective unit with different pollution levels. A: Porcelain insulator string, B: Glass insulator string

pronounced under medium pollution. It should be noted that under heavy pollution in both porcelain and glass insulator strings, the leakage current remains constant at 7270 μA, and therefore, it is not shown in Figure 9.

3. 4. The Effect of String Deviation Angle, Pollution Levels, and Defective Units (Scenario 4) In this scenario, the combined effects of surface pollution levels (clean, light, medium, heavy), the presence of a defective insulator unit, and various deviation angles (−30°, −15°, 0°, +15°, +30°, +70°) on the voltage distribution across insulator strings are studied. These analyses are conducted for porcelain and glass insulator strings, which exhibit different dielectric properties. When there are deviation angle of the insulator string from the vertical position, and also the presence of a defective unit simultaneously with surface pollution, the electrical behavior of the string becomes significantly more complex, and the overall efficiency of the string is heavily affected by the combination of these factors.

Figures 10 and 11 illustrate the voltage distribution across the end units (units 1 and 5) of porcelain and glass insulator strings under the influence of various deviation angles and pollution levels. In Figure 10, the analysis is performed assuming no defects in the string, whereas

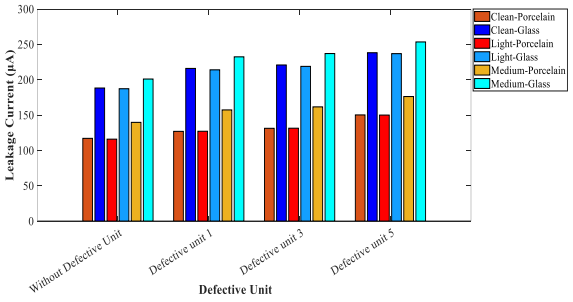


Figure 9. Leakage current in a 5-unit porcelain and glass insulator strings under a defective unit and different pollution levels

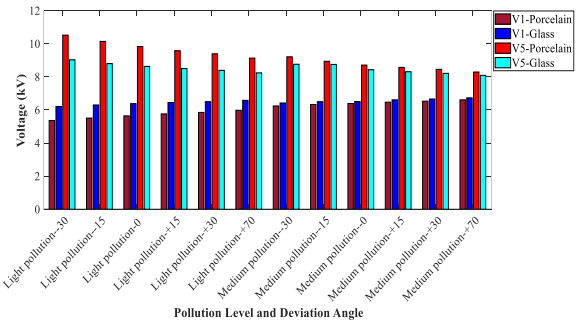


Figure 10. Voltage distribution on units No. 1 and 5 (V1 ,V5) of porcelain and glass strings under the effects of string deviation angles, pollution levels, and without considering a defective unit

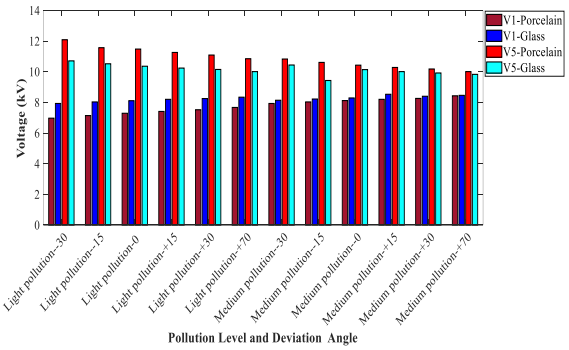


Figure 11. Voltage distribution on insulator units No. 1 and 5 (V1 ,V5) of porcelain and glass strings under the effects of string deviation angles, pollution levels, and a defective unit at position 3

Figure 11 examines the combined effects of deviation angle, pollution level, and the presence of a faulty unit located at position 3 within the string. The simulation results show that this combination leads to the most non-uniform voltage distribution among all the studied scenarios. In this condition, the electrical stress on healthy units, especially those near the energized conductor, increases significantly. This phenomenon is further intensified at negative deviation angles (towards the tower), and when combined with surface pollution and mechanical defects, it causes unstable leakage currents and a notable reduction in the string's overall efficiency.

4. DISCUSSION

According to the conducted studies, the results show that the voltage distribution along the glass insulator string is significantly more uniform than that of the porcelain insulator string, leading to an overall improvement in insulation efficiency. This desirable behavior is mainly due to the higher self-capacitance of glass insulators compared to porcelain ones, which results in a more uniform voltage distribution among the units in the string. However, a notable characteristic of glass insulators is their higher leakage current passing through the string compared to porcelain insulators. Under certain conditions, such as high humidity or severe surface contamination, this can cause increased power loss, localized heating, and ultimately a higher risk of surface discharges and damage to the insulator structure.

In addition, the presence of a defective unit, particularly when positioned in the middle of the string, introduces a substantial imbalance in the voltage profile. The highest voltage drop is observed on the unit closest to the energized conductor, resulting in a noticeable reduction in string efficiency. This localized stress increases the risk of flashover and underscores the

importance of periodic inspection and early detection of faulty insulator units in polluted or mechanically stressed environments. Overall, these findings emphasize that optimal string design must consider not only environmental pollution levels but also the mechanical configuration and condition of individual insulator units. Designing for worst-case combined conditions—heavy pollution, strong deviation, and internal defects—is essential for ensuring long-term system reliability. When the insulator surface is contaminated and the insulator string has a defective unit with angle deviation, the electrical behavior of the string becomes significantly more complex, and the overall string efficiency is impacted by the combined effects of these factors. The simulation results indicate that this state leads to the most non-uniform voltage distribution among all examined scenarios. In this condition, the healthy units adjacent to the defective unit (particularly those closer to the energized conductor) are subjected to higher electrical stress. As a result, the voltage imposed on them increases, and the occurrence of surface discharges accelerates. This effect becomes more severe under negative deviation angles (toward the tower), as the voltage tends to concentrate on specific parts of the string. Furthermore, the presence of heavy surface pollution, along with a defective unit and mechanical deviation, causes surface leakage current to distribute non-uniformly across the string. Consequently, the performance of the entire string becomes highly dependent on the physical configuration and health status of individual units. Under these combined stress conditions, string efficiency can drop to its lowest possible value. This highlights the critical importance of environmental conditions, installation angle, and real-time condition monitoring of insulator strings in high-voltage networks.

In summary, the simultaneous presence of these critical factors not only leads to a considerable decrease in string efficiency but also greatly increases the risk of surface discharges and eventual dielectric failure. Therefore, it is recommended that insulator string design and operation be conducted with consideration of such worst-case combined scenarios to ensure system safety and long-term reliability.

5. CONCLUSION

In this study, the performance of five-unit porcelain and glass insulator strings in 63 kV overhead transmission lines was investigated under combined conditions, including various levels of surface pollution, deviation

TABLE 3. Efficiency of a 5-unit porcelain insulator string under different pollution levels, with deviation angle, and a defective unit at position 3

Deviation Angle (θ)	Clean		Light Pollution		Medium Pollution		Heavy Pollution	
	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3
-30	69.28	60.23	69.09	69.13	78.96	67.12	100	100
-15	71.89	61.97	71.69	61.87	81.33	68.52	100	100
0	74.10	63.43	73.96	63.36	83.43	69.70	100	100
+15	75.98	64.62	75.87	64.56	84.81	70.71	100	100
+30	77.59	65.61	77.41	65.55	86.09	71.41	100	100
+70	79.83	67.17	79.70	67.03	87.72	72.57	100	100

TABLE 4. Efficiency of a 5-unit glass insulator string under different pollution levels, with deviation angle, and a defective unit

Deviation Angle (θ)	Clean		Light Pollution		Medium Pollution		Heavy Pollution	
	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3	Without Defective Units	Defective in Unit 3
-30	80.69	67.97	80.56	67.92	82.97	69.67	100	100
-15	82.63	69.2	82.63	69.15	84.88	70.77	100	100
0	84.30	70.24	84.25	70.17	86.23	71.74	100	100
+15	85.66	71.06	85.59	71.06	87.48	72.67	100	100
+30	86.74	71.76	86.67	71.71	88.53	73.34	100	100
+70	88.32	72.67	88.25	72.67	89.94	73.97	100	100

angles from the vertical alignment, and the presence of a defective unit. This analysis was conducted using circuit-based modeling and simulation in the MATLAB environment.

The simulation results demonstrated that increasing surface pollution, particularly under severe conditions, led to a more uniform voltage distribution across the insulator units, which in turn improved the overall efficiency of the string. For example, the efficiency of the insulator string under clean conditions was approximately 74% for porcelain insulators and 84% for glass insulators, while under severe pollution conditions, the efficiency of both strings reached 100%. However, this improvement was accompanied by higher leakage currents, potentially increasing the risk of surface discharges and electrical failure. The effect of deviation angle showed that deviation away from the tower resulted in a more uniform voltage distribution and reduced leakage current compared to deviation toward the tower, thereby enhancing the string's efficiency .For example, under medium pollution , the efficiency of the glass insulator was 82.9% at a -30° angle, 86.23% at 0°, and 88.53% at +30°. These results indicate that angular deviations away from the tower can lead to improved performance of the glass insulator.

When a defective unit was present, especially in the middle position of the string, the voltage distribution became significantly unbalanced, subjecting the neighboring healthy units to higher electrical stress. The combined presence of a defective unit, negative deviation angle (toward the tower), and heavy pollution created a critical condition that led to a severe drop in efficiency and localized voltage concentration within the string .For the porcelain insulator, when examining three different conditions of pollution, defective units, and deviation angles, the worst performance was observed at a -30° angle under clean conditions, with an efficiency of 60.23% when the defective unit was located in unit 3. In contrast, when there was no defective unit and the angle was +30°, the efficiency under clean conditions increased to 77.6%, resulting in an improvement of approximately 22% in efficiency.

To ensure greater reliability and safety in power transmission systems, the design and operation of insulator strings must account for worst-case combined scenarios. This highlights the necessity of implementing preventive strategies such as periodic inspections, continuous monitoring, regular cleaning, and mechanical design improvements, particularly in areas exposed to heavy pollution.

Acknowledgements

The authors acknowledge the scientific support of Babol Noshirvani University of Technology.

Funding

This research did not receive any specific grant from any funding agencies in the public, commercial, or non-profit sectors.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

This article has no conflict with studies by other authors in this field.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used Grammarly exclusively for minor language editing to improve readability. After using this tool, the authors carefully reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Author(s) Biosketches

M. Mirzaie received Ph.D. degree from Iran University of Science and Technology in Electrical Engineering (2007). He is currently as professor in Babol Noshirvani University of Technology. His research interests include high-voltage engineering and also operation problems, and condition assessment/monitoring of high-voltage equipments.

S. J. Ghasemnezhad is a master's student in Electrical Engineering (Power Systems) at Babol Noshirvani University of Technology, Babol, Iran.

A. Gholami is a Ph.D. student in Electrical Engineering (Power Systems) at Babol Noshirvani University of Technology, Babol, Iran. Her research interests are electric and magnetic field analysis and high-voltage engineering.

REFERENCES

1. Dhalaan S, Elhirbawy M, editors. Simulation of voltage distribution calculation methods over a string of suspension

insulators. 2003 IEEE PES Transmission and Distribution Conference and Exposition (IEEE Cat No 03CH37495); 2003: IEEE. 10.1109/TDC.2003.1335366

2. Salthouse E. Computation of the Voltage Distribution across a String of Suspension Insulators Using Ecap. International Journal of Electrical Engineering Education. 1974;11(1):78-81. 10.1177/002072097401100109
3. Kontargyri V, Plati L, Gonos I, Stathopulos I. Measurement and simulation of the voltage distribution and the electric field on a glass insulator string. Measurement. 2008;41(5):471-80. 10.1016/j.measurement.2007.07.002
4. Topalis F, Gonos I, Stathopulos I. Dielectric behaviour of polluted porcelain insulators. IEE Proceedings-Generation, Transmission and Distribution. 2001;148(4):269-74. 10.1049/ip-gtd:20010258
5. Xueming Z, Danhui H, Zhiqiang F, Ding Z, Xiaopo M, editors. Simulation Analysis of Voltage Distribution of 500kV Degraded Magnetic Insulator String. IOP Conference Series: Earth and Environmental Science; 2020: IOP Publishing. 10.1088/1755-1315/617/1/012017
6. Pattanadech N, editor The measurement technique for distributed voltage of a string insulator using a standard sphere gap. 2004 International Conference on Power System Technology, 2004 PowerCon 2004; 2004: IEEE. 10.1109/ICPST.2004.1460155
7. Zhang B, Han S, He J, Zeng R, Zhu P. Numerical analysis of electric-field distribution around composite insulator and head of transmission tower. IEEE Transactions on power delivery. 2006;21(2):959-65. 10.1109/TPWRD.2005.859293
8. Khan S, Alam S, Saleem MZ. Analysis of electric field and leakage current of glass and porcelain insulators under clean and polluted conditions: A comparative study of three profiles. Electric Power Systems Research. 2025;239:111283. 10.1016/j.epsr.2024.111283
9. Fauziah D, Khaidir IM. The evaluation of daily comparative leakage currents on porcelain and silicone rubber insulators under natural environmental conditions. IEEE Access. 2021;9:27451-66. 10.1109/ACCESS.2021.3057626
10. İzgi E, İnan A, Ay S. The analysis and simulation of voltage distribution over string insulators using Matlab/Simulink. Electric power components and systems. 2008;36(2):109-23. 10.1080/15325000701548923
11. Shen W, editor Simulation Research on Electric Field Distribution of Insulator String in ULTRA High Voltage Transmission Line. Journal of Physics: Conference Series; 2021: IOP Publishing. 10.1088/1742-6596/2108/1/012024
12. Mohamed I, Aramugam K, Khan MA, editors. Simulation and Measurement of the Voltage Distribution on Porcelain Insulator String under Polluted Condition. 2018 IEEE 4th International Symposium in Robotics and Manufacturing Automation (ROMA); 2018: IEEE. 10.1109/ROMA46407.2018.8986724
13. Dhalaan S, Elhirbawy M, editors. Investigation on the characteristics of a string of insulator due to the effect of dirt. 2003 IEEE PES Transmission and Distribution Conference and Exposition (IEEE Cat No 03CH37495); 2003: IEEE. 10.1109/TDC.2003.1335059
14. Yong C, Feng H, Yizheng D, editors. Study on Withstand Voltage Characteristics and Surface Electrical Field Distribution along Polluted Insulators. 2008 International Conference on High Voltage Engineering and Application; 2008: IEEE. 10.1109/ICHVE.2008.4773873
15. Jiang Y, Li L, Wang R, Lu M, Jiang Y, Liu Z, editors. Simulation Study on the Cause of Diameter Distribution of Contamination Particle on Porcelain Insulator Surface in High Voltage Electrostatic Field. 2018 IEEE International Conference on High Voltage Engineering and Application (ICHVE); 2018: IEEE. 10.1109/ICHVE.2018.8642177
16. Masinga R, Okojie DE, Nnachi AF, Akumu AO, Tshubwana PR, editors. Analysis of Potential Distribution Efficiency of Polluted 132KV AC Transmission Line Insulators. 2020 6th IEEE International Energy Conference (ENERGYCon); 2020: IEEE. 10.1109/ENERGYCon48941.2020.9236552
17. He J, Gorur RS. Charge simulation based electric field analysis of composite insulators for HVDC lines. IEEE Transactions on Dielectrics and Electrical Insulation. 2015;21(6):2541-8. 10.1109/TDEI.2014.004541
18. Stevenson H, Alternator AER, Woodruff R. 1935 Index—Electrical engineering and transactions. 10.1109/EE.1935.6538846
19. Zhao X, Yang X, Hu J, Wang H, Yang H, Li Q, et al. Grading of electric field distribution of AC polymeric outdoor insulators using field grading material. IEEE Transactions on Dielectrics and Electrical Insulation. 2019;26(4):1253-60. 10.1109/TDEI.2019.007989
20. Rosli H, Othman N, Jamail N, Ismail M. Potential and Electric Field Characteristics of Broken Porcelain Insulator. International Journal of Electrical & Computer Engineering (2088-8708). 2017;7(6). 10.11591/ijece.v7i6.pp3114-3123
21. Unahalekhaka P, Phonkaphon S. Simulation of Potential and Electric Field Due to Defective Insulator in 115 kV Transmission Line. GMSARN International Journal. 2014;8(1):1-6.
22. Reddy BS, Naik BS, Kumar U, Satish L, editors. Potential and electric field distribution in a ceramic disc insulator string with faulty insulators. 2012 IEEE 10th international conference on the properties and applications of dielectric materials; 2012: IEEE. 10.1109/ICPADM.2012.6318928
23. Wei H, Fan Y, Jingang W, Hao Y, Minyou C, Degui Y. Inverse application of charge simulation method in detecting faulty ceramic insulators and processing influence from tower. IEEE transactions on magnetics. 2006;42(4):723-6. 10.1109/TMAG.2006.871393
24. Cho H, Han S, Park K, Han D, Lee D, Choi I, et al., editors. Ageing Characteristics of porcelain suspension insulators for transmission line by cool and heat accelerate method. Proceedings of the 7th International Conference on Properties and Applications of Dielectric Materials (Cat No 03CH37417); 2003: IEEE. 10.1109/ICPADM.2003.1218439
25. Zhang B, He J, Zeng R, Liang X. Voltage distribution along a long ceramic insulator string in a high-voltage tower window. COMPEL-The international journal for computation and mathematics in electrical and electronic engineering. 2010;29(3):811-23. 10.1108/03321641011028323
26. El-Kishky H, Gorur R. Electric potential and field computation along AC HV insulators. IEEE Transactions on Dielectrics and Electrical Insulation. 2002;1(6):982-90. 10.1109/94.368665
27. Forrest JS. The characteristics and performance in service of high-voltage porcelain insulators. Journal of the Institution of Electrical Engineers-Part II: Power Engineering. 1942;89(7):60-80. 10.1049/ji-2.1942.0008
28. Karimzadeh A, Akbari M, Roohi R, Amiri MJ. Dynamic behavior of galloping micro energy harvester with the elliptical bluff body using CFD simulation. Sustainability. 2023;15(16):12187. https://doi.org/10.3390/su151612187
29. Li HH, Zheng Z, Chen HB, Bai H, Zhang HZ, Ruan JJ, et al. Potential and Electric Field Distribution of 220kV Insulator String with a Faulty Insulator under Windage Yaw Condition. Applied Mechanics and Materials. 2014;521:317-20. 10.4028/www.scientific.net/AMM.521.317
30. Dadashizadehsamakosh J, Mirzaie M. simulation and analysis of uniform and non-uniform (longitudinal and fan-shaped) pollutions effect on the potential and electric field distribution of polymeric insulator using finite element method. Journal of

Modeling in Engineering. 2019;17(56):1-12. 10.22075/jme.2018.14237.1394

31. Zhang Z, Liu X, Jiang X, Hu J, Gao DW. Study on AC flashover performance for different types of porcelain and glass insulators with non-uniform pollution. IEEE Transactions on Power Delivery. 2013;28(3):1691-8. 10.1109/TPWRD.2013.2245153
32. Xiao H, Hao L, Zhen D, Kai Q, Jian L, Biwu Y, et al., editors. A Novel Algorithm of Insulator String Distributed Voltage Based on Stray Capacitance. 2018 International Conference on Power System Technology (POWERCON); 2018: IEEE. 10.1109/POWERCON.2018.8601621
33. Taghizadeh F, Jafari AJ, Gholami M, Kermani M, Arfaeinia H, Mohammadi S, et al. Monitoring of airborne asbestos fibers in an urban ambient air of Shahryar City, Iran: levels, spatial distribution, seasonal variations, and health risk assessment. Environmental Science and Pollution Research. 2019;26(7):6450-9. https://doi.org/10.1007/s11356-018-4029-0
34. Tonmitr N, Tonmitr K, Kaneko E. The Comparison of the String Suspension Porcelain (5-6-7 insulators)/string Efficiency in Case of Dry and Wetted water Contamination Condition. Applied Mechanics and Materials. 2015;781:280-3. 10.4028/www.scientific.net/AMM.781.280
35. Goudarzi M GA, Shahrtash SM. . Experimental evaluation and finite element modelling of the effect of aging on dry-band discharge intensity in polluted polymer insulators. International Journal of Engineering Transactions A: Basics. 2026;39(1):265-80. 10.5829/ije.2026.39.01a.20
36. Ahmad F, Khan Q, Alam A. Analysis of flashover voltages of disc type insulator under artificial pollution condition. International Journal of Engineering Transactions C: Aspects. 2016;29(6):762-8. 10.5829/idosi.ije.2016.29.06c.05
37. Tonmitr N, Tonmitr K, Kaneko E. The effect of controlling stray and disc capacitance of ceramic string insulator in the case of clean and contaminated conditions. Procedia Computer Science. 2016;86:333-6. 10.1016/j.procs.2016.05.090
38. Zhao T, Comber MG. Calculation of electric field and potential distribution along nonceramic insulators considering the effects of conductors and transmission towers. IEEE Transactions on Power Delivery. 2000;15(1):313-8. 10.1109/61.847268
39. Araya J, Montaña J, Schurch R. Electric field distribution and leakage currents in glass insulator under different altitudes and pollutions conditions using FEM simulations. IEEE Latin America Transactions. 2021;19(8):1278-85. 10.1109/TLA.2021.9475858
40. Lv Y, Wang Y, Wang J, Wang J, Yang J. Simulation of electric field distribution along insulator surface in polluted environments. Energy Engineering: Journal of the Association of Energy Engineers. 2021;118(3):631. 10.32604/EE.2021.014578
41. Salem AA, Abd-Rahman R, Al-Gailani SA, Kamarudin MS, Ahmad H, Salam Z. The leakage current components as a diagnostic tool to estimate contamination level on high voltage insulators. IEEE Access. 2020;8:92514-28. 10.1109/ACCESS.2020.2993630
42. Asadpoor M, Mirzaie M. Simulation and measurement of the voltage distribution on high voltage suspension Porcelain insulator string under pollution condition. International Journal of Applied Science and Engineering Research. 2012;1(1):165-75.
43. Chowdhury S, Lahiri A, Chakravorti S. Surface resistance modified electric field computation in asymmetric configuration using surface charge simulation method: a new approach. IEEE Transactions on Dielectrics and Electrical Insulation. 2012;19(3):1068-75. 10.1109/TDEI.2012.6215114
44. Khodsuz M, Haghgoo Rostami R. The investigation and simulations of the suspension insulator string offset effects on the voltage and electric field distributions using the finite element method. Applied Electromagnetics. 2022;10(1):91-7.

Persian Abstract

چکیده

زنجیره مقره‌ها یکی از اجزای مهم و حیاتی خطوط انتقال هوایی هستند که وظیفه‌ی عایق‌بندی و جلوگیری از جریان‌های نشتی را بر عهده دارند. در این مقاله، تأثیر عوامل مختلف بر توزیع ولتاژ و جریان نشتی زنجیره مقره‌ها مورد بررسی قرار گرفته است. توزیع غیریکنواخت ولتاژ بین واحدهای مقره در زنجیره می‌تواند منجر به تخریب مقره و کاهش قابلیت اطمینان خطوط هوایی و شبکه قدرت شود. در این مقاله، تحلیل دقیقی از زنجیره مقره‌های چینی و شیشه‌ای با پنج واحد مقره، تحت شرایط گوناگون، با استفاده از مدل‌سازی مداری و شبیه‌سازی در محیط نرم‌افزار MATLAB، انجام شده است. سناریوهای موردبررسی شامل آلودگی‌های سطحی مختلف، انحراف زاویه از حالت عمودی و وجود یک واحد معیوب می‌باشد. نتایج نشان داده است که با افزایش میزان آلودگی سطحی، توزیع ولتاژ بین واحدهای زنجیره یکنواخت‌تر شده و در نتیجه راندمان کلی زنجیره بهبود می‌یابد، اما هم‌زمان جریان نشتی نیز افزایش می‌یابد. همچنین، در حالت انحراف زنجیره به سمت بیرون (دور از دکل) نسبت به انحراف به سمت داخل، توزیع ولتاژ یکنواخت‌تر شده و کاهش قابل‌توجهی در جریان نشتی مشاهده شده است. در سناریوی وجود واحد معیوب، تنش ولتاژ بر واحد مقره‌ها کاهش یافته و سهم ولتاژ آن بسته به موقعیت نسبی، میان واحدهای سالم توزیع می‌شود. همچنین نتایج نشان داده است که زنجیره مقره شیشه‌ای نسبت به چینی، توزیع ولتاژ یکنواخت‌تر و راندمان بالاتری را در شرایط تمیز و همچنین در آلودگی‌های کم تا متوسط ارائه می‌دهد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2395-2406

International Journal of Engineering

Journal Homepage: www.ije.ir

Valorization of Polypropylene Plastic Waste via Slow Pyrolysis: Pathways to Pyro-Oil Production

R. I. Mainil*[a], J. J. Panjaitan[a], A. W. Sari[b], A. Susilawati[b], A. K. Mainil[c], O. Farobie[d], A. Amrullah[e], A. Aziz[a]

[a] *Thermal Engineering Laboratory, Department of Mechanical Engineering, Universitas Riau, Kampus Bina Widya, Jl. HR. Soebrantas Km 12.5, Pekanbaru 28293, Riau, Indonesia*
[b] *Department of Mechanical Engineering, Universitas Riau, Kampus Bina Widya, Jl. HR. Soebrantas Km 12.5, Pekanbaru 28293, Riau, Indonesia*
[c] *Department of Mechanical Engineering, Universitas Bengkulu, Bengkulu, 38371A, Indonesia*
[d] *Department of Mechanical and Biosystem Engineering, IPB University, IPB Darmaga Campus, Bogor, West Java 16002, Indonesia*
[e] *Department of Mechanical Engineering, Lambung Mangkurat University, Banjarmasin, South Kalimantan, Indonesia*

PAPER INFO

Paper history:
Received 24 August 2025
Received in revised form 21 September 2025
Accepted 24 October 2025

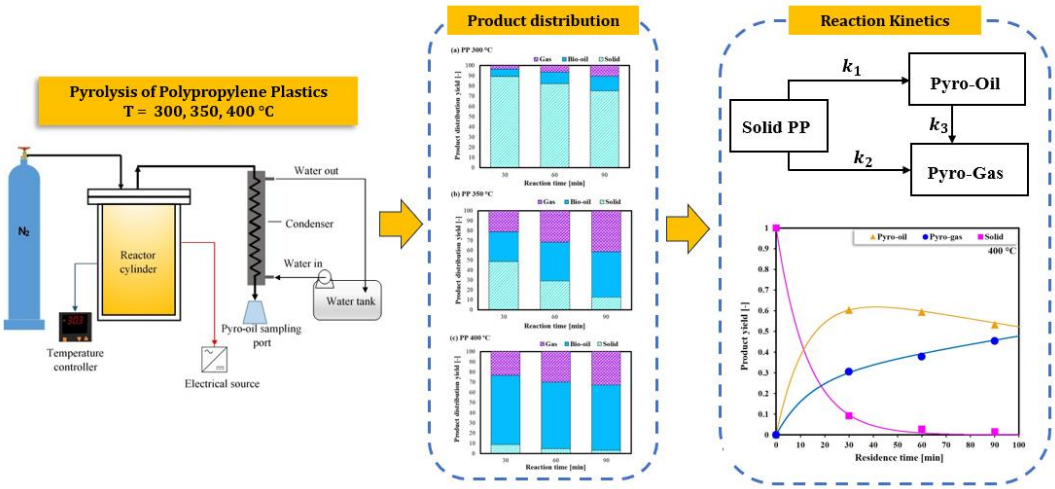
Keywords:
Energy
Pyro-Oil
Pyrolisis
Polypropylen
Kinetics

ABSTRACT

This research examines the pyrolysis of polypropylene (PP) plastic waste as a feasible method for pyro-oil generation. The primary objective is to convert non-biodegradable polymer waste into value-added liquid fuel while addressing environmental concerns associated with plastic pollution. The pyrolysis of PP was conducted in a batch reactor at regulated temperatures of 300, 350, and 400 °C with reaction times of 30, 60, and 90 minutes to examine the thermal degradation behavior, product distribution, and quality of the obtained pyro-oil. The highest pyro-oil yield of 68% was achieved at 400 °C, indicating optimal thermal cracking of polypropylene chains. Under these conditions, the pyro-oil exhibited a hydrocarbon content of 49.3%, emphasizing its viability as an alternative energy source. The calorific value of the pyro-oil ranged from 48 MJ kg^{-1} to nearly 49 MJ kg^{-1}. The possible reaction pathways were determined in this study, as indicated by the pre-exponential factor and activation energy derived from the Arrhenius equation. These findings demonstrate the feasibility of valorizing polypropylene waste through slow pyrolysis, offering a promising pathway for resource recovery, circular economy implementation, and reduction of the environmental impact of plastic waste.

doi: *10.5829/ije.2026.39.10a.04*

Graphical Abstract



*Corresponding Author Email: rahmat.iman@lecturer.unri.ac.id (R. I. Mainil)

Please cite this article as: Mainil RI, Panjaitan JJ., Sari AW, Susilawati A, Mainil AK, Farobie O, Amrullah A, Aziz A. Vaorization of Polypropylene Plastic Waste via Sow Pyrolysis: Pathways to Pyro-Oil Production. International Journal of Engineering, Transactions A: Basics. 2026;39(10): 2395-406.

1. INTRODUCTION

Indonesia is currently faced with a critical challenge stemming from plastic waste generation. Daily, the country produces over 3.22 million metric tons of plastic waste, making it the second-largest contributor to global plastic waste, surpassed only by China (1). A substantial portion of this plastic waste, estimated between 0.48 and 1.29 million metric tons annually, is discharged into the marine environment, causing serious ecological degradation and disruption of marine ecosystems. The generation of plastic waste continues to escalate in line with global population growth. This trend is further driven by the increasing demand for plastics, which is due to their high versatility, low density, exceptional durability, resistance to corrosion, and overall cost-effectiveness (2-4).

The Indonesian authorities have set an ambitious goal to reduce plastic waste generation by 70% by 2025 (5), reflecting the urgent need for sustainable and effective waste management practices. Conventional landfilling presents numerous environmental and health concerns, including the extremely slow biodegradation of plastics, risks to terrestrial and aquatic organisms, and various forms of pollution (6). While recycling remains an important waste management strategy, it faces significant challenges such as high operational costs and the demand for a large labor force (7). Converting plastic waste into energy offers a promising alternative; however, direct incineration for thermal energy production has critical limitations, including greenhouse gas emissions that exacerbate climate change and the release of toxic pollutants (8).

Pyrolysis offers a viable alternative for the conversion of plastic waste. This thermochemical process operates under oxygen-deficient conditions, utilizing heat to decompose feedstocks into pyrolytic oil, solid residues, and gaseous products, generally observed in the temperature interval of 250–600 °C (9-12). Recent research has extensively investigated the production of pyro-oil from diverse plastic waste feedstocks via pyrolysis (13-15).

Polypropylene (PP) plays a significant role in the global plastic waste challenge, as it is the most extensively produced thermoplastic, accounting for approximately 17% of annual global plastic output, 18% of plastic packaging, and 24% of plastics found in municipal solid waste (16-18). Its chemical structure, characterized by a high volatile content, offers significant potential for conversion into fuel through pyrolysis (4).

Pasae et al. (19) investigated the pyrolysis of polypropylene (PP) plastic waste at temperatures ranging from 250 °C to 350 °C, with a constant holding time of 120 minutes. The highest oil yield of 62.56% was obtained at 350 °C. In a related study, Muchammad (20) pyrolyzed PP waste at three temperatures—250 °C, 300 °C, and 350 °C—yielding oil volumes of 420 mL, 480 mL, and 500 mL, respectively. Corresponding densities were 0.75 g/mL, 0.76 g/mL, and 0.77 g/mL, while calorific values were 39,221 kJ/kg, 38,870 kJ/kg, and 38,301 kJ/kg. The resulting pyrolysis oil exhibited an octane rating of 83.5 and a viscosity of 0.034 Poise.

Adoe et al. (21) employed a simple reactor for PP pyrolysis at a constant temperature of 250 °C, varying the reaction time between 1, 2, and 3 hours. The resulting pyrolysis oil demonstrated calorific values ranging from 10,296 to 11,670 cal/g, exceeding that of gasoline (8,356 cal/g). Its density (0.760–0.780 kg/L) was also higher than gasoline (0.680 kg/L), while the viscosity (0.65–0.78 cP) was comparable to that of gasoline (0.652 cP). Further studies have compared the pyrolysis conversion performance of PP with that of low-density polyethylene (LDPE) (22, 23), highlighting differences in product yield and composition between the two plastic types. To improve the efficiency of PP pyrolysis, several investigations have also examined the role of catalysts in enhancing conversion rates, optimizing product distribution, and reducing undesirable by-products (24, 25).

The ability to alter the solid structure of PP through thermal treatment is of significant importance. Based on analogous observations in plastics pyrolysis, the process can be generally described as follows: upon heating, the solid PP structure begins to break down, releasing oil and increasing the hydrocarbon content in the resulting pyrolysis oil. Chemically, pyrolysis oil consists of a complex mixture of various hydrocarbon compounds. Although several studies have explored how pyrolysis temperature affects the characteristics of products obtained from PP, research on the decomposition kinetics of PP into different products under varying temperatures and residence times remains limited. Therefore, this study aims to investigate the effects of temperature and reaction time on the decomposition behavior of PP plastic, with a particular focus on optimizing pyrolysis oil production.

2. MATERIALS AND METHODS

2. 1. Preparation of Feedstock The PP plastics were obtain from waste collection of Mechanical Engineering Department of Riau University. The PP waste was crushed and screened to a particle size of 1 mm.

2. 2. Experimental Details The experiment was conducted in a stainless-steel batch vessel equipped with an electric heater, condenser, thermocouple, and a temperature controller (Figure 1).

The reactor has an inner diameter (ID) of 0.15 m, OD of 0.16 m, and height of 0.36 m. An electric band heater

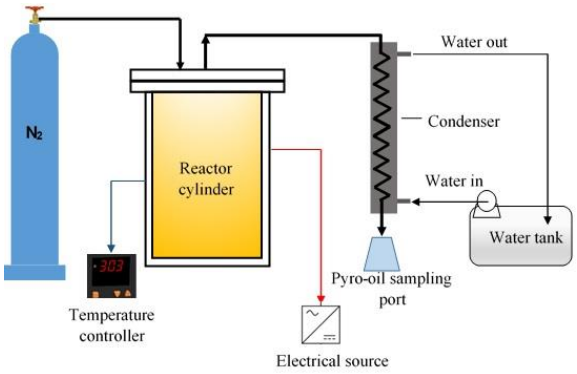


Figure 1. Schematic of experimental apparatus (pyrolisis reactor)

covers the outside of the vessel to heat and maintain a uniform temperature inside the reactor chamber during the experiments. The vessel was loaded with 500 g of PP plastics. Prior to experimental procedure, nitrogen was introduced into the reactor at 400 mL·min^{-1} for 15 minutes to displace residual oxygen. The quality of the resulting pyro-oil is affected by the presence of oxygen in the reactor. Reduced oxygen levels improve pyrolysis efficiency, thereby increasing the production of condensable volatiles (26). The pyrolisis was conducted at temperatures of 300, 350, and 400 °C and was timed at different intervals of 30, 60, and 90 min. As pyrolysis progressed, the vapors were condensed and routed to a liquid sampling port for collection. The reactor was cooled after the reaction, and the formed char was subsequently collected and measured using a digital scale. All experiments were conducted in duplicate to ensure the accuracy. The yield of each product ($Y_{pyro-oil}$ and Y_{char}) was calculated as a ratio of product mass ($m_{pyro-oil}$ and m_{char}) to the original mass of feedstock ($m_{feedstock}$). The yield of non-condensable gases was quantified indirectly, calculated by difference between the initial feedstock and the combined yield of the collected liquid and solid, as expressed in Equations 1-3:

$$Y_{pyro-oil} = \frac{m_{pyro-oil}}{m_{feedstock}} x\ 100\ \% \quad (1)$$

$$Y_{char} = \frac{m_{char}}{m_{feedstock}} x\ 100\ \% \quad (2)$$

$$Y_{gas} = 100\ \% - (Y_{pyro-oil} + Y_{char}) \quad (3)$$

2. 3. Product Characterization Calorific value analysis of the pyro-oil was conducted using a Bomb Calorimeter (Shandong Co. Ltd, BCY-1A). The density of each liquid product was determined using a pycnometer based on a comparative mass-to-volume approach. Initially, the volume of the pycnometer was established by filling it with distilled water and recording the mass at a controlled temperature of 25 °C, at which the density of water is well defined. Using this reference, the actual volume of the pycnometer was calculated. Subsequently, the pycnometer was filled with the liquid sample, and the mass was measured. Sample density was calculated from the measured mass and the calibrated volume of the pycnometer. The density (ρ) was computed using the following equation (Equation 4):

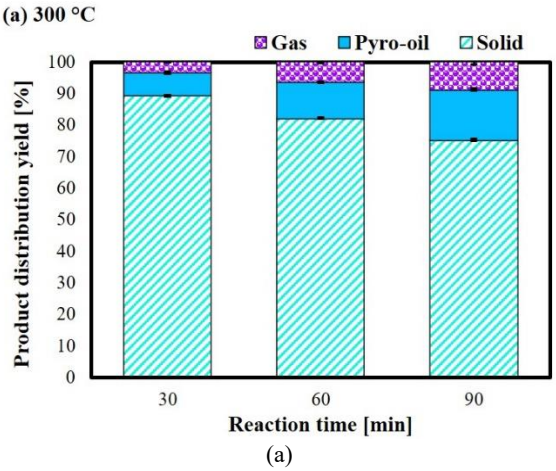
$$\rho = \frac{\text{mass of pyro oil in picnometer}}{\text{volume of pyro oil in picnometer}} \quad (4)$$

A GC-MS system equipped with an RTx-5MS capillary column was employed to analyze the composition of the pyro-oil. The gas chromatograph was initially held at 150 °C for 5 min, followed by a temperature increase of 10 °C per min until reaching 300 °C. The final temperature was then maintained at 300 °C for 26 min in the oven. To determine the pyro-oil components, the acquired mass spectra were compared against the NIST 2008 c2.0/X-Calibur data library, and compound identification was confirmed using standard reference solutions.

3. RESULTS AND DISCUSSION

3. 1. Effect of Reaction Temperature and Time on Product Composition To assess the effect of temperature variation and reaction time on pyrolysis product fractions of polypropylene (PP), a set of temperatures of 300, 350, and 400 °C and residence times of 30, 60, and 90 min were employed.

A clear pattern was observed, indicating a substantial rise in oil yield with increasing temperature between 300 and 400 °C (Figure 2). At elevated temperatures, the thermal degradation of PP is predominantly governed by random chain scission mechanisms, whereby the C–C bonds cleavage along the polymer chain is significantly accelerated. This enhanced rate of chain scission



(a)

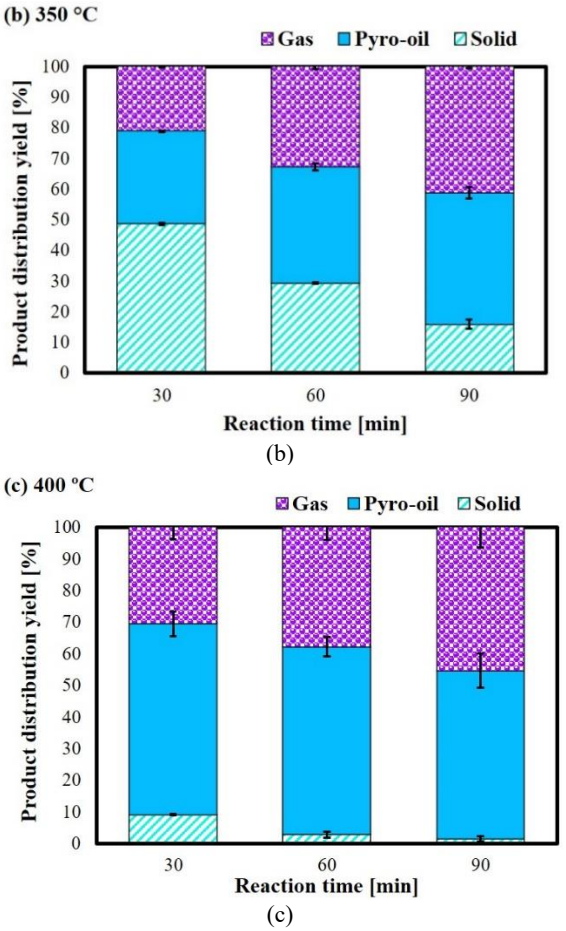


Figure 2. Effect of temperature and reaction time on pyrolisis product of polypropylene plastic at (a) 300 °C, (b) 350 °C, and (c) 400 °C

facilitates the breakdown of long polymer chains into lower molecular weight hydrocarbon fragments. These smaller, more volatile compounds readily undergo vaporization under pyrolysis conditions and subsequently condense into liquid oil fractions upon cooling (27). Consequently, higher pyrolysis temperatures not only promote greater conversion efficiency but also favor the production of oil over solid residues. Conversely, char formation declined markedly with increasing temperature. At low temperatures, incomplete depolymerization leads to heavy residues and char due to insufficient energy for complete fragmentation. These findings are consistent with previous observations of products obtained from pyrolysis of PP (28-30).

Pyro-oil generation at 300 °C resulted in comparatively low yields, with values of 7.2%, 11.4%, and 16.09% obtained at reaction times of 30, 60, and 90 min, respectively, suggesting that pyrolysis of polypropylene was incomplete under these conditions. At a higher temperature of 350 °C, pyro-oil production increased significantly (30.20 %, 38.06 %, and 43.00 % at reaction times of 30, 60, and 90 min, respectively). This result indicates that higher temperatures favor full decomposition to volatiles rather than partial carbonization residues. Longer reaction times provide sufficient duration for heat transfer and thermal cracking reactions to proceed toward completion. Polypropylene chains require both adequate temperature and time to break down into smaller, volatile hydrocarbon fragments. However, a substantial increase in gas formation was also observed at this temperature, with gas yields reaching 41.10 % at a reaction time of 90 min. This behavior is attributed to the increased intensity of thermal cracking reactions under these conditions. Elevated temperatures provide sufficient energy not only to depolymerize polypropylene into liquid-range hydrocarbons but also to further break down these intermediates into lighter gaseous compounds. Moreover, extended residence times facilitate secondary cracking reactions, wherein the initially formed liquid products undergo additional thermal decomposition, thereby increasing gas yields. The greater generation and stability of free radicals at higher temperatures also promote extensive chain scission and fragmentation, favoring the formation of volatile gases (31). Consequently, the thermodynamic equilibrium of the pyrolysis process shifts toward the production of permanent gases as heavy and intermediate fractions continue to degrade with prolonged exposure.

Interestingly, at the highest temperature, char formation was suppressed by 9.21 %, 2.87 %, and 1.52 % at residence times of 30, 60, and 90 min. This was accompanied by a significant increase in bio-oil of 60.29 %, 59.33 %, and 53.18 % at reaction times of 30, 60, and 90 min. This phenomenon can be explained by temperature-dependent reaction pathways during polypropylene pyrolysis. At very high temperatures, thermal degradation is sufficiently intense to promote almost complete depolymerization of the polymer chains into volatile products, thereby minimizing the formation of thermally stable char residues. The elevated temperature favors rapid chain scission and volatilization without allowing sufficient time for cross-linking and aromatization reactions that typically stabilize carbonaceous char structures (32).

Although char formation decreased and pyro-oil yield increased markedly at 400 °C, the non-condensable gas fraction exhibited only slight variation compared to 350 °C. This outcome can be explained by the balance between vapor-phase reactions and reactor conditions. At elevated temperatures, primary chain scission generates a greater quantity of volatile hydrocarbons; however, under the residence times employed, these intermediates predominantly condense as liquid products rather than undergoing further secondary cracking into permanent gases. The slight increase in gas yield therefore indicates a kinetic limitation, wherein the rate of secondary

cracking requiring higher activation energies and extended vapor-phase contact is insufficient to compete with rapid volatilization and condensation processes.

The results observed were compared to those of other studies as shown in Table 1.

3. 2. Effect of Reaction Temperature and Time on Pyro-oil Density Pyro-oil produced from slow pyrolysis of polypropylene plastic was examined for physicochemical characteristics. Figure 3 illustrates the density of pyrolysis oil obtained from PP at temperatures of 300°C, 350°C, and 400°C with reaction times of 30, 60, and 90 min, compared to biodiesel as a reference. The results show that at 300°C, the oil density remains consistently at approximately 730 kg m^{-3} across all reaction times. Increasing the temperature to 350°C yields a slightly higher and stable density of around 740 kg m^{-3}, while at 400°C, the density further increases to approximately 760 kg m^{-3}.

Variation in reaction time within the studied range (30–90 min) does not significantly influence oil density at each temperature, suggesting that temperature is the key factor controlling pyrolysis oil density. When compared to biodiesel, which has a density of about 850 kg m^{-3}, the pyrolysis oils exhibit lower densities, indicating a prevalence of lighter hydrocarbon fractions in the oil products. The observed increase in density with temperature implies that higher pyrolysis temperatures favor secondary cracking reactions, leading to the formation of heavier hydrocarbon chains or aromatic compounds in the oil. This finding indicates that while pyrolysis oil derived from PP has potential as an alternative fuel source, its lower density relative to biodiesel may necessitate further upgrading or blending to meet standard fuel specifications, though its intermediate density profile suggests suitability as a feedstock for further refining into lighter fuel fractions or chemical intermediates.

3. 3. Effect of Reaction Temperature and Time on Calorific Value The calorific value characteristics of pyro-oil from PP pyrolysis are shown in Figure 4. Overall, the data indicate a clear trend where increasing temperature results in higher calorific values. At 300°C, the calorific value ranges from approximately 44 MJ kg^{-1} at 30 min to about MJ kg^{-1} at 90 min, indicating a moderate increase with extended reaction time. At350°C, the calorific value is consistently higher, remaining around 47 MJ kg^{-1} across all reaction times with minimal variation, suggesting that reaction time has little influence at this temperature.

At 400°C, the highest calorific values are observed, ranging from approximately 48 MJ kg^{-1} at 30 min to nearly 49 MJ kg^{-1} at 90 min, demonstrating both the beneficial impact of elevated temperature and a slight increase with longer reaction time. This trend suggests

TABLE 1. Oil yields obtained from PP pyrolysis

Plastic type	Reactor	Catalyst	Temperature (°C)	Yield of pyrolysis product (wt%)			Reference
				Oil	Solid	Gas	
PP	Batch	-	400	60.29	9.21	30.5	This Study
PP	Batch	-	500	70	5.7	24.3	Eldahshory et al. (33)
PP	Batch	Kaolin	500	80.75	1.7	17.55	Eldahshory et al. (33)
PP	Micropyrolyzer	-	600	68	1.64	30.36	Supriyanto et al. (34)

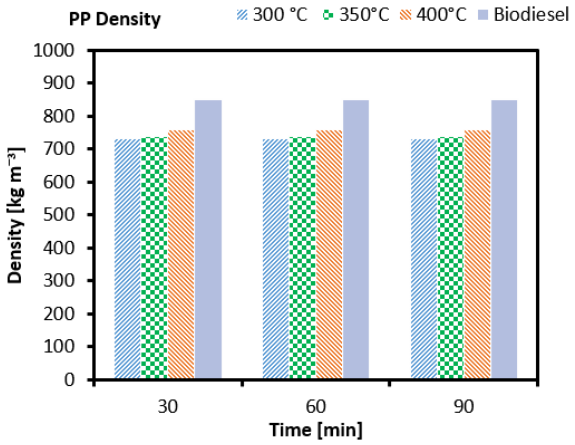


Figure 3. Density of pyro-oil compared to biodoesel

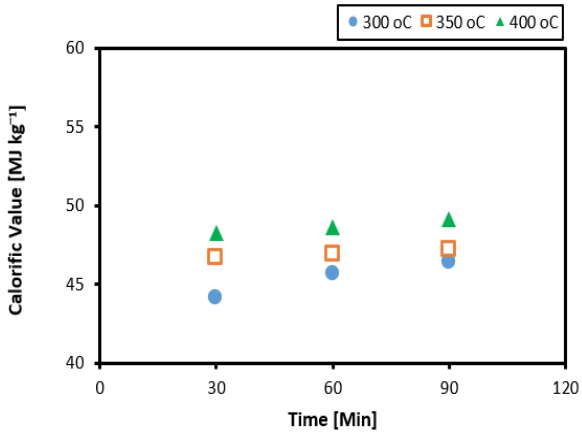


Figure 4. Calorific value of pyrolisis oil from PP

that higher temperatures favor the formation of oil with greater energy content, likely due to enhanced cracking and deoxygenation reactions producing hydrocarbons with higher heating values. Furthermore, the minimal influence of reaction time at 350°C demonstrates that temperature has a greater influence in determining calorific value than prolonged heating within the studied duration. These findings imply that conducting pyrolysis at 400°C is optimal for maximizing the energy density of PP-derived pyrolysis oil, enhancing its potential application as a fuel alternative. These findings are consistent with those reported in earlier studies (35, 36).

When compared to conventional diesel fuel, which typically has a calorific value in the range of 42–46 MJ kg^{-1} (37), the pyrolysis oil derived from PP in this study demonstrates comparable or even higher energy content. Notably, at 400°C, the calorific value of pyrolysis oil further increases, ranging from approximately 48 MJ kg^{-1} to nearly 49 MJ kg^{-1}, which is higher than the standard calorific value of diesel fuel. This indicates that PP-derived pyrolysis oil, especially at 400°C, possesses an energy density that exceeds that of conventional diesel, highlighting its potential as a high-energy alternative fuel. The elevated calorific values at higher temperatures suggest that increased thermal cracking enhances hydrocarbon quality by producing lighter and more energy-dense fractions. However, it is essential to consider that despite these promising calorific values, the direct utilization of pyrolysis oil as a diesel substitute may still require further upgrading to address issues such as viscosity, stability, and contaminant removal to meet fuel standards.

3. 4. Pyro-Oil Composition The pyro-oil elucidated from pyrolysis of PP plastics was quantified through GC/MS analysis. The significant differences in pyro-oil composition derived from pyrolysis of polypropylene were observed across different temperatures, as presented in Table 2.

Hydrocarbon and alcohol were the two most prevalent compounds found in pyro-oil of polypropylene. The formation of hydrocarbons is predominantly driven by random chain scission and radical reactions in which the C–C bonds in the polypropylene backbone undergo homolytic cleavage, generating alkyl radicals. These radicals undergo β-scission, resulting in the generation of lower-molecular-weight hydrocarbons, primarily alkanes and alkenes (38). For example, the scission of a polypropylene chain can yield molecules such as propene, hexene, or nonane. The primary and secondary radicals generated may undergo recombination to form branched or cyclic hydrocarbons. Alternatively, disproportionation reactions can produce an alkane and an alkene pair (39). Alcohols are generated via side reactions involving oxygen-containing species or radical recombination with trace hydroxyl sources.

Figure 5 reveals that hydrocarbons constituted the major fraction across all temperatures. This trend suggests that at lower temperatures, PP primarily undergoes random β-scission, generating aliphatic and cyclic hydrocarbons as dominant products. However, with increasing temperature, these hydrocarbons are likely subjected to secondary cracking or oxidation reactions, reducing their relative abundance.

TABLE 2. Major chemical constituents of pyro-oil from PP pyrolysis

No	Compounds	R.Time	Temperature (°C)			Chemical Formula	Group
			300	350	400		
			% Area				
1	1,5-Hexadiene, 2,5-dimethyl (CAS)	3.096	1.01	na	na	C_8H_{14}	Hydrocarbon
2	Cyclopentane, 1,1,3,4-tetramethyl-, cis-(CAS)	3.67	1.49	1.22	na	C_9H_{18}	Hydrocarbon
3	DIMETHYL-2,3 HEPTENE-3	4.066	4.5	4.9	1.51	C_9H_{18}	Hydrocarbon
4	1-Undecene (CAS)	4.418	15.94	13.71	6.79	$C_{11}H_{22}$	Hydrocarbon
5	Cyclohexane, 1,2,3-trimethyl-(CAS)	4.727	2.88	2.66	2.46	C_9H_{18}	Hydrocarbon
6	3-Octene, 2,2-dimethyl-(CAS)	5.686	2.79	2.31	1.81	$C_{10}H_{20}$	Hydrocarbon
7	trans-3,8-undecadiene	6.444	1.21	1.18	na	$C_{11}H_{20}$	Hydrocarbon
8	NONANE, 2,6-DIMETHYL	11.078	1.57	1.27	na	$C_{11}H_{24}$	Hydrocarbon
9	DECANE, 4-METHYL-	11.296	1.72	1.38	na	$C_{11}H_{24}$	Hydrocarbon
10	2-DECENE, 2,4-DIMETHYL-	13.552	na	1.2	1.3	$C_{12}H_{24}$	Hydrocarbon
11	3-OCTADECENE	14.593	7.25	6.34	5.15	$C_{18}H_{36}$	Hydrocarbon
13	1-Tridecanol	16.665	1.31	na	na	$C_{13}H_{28}O$	Alcohol
14	(2,4,6-Trimethylcyclohexyl) methanol	17.748	1.9	1.42	2.05	$C_{10}H_{20}O$	Alcohol

15	3-DECENE, 2,2-DIMETHYL-, (E)-	22.677	1.86	2.06	1.73	$C_{12}H_{24}$	Hydrocarbon
16	Cyclohexane, eicosyl- (CAS)	23.592	1.94	1.95	1.62	$C_{26}H_{54}$	Hydrocarbon
17	1-Nonadecene (CAS)	23.91	11.08	10.71	9.39	$C_{19}H_{38}$	Hydrocarbon
18	Tridecanol (CAS)	24.181	5.98	5.79	4.77	$C_{13}H_{28}O$	Alcohol
19	1-Nonadecene (CAS)	24.524	8.89	8.66	7.4	$C_{19}H_{38}$	Hydrocarbon
20	1-Hexadecanol, 3,7,11,15-tetramethyl- (CAS)	25.358	1.16	1.25	1.23	$C_{20}H_{42}O$	Alcohol
21	hydroxymethylcyclododecane	26.314	1.72	2.09	2.73	$C_{13}H_{26}O$	Alcohol
22	2-Cyclohexylnonadecane	29.975	0.99	na	1.35	$C_{25}H_{50}$	Hydrocarbon
23	1-Hexadecanol, 3,7,11,15-tetramethyl- (CAS)	31.037	3.52	4.28	4.33	$C_{20}H_{42}O$	Alcohol
25	1-Hexadecanol, 3,7,11,15-tetramethyl- (CAS)	31.873	1.98	2.25	2.38	$C_{20}H_{42}O$	Alcohol
26	1-Eicosanol (CAS)	32.196	1.31	1.73	2.43	$C_{20}H_{42}O$	Alcohol
27	(2,4,6-Trimethylcyclohexyl) methanol	33.202	1.52	2.32	3.27	$C_{10}H_{20}O$	Alcohol
28	2-Cyclohexylnonadecane	36.267	na	na	1.84	$C_{25}H_{50}$	Hydrocarbon
29	1-Hentetracontanol (CAS)	37.225	2.91	na	na	$C_{41}H_{84}O$	Alcohol
33	11-Dodecen-1-ol, 2,4,6-trimethyl-, (R,R,R)- (CAS)	39.181	1.06	1.87	3.02	$C_{15}H_{30}O$	Alcohol
34	2-Cyclohexylnonadecane	41.853	na	na	1.22	$C_{25}H_{50}$	Hydrocarbon
35	Cyclohexane, 1,2,3,5-tetraisopropyl-	44.493	na	1.46	na	$C_{18}H_{36}$	Hydrocarbon
36	1-Hentetracontanol (CAS)	42.72	na	na	2.78	$C_{41}H_{84}O$	Alcohol
39	Cyclohexane, 1,2,3,5-tetraisopropyl-	44.526	na	na	2.7	C_6H_{12}	Hydrocarbon
40	Tetrapentacontane, 1,54	47.709	na	na	1.23	$C_{54}H_{11}$	Hydrocarbon
42	Cyclohexane, 1,2,3,5-tetraisopropyl-	49.369	na	na	1.79	C_6H_{12}	Hydrocarbon

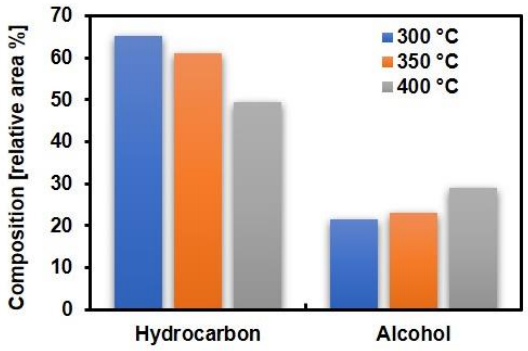


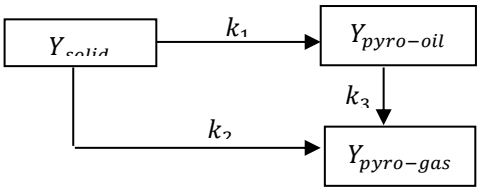
Figure 5. General composition of pyro-oil at different temperatures

Interestingly, the alcohol content exhibits a marked increase with temperature, rising from 24.1% at 250 °C to 28.7% at 350 °C. This trend supports the hypothesis that alcohols are primarily formed through secondary radical recombination reactions, possibly involving hydrogen abstraction or hydroxylation of intermediate alkyl radicals. The increasing alcohol yield at elevated temperatures may also be ascribed to the decomposition of peroxides or hydroperoxides formed from trace oxygen or water contaminants during pyrolysis.

Overall, the data demonstrate that increasing pyrolysis temperature shifts the product distribution from hydrocarbons to more oxygenated species, particularly alcohols. This shift reflects the thermochemical transformation pathways of polypropylene, wherein radical-mediated reactions and oxidative fragmentations become more prominent with temperature.

3. 5. Reaction Kinetics of Polypropylene Degradation A thorough understanding of reaction rate constants is essential for studying the decomposition behavior of plastics under thermochemical treatment. In many studies, plastic degradation has been described by first-order kinetics to determine the reaction rate constant. Safavi et al. (40) developed reaction models aimed at optimizing the plastic pyrolysis process. Three different lumped models, integrating both parallel and competing reaction pathways, were assessed against experimental results obtained across a wide range of temperatures. They explicitly applied a first-order kinetic model to determine activation energies for most plastics. The behavior of polypropylene under pyrolysis can be formulated by a model representing solid degradation as follows:

In this context, Y_{solid} represents the initial mass fraction of polypropylene in its solid form. The variable



Scheme 1. PP pyrolsis reaction kinetic model

$Y_{pyro-oil}$ and $Y_{pyro-gas}$ denote the fractions of the solid polymer that have been converted into pyrolytic oil and gas, respectively The proposed reaction pathway further indicates that a portion of the pyro-oil undergoes secondary conversion to gaseous products, while a fraction of the produced gas may subsequently condense and contribute to the formation of pyro-oil.

Assuming a first-order reaction, the rate of product change can be expressed using the following formula (Equations 5-7):

$$\frac{dY(solid)}{dt} = -(k_1 + k_2)[solid] \quad (5)$$

$$\frac{dY(pyro-oil)}{dt} = k_1[solid] - k_3[pyro-oil] \quad (6)$$

$$\frac{dY(pyro-gas)}{dt} = k_2[solid] + (k_3)[pyro-gas] \quad (7)$$

where, k denotes the reaction rate constant [s^{-1}], t represents the reaction time [s], and Y(X) corresponds to the product yield.

The rate constants were determined via non-linear regression using the least-squares-error (LSE) approach, a standard method that reduces the discrepancy between experimental data and model predictions to achieve optimal fitting. The fitting curve is presented in Figure 6, while the corresponding rate constant values are listed in Table 2.

Close correlation was observed between the experimental results and the calculated values. Figure 7 displays the parity plots for each product yield, confirming a strong correlation between the calculated and experimental results.

The Arrhenius formula (Equation 8) ws used to determined activation energy (E_a) and pre-exponential factor (A).

$$k = A\, e^{\left(\frac{-E_a}{RT}\right)} \quad (8)$$

where T is the reaction temperature [K], R is the universal gas constant [8.314 J/mol·K]. The Arrhenius plots provide valuable insight into the influence of temperature on the reaction rate constants during PP pyrolysis.

The fitted activation energies and pre-exponential factors provide mechanistic insight into PP conversion pathways. The relatively high E_a for solid → pyro-oil (99 kJ·mol^{-1}) indicates that formation of stable liquid-range

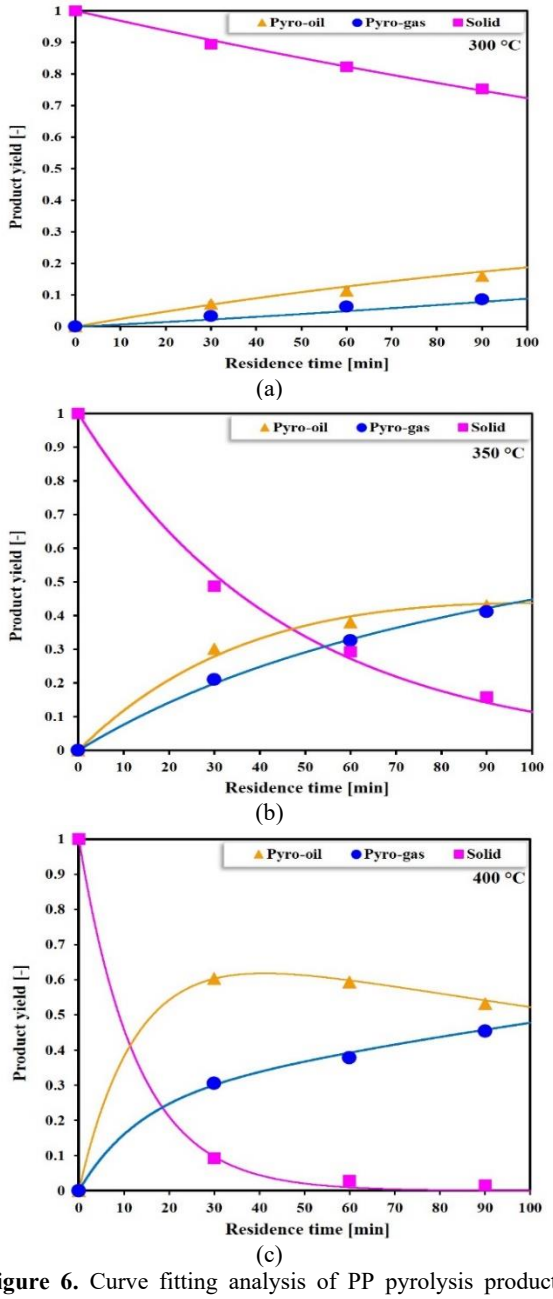


Figure 6. Curve fitting analysis of PP pyrolysis product yields at varying temperatures (a) 300 °C, (b) 350 °C, (c) 400 °C

fragments requires substantial energetic input, whereas lower E_a values for secondary cracking reflect more facile fragmentation pathways.

Despite its potential as a renewable energy source, PP-derived pyro-oil cannot be utilized directly as a drop-in transportation fuel due to several limitations such as physicochemical properties and chemical impurities. To overcome these limitations, upgrading strategies are essential. Distillation can fractionate the oil into lighter and middle distillates, improving homogeneity.

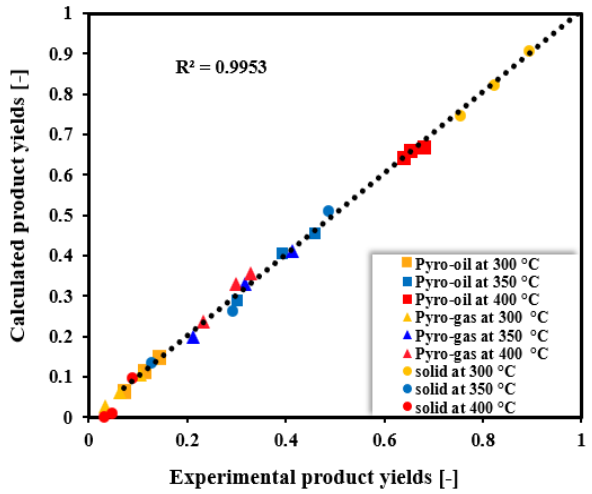


Figure 7. Parity plot of PP pyrolysis

Hydrotreating and hydrocracking effectively remove oxygenates and saturate olefins, producing stable paraffinic fuels with higher calorific value and improved cetane number. Alternatively, catalytic cracking or reforming can convert heavy fractions into lighter hydrocarbons or aromatic compounds suitable for gasoline or jet-fuel blending. Solvent extraction and adsorption methods may also be applied to reduce corrosive oxygenates and chlorinated compounds. In the short term, controlled blending with diesel or biodiesel could allow partial utilization.

These considerations highlight that, while PP pyro-oil shows promise as an energy carrier, its direct use is constrained by fuel property limitations. Consequently, coupling pyrolysis with appropriate upgrading steps is indispensable to realize its potential as a sustainable fuel alternative.

TABLE 2. Kinetic Parameters, Activation Energy, and Pre-Exponential Factor of Each Reaction of Polypropylene conversion via pyrolysis (Experimental Conditions: 300−400 °C, residence time 30-90 min)

kinetic parameter	Reaction	k [s^{-1}]			Activation energy, E_a [kJ mol^{-1}]	Pre-exponential factor, A [s^{-1}]
		300 °C	350°C	400°C		
k_1	solid → pyro-oil	2.5 x 10^{-3}	13.4 x 10^{-3}	5.60 x 10^{-2}	99.12	4.57 x 10^{4}
k_2	solid → pyro-gas	0.7 x 10^{-3}	8.3 x 10^{-3}	21.9 x 10^{-3}	111.85	2.18 x 10^{5}
k_3	pyro-oil → pyro-gas	2.9 x 10^{-3}	3.40 x 10^{-3}	3.70 x 10^{-3}	8.23	2.71 x 10^{-4}

4. CONCLUSION

These findings suggest that increasing the pyrolysis temperature significantly enhances the conversion efficiency of polypropylene by promoting extensive thermal degradation and reducing char formation, thereby maximizing pyro-oil yields. However, while moderate temperature increases (e.g., to 350°C) enhance both oil and gas production due to intensified cracking reactions, further temperature elevation results in a preferential formation of liquid hydrocarbons over gases, indicating an optimal temperature window where oil yield is maximized while gas formation slightly increases. Additionally, the suppression of char formation at higher temperatures highlights the importance of sufficient thermal energy to overcome cross-linking and aromatization pathways responsible for char stabilization. A first-order reaction with an Arrhenius-type rate constant was employed to elucidate the waste decomposition mechanism. Overall, these results indicate that both temperature and reaction time are critical parameters in steering product distribution toward desirable liquid fuels in polypropylene pyrolysis processes.

Acknowledgments

The authors gratefully acknowledge the Thermal Engineering Laboratory, Riau University for providing experimental apparatus, and the Directorate of Research and Community Service, Ministry of Higher Education, Science and Technology (DPPM, Kemendiktisaintek 2025) for funding this research [Grant no. 102/C3/DT.05.00/PL/2025, no. 19571/UN19.5.1.3/AL.04/2025].

Funding

This work was supported by Directorate of Research and Community Service, Ministry of Higher Education, Science and Technology (DPPM, Kemendiktisaintek 2025) [Grant no. 102/C3/DT.05.00/PL/2025, no. 19571/UN19.5.1.3/AL.04/2025].

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing to improve readability. After using this tool, the author carefully reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Author(s) Biosketches

Rahmat Iman Mainil is an Associate Professor in Mechanical Engineering, Riau University with a focus on renewable energy systems, thermal management, and energy conversion technologies. His research interests include thermochemical conversion, photovoltaic–thermal systems, heat transfer, and sustainable energy applications.

Junior Jonathan Panjaitan is a former Mechanical Engineering student at Riau University.

Annisa Wulan Sari is a lecturer in the Department of Mechanical Engineering at Riau University. She is involved in several research projects related to energy conversion.

Anita Susilawati is an Associate Professor in Mechanical Engineering, Riau University specializing in manufacturing technology. Her research interests include advanced manufacturing processes, materials processing, and sustainable manufacturing systems.

Afdhal Kurniawan Mainil is an Associate Professor in Mechanical Engineering at the University of Bengkulu, Indonesia. His research interests include heat transfer, refrigeration and air-conditioning, and thermal systems, with contributions to international scientific publications.

Obie Farobie is a Lecturer and Researcher at IPB University. His research focuses on renewable energy, biomass conversion, and biofuel production, with publications in international journals.

Apip Amrullah is an Associate Professor in Lambung Mangkurat University specializing in biomass utilization and energy conversion. His work focuses on kinetic analysis and thermal processes for renewable energy applications.

Azridjal Aziz is a Professor of Mechanical Engineering, Riau University specializing in cooling and refrigeration systems, solar energy, and thermal systems. His research includes heat transfer, energy efficiency, and sustainable energy technologies.

REFERENCES

1. Jambeck JR, Geyer R, Wilcox C, Siegler TR, Perryman M, Andrady A, et al. Plastic waste inputs from land into the ocean. science. 2015;347(6223):768-71.
2. Syamsiro M, Saptoadi H, Norsujianto T, Noviasri P, Cheng S, Alimuddin Z, et al. Fuel oil production from municipal plastic wastes in sequential pyrolysis and catalytic reforming reactors. Energy procedia. 2014;47:180-8. https://doi.org/10.1016/j.egypro.2014.01.212
3. Wong S, Ngadi N, Abdullah TAT, Inuwa IM. Current state and future prospects of plastic waste as source of fuel: A review. Renewable and sustainable energy reviews. 2015;50:1167-80. https://doi.org/10.1016/j.rser.2015.04.063
4. Sharuddin SDA, Abnisa F, Daud WMAW, Aroua MK. A review on pyrolysis of plastic wastes. Energy conversion and management. 2016;115:308-26. https://doi.org/10.1016/j.enconman.2016.02.037
5. Hendiarti N. Combating marine plastic debris in Indonesia. Dipresentasikan pada Science to Enable and Empower Asia Pacific for SDGs (Jakarta, 30 Juli 2018). 2018. http://www.unesco.or.id/publication/SC_Retreat/4_MarineDebrisIndonesia.pdf
6. Rafey A, Siddiqui FZ. A review of plastic waste management in India–challenges and opportunities. International Journal of Environmental Analytical Chemistry. 2023;103(16):3971-87. https://doi.org/10.1080/03067319.2021.1917560
7. Panda AK, Singh RK, Mishra D. Thermolysis of waste plastics to liquid fuel: A suitable method for plastic waste management and manufacture of value added products—A world prospective. Renewable and Sustainable Energy Reviews. 2010;14(1):233-48. https://doi.org/10.1016/j.rser.2009.07.005
8. Miandad R, Barakat M, Aburiazaiza AS, Rehan M, Ismail I, Nizami A. Effect of plastic waste types on pyrolysis liquid oil. International biodeterioration & biodegradation. 2017;119:239-52. https://doi.org/10.1016/j.ibiod.2016.09.017
9. Yahaya H, Hamdan H, Dolah R, Muhid N, Rashidi KA, Rashidi SB. Catalytic fast-pyrolysis process design and equipment setup for converting palm oil empty fruit bunch biomass to bio-oil. Journal of Advanced Research in Fluid Mechanics and Thermal Sciences. 2017;32:1-8.
10. Dai L, Wang Y, Liu Y, Ruan R, He C, Yu Z, et al. Integrated process of lignocellulosic biomass torrefaction and pyrolysis for upgrading bio-oil production: A state-of-the-art review. Renewable and Sustainable Energy Reviews. 2019;107:20-36. https://doi.org/10.1016/j.rser.2019.02.015
11. Mainil RI, Panjaitan JJ, Mainil AK, Putra H, Aziz A, editors. Oil production from polypropylene plastic via pyrolysis. AIP Conference Proceedings; 2025: AIP Publishing LLC.
12. Amrullah A, Farobie O, Bayu A, Syaftika N, Hartulistiyoso E, Moheimani NR, et al. Slow pyrolysis of ulva lactuca (chlorophyta) for sustainable production of bio-oil and biochar. Sustainability. 2022;14(6):3233. https://doi.org/10.3390/su14063233
13. Jiang L, Zhang D, Li M, He J-J, Gao Z-H, Zhou Y, et al. Pyrolytic behavior of waste extruded polystyrene and rigid polyurethane by multi kinetics methods and Py-GC/MS. Fuel. 2018;222:11-20. https://doi.org/10.1016/j.fuel.2018.02.143

14. Chen R, Li Q, Xu X, Zhang D. Comparative pyrolysis characteristics of representative commercial thermosetting plastic waste in inert and oxygenous atmosphere. Fuel. 2019;246:212-21. https://doi.org/10.1016/j.fuel.2019.02.129
15. Honus S, Kumagai S, Fedorko G, Molnár V, Yoshioka T. Pyrolysis gases produced from individual and mixed PE, PP, PS, PVC, and PET—Part I: Production and physical properties. Fuel. 2018;221:346-60. https://doi.org/10.1016/j.fuel.2018.02.074
16. Rabnawaz M, Wyman I, Auras R, Cheng S. A roadmap towards green packaging: the current status and future outlook for polyesters in the packaging industry. Green Chemistry. 2017;19(20):4737-53. https://doi.org/10.1039/c7gc02521a
17. Geyer R, Jambeck JR, Law KL. Production, use, and fate of all plastics ever made. Science advances. 2017;3(7):e1700782. https://doi.org/10.1126/sciadv.1700782
18. Geyer R. Production, use, and fate of synthetic polymers. Plastic waste and recycling: Elsevier; 2020. p. 13-32.
19. Pasae Y, Bulo L, Lande C, Allo EL. Pyrolysis of Polypropylene Plastic Waste: Overview of Effect of Temperature on Pyrolysis Reactors in Capacity of 1 Kg/batch. Journal Dynamic Saint. 2020;5(1):939-44.
20. Muchammad M. Analisis pemanfaatan limbah sampah plastik jenis polypropylene menjadi bahan bakar alternatif. Majalah Ilmiah Momentum. 2018;14(1). https://doi.org/10.36499/jim.v14i1.2189
21. Adoe DG, Bunganaen W, Krisnawi IF, Soekwanto FA. Pirolisis Sampah Plastik PP (Polyprophylene) menjadi Minyak Pirolisis sebagai Bahan Bakar Primer. LONTAR Jurnal Teknik Mesin Undana. 2016;3(1):17-26.
22. Uthpalani P, De Silva D, Weerasinghe V, Premachandra J, Sarathchandra T. Pyrolysis of waste LDPE and waste PP plastics into fuel oil in a low-cost, lab-scale pyrolyzing unit. Journal of Science of the University of Kelaniya. 2023;16(1). https://repositorio.ufsc.br/xmlui/bitstream/handle/123456789/167638/341506.pdf?sequence=1&isAllowed=y%0Ahttps://repositorio.ufsm.br/bitstream/handle/1/8314/LOEBLEIN%2C
23. Alit IB, Susana IGB, Mara IM. Conversion of LDPE and PP plastic waste into fuel by pyrolysis method. Global Journal of Engineering and Technology Advances. 2022;10(3):073-8. https://doi.org/10.30574/gjeta.2022.10.3.0055
24. Santoso A, Sholikhah A, Sumari S, Asrori M, Wijaya A, Retnosari R, et al. Effect of active zeolite in the pyrolysis of polypropylene and low density polyethylene types of plastic waste. Journal of Renewable Materials. 2022;10(11):2781. https://doi.org/10.32604/jrm.2022.021401
25. Morina M, Sidjabat O. CATALYTIC PYROLYSIS OF POLYPROPYLENE PLASTIC FROM MUNICIPAL SOLID WASTES INTO GASOLINE FRACTION COMPOUNDS USING THE MIXED NICKEL (Ni) AND CHROME (Cr) METALS AS THE CATALYST. Scientific Contributions Oil and Gas. 2014;37(3):129-40. https://doi.org/10.29017/scog.37.3.634
26. Varma AK, Thakur LS, Shankar R, Mondal P. Pyrolysis of wood sawdust: Effects of process parameters on products yield and characterization of products. Waste Management. 2019;89:224-35. https://doi.org/10.1016/j.wasman.2019.04.016
27. Harussani M, Sapuan S, Rashid U, Khalina A, Ilyas R. Pyrolysis of polypropylene plastic waste into carbonaceous char: Priority of plastic waste management amidst COVID-19 pandemic. Science of The Total Environment. 2022;803:149911. https://doi.org/10.1016/j.scitotenv.2021.149911
28. Cai W, Kumar R, Zhu Z, Varjani S, Hu Z-T, Wong JW-C, et al. Catalytic pyrolysis of polypropylene waste for liquid fuels production using Ni/Al-MOF-derived catalysts. Next Sustainability. 2024;4:100059. https://doi.org/10.1016/j.nxsust.2024.100059
29. Miandad R, Rehan M, Barakat MA, Aburiazaiza AS, Khan H, Ismail IM, et al. Catalytic pyrolysis of plastic waste: moving toward pyrolysis based biorefineries. Frontiers in energy research. 2019;7:437000. https://doi.org/10.3389/fenrg.2019.00027
30. Martynis M, Winanda E, Harahap A, editors. Thermal pyrolysis of polypropylene plastic waste into liquid fuel: reactor performance evaluation. IOP Conference Series: Materials Science and Engineering; 2019: IOP Publishing.
31. Sidhu N, Mastalski I, Zolghadr A, Patel B, Uppili S, Go T, et al. On the intrinsic reaction kinetics of polypropylene pyrolysis. Matter. 2023;6(10):3413-33. https://doi.org/10.1016/j.matt.2023.07.020
32. Aboulkas A, El Harfi K, El Bouadili A, Chanâa MB, Mokhlisse A. Pyrolysis kinetics of polypropylene: Morocco oil shale and their mixture. Journal of thermal analysis and calorimetry. 2007;89(1):203-9. https://doi.org/10.1007/s10973-007-7398-z
33. Eldahshory A, Emara K, Abd-Elhady M, Ismail M. Catalytic pyrolysis of waste polypropylene using low-cost natural catalysts. Scientific Reports. 2023;13(1):11766. https://doi.org/10.1038/s41598-023-37769-8
34. Ylitervo P, Richards T. Gaseous products from primary reactions of fast plastic pyrolysis. Journal of Analytical and Applied Pyrolysis. 2021;158:105248. https://doi.org/10.1016/j.jaap.2021.105248
35. Khair H, Listiany B, Utami R, editors. Pyrolysis of polypropylene plastic waste: An analysis of oil quantity, density, viscosity, and calorific value. IOP Conference Series: Earth and Environmental Science; 2023: IOP Publishing.
36. Ridwan A, Istana B, Abdurrahman R, Aprio T. Investigating Bio-Oil Yield from Co-pyrolysis of Mixed Medical Waste and Biomass under Varying Temperature and Residence Time. International Journal of Automotive and Mechanical Engineering. 2025;22(2):12293-303. https://doi.org/10.15282/ijame.22.2.2025.4.0941
37. Osueke C, Olayanju T, Ezugwu C, Onokwai A, Ikpotokin I, Uguru-Okorie D, et al. Comparative calorific evaluation of biomass fuel and fossil fuel. International Journal Civil Engineering and Technology. 2018;9(13):1576-90.
38. Matsushita M, Sakai M, Miura E, Omata T, Kamo T, Matsukata M. Polypropylene Decomposition in Hydrocarbon Solvents: Zeolite-Catalyzed Chemical Recycling. ACS Sustainable Resource Management. 2024;1(12):2539-46. https://doi.org/10.1021/acssusresmgt.4c00217
39. Ojha DK, Vinu R. Fast co-pyrolysis of cellulose and polypropylene using Py-GC/MS and Py-FT-IR. RSC Advances. 2015;5(82):66861-70. https://doi.org/10.1039/c5ra10820a
40. Safavi A, Richter C, Unnthorsson R. A study of parallel and competitive reaction schemes in kinetic modeling of plastic pyrolysis. ACS omega. 2024;9(4):4811-8. https://doi.org/10.1021/acsomega.3c08306

Persian Abstract

چکیده

این تحقیق، پیرولیز آهسته ضایعات پلاستیکی پلی‌پروپیلن (PP) را به عنوان یک روش عملی برای تولید پیرو–اویل بررسی می‌کند. هدف اصلی، تبدیل ضایعات پلیمری غیر قابل تجزیه زیستی به سوخت مایع با ارزش افزوده و در عین حال پرداختن به نگرانی‌های زیست‌محیطی مرتبط با آلودگی پلاستیک است. آزمایش‌های پیرولیز آهسته در یک راکتور ناپیوسته تحت دماهای تنظیم‌شده ۳۰۰، ۳۵۰ و ۴۰۰ درجه سانتیگراد با زمان‌های واکنش ۳۰، ۶۰ و ۹۰ دقیقه انجام شد تا رفتار تخریب حرارتی، توزیع محصول و کیفیت پیرو–اویل به‌دست‌آمده بررسی شود. بالاترین میزان تولید پیرو–اویل، معادل ۶۸ درصد، در دمای ۴۰۰ درجه سانتیگراد حاصل شد که نشان‌دهنده ترک‌خوردگی حرارتی بهینه زنجیره‌های پلی‌پروپیلن است. در این شرایط، پیرو–اویل محتوای هیدروکربنی ۴۹.۳ درصد را نشان داد که بر قابلیت آن به عنوان منبع جایگزین انرژی تأکید دارد. پیرو–اویل ارزش حرارتی در محدوده ۴۸ مگاژول بر کیلوگرم تا نزدیک به ۴۹ مگاژول بر کیلوگرم دارد. مسیرهای واکنش ممکن در این مطالعه تعیین شده‌اند، همانطور که توسط ضریب پیش نمایی و انرژی فعال‌سازی به‌دست‌آمده از معادله آرنیوس نشان داده شده است. این یافته‌ها امکان‌سنجی ارزش‌آفرینی ضایعات پلی‌پروپیلن از طریق پیرولیز آهسته را نشان می‌دهند و مسیری امیدوارکننده برای بازیابی منابع، اجرای اقتصاد چرخشی و کاهش اثرات زیست‌محیطی ضایعات پلاستیکی ارائه می‌دهند.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2407-2418

International Journal of Engineering

Journal Homepage: www.ije.ir

Condition Monitoring of Silicone Rubber Insulators Using Criteria Established through Wavelet Transform Analysis

F. Karamoddin, S. M. Seyyedbarzegar*, A. Samani

Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood, Iran

PAPER INFO

Paper history:
Received 04 July 2025
Received in revised form 15 October 2025
Accepted 25 December 2025

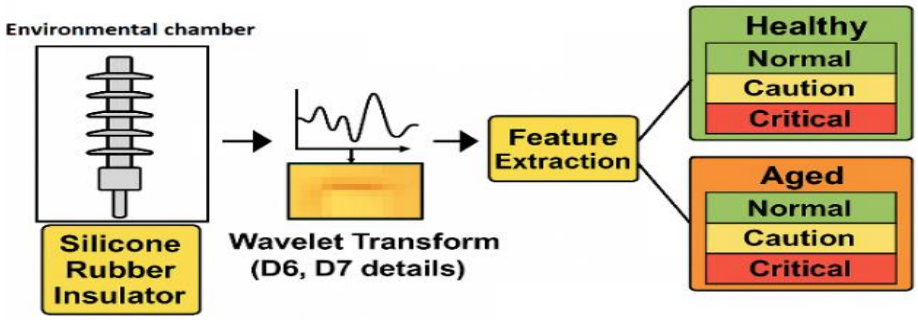
Keywords:
Aging
Leakage Current
Polymer Insulator
Uniform Pollution
Wavelet Transform

ABSTRACT

The condition assessment of silicon rubber insulators has always been a crucial requirement in power systems. Utilizing leakage current as a fast and online method has played a significant role in this regard. However, variations in operating conditions can lead to substantial errors in decisions regarding insulator condition. Under the influence of environmental conditions on the leakage current, it is difficult to define criteria capable of distinguishing different operating conditions. In this regard, achieving a model with minimal complexity for decision-making is considered as a suitable solution. In this study, an index based on leakage current is presented to categorize the performance of insulators into three states: normal, caution, and critical. For this purpose, experimental data were practically collected under light, medium, and heavy pollution conditions in varying humidity levels up to 90%. All tests were performed under different environmental conditions on healthy and aged insulators to investigate the effect of surface degradation on leakage current and their harmonic components. To analyze the leakage current, wavelet transform was employed, and the standard deviation of wavelet detail coefficients was used. Also, distinguishing a healthy insulator from one that has been aged by UV radiation based on detail coefficients is introduced in this paper. From the wavelet transform results, it appears that a D6 standard deviation above 0.01 reliably separates healthy insulators from aged ones. Furthermore, by using the standard deviation of D7 and defining threshold limits for healthy and aged insulators, it became possible to differentiate their performance. In order to evaluate the accuracy and sensitivity of the proposed indicators, the confusion matrix was used, which is able to distinguish different operating conditions with an accuracy of 89.58%.

doi: *10.5829/ije.2026.39.10a.05*

Graphical Abstract



1. INTRODUCTION

Insulators play a critical role in electrical power transmission systems by ensuring adequate insulation distances and withstanding the electromechanical stresses inherent in power systems. Given their direct influence on system reliability and operational stability, the monitoring and assessment of insulator conditions

* Corresponding author email: seyyedbarzegar@shahroodut.ac.ir (S. M. Seyyedbarzegar)

Please cite this article as: Karamoddin F, Seyyedbarzegar SM, Samani A. Condition Monitoring of Silicone Rubber Insulators Using Criteria Established through Wavelet Transform Analysis. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2407-18.

have attracted considerable research interest (1, 2). Among the various factors affecting insulator performance, environmental conditions are particularly significant, as they can compromise the operational integrity of insulators under both steady-state and fault scenarios. Humidity, pollution, solar radiation, and wind are key environmental parameters that markedly impact the functional behavior of insulators.

Under normal operating conditions, the coexistence of humidity and surface contamination can initiate leakage currents along the insulator surface. These currents can subsequently lead to surface discharges, accelerating degradation processes and potentially triggering electrical discharges. Such phenomena not only deteriorate the physical integrity of insulators but also pose severe risks to grid reliability, including short circuits and service interruptions (3-6). Moreover, partial discharges occurring on the insulator surface, influenced by the material composition, contribute to progressive aging and deterioration of dielectric properties, which adversely affect long-term performance (7, 8). Notably, the magnitude of leakage current tends to increase with surface degradation and the worsening of environmental factors, which promote the retention of moisture and pollutants (9, 10).

Numerous studies have demonstrated that leakage current characteristics can serve as diagnostic indicators for assessing the operational state of insulators. Consequently, the analysis of leakage current has been widely recognized as an effective diagnostic tool (11-14). Beyond the magnitude of leakage current, its harmonic components have also been shown to vary under changing environmental and operational conditions. Researchers have employed Fourier Transform (FT) and Fast Fourier Transform (FFT) techniques to investigate these harmonic variations (15-17). Such analyses have revealed strong correlations between harmonic behavior and environmental stressors such as humidity and pollution levels (18). For instance, literature review (19-22) evaluated the third, fifth, and seventh harmonic components under diverse contamination levels, establishing the potential of odd harmonic signatures to discriminate between different insulator conditions. Similarly, Total Harmonic Distortion (THD) has been utilized to assess the influence of environmental variations on harmonic content (23). While contamination and humidity remain critical factors influencing leakage current behavior, their impact is strongly modulated by the deposition pattern of pollutants on the insulator surface. Silicone rubber insulators (SIR) exhibit inherent hydrophobicity, which initially resists pollution adhesion. However, prolonged exposure to ultraviolet (UV) radiation results in surface aging and a gradual loss of hydrophobicity (24). This degradation facilitates pollutant accumulation, elevates leakage current levels, and significantly reduces flashover voltage (25, 26). Structural changes in the silicone rubber due to aging particularly alter the progression of harmonic components, notably the third and fifth harmonics (27). The study conducted by Seyyedbarzegar (28) highlighted how these harmonic changes can effectively differentiate aged insulators from their healthy counterparts, proposing relevant diagnostic indicators.

Although leakage current-based diagnostics are regarded as reliable and cost-effective, their performance is highly dependent on the employed signal processing methodologies (29). In this context, wavelet transform (WT) has emerged as a powerful analytical tool, offering simultaneous time–frequency localization capabilities that are particularly suited for non-stationary signals such as leakage currents. Several studies have demonstrated the effectiveness of wavelet-based techniques in characterizing contamination severity and discharge phenomena on insulator surfaces. For example, the STD-MRA method derived from wavelet analysis was employed by Douar et al. (16) to assess pollution severity on ceramic and polymeric insulators. Wavelet-based analyses were further used by several researchers (30-33) to investigate discharge behavior and extract relevant frequency components, enabling the detection of localized discharges and their spatial characteristics. Additionally, Narayanan et al. (34) proposed an advanced diagnostic system integrating Discrete Wavelet Transform (DWT) with fuzzy data mining for the assessment of pollution severity in high-voltage transmission systems.

This paper extends prior work by simultaneously investigating the influence of environmental factors such as humidity and pollution levels and UV-induced aging on insulator performance, utilizing wavelet-based analysis. We proposed a novel diagnostic indicator based on the standard deviation of wavelet detail coefficients and validate it through controlled laboratory experiments. Results show that this indicator effectively categorizes insulators as healthy or aged and further differentiates three operational states: normal, caution, and critical. This provides a robust framework for condition monitoring and preventive maintenance in power systems. The confusion matrix was employed to evaluate the accuracy and sensitivity of the proposed indicators. Figure 1 illustrates the overall method for insulator condition assessment based on features extracted from the wavelet transform.

2. EXPRIMENTAL SETUP

The high-voltage laboratory setup includes a 220 V / 100 kV transformer, with an output adjustable between 0 and 100 kV. To protect the transformer against short-circuit currents, a 10 MΩ resistor is connected in series with the

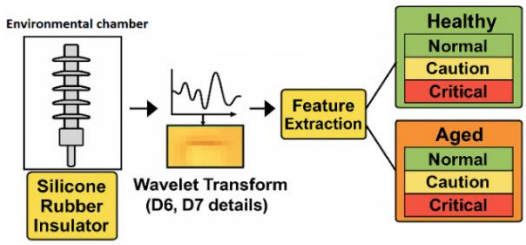


Figure 1. Overview of the insulator condition assessment method based on wavelet transform

transformer in the circuit. Additionally, to measure the leakage current, a pure resistor is connected in series with the insulator. The voltage across the resistor is measured and recorded using a digital oscilloscope with a sampling frequency of 10 kHz. Finally, the current is calculated by dividing the voltage by the resistance.

Figure 2 illustrates the high-voltage experimental setup and the equipment used. To investigate the effect of environmental conditions on leakage current, a standard chamber with dimensions of 2×2×2 meters is employed. The required humidity is provided using a fog generator and regulated by a dedicated controller to ensure precise control of the experimental conditions. To create a pollution layer on the insulator, the solid layer method based on IEC 60507 standard was used. In this method, a solution made from salt (NaCl), kaolin, and high-purity distilled water is sprayed onto the insulator.

To evaluate the electrical conductivity of the pollution layer on the insulator, the conductivity of the pollution solution was measured using a conductivity meter, and the equivalent salt deposit density (ESDD) was calculated based on IEC 60507 standards and according to Equation 1. Based on the obtained ESDD values, different pollution levels can be achieved, as outlined in Table 1.

$$ESDD = (Sa \times V)/A \quad (mg/cm^2) \tag{1}$$

$$Sa = (0.57\ \sigma_{20})\ .1.03 \quad (mg/cm^3) \tag{2}$$

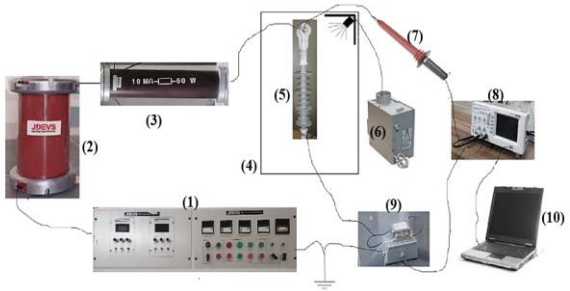


Figure 2. High voltage laboratory setup: 1) Control Desk, 2) Transformer, 3) Resistor, 4) Fog chamber, 5) Insulator, 6) Fog maker, 7) High voltage probe, 8) Digital oscilloscope, 9) Leakage current meter, 10) Computer

$$\sigma_{20}=\sigma_{\theta}\ (\ 1-b\ (\ \theta-20\ ))\quad S/cm \tag{3}$$

where S_a is the salinity of the distilled water, V is the volume of distilled water, A is the cleaned surface area (cm^2), σ_{20} is the conductivity at 20°C (S/m), σ is the measured conductivity at temperature θ, and b is the temperature coefficient.

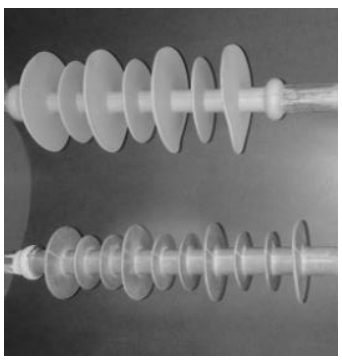
Figure 3 shows the appearance of the insulators used under light, medium, and heavy pollution levels. The structural specifications of these insulators are provided in Table 2.

For aging the insulator surface, a wooden chamber with dimensions of 2×2×2 meters and internal walls covered with aluminum sheeting was designed and constructed. The aging process considered the effects of UV radiation, humidity, voltage, and heat. A 40 W/m^2 UV lamp was used to simulate the effects of solar UV radiation, two heaters were used for heat generation, and a fog generator was utilized to produce fog and humidity during the aging process. Additionally, an automatic control system was developed to schedule the on/off times for each aging factor, such as humidity, heat, and UV. The tested insulators were placed inside the chamber for 3000 hours during the aging process. Table 3 details the aging procedure for the insulators under investigation.

Also, in order to show the effect of UV radiation on the insulator and the aging of its surface, Figure 4 has been prepared. As can be seen in this figure, UV radiation has greatly reduced the hydrophobicity of the insulator surface.

TABLE 1. Amount of kaolin, salt, and distilled water for different pollution levels

Pollution level	SDD (mg/cm^2)	Distilled water (l)	Salt (g/l)	Kaolin (g/l)
Light (LP)	0.01-0.04	1	20	40
Medium (MP)	0.04-0.15	1	40	40
Heavy (HP)	0.15-0.4	1	80	40



Type A

Type B

Figure 3. Shape of studied insulators for 20kV distribution power system

TABLE 2. Timing of factors affecting the aging of insulators in a 24-hour period

Duration (Hour)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Voltage																								
UV																								
Heater																								
Humidity																								

TABLE 2. Features of distribution class silicon rubber insulators

Insulator Type	A	B
Voltage Level	20 kV	20 kV
Height	270 mm	300 mm
Creepage Distance	698 mm	601 mm
Shed Number	7	10

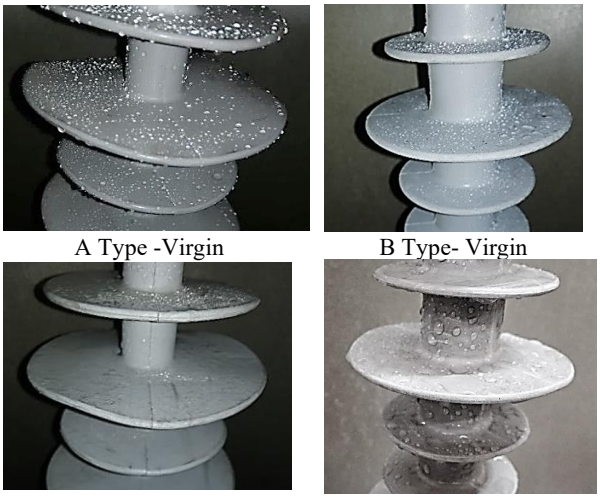


Figure 4. Comparison of new and old insulation based on the surface hydrophobicity of the polymer material

Figures 5 and 6 show the variations in leakage current passing through insulators of type A and B in both healthy and aged states at 80% humidity, under clean conditions and three levels of pollution: light, medium, and heavy. By comparing these waveforms, it can be seen that the increase in the level of pollution in constant humidity has caused an increase in the magnitude of the leakage current in all cases. In addition, aging also leads to the increase of the main component and the change in the waveform of the leakage current, which is caused by the change in the harmonic components.

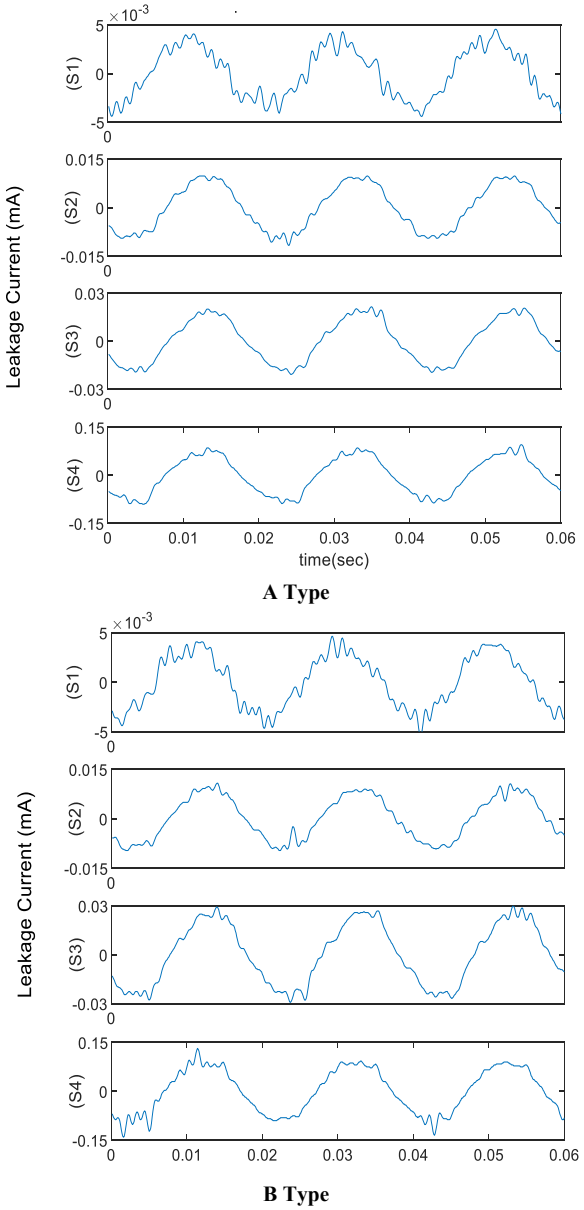


Figure 5. Leakage current waveform for virgin insulator in 80% humidity: (S1) clean surface, (S2) light pollution, (S3) medium pollution; (S4) heavy pollution

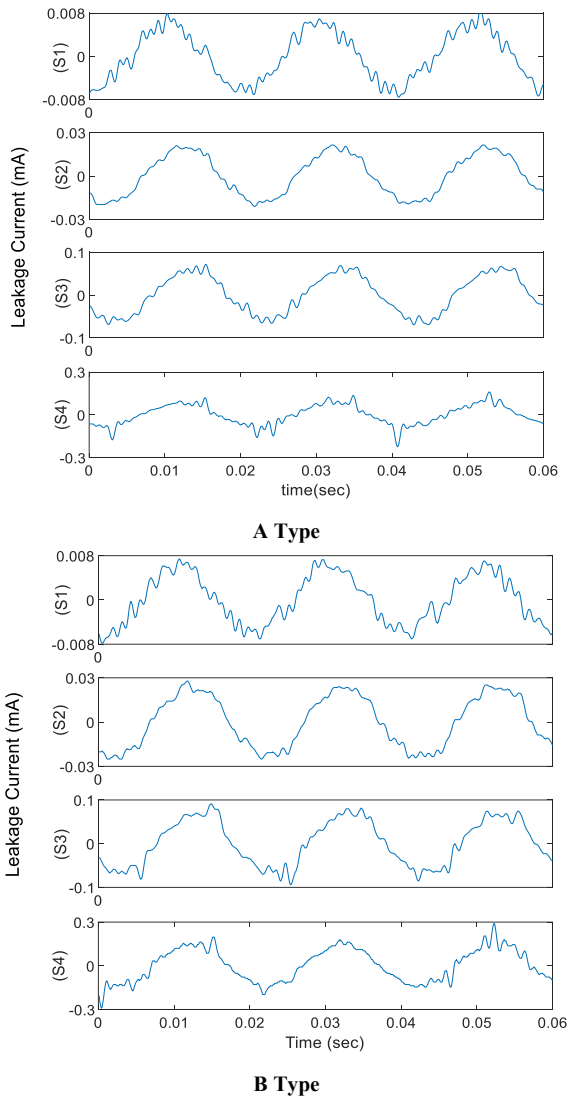


Figure 6. Leakage current waveform for aged insulators in 80% humidity: (S1) clean surface, (S2) light pollution, (S3) medium pollution; (S4) heavy pollution

3. WAVELET TRANSFORM

Wavelet Tansform is a powerful tool widely used for processing transient and non-transient signals. This transformation provides the possibility of simultaneous analysis of time and frequency of the signal. Also, by using this transformation, the characteristics of different signals can be extracted or their noise can be removed. Wavelet analysis involves dividing a signal into modified and scaled versions of a prototype function called the original or mother wavelet, which itself has a finite length and is a highly damped oscillation. Two main types of wavelet transforms are generally used for signal analysis: Continuous Wavelet Transform (CWT) and Discrete Wavelet Transform (DWT). The CWT requires more computational time and utilizes continuous scales and shifts. The mathematical expression for the Continuous Wavelet Transform is given as:

$$CWT_{\omega}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\psi^{*}\left(\frac{t-b}{a}\right) dt \quad (4)$$

where, $x(t)$ is a continuous signal and ψ is the mother wavelet scaled by parameter a and shifted by a factor b. Also, the values $a > 1$ and $b > 0$.

The DWT has a shorter computational time than the continuous one and uses a certain range of scales and shifts. The mathematical expression for DWT is:

$$DWT_{\chi}^{\psi}(m,n) = \frac{1}{\sqrt{a}} \int x(t)\psi^{*}(\frac{t-nba^{m}}{a^{m}})dt \quad (5)$$

where $x(t)$ is the discrete signal and ψ represents the mother wavelet scaled by a^{m} and shifted by a factor nba^{m}. The values of m and n are positive integers and the values of $a > 1$ and $b > 0$.

The Multi-Resolution Analysis (MRA) using DWT allows decomposition of a discrete-time signal $x(t)$ at various decomposition and shift levels. and the studied signal is separated in different frequency bands using MRA analysis, and this process continues until the required resolution is achieved. The time domain signal is decomposed by passing through the successive high-pass and low-pass filters. If h and g represent the values of the high-pass and low-pass filters, respectively, and $c[n]$ is the signal sequence, the decomposition starts with the convolution of $c_0[n]$ with h and g, resulting in the components $c_1[n]$ and $d_1[n]$. Subsequently, the next component is decomposed into $c_2[n]$ and $d_2[n]$. This process continues until the desired resolution is reached. Mathematically, this process is obtained through the following equations:

$$c_m[n] = \sum h[k-2n]\, c_{m-1}[k] \quad (6)$$

$$d_m[n] = \sum g[k-2n]\, c_{m-1}[k] \quad (7)$$

This decomposition process enables the signal to be analyzed without loss of information. In addition, the original signal can be reconstructed using the inverse DWT. By employing multiresolution signal decomposition, two important features of the leakage current can be calculated. The first feature is time localization, which manifests as a large coefficient at the time of the disturbance. The second feature is the ability to distinguish the signal energy in different frequency bands. According to the sampling frequency, the leakage current signals are decomposed into 7 levels using the Daubechies 4, which is shown in Figure 7. Also, the frequency bands corresponding to the detailed components are shown in Table 4.

4. STANDARD DEVIATION

Standard deviation can be expressed as a significant

parameter in the analysis of leakage current signals for monitoring of insulators. This parameter can be considered as a measure of the energy present in a zero-mean signal. Therefore, the standard deviation values for the detailed components are calculated from D1 to D7 in order to identify the transient energy present in the signal at different levels of decomposition. The standard deviation of the n^{th} level of the signal can be calculated using Equation 8:

$$STD = \sqrt{\frac{1}{N_n - 1} \sum_{j=1}^{N_n} [d_n(j) - \mu_n]^2} \quad (8)$$

where d_n is the signal detail vector and μ_n and N_n are the mean and length of the signal detail vector in different frequency bands, respectively.

5. RESULTS

In order to analyze the insulator leakage current under the influence of humidity, contamination and UV aging for providing a suitable monitoring index, laboratory results were extracted and examined. Accordingly, the independent and simultaneous effect of each parameter on the D1 to D7 of the wavelet transform was evaluated. Finally, an appropriate index based on D6 and D7 was presented along with the definition of threshold limits for distinguishing operating conditions. In general, the main logic in choosing D6 and D7 as condition detection indicators is related to their frequency band. Since environmental conditions and aging have the greatest impact on the main and the third harmonic component, the selection of these two parts of the wavelet transform details can be very effective.

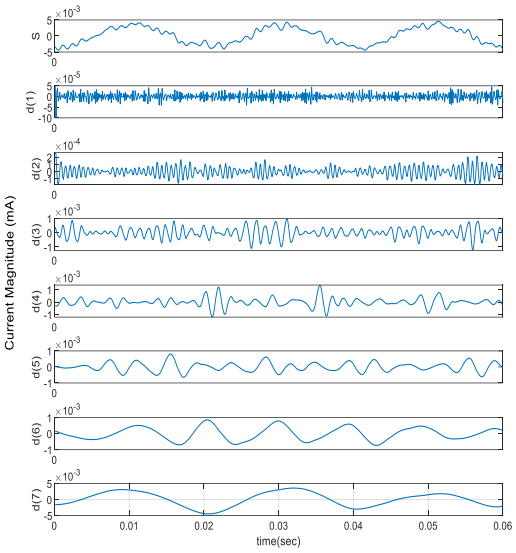


Figure 7. DWT: (S) Main signal, (d1) -(d7) Seven level detailed coefficients

TABLE 4. Frequency Band of DWT Detailed Components of leakage current

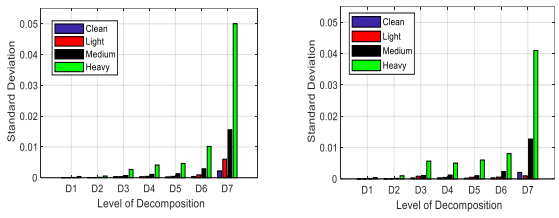
Detailed Components of DWT	Frequency Band (Hz)
D1	2500 - 5000
D2	1250 - 2500
D3	625 - 1250
D4	312.5 - 625
D5	156.25 - 312.5
D6	78.125 - 156.25
D7	39.0625 - 78.125

5. 1. Effect of Pollution on Standard Deviation

Figure 8 shows the impact of ESDD on the standard deviation across different pollution levels under 80% humidity for insulators A and B. Examining the changes of standard deviation in two insulators A and B showed that with the increase in pollution level, the standard deviation from D1 to D7 in both insulators had an increasing trend. In both insulators, the most changes were observed in the conditions of heavy pollution, which reached its highest value in D7. Also, the studies show that the changes of D7 show an acceptable trend compared to other detail components in medium pollution. The results of this level of separation of harmonic components can be considered as a suitable indicator for different pollution levels. In addition, it can be seen that insulator A is more sensitive to the increase in the level of pollution compared to insulator B, and its standard deviation has increased more than insulator B. However, the trend of changes in the standard deviation of both insulators is similar for the pollution level changes, and by defining the same threshold limits, the severity of pollution can be completely recognized for both insulators.

5. 2. Effect of Humidity on Standard Deviation

Figure 9 illustrates the effect of humidity on the standard deviation under heavy pollution conditions for both Insulator A and Insulator B. The analysis reveals that as the humidity level increases from 50% to 90%, the standard deviation correspondingly rises in both



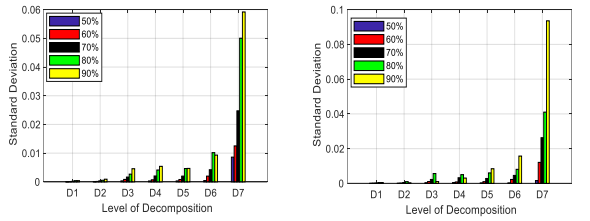
A Type **B Type**

Figure 8. The effect of pollution on the standard deviation of the wavelet transform at 80% humidity and different pollution

insulators. This increasing trend is particularly pronounced under heavy pollution, where the highest standard deviation is recorded in D7 at 90% humidity. Consistent with the observations in the previous section, the standard deviation in D7 is significantly higher than in other components, indicating its potential as a key parameter for distinguishing the operational conditions of insulators.

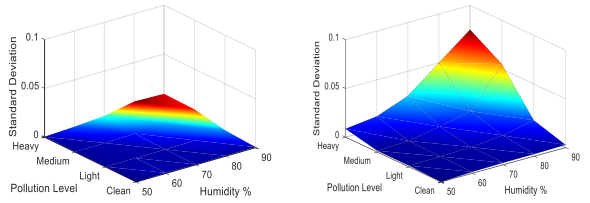
5. 3. Combined Effect of Pollution and Humidity on Standard Deviation In order to provide a comprehensive analysis of the effect of pollution and humidity on the standard deviation of the wavelet transform detail components, the combined effect of these factors is shown in this section. In this regard, Figure 10 displays the combined effect of humidity and pollution on the standard deviation of details D6 and D7 for the two insulators.

The results show that with the simultaneous increase in humidity and contamination levels, the standard deviation increases continuously in all tests. As shown, the D7 component consistently serves as a better indicator compared to D6, and its increasing trend is such that it enables the use of D7 for diagnosing the condition of insulators. Although the increase of humidity and pollution increases the harmonic components of the leakage current, however, the range of changes of the D6 component is very small, and this behavior makes it difficult to use it as an indicator to evaluate the condition of insulators.

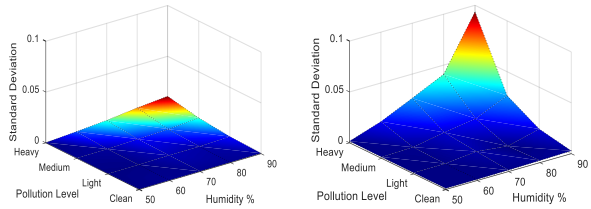


A Type **B Type**

Figure 9. The effect of humidity on the standard deviation of the wavelet transform at heavy pollution and different humidity levels



A Type – D6 **A Type - D7**



B Type – D6 **B Type - D7**

Figure 10. Simultaneous effect of humidity and pollution on D6 and D7 in healthy condition of insulator

5. 4. Performance Criterion Based on the Impact of Humidity and Pollution Based on the results of the effect of pollution and humidity on STD and the comparison of the values obtained for D7, the operating conditions of the insulator can be classified into three categories: normal, caution, and critical. This classification is based on the insulator's need for immediate washing and periodic washing to reduce the probability of electrical discharge occurring on the insulator surface. Table 5 shows the range of variations in D7 with respect to the effects of humidity and pollution on the classification of insulator operating conditions. The threshold limits of this classification were determined based on the results obtained from the analysis of the laboratory test samples. The results showed that the values obtained for D7 follow a normal distribution and there is no scattered data among the results obtained. Accordingly, the [mean ±2 standard deviation $(\bar{X} + 2S)$] method was used to determine the threshold limits.

The results of the investigations have shown that heavy pollution under humidity conditions of 70% to 90% is consistently considered as critical conditions. Under these circumstances, insulators are exposed to the most significant fluctuations in leakage current, and the probability of electrical discharges occurring increases considerably. On the other hand, medium pollution within the humidity range of 80% and 90%, and heavy pollution under humidity conditions of 60% to 70% are categorized as caution conditions. In this case, there is a possibility of increased leakage current, but its magnitude is lower compared to critical conditions. All other conditions in terms of pollution level and humidity will be classified as normal. It is worth noting that when the ambient humidity is less than 60%, there is no noticeable effect on the leakage current and the insulator performance is not affected by pollution. Accordingly, all pollution levels under humidity less than 60% are classified as normal conditions.

Examining the results for the D6 component compared to the D7 component has shown that this component does not have the ability to distinguish between the operating conditions. The reason for this is that the harmonic components have less influence on the standard deviation than the fundamental component. In other words, the fundamental component of the leakage current exhibits a greater increase with rising humidity and pollution levels than the other harmonic components.

TABLE 3. The criterion for diagnosing the condition of insulation by considering the effect of pollution and humidity

Performance criteria	Range
Normal	$STD_{D7} < 0.02$
Caution	$0.02 < STD_{D7} < 0.03$
Critical	$STD_{D7} > 0.03$

TABLE 4. The criterion for diagnosing the condition of insulation considering aging effect

Performance criteria	Range
Normal	$STD_{D7} < 0.04$
Caution	$0.04 < STD_{D7} < 0.08$
Critical	$STD_{D7} > 0.08$

5. 5. Effect of Aging on Standard Deviation and Performance Criterion As stated and shown in Figure 6, aging as a factor influencing the performance of the insulator increases the range of the leakage current. Accordingly, in this article, this factor has been investigated along with humidity and pollution and its effect on standard deviation has been evaluated. Figure 11 also shows the results of the effect of aging on the standard deviation index for D6 and D7. As can be seen, aging has a significant effect on the increase in the standard deviation. These conditions lead to a change in the defined threshold limits for detecting the insulator condition, so that their values will exceed the thresholds set for a healthy insulator.

The criterion for diagnosing the condition of the insulator considering the effect of aging is shown in Table 6. According to this criterion, if the standard deviation is greater than 0.08, the insulator is in a critical state of its performance. This condition typically occurs under heavy pollution with humidity ranging from 70 to 90%, as well as under medium pollution with humidity ranging from 80 to 90%. Furthermore, under heavy pollution with 60% humidity, medium pollution with 60 and 70% humidity, and light pollution with 90% humidity, the standard deviation falls within the range of 0.04 to 0.08, which is considered a caution condition. Finally, considering the various classifications of humidity and pollution for an aged insulator, the remaining regions are classified as normal conditions, in which case the standard deviation will be less than 0.04.

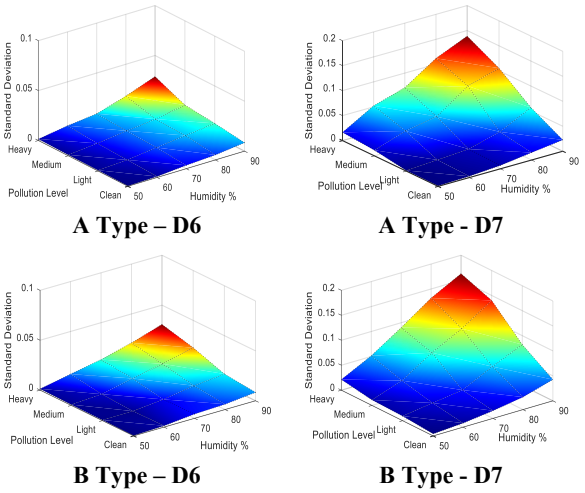


Figure 11. Simultaneous effect of humidity and pollution on standard deviation D6 and D7 in aged condition

Considering the distinctly different threshold limits established for healthy and aged states, a clear separation of these two conditions at the initial stage is crucial. Based on the results presented by Ahmadi-Veshki et al. (1), it has been observed that under the influence of the aging process, the third harmonic component of the leakage current passing through the insulator increases and its fifth component decreases. This condition leads to a change in the D6 component of the wavelet transform, distinguishing it from the healthy state. Therefore, examining the results for the D6 component along with the D7 component allows the detection of the aged state and the determination of the appropriate threshold limits. The conducted investigations show that the D6 standard deviation values greater than 0.01 indicate aging of the insulator, while the standard deviation of D6 for a healthy insulator is significantly less than 0.01 under all defined operating conditions. Overall, by having the standard deviation values of D6 and D7 and considering the defined threshold limits, it will be possible to detect the insulator condition.

Based on the results obtained, a flowchart can be presented for monitoring the condition of silicon rubber insulators. This flowchart, illustrated in Fig. 12, utilizes the standard deviation index extracted from the harmonic analysis of the insulator leakage current through wavelet transform.

5. 6. Determination of Indices Performance The performance of the suggested indicators to accurately estimate the insulator state from the 48 test observations was investigated. All tests were performed under new conditions for humidity levels of 65%, 75%, 85%, and 95% and SDD of 0.025, 0.095, and 0.25 mg/cm^2 on healthy and aged insulators, and the results were evaluated based on D6 and D7 indices.

The confusion matrix illustrated in Figure 13 was used to compute the sensitivity and accuracy of these indices. The values of the confusion matrix were determined based on the capability of the insulator indicator to predict the correct condition (35).

Accuracy measures the overall correctness of the model across all classes. It is the ratio of correctly predicted instances to the total number of instances. In

this case the overall accuracy of the model is approximately 89.58%.

For multi-class analysis, Precision and Recall are calculated for each class, treating that class as the "Positive" class and all other classes as "Negative". Table 7 shows the results of the multi-class analysis.

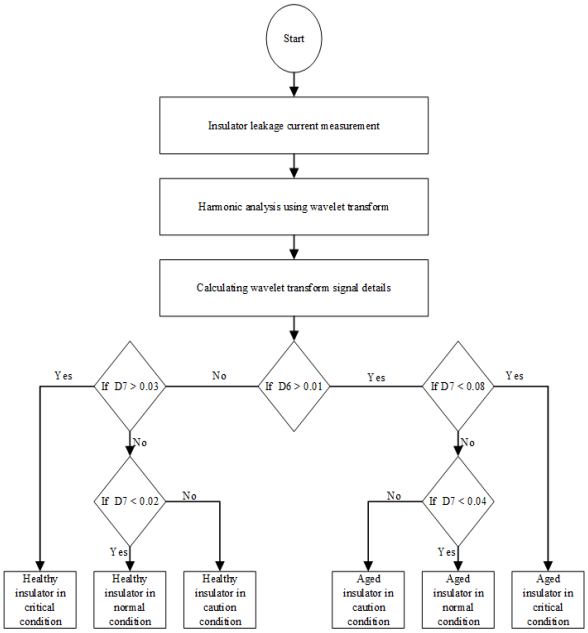


Figure 12. Flowchart of monitoring the condition of silicon rubber insulators

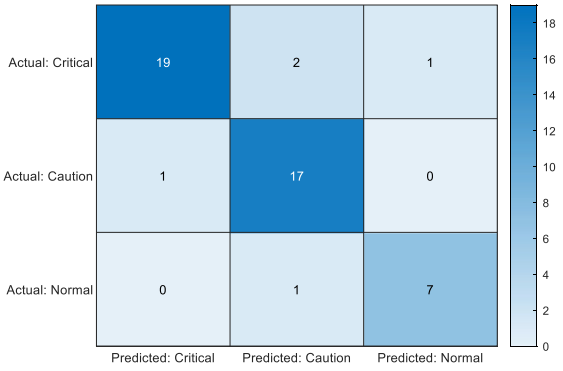


Figure 13. Confusion matrix for determining indices' performance

TABLE 7. Precision and Recall are calculated for each class in three-class analysis

Class	Precision	Recall
Critical	0.95	0.86
Caution	0.85	0.94
Normal	0.88	0.88

According to this results the model exhibits its highest precision for the Critical state. This is a strong positive, meaning that when the model issues a "Critical" alert, it is highly reliable (95% of the time, the instance is truly critical). The model is most effective at identifying all actual Caution cases, with a Recall of 0.94. In fact, it only missed 1 Caution case. Also, the model's primary weakness is missing Critical cases. Specifically, 2 truly Critical cases were misclassified as Caution, and 1 was misclassified as Normal. Depending on the application, missing a Critical state (a False Negative) could be a severe and costly error. The model achieves a high overall accuracy (89.58%), demonstrating a generally robust performance across all three classes, although there is a trade-off between the Precision and Recall values for the Critical and Caution classes.

6. CONCLUSION

In this paper, a criterion for evaluating and diagnosing the condition of polymer insulators based on the analysis and investigation of leakage current using wavelet transform has been introduced. In this research, factors such as humidity, pollution and aging were considered as effective factors on leakage current. These factors can affect the operating conditions, the magnitude of the leakage current, and ultimately the diagnostic criterion for the insulator's condition. Research was done in a laboratory on a 20 kV polymer insulator. Accordingly, the leakage current was measured and analyzed for variations in humidity at 5 levels and ESDD at 4 levels for two types of insulators in both healthy and aged states. The results of this study can be presented as follows:

1. Among the parameters extracted from the wavelet transform, the standard deviation has shown the highest sensitivity to changes in environmental conditions and aging.
2. Considering the frequency range of wavelet transform, the standard deviation of D7 has been chosen as an indicator for detecting the condition of the insulator.
3. Examining the standard deviation in different conditions has shown that humidity has a greater effect than pollution on changing operating conditions. In other words, the changes in leakage current and wavelet transform components due to humidity changes at a constant level of ESDD are consistently greater than the changes in ESDD at a constant humidity level.
4. The threshold limit for the standard deviation of D7 for classifying insulator operating conditions, without considering the effect of aging, is less than 0.02 in the normal state. Conversely, values greater than 0.03 will place the insulator in a critical state.
5. Considering the effect of aging, the performance threshold in the normal state has been changed from

0.02 to 0.04 and from 0.03 to 0.08 in the critical state. This change indicates the significant effect of aging on the leakage current and wavelet transform components.

6. In order to distinguish between healthy and aged conditions, the standard deviation of D6 was utilized. Values greater than 0.01 for this component indicate an aged state and are considered a criterion for the change in the insulator's condition.
7. Evaluation of the method presented in this paper shows the optimal performance of the standard deviation as a unique indicator, in addition to distinguishing healthy insulators from aged ones, for determining the level of insulation contamination into three states: normal, precautionary, and critical.
8. The proposed classification model (based on indices D6 and D7) demonstrated strong overall performance in differentiating the three states achieving an overall accuracy of 89.58%.
9. This model has a very high reliability in predicting Critical states, where its detection is very important for power system stability and preventing interruptions in its continuous operation. The Precision for the Critical class was calculated to be 95%.

Acknowledgements

The authors acknowledge the support of Shahrood university of Technology Research Center, which provided laboratory facilities for this study.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors did not use Generative AI and AI-assisted Technologies in the Writing Process.

REFERENCES

1. reza Ahmadi-veshki M, Mirzaie M, Sobhani R. Reliability assessment of aged SiR insulators under humidity and pollution conditions. International Journal of Electrical Power & Energy Systems. 2020;117:105679. https://doi.org/10.1016/j.ijepes.2019.105679
2. Salem AA, Lau KY, Rahiman W, Abdul-Malek Z, Al-Gailani SA, Mohammed N, et al. Pollution flashover voltage of transmission line insulators: Systematic review of experimental works. IEEE Access. 2022;10:10416-44. https://doi.org/10.1109/ACCESS.2022.3143534
3. Sopelsa Neto NF, Stefenon SF, Meyer LH, Bruns R, Nied A, Seman LO, et al. A study of multilayer perceptron networks applied to classification of ceramic insulators using ultrasound. Applied Sciences. 2021;11(4):1592. https://doi.org/10.3390/app11041592
4. Salem AA, Abd Rahman R, Kamarudin M, Othman N, editors. Factors and models of pollution flashover on high voltage outdoor insulators. 2017 IEEE Conference on Energy Conversion (CENCON); 2017: IEEE.
5. Ren A, Liu H, Wei J, Li Q. Natural contamination and surface flashover on silicone rubber surface under haze–fog environment. Energies. 2017;10(10):1580. https://doi.org/10.3390/en10101580
6. Sopelsa Neto NF, Stefenon SF, Meyer LH, Ovejero RG, Leithardt VRQ. Fault prediction based on leakage current in contaminated insulators using enhanced time series forecasting models. Sensors. 2022;22(16):6121. https://doi.org/10.3390/s22166121
7. Yasuda K, Arazoe S, Igarashi T, Yanabu S, Ueta G, Okabe S. Comparison of the insulation characteristics of environmentally-friendly oils. IEEE Transactions on Dielectrics and Electrical insulation. 2010;17(3):791-8. https://doi.org/10.1109/TDEI.2010.5492252
8. Salem AA, Abd-Rahman R, Al-Gailani SA, Kamarudin MS, Ahmad H, Salam Z. The leakage current components as a diagnostic tool to estimate contamination level on high voltage insulators. IEEE Access. 2020;8:92514-28. https://doi.org/10.1109/ACCESS.2020.2993630
9. Salem AA, Abd-Rahman R, Ahmad H, Kamarudin M, Jamal N, Othman N, et al., editors. A new flashover prediction on outdoor polluted insulator using leakage current harmonic components. 2018 IEEE 7th International Conference on Power and Energy (PECon); 2018: IEEE.
10. Abd-Rahman R, Haddad A, Harid N, Griffiths H. Stress control on polymeric outdoor insulators using zinc oxide microvaristor composites. IEEE Transactions on Dielectrics and Electrical Insulation. 2012;19(2):705-13. https://doi.org/10.1109/TDEI.2012.6180258
11. Gao S, Jia Y, Bi X, Cao B, Li X, Yang D, et al., editors. Prediction method of leakage current of insulators on the transmission line based on BP neural network. 2018 IEEE 2nd International Electrical and Energy Conference (CIEEC); 2018: IEEE.
12. Palangar MF, Amin U, Bakhshayesh H, Ahmad G, Abu-Siada A, Mirzaie M. Identification of composite insulator criticality based on a new leakage current diagnostic index. IEEE Transactions on Instrumentation and Measurement. 2021;70:1-10. https://doi.org/10.1109/TIM.2021.3052910
13. Seman LO, Stefenon SF, Mariani VC, dos Santos Coelho L. Ensemble learning methods using the Hodrick–Prescott filter for fault forecasting in insulators of the electrical power grids. International Journal of Electrical Power & Energy Systems. 2023;152:109269. https://doi.org/10.1016/j.ijepes.2023.109269

14. Jia Z, Chen C, Wang X, Lu H, Yang C, Li T. Leakage current analysis on RTV coated porcelain insulators during long term fog experiments. IEEE transactions on dielectrics and electrical insulation. 2014;21(4):1547-53. https://doi.org/10.1109/TDEI.2014.6879943
15. Joneidi I, Shayegani A, Mohseni H, Mohseni S, Jebeli-Javan M. Leakage current analysis and FFT calculation on polluted polymer insulator. International Journal of Computer and Electrical Engineering. 2013;5(1):133. https://doi.org/10.7763/IJCEE.2013.V5.676
16. Douar M, Mekhaldi A, Bouzidi M. Flashover process and frequency analysis of the leakage current on insulator model under non-uniform pollution conditions. IEEE Transactions on Dielectrics and Electrical Insulation. 2010;17(4):1284-97. https://doi.org/10.1109/TDEI.2010.5539685
17. Villalobos RJ, Moran LA, Huenupán F, Vallejos F, Moncada R, Pesce C. A new current transducer for on-line monitoring of leakage current on HV insulator strings. IEEE Access. 2022;10:78818-26. https://doi.org/10.1109/ACCESS.2022.3191349
18. Deb S, Ghosh R, Dutta S, Dalai S, Chatterjee B, editors. Effect of humidity on leakage current of a contaminated 11 kV Porcelain Pin Insulator. 2017 6th International Conference on Computer Applications In Electrical Engineering-Recent Advances (CERA); 2017: IEEE.
19. Salem AA, Lau KY, Rahiman W, Abdul-Malek Z, Al-Gailani SA, Rahman RA, et al. Leakage current characteristics in estimating insulator reliability: experimental investigation and analysis. Scientific Reports. 2022;12(1):14974. https://doi.org/10.1038/s41598-022-17792-x
20. Salem AA, Abd-Rahman R, Al-Gailani SA, Salam Z, Kamarudin MS, Zainuddin H, et al. Risk assessment of polluted glass insulator using leakage current index under different operating conditions. Ieee Access. 2020;8:175827-39. https://doi.org/10.1109/ACCESS.2020.3027226
21. Li J, Sima W, Sun C, Sebo SA. Use of leakage currents of insulators to determine the stage characteristics of the flashover process and contamination level prediction. IEEE Transactions on Dielectrics and Electrical Insulation. 2010;17(2):490-501. https://doi.org/10.1109/TDEI.2010.5442323
22. Teymourian MH SS. Proposing surge arrester monitoring indicators and their evaluation using correlation coefficient method: considering the effect of longitudinal and fan-shaped type non-uniform pollution. International Journal of Engineering Transactions C: Aspects. 2026;39(3):789–801. https://doi.org/10.5829/ije.2026.39.03c.17
23. Suda T. Frequency characteristics of leakage current waveforms of a string of suspension insulators. IEEE Transactions on power delivery. 2005;20(1):481-7. https://doi.org/10.1109/TPWRD.2004.838650
24. Ahmadi-Joneidi I, Majzoobi A, Shayegani-Akmal AA, Mohseni H, Jadidian J. Aging evaluation of silicone rubber insulators using leakage current and flashover voltage analysis. IEEE Transactions on Dielectrics and Electrical Insulation. 2013;20(1):212-20. https://doi.org/10.1109/TDEI.2013.6451437
25. Khodsuz M, Zamani SA. Novel criteria for silicone rubber insulators condition monitoring based on leakage current analysis: considering asymmetric aging and pollution. Engineering Applications of Artificial Intelligence. 2024;133:108175. https://doi.org/10.1016/j.engappai.2023.108175
26. Salem AA, Al-Gailani SA, Amer AAG, Alsharef M, Bajaj M, Zaitsev I, et al. Classification of RTV-coated porcelain insulator condition under different profiles and levels of pollution. Scientific Reports. 2024;14(1):22759. https://doi.org/10.1038/s41598-024-73520-7
27. Esmaieli M, Khodsuz M, Nouri H. Condition assessment criteria evaluation of asymmetric aged and fully aged silicone rubber insulators based on leakage current harmonics. IET Science, Measurement & Technology. 2024;18(4):163-81. https://doi.org/10.1049/smt2.12198
28. Seyyedbarzegar S. The evaluation of asymmetrically aging effect on silicon rubber insulators flashover voltage. World Journal of Engineering. 2025. https://doi.org/10.1108/WJE-09-2024-0535
29. Shaukat S, Arshad A, Shah WA, editors. Use of wavelet transform to analyze leakage current of silicone rubber insulators under polluted conditions. 2022 2nd International Conference of Smart Systems and Emerging Technologies (SMARTTECH); 2022: IEEE.
30. Algeelani NA, Piah MAM, editors. Characterization of leakage current on high voltage glass insulators using wavelet transform technique. 2012 4th International Conference on Intelligent and Advanced Systems (ICIAS2012); 2012: IEEE.
31. Li G, Huang L. Research on wavelet-based sparse representation of insulator leakage current signal. International Transactions on Electrical Energy Systems. 2015;25(1):72-88. https://doi.org/10.1002/etep.1800
32. Maadjoudj D, Mekhaldi A, Teguar M, editors. Wavelet transform of the leakage current to monitoring the insulator surface under non-uniform desert pollution. 2018 International Conference on Electrical Sciences and Technologies in Maghreb (CISTEM); 2018: IEEE.
33. Maadjoudj D, Kherif O, Mekhaldi A, Teguar M. Features characterizing the surface state of HV insulator glass model under desert pollution. IEEE Transactions on Dielectrics and Electrical Insulation. 2022;28(6):1964-72. https://doi.org/10.1109/TDEI.2021.009739
34. Narayanan VJ, Sivakumar M, Karpagavani K, Chandrasekar S. Prediction of flashover and pollution severity of high voltage transmission line insulators using wavelet transform and fuzzy C-means approach. Journal of Electrical Engineering and Technology. 2014;9(5):1677-85. https://doi.org/10.5370/JEET.2014.9.5.1677
35. Khodsuz M, Teymourian MH, Seyyedbarzegar S. New criteria for metal oxide surge arrester condition monitoring based on leakage current analysis: Considering non-uniform pollution effect. IET Generation, Transmission & Distribution. 2024;18(5):1072-89. https://doi.org/10.1049/gtd2.13123

Persian Abstract

چکیده

ارزیابی وضعیت مقره‌های لاستیک سیلیکونی یکی از الزامات اساسی در سامانه‌های قدرت است. در این زمینه، استفاده از جریان نشتی به عنوان یک روش سریع و برخط نقش مهمی ایفا کرده است. با این حال، تغییرات در شرایط بهره‌برداری می‌تواند منجر به بروز خطاهای قابل توجهی در تصمیم‌گیری‌ها در مورد وضعیت مقره شود .تحت تأثیر شرایط محیطی بر جریان نشتی، تعریف معیارهایی که قادر به تمایز شرایط مختلف بهره‌برداری باشند، دشوار است. در همین راستا، دستیابی به یک مدل با کمترین پیچیدگی برای تصمیم‌گیری، به عنوان یک راه‌حل مناسب در نظر گرفته می‌شود. در این پژوهش، یک شاخص مبتنی بر جریان نشتی ارائه شده است تا عملکرد مقره‌ها را در سه حالت عادی، احتیاط و بحرانی طبقه‌بندی کند. بدین منظور، داده‌های آزمایشگاهی به صورت عملی تحت شرایط آلودگی سبک، متوسط و سنگین در سطوح رطوبت متغیر تا ۹۰ درصد جمع‌آوری شدند. تمام آزمون‌ها تحت شرایط محیطی متفاوت بر روی مقره‌های سالم و پیر شده انجام شدند تا تأثیر تخریب سطح بر جریان نشتی و مؤلفه‌های هارمونیکی آن‌ها بررسی شود. برای تجزیه و تحلیل جریان نشتی، از تبدیل موجک استفاده شد و انحراف معیار ضرایب جزئیات موجک مورد بهره‌برداری قرار گرفت. همچنین، در این مقاله، روشی برای تمایز مقره سالم از مقره‌ای که بر اثر اشعه فرابنفش پیر شده است، بر اساس ضرایب جزئیات معرفی شده است.بررسی نتایج به دست آمده از تبدیل موجک نشان داده است که انحراف معیار D6 بزرگتر از ۰.۰۱ از دقت بسیار خوبی برای تمایز مقره‌های سالم و پیر شده برخوردار است. علاوه بر این، با استفاده از انحراف معیار D7 و تعریف حدود آستانه‌ای برای مقره‌های سالم و پیر شده، تمایز عملکرد آن‌ها نیز ممکن شد. با هدف ارزیابی دقت و حساسیت شاخص‌های پیشنهاد شده از ماتریس سردرگمی استفاده شده است که با دقت ۸۹/۵۸ ٪ قادر به تفکیک شرایط مختلف بهره‌برداری است

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2419-2433

International Journal of Engineering

Journal Homepage: www.ije.ir

Geological and Economic Assessment and Comparative Analysis of Uranium ore Deposits in Russia and Iran

O. A. Marinina, E. V. Kildiushov*

Department of Industrial Economics, Saint Petersburg Mining University, Saint Petersburg, Russia

PAPER INFO

Paper history:
Received 25 July 2025
Received in revised form 15 September 2025
Accepted 12 December 2025

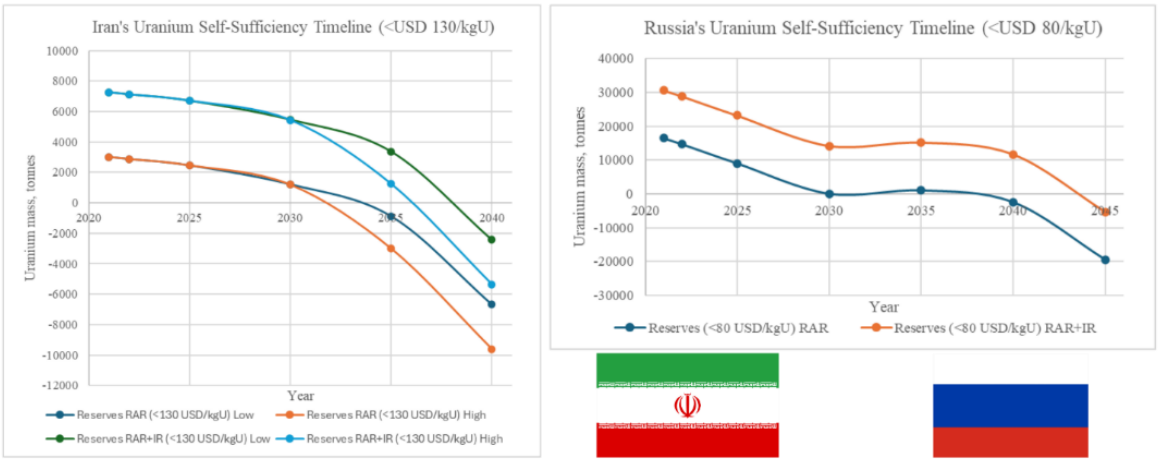
Keywords:
Uranium Resources
Uranium Self-sufficiency
Uranium Production
Uranium Market Analysis

ABSTRACT

This study conducts a comparative geological and economic assessment of uranium resources in Russia and Iran, with a focus on their ability to support the development of domestic nuclear energy. Unlike previous research, which has predominantly examined Russian–Iranian cooperation from political or institutional perspectives, this work evaluates uranium availability using geological classifications and economic cost categories of reserves. It distinguishes between recoverable resources reported for Russia and in situ resources reported for Iran, with the latter converted to a comparable basis through the application of appropriate recovery factors. In addition to analyzing the structure and potential of the resource base, the study takes into account historical and projected uranium production as well as consumption scenarios for both countries. On this basis, long-term self-sufficiency horizons are estimated through 2045. The findings show that Russia possesses a vast and geologically diverse uranium base, capable of ensuring long-term supply if new mining projects are developed to offset the expected decline in production after 2030. Iran, by contrast, holds more limited and higher-cost reserves, concentrated mainly in metasomatite deposits, which makes the achievement of sustained self-sufficiency unlikely without external cooperation or technological breakthroughs. At the same time, geological similarities between the two countries create a basis for potential collaboration in uranium exploration and development. By addressing a relatively underexplored aspect of the literature, this study provides a structured assessment of uranium supply sustainability and introduces a comparative perspective relevant not only to Russia and Iran but also to other countries with emerging nuclear programs.

***doi**: 10.5829/ije.2026.39.10a.06*

Graphical Abstract



*Corresponding Author Email: evgenii.kildiushov@gmail.com (E. V. Kildiushov)

Please cite this article as: Marinina OA, Kildiushov EV. Geological and Economic Assessment and Comparative Analysis of Uranium ore Deposits in Russia and Iran. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2419-33.

NOMENCLATURE

R_{rec}	Recoverable uranium reserves (tonnes U)	P	Annual uranium production (tonnes/year)
$R_{in\ situ}$	In situ uranium reserves (tonnes U)	C	Annual uranium consumption (tonnes/year)
RF	Recovery factor (%)	X	Target year
$R_{X/A}$	Uranium reserves in year X/A (tonnes U)	A	Base year
R_{last}	Last known positive level of reserves (tonnes U)	Y	Number of years remaining before exhaustion

1. INTRODUCTION

Uranium continues to play a key role in the global energy system, valued for its high energy density and its strategic importance in nuclear power generation (1). As global energy demand rises and climate targets become more urgent, nuclear energy is increasingly seen as an essential part of sustainable and secure energy strategies (2, 3). Many countries are expanding investments in nuclear capacity to ensure stable baseload electricity alongside decarbonization goals (4, 5).

In recent years, the uranium market has exhibited notable volatility. After a prolonged period of low prices, uranium experienced a sharp surge between 2021 and early 2024, driven by renewed nuclear interest, energy security concerns, and speculative restocking. This was followed by a correction, with prices declining toward US$ 140/kgU by early 2025. These fluctuations highlight the sector's sensitivity to short-term disruptions and its ongoing dependence on broader geopolitical and economic trends (6).

Understanding the availability, economic accessibility, and geological distribution of uranium resources is critical for anticipating future market dynamics (7). However, cross-country comparisons are complicated by differences in exploration maturity, reporting standards, and evolving production costs. Technological advances—such as recycling, fuel reprocessing, and advanced reactor designs—may also reshape both the quantity and the quality of economically viable uranium resources with respect to time (8, 9).

This study presents a comparative analysis of uranium resources in Russia and Iran—two countries with contrasting positions in the nuclear sector. Russia possesses one of the largest uranium resource bases globally, along with an integrated fuel cycle and substantial technological expertise. Iran, by contrast, has a smaller and less diversified uranium endowment but growing nuclear ambitions. Similar development barriers—such as dependence on external technology and institutional limitations—have been noted in emerging energy sectors like Mozambique's gas industry (10). By examining geological structures, production and consumption balances, and long-term self-sufficiency horizons, this work aims to assess their respective uranium security outlooks within a changing global nuclear landscape.

The aim of this research is to conduct a geological and economic assessment of uranium resources in Russia and Iran, with emphasis on analyzing absolute volumes and the potential of different deposit types and economic cost categories, determining their capacity for long-term self-sufficiency, and identifying complementarities that could serve as a basis for bilateral cooperation.

To achieve this aim, the study defines the following objectives:

1. To analyze the structure of uranium resources in Russia and Iran by type (RAR and IR) and by economic cost categories (<40, <80, <130, and <260 US$/kgU).
2. To harmonize differences in resource reporting and to analyze reserves by deposit type and mining method, applying recovery factors specific to each extraction technology.
3. To assess historical and projected uranium production, along with consumption scenarios for both countries, in order to evaluate the balance between supply and demand.
4. To estimate long-term self-sufficiency prospects and to identify complementarities that may provide a basis for bilateral cooperation

2. LITERATURE REVIEW

2. 1. Typology and Global Distribution of Economically Recoverable Uranium Resources

As of June 2025, uranium prices remain slightly above US$ 140/kgU, making it especially relevant to focus on deposits recoverable at prices below US$ 130/kgU. Based on January 2023 data, global reasonably assured and inferred resources at this threshold total approximately 5.93 million tonnes, distributed across a wide spectrum of uranium deposit types. The most significant category is sandstone-hosted deposits, contributing around 1.90 million tonnes. These occur in permeable fluvial and marginal marine sandstones, where uranium precipitates through redox reactions with organic matter or sulfides, and they are particularly amenable to in situ recovery. Key production regions include Kazakhstan, the United States (notably Wyoming and Texas), Uzbekistan, Nigeria and Mongolia (11).

Proterozoic unconformity-related deposits, although geographically limited, represent the richest known uranium ores by grade, with 699,000 tonnes recoverable at this threshold (12). Found primarily in the Athabasca Basin of Canada and the Alligator Rivers Province in Australia, these deposits form at the contact between sedimentary basins and older crystalline basement rocks

and typically exhibit uranium concentrations well over 1% (13). Polymetallic iron oxide breccia complex deposits, typified by Olympic Dam in South Australia, contribute 1.35 million tonnes. Although the uranium content is low, it is economically viable as a by-product of copper, gold, and silver mining due to the vast size of the orebody (14). Ancient placer-type paleo-quartz-pebble conglomerates, mainly in South Africa's Witwatersrand Basin and Canada's Elliot Lake District, contain 311,000 tonnes of uranium; despite modest grades, they are valued for their scale and geological stability (15). Metasomatite deposits—linked to large-scale sodium or potassium metasomatism in Precambrian shields—yield about 573,500 tonnes and are prominent in Ukraine, Brazil, and Russia (16). Intrusive deposits, totaling 372,200 tonnes, are found in peralkaline complexes, anatectic granites, or carbonatites in regions like Namibia (Rössing, Husab), Greenland (Kvanefjeld), and Brazil (Poços de Caldas) (17-19). Volcanic-related deposits provide 233,000 tonnes and are concentrated in pyroclastic and caldera settings such as Russia's Streltsovska Caldera, Peru's Macusani Plateau, and areas of China (20, 21). Granite-related uranium systems, mainly in France, Germany, the Czech Republic, and China, add 102,900 tonnes and are associated with hydrothermal processes around peraluminous leucogranites (22, 23). Metamorphite deposits, hosted in high-grade metamorphic rocks, offer 87,500 tonnes and are especially notable in India, Canada, and Kazakhstan. Surficial uranium deposits contribute 209,100 tonnes, forming in calcrete-rich valley-fill and playa environments under arid conditions, with Namibia and Australia hosting the largest examples. Smaller but relevant contributions come from carbonate-hosted uranium deposits (19,800 tonnes), such as India's Tummalapalle, and phosphate-bearing formations (7,100 tonnes), with recoverable uranium found as a by-product in Morocco and Florida. Lignite- and coal-associated deposits, mainly in Kazakhstan, the USA, and South Africa, offer 30,900 tonnes through uranium accumulation in organic-rich layers (24). At the current price threshold, black shale and collapse breccia pipe deposits provide no recoverable uranium, though large low-grade resources in Sweden, the USA, and parts of Central Asia could gain relevance with technological advances or further price increases (25, 26).

The distribution of these deposit types is shown in Figure 1. The uranium resource base is not only geologically diverse, but also unevenly distributed in terms of economic recoverability. This distribution highlights both the geological complexity and geographic diversity of the world's uranium resource base, reinforcing the need for tailored mining and processing strategies aligned with deposit type and economic context.

2. 2. Mining Methods and Geological Determinants of Uranium Recovery Uranium can be extracted using a range of mining and recovery techniques, each of which is selected based on the depth, grade, geological setting, and geotechnical properties of the deposit. The primary methods include open pit mining, underground mining, in situ leaching (ISL), and co- or by-product recovery. The selection of method is closely linked to geological and hydrogeological factors, which influence not only operational feasibility but also the efficiency and economics of uranium recovery (27). Table 1 provides an overview of these major extraction methods, their typical ore types, and key geological and operational characteristics.

Open pit mining is applied to ore bodies that are close to the surface and of sufficient lateral extent to economically justify the removal of overburden. The method is common for surficial, intrusive, and shallow sandstone deposits, as well as some metasomatite and lignite-associated uranium occurrences. The feasibility of open pit mining depends on factors such as orebody geometry, stripping ratio, and the mechanical properties of host rocks, which determine slope stability and excavation methods. While large-scale operations such as the Rössing mine in Namibia demonstrate the potential size of open pits, this method becomes uneconomical at greater depths unless justified by high ore grades or the presence of valuable co-products (29).

Underground mining is chosen when ore bodies are located at greater depths or when surface impact must be minimized. It is typical for high-grade unconformity-related deposits, granite-hosted veins, metamorphic or volcanic systems, and compact carbonate or conglomerate formations. A range of mining techniques is used depending on ore geometry and ground stability, including room and pillar, cut-and-fill, and jet boring.

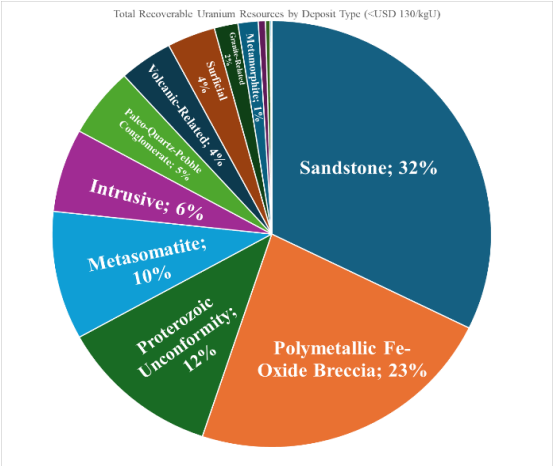


Figure 1. Total Recoverable Uranium Resources by Deposit Type (<US$ 130/kgU) (11)

TABLE 1. Mining Methods and Associated Ore Types (11, 27, 28)

Mining Method	Ore Types	Geological/Technical Features	Overall recovery factor, %
Open Pit Mining	Sandstone, Surficial, Intrusive, Metasomatite (shallow), Lignite – Coal	Applied to shallow deposits that are laterally extensive and near surface. Used in calcrete-hosted surficial deposits (e.g., in Namibia and Australia), intrusive anatectic uranium deposits (e.g., Rössing), and some metasomatite or coal-related deposits when conditions allow. Preferred where stripping ratio is favorable and ore grade justifies large-scale excavation.	80
Underground Mining	Proterozoic Unconformity, Volcanic-Related, Granite-Related, Metamorphite, Paleo-Quartz-Pebble Conglomerate, Carbonate, Collapse Breccia, Metasomatite (deep)	Used for deep or structurally complex deposits with high grade or steep geometry. Ground control is essential, particularly in water-saturated zones (e.g., McArthur River, Cigar Lake). Common for vein-hosted, stratabound, or breccia-type mineralization. Also used in steeply dipping granite-related or carbonate strata-bound.	75
In Situ Leaching (ISL)	Sandstone	Used exclusively for certain sandstone deposits with suitable hydrogeological settings—permeable, saturated, confined aquifers located below the water table, and uranium in mineral forms amenable to leaching (e.g., uraninite, coffinite). Most common in roll-front and tabular deposits with redox interfaces.	85 (Acid) 70 (Alkaline)
Co-/By-product	Polymetallic Fe-Oxide Breccia, Phosphate	Uranium is not the primary commodity. Extracted as a by-product in polymetallic operations like Olympic Dam (Cu-Au-U) or potentially during phosphate processing. Recovery is technically feasible but rarely commercialized for phosphate at present. Economics depend heavily on the primary mineral product.	65

Ground conditions are critical: saturated or weak host rocks require grouting or artificial ground freezing, as was done at McArthur River (30). Emerging monitoring solutions—such as UAV-based image segmentation systems used in infrastructure diagnostics—could be adapted to enhance site safety and environmental oversight in such geotechnically demanding settings (31). In such settings, ventilation design and dust and radon management also become major cost and safety considerations. The method's economic success is tied closely to geological complexity, rock strength, and access conditions (32-34).

In situ leaching, is used primarily for uranium hosted in permeable, saturated, and confined sandstone aquifers. The process involves circulating a lixiviant—typically acidic or alkaline—through injection and recovery wells to dissolve uranium in place. ISL avoids many of the environmental impacts of conventional mining and is especially prevalent in Kazakhstan and parts of the United States. However, it requires specific conditions: the host rock must be chemically and physically suited to leaching, the mineralization must be accessible to fluids, and the surrounding formation must effectively confine the lixiviant. Geological heterogeneity, carbonate content, and permeability variations can all affect recovery efficiency and cost (28).

Co-product and by-product recovery occurs when uranium is extracted alongside other minerals, often during the processing of copper, gold, or phosphate ores. At Olympic Dam in Australia, uranium is a by-product of a large-scale polymetallic mining operation. In phosphate-rich regions like Morocco and Florida, uranium recovery has been studied and piloted but is not practiced commercially today. In these settings, uranium is not the primary target, and its economic viability depends on integration with the main processing flow and the market value of the recovered uranium. Grades are typically low, and process design must be tailored to extract uranium without disrupting the primary product stream (35, 36).

Geological characteristics play a decisive role in determining both the cost and the method of uranium extraction. Critical factors include orebody depth, thickness, geometry, host rock strength, permeability, chemical composition, and the mineralogical form of uranium (37). For instance, carbonates and clay-rich rocks can increase reagent consumption, and refractory uranium minerals complicate leaching (38-40). Operational parameters such as overburden removal, ground control, and groundwater chemistry also shape project design and economics. As a result, the recovery factor in uranium production varies significantly across deposits and depends strongly on site-specific geological and technical conditions (41, 42).

2. 3. Global Leaders in Uranium Resources: Distribution, Costs, and Geological Profiles

Uranium resources are categorized internationally according to a two-dimensional classification system developed by the International Atomic Energy Agency (IAEA), which evaluates both the geological confidence and the estimated production cost of known deposits. The

classification distinguishes between Reasonably Assured Resources (RAR) and Inferred Resources (IR), which together form the category of Identified Resources. RAR are defined by a high degree of geological confidence, based on detailed sampling and evaluation, and are considered recoverable under current economic and technological conditions. Inferred Resources are supported by more limited data and carry a greater degree of uncertainty. Resource categories are also grouped by cost of recovery, with thresholds set at less than US$ 40/kgU, US$ 40–80/kgU, US$ 80–130/kgU, and US$ 130–260/kgU. These are not market prices, but estimates that incorporate mining, processing, infrastructure, environmental management, and other associated expenses (11, 43).

Using this classification, the global inventory of identified uranium resources currently amounts to about 5.93 million tonnes recoverable at production costs below US$ 130/kgU. Several countries stand out as having particularly large and significant uranium endowments. Figure 2 presents the distribution of identified recoverable uranium resources by country, illustrating both the concentration of uranium stocks in a few leading countries and the broader distribution among other resource holders (44). Based on the most recent and detailed data, the leading countries by total identified uranium resources recoverable at costs below US$ 130/kgU—which includes both Reasonably Assured Resources (RAR) and Inferred Resources (IR)—illustrate a diverse global distribution shaped by geology, extraction methods, and national policy (45).

Australia holds the world's largest identified uranium resource base, with 1,236,200 tonnes of recoverable uranium in the < US$ 130/kgU category. This represents approximately 28% of global identified resources at this cost level. A major share of these resources—around 68%—is concentrated in a single site: the Olympic Dam deposit. This polymetallic iron oxide breccia complex is primarily mined for copper, with uranium recovered as a co-product. While vast in scale, the uranium grades at Olympic Dam are low, and production is shaped by the economics of other metals. Other deposits, such as Ranger and Four Mile, contribute smaller shares to Australia's overall total.

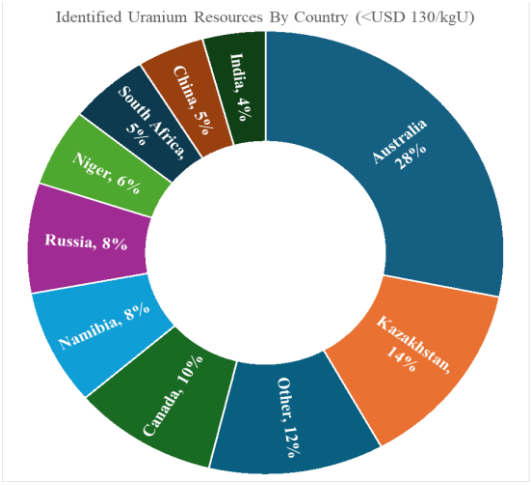


Figure 2. Identified Uranium Resources by Country (11)

Kazakhstan ranks second globally, with 814,000 tonnes of identified recoverable uranium in this cost category, corresponding to roughly 14% of the global total. The country's resources are dominated by roll-front sandstone deposits hosted in permeable aquifers, enabling extraction via ISL. These deposits are located mainly in the Chu-Sarysu and Syrdarya basins. Kazakhstan accounts for the majority of the world's lowest-cost uranium, with 272,000 tonnes in the <US$ 40/kgU bracket alone. State-owned Kazatomprom has optimized ISL techniques, allowing Kazakhstan to become the world's largest uranium producer.

Canada follows with a total of 582,000 tonnes of recoverable uranium below US$ 130/kgU, representing about 10% of global identified resources. These are primarily located in the Athabasca Basin in northern Saskatchewan and consist of high-grade Proterozoic unconformity-related deposits, including McArthur River and Cigar Lake. Although extraction is technically complex due to high water pressure and unstable ground conditions, the ore grades—often exceeding 10%—make these some of the most economically valuable uranium deposits in the world.

Russia possesses 652,500 tonnes of identified recoverable uranium, of which about 73% is recoverable at costs below US$ 130/kgU—this corresponds to roughly 477,000 tonnes. Russia's resource base is geologically diverse, comprising volcanic-related, metasomatic, and sandstone-hosted deposits. Volcanic and metasomatic types dominate in regions such as the Elkon uranium district in Yakutia and the Vitim region in Buryatia. Most of the country's uranium is expected to be extracted via underground mining, with a smaller share suitable for ISL. Russia's vertically integrated fuel cycle and domestic nuclear demand anchor its long-term production strategy.

Namibia has also increased its resource base significantly in recent years. As of January 2023, it holds roughly 498,000 tonnes of recoverable uranium, with 322,000 tonnes in the RAR category and 176,000 tonnes in the IR category at < US$ 130/kgU. Most of the country's uranium comes from large open-pit operations, such as Rössing and Husab, where ore grades are low but economies of scale make mining feasible. Recent discoveries at Tumas, Wings, and Koppies have added to Namibia's endowment.

In summary, five countries—Australia, Kazakhstan, Canada, Russia, and Namibia—collectively hold the majority of the world's identified recoverable uranium resources at economically competitive cost levels. Each

country's position reflects a unique interplay of geological endowment, extraction technology, and development strategy, shaping its role in the global nuclear fuel supply chain (46).

2. 4. Iran's Uranium Sector: Development, Challenges, and Strategic Outlook It should be noted that Russia's uranium sector, as one of the global leaders in terms of reserves and production, has been examined in the preceding sections of this study. This subsection is devoted to Iran in order to explain the reasons for its inclusion in the analysis and to consider its specific challenges and strategic outlook. Iran's uranium sector has undergone substantial development in recent decades, but its current position in the global uranium landscape remains limited, both in terms of scale and maturity. The country reports a total of 9,851 tonnes of identified conventional uranium resources recoverable at production costs below US$ 130/kgU. This includes 4,316 tonnes of reasonably assured resources (RAR) and 5,535 tonnes of inferred resources (IR) (11). Iran has developed specific elements of the uranium supply chain, including exploration, mining, and processing.

Iran's resource base is geologically diverse, encompassing metasomatic, granite-related, metamorphite, and volcanic-related deposit types, with metasomatic deposits accounting for the overwhelming majority of both RAR and IR. However, no resources are reported in < US$ 40/kgU or < US$ 80/kgU cost categories, indicating that none of Iran's current resources qualify as low-cost. Instead, all identified uranium resources are grouped in < US$ 130/kgU band, without further cost breakdowns. This cost profile, combined with a reliance on underground mining and relatively low ore grades (e.g., Saghand: 0.0552% U), highlights the technical and economic constraints facing the sector.

Iran's main production infrastructure consists of the Saghand uranium mine (mines No. 1 and No. 2) and the Ardakan processing plant, which began operating in 2017 with a nominal capacity of 50 tonnes U per year. The Saghand operation uses both open pit and underground mining, depending on ore depth, and employs acid leaching in combination with ion exchange and solvent extraction for processing. A second mine, Gachin, which supplied uranium to the Bandar Abbas plant, ceased operations in 2016. Total recoverable resources linked to Saghand are estimated at 500 tonnes, and those from Gachin at 84.1 tonnes, illustrating the constrained scale of Iran's currently active resource base.

Plans for the development of additional production centres are underway. The Narigan and Khoshumi deposits have reached the conceptual and detailed design stages, respectively, but have not yet transitioned into active production. Their feasibility reflects an effort to diversify the portfolio beyond Saghand, though their technical parameters and expected outputs remain limited.

Iran continues to explore a range of metallogenic trends across the country, including Kerman-Sistan, Naiin-Jandagh, Birjand-Kashmar, and Hamedan-Marand. These areas are under evaluation for granite-related, volcanic, sedimentary, and polymetallic uranium potential. Most of the work remains at the surface or early drilling stage. While official documents cite 9,800 tonnes of prognosticated resources and 48,100 tonnes of speculative conventional resources, these remain undeveloped. In addition, 53,000 tonnes of speculative unconventional uranium resources have been identified in phosphate rocks, black shales, and other non-traditional hosts. Their extraction would require significant technological advances and long-term development.

Details regarding recent exploration and development expenditures, including drilling activity and physical effort, are not available in publicly accessible sources. This absence of open data makes it difficult to assess the momentum or continuity of Iran's investment in uranium development. While it has been reported that eight mines were operating in 2023, and that six more are planned by early 2024, only Saghand appears to have a defined and active production pipeline. Iran's production forecast for 2025 stands at 71 tonnes of uranium, a figure that remains modest by international standards.

The sector is entirely state-owned, with the Atomic Energy Organization of Iran (AEOI) serving as both regulator and operator. Employment data for the uranium sector suggest a stable but limited industrial footprint, with around 280 personnel involved in production-related activities in recent years.

In conclusion, Iran's uranium industry is characterized by a strategic intent to build self-sufficiency, but remains constrained by modest geological reserves, relatively high extraction costs, and limited production capacity. Although exploration continues and speculative resources appear substantial, most of Iran's uranium remains either uneconomic or undeveloped. The country's ability to expand its role in the global uranium market will depend on the outcome of ongoing exploration efforts and the successful transition of known deposits into active, efficient production. For now, Iran remains a small but steadily consolidating player in the international uranium landscape (47).

Current assessments of uranium resources face several methodological limitations. International comparisons often treat reported figures as directly comparable, despite differences in geological data quality, classification systems, and levels of exploration between countries. Additionally, resource estimates are typically presented as static snapshots, with limited integration of long-term trends in production, consumption, and pricing. As a result, the strategic

significance of uranium endowments—and the feasibility of turning resources into production—can be difficult to assess consistently across national contexts.

2. 5. Comparative Position of This Study in the Existing Literature Research on nuclear energy and uranium resources has been conducted from various perspectives, yet several significant gaps remain. El Khalfi analyzed the Joint Comprehensive Plan of Action (JCPOA) and highlighted the geopolitical and diplomatic implications of Iran's nuclear program, emphasizing the challenges of international negotiations and the impact of sanctions. However, this study focused mainly on political processes and did not address the quantitative assessment of uranium reserves or Iran's ability to maintain long-term self-sufficiency in nuclear fuel (48).

Kalbasi (49) examined Russian–Iranian cooperation in the nuclear sector, emphasizing the influence of sanctions, bilateral agreements, and foreign policy discourses on the development of these relations. The work effectively illustrated the political context but did not include an evaluation of geological and economic indicators, particularly the structure and recoverability of uranium resources, which are essential for assessing the sustainability of nuclear programs.

Mirgorod and Parubochaya (50) assessed Rosatom as a key instrument of Russian foreign policy in the Middle East. They emphasized the corporation's competitive advantages, including technological leadership, flexible financing models, and the capacity to enhance Russia's diplomatic influence through nuclear projects in countries such as Egypt, Iran, and Turkey. Nevertheless, their analysis remained focused on geopolitical leverage and external influence, without considering the comparative economic value of uranium reserves or the long-term self-sufficiency prospects of Russia and its partners.

Siddi and Silvan (51) analyzed Rosatom's role in international relations through the lenses of realism, liberalism, and dependency theory. Their study demonstrated the complexity of global nuclear supply chains and identified Rosatom's dominant position in uranium enrichment and reactor construction. However, their analysis was largely theoretical and institutional, omitting country-specific assessments of resource bases and calculations of self-sufficiency horizons.

Kharitonova (52) discussed the prospects for Russian–Iranian energy cooperation, focusing primarily on oil, gas, and related infrastructure projects. While the study provides valuable insights into the broader framework of bilateral energy relations, it does not examine uranium resources or the nuclear fuel cycle. In this regard, the present study contributes by extending the analysis of Russian–Iranian energy cooperation to the strategic dimension of uranium, thereby addressing a gap in the literature.

Taken together, these works highlight the gap that defines the novelty of the present study, which lies in its integrated geological and economic perspective. Unlike previous research, which has predominantly emphasized political, diplomatic, or institutional aspects, this analysis consolidates heterogeneous data into a single framework by converting recoverable resources for Russia and in situ resources for Iran to a comparable basis through the application of recovery factors. Beyond this adjustment, the study examines the differentiation of uranium resources by deposit type and their classification by economic cost categories, which indicate the share of reserves feasibly developed under specific price levels. Previous studies have not addressed the issue of long-term self-sufficiency horizons, leaving a gap in understanding how resource availability translates into future supply security. While the existing literature has discussed Russian–Iranian cooperation primarily in the fields of oil and gas, uranium resources have not been analyzed in this context. By introducing this dimension and linking geological characteristics with economic and energy policy considerations, the present study enriches the literature and provides a more comprehensive view of the sustainability of Russia's and Iran's nuclear energy programs, including the potential for bilateral cooperation in the uranium sector.

3. METHODS

This study is based on a detailed comparative analysis of uranium reserves, production, and consumption in Russia and Iran. The methodology follows a structured sequence of stages to ensure the thoroughness and validity of the analysis.

First, the uranium reserves were examined by type: RAR's and IR's. For each type, reserves were considered across different economic categories, based on cost ranges of uranium production (<40, <80, <130, and < US$ 260/kgU). Russia's reserves were provided as recoverable tonnes, while Iran's were reported as in situ tonnes, which required special attention during interpretation and comparison.

To adjust for the difference in reporting standards, a standard formula was used to calculate the recoverable reserves for Iran:

$$R_{rec} = R_{in\ situ} \cdot \left(\frac{RF}{100}\right) \quad (1)$$

The recovery factor reflects the efficiency of uranium extraction depending on the mining method (53, 54). For example, for underground mining, recovery rates between 80 and 90% were assumed.

Second, reserves were broken down by mining methods, namely underground mining, open-pit mining, in situ leaching (ISL), and co-product or by-product mining. Each method was analyzed separately, with

adjustments made according to the specific recovery factor applicable to each mining technology.

Third, historical uranium production was assessed, based on records by mining method and by year through 2023. In addition, mid-term production projections up to 2045 were incorporated to reflect planned production dynamics for both countries.

Fourth, uranium consumption was evaluated. For Russia, a relatively stable forecast of annual reactor-related uranium requirements was used. For Iran, two consumption scenarios—low and high—were considered to capture uncertainty linked to the potential expansion of nuclear energy use.

To assess the dynamic change in available reserves over time, the following formula was applied:

$$R_x = R_A + (P - C) \cdot (X - A) \quad (2)$$

Additionally, to estimate the number of years until uranium reserves are exhausted once a deficit occurs, the following formula was used:

$$Y = \frac{R_{last}}{C-P} \quad (3)$$

Throughout the analysis, reserves for each country were considered from multiple perspectives—by type of resource and by mining method—to identify commonalities, differences, and potential grounds for cooperation. Special attention was paid to finding comparable structures in resource classification and mining practices, which could support further strategic alignment.

Finally, based on the data obtained, initiatives were proposed for developing bilateral relations in the field of uranium exploration and mining between Russia and Iran. Possible points of contact for technological cooperation, joint ventures, and knowledge sharing were outlined, recognizing the complementarity of the two countries' resource bases and mining experiences. Future studies may consider binary discounting approaches, as used in hydrogen project evaluations, to better reflect cost-revenue risk asymmetry in uranium economics (55).

4. RESULTS

4. 1. Uranium Price Scenarios: 2025–2045 This section does not provide an original econometric forecast but rather summarizes findings from authoritative sources. The analysis is broadly consistent with the IAEA/NEA Red Book, where uranium prices are considered in relation to supply–demand adequacy without quantitative projections, and it additionally draws on recent analytical work, which employs a mathematical model to forecast the cost and volume of uranium production under various global nuclear energy development scenarios (56). The price ranges presented here should therefore be understood as illustrative scenarios that reflect different combinations of demand growth, supply responses, and project economics, rather than strictly deterministic forecasts.

Over the past two decades, uranium prices have exhibited cyclical behavior, with sharp peaks followed by long correction phases. As shown in Figure 3, the most recent surge in both spot and long-term prices peaked in early 2024, reaching levels near US$ 190/kgU, before declining again toward US$140/kgU range by early 2025. While this increase was significant, it remains unclear whether it reflects a durable structural shift or a short-term correction driven by speculative activity, restocking cycles, or temporary supply disruptions. What is clear, however, is that the current price level—slightly above US$ 130/kgU—brings a much broader segment of the global uranium resource base into potential economic viability, particularly those deposits situated in < US$ 130 and < US$ 260/kgU cost categories.

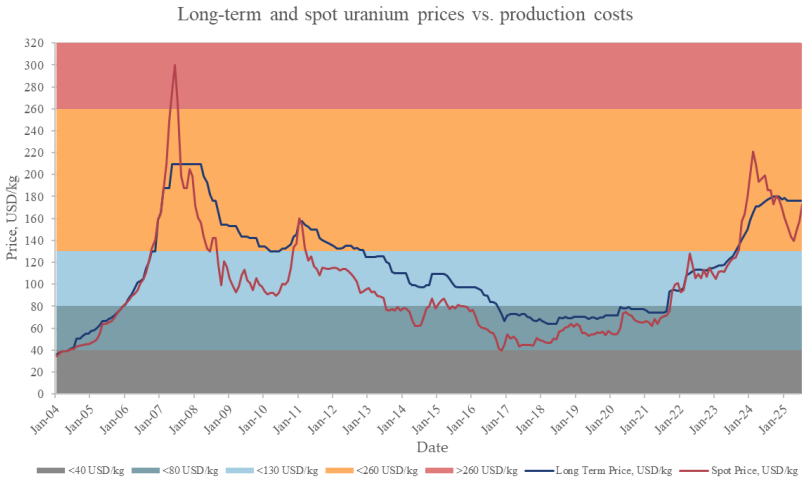


Figure 3. Long-term and spot uranium prices vs. production costs (57)

Looking ahead, several price scenarios may unfold depending on policy, supply chain dynamics, and technological developments. In a moderate-demand scenario, where global nuclear expansion continues at a measured pace and supply steadily recovers, prices are expected to stabilize in the range of US$ 150–170/kgU by 2030. This would be supported by gradual restarts of idled production, new capacity in countries like Kazakhstan and Canada, and improvements in secondary supply. In a high-demand scenario, driven by accelerated nuclear deployment for decarbonization and geopolitical competition over energy security, prices could rise more sharply, reaching US$ 180–200/kgU by 2030 and exceeding US$ 220/kgU by 2035. This would reflect constrained project pipelines and longer permitting timelines for new mines.

A low-demand or oversupply scenario—resulting from delayed reactor construction, wider adoption of alternative fuels, or faster development of small modular reactors (SMRs) with lower uranium requirements—could lead to a price retreat to US$ 100–120/kgU, particularly if new supply enters the market too quickly. However, even under such conditions, long-term prices are unlikely to return to pre-2021 lows, given higher input costs and a more risk-aware investment environment.

For the 2040–2045 period, prices will likely converge in the US$ 160–200/kgU range under most realistic scenarios. At that point, many high-cost deposits in < US$ 260/kgU band may become competitive, especially in countries with integrated fuel cycles or stable mining environments. Overall, future pricing will be shaped not only by physical supply and demand, but also by project economics, geopolitical alignments, and the willingness of states and utilities to support uranium procurement as a strategic priority.

4. 2. Uranium Deposit Types in Russia and Iran by Production Cost To compare the geological composition of uranium resources in Russia and Iran, Figure 4 presents the distribution of deposit types in relative terms (%), grouped by production cost categories. Since Russia's total uranium endowment is significantly larger than Iran's, the data are shown proportionally to highlight structural differences rather than absolute quantities. Only < US$ 130/kgU and < US$ 260/kgU categories are included, as no information has been found indicating that either country holds reserves in < US$ 40/kgU cost band. Although Russia does hold a small volume of sandstone-hosted resources in < US$ 80/kgU category, Iran has none in this cost range, and therefore it was decided to limit the comparison to the two broader thresholds where both countries have meaningful data.

An additional feature worth noting is that Iran's

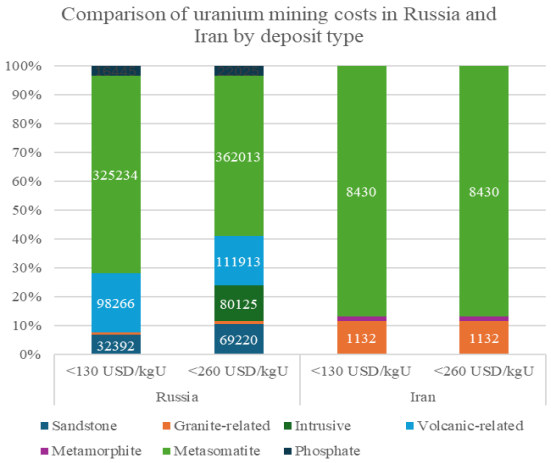


Figure 4. Comparison of uranium mining costs in Russia and Iran by deposit type (11)

uranium resource composition does not change between < US$ 130/kgU and < US$ 260/kgU categories, indicating that all currently identified reserves fall within the lower-cost band. This may suggest that either Iran's resources are uniformly moderate in extraction cost, or that higher-cost reserves have yet to be discovered or disclosed. In contrast, Russia's profile expands significantly in < US$ 260/kgU range, incorporating a wider array of deposit types and highlighting the economic scalability of its uranium base.

The chart reflects that Iran's uranium resources are overwhelmingly metasomatite in origin, while Russia shows a more diverse distribution that includes not only metasomatite, but also significant shares of volcanic-related, intrusive, and sandstone-hosted deposits. This contrast highlights the broader geological base supporting Russia's uranium sector, in contrast to Iran's heavier reliance on a single deposit type.

4. 3. Russia and Iran: Uranium Supply and Demand Outlook In Russia, uranium consumption is projected to increase only slightly over the coming decades—from 4,400 tonnes in 2021 to 4,900 tonnes by 2045—indicating a stable demand outlook based on long-term reactor operation and planned capacity additions. In contrast, production follows a non-linear path. It rises from 2,635 tonnes in 2021 to a peak of 4,900 tonnes by 2030, achieving full domestic supply coverage at that point. However, this is followed by a sharp decline to 1,500 tonnes by 2040, a level that remains unchanged through 2045 (11). This creates a projected shortfall of over 3,000 tonnes per year by the end of the period. The pattern reflects the expected depletion of existing deposits and the absence of new capacity being added in the model. The corresponding production and consumption figures are shown in Figure 5.

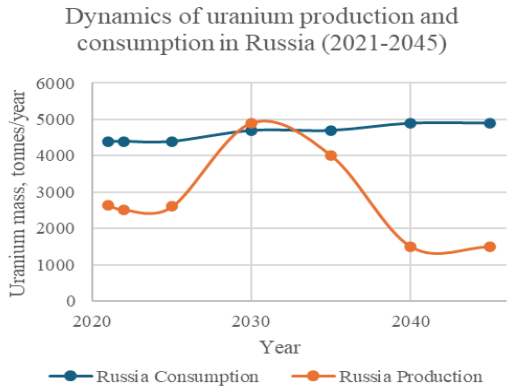


Figure 5. Dynamics of uranium production and consumption in Russia (2021-2045)

Iran, by contrast, faces a different type of imbalance. Uranium consumption is projected under two alternative demand scenarios, depending on the pace of reactor development. In the high scenario, consumption rises from 160 tonnes in 2021 to 1,390 tonnes by 2040. In the low scenario, it reaches 1,230 tonnes over the same period. Meanwhile, production increases only modestly—from 21 tonnes per year in 2021–2022 to 71 tonnes from 2025 onward—and then remains flat (11). In both scenarios, the gap between demand and domestic supply widens steadily, indicating that Iran will remain heavily dependent on imports or cooperation to meet its future needs. These projections are presented in Figure 6. Taken together, the Russian and Iranian cases illustrate two distinct challenges. Russia risks a supply deficit after 2035 despite currently strong production, while Iran will face chronic underproduction relative to even modest growth in demand. These dynamics underscore the strategic value of cooperation in uranium development, particularly in areas such as reserve expansion, exploration technology, and long-term supply agreements.

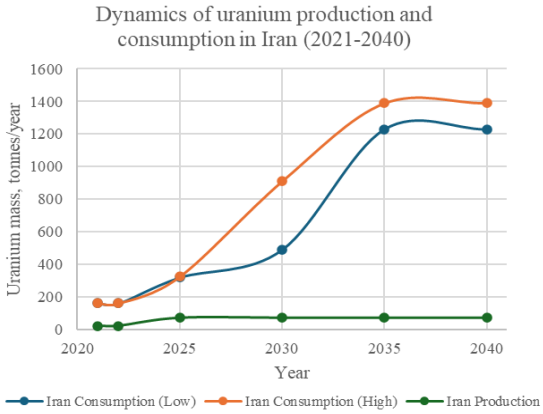


Figure 6. Dynamics of uranium production and consumption in Iran (2021-2045)

4. 4. Uranium Self-Sufficiency Horizons in Russia and Iran An important outcome of this study is the estimation of how long Russia and Iran can meet their domestic uranium demand using their own identified resources, differentiated by cost categories and resource classifications (RAR and RAR+IR). The results provide a direct timeline for potential resource depletion under current consumption projections.

For Russia, two cost-based resource thresholds were considered. In the < US$ 80/kgU category, where the country holds a moderate reserve (18,224 tonnes RAR and 14,168 tonnes IR), calculations show that reasonably assured resources (RAR) alone will be depleted by 2036, while RAR+IR combined will last until 2041. These dynamics are illustrated in Figure 7, which shows a steep decline in both reserve types over time, especially beyond 2030. When evaluating the broader < US$ 130/kgU category—which includes a significantly larger resource base (202,304 tonnes RAR and 274,269 tonnes IR)—the self-sufficiency horizon expands substantially.

As shown in Figure 8, reserves in this category remain relatively stable across the period 2021–2045. In this case, RAR would suffice until 2093, and RAR+IR would extend the timeline until 2174. Since the < US$ 130/kgU band already provides such long-term coverage, additional analysis of higher-cost categories was deemed unnecessary for Russia.

For Iran, no reserves exist in the < US$ 80/kgU range, and the < US$ 260/kgU category does not add any new resources beyond what is already classified under < US$ 130/kgU. Therefore, only the < US$ 130/kgU category was analyzed. Given Iran's much smaller resource base and growing projected demand, self-sufficiency timelines are significantly shorter. As illustrated in Figure 10, under the low consumption scenario, Iran can meet its needs using RAR alone until 2033, and with RAR+IR until 2039. Under the high consumption scenario, these timelines shorten further: until 2031 for RAR and 2038 for RAR+IR.

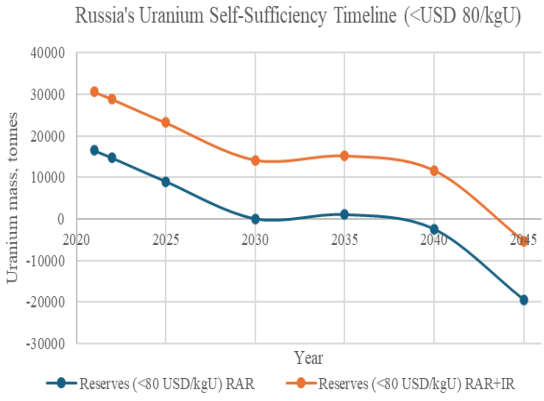


Figure 7. Russia's uranium self-sufficiency timeline (<US$ 80/kgU)

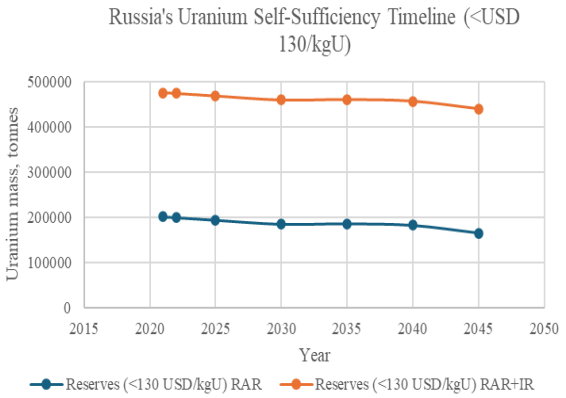


Figure 8. Russia's uranium self-sufficiency timeline (<US$ 130/kgU)

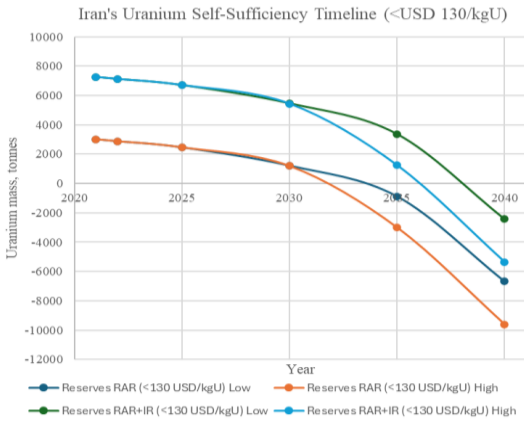


Figure 9. Iran's uranium self-sufficiency timeline (<US$ 130/kgU)

These results highlight a key structural difference between the two countries. While Russia has sufficient domestic reserves to support its nuclear program for decades—particularly at cost levels under US$ 130/kgU—Iran's limited resource base, especially in the context of rising demand, leads to a much earlier depletion horizon. This calculation of specific depletion years represents a core methodological feature of the study and provides a clear benchmark for assessing long-term uranium security under varying economic and policy scenarios.

5. DISCUSSION

The analysis of uranium price dynamics and resource self-sufficiency provides a clear framework for assessing national vulnerability and long-term planning needs. However, the practical impact of future price changes—despite their potential volatility—is likely to be limited for both Russia and Iran in the near to medium term. For Iran, the reason is straightforward: the country's uranium reserves are modest in size, concentrated in a narrow geological category, and largely insufficient to meet future demand under either low or high consumption scenarios. As such, even if uranium prices were to shift significantly, the overall strategic position of Iran's domestic supply would remain largely unchanged. The issue is not one of cost, but of scale.

In contrast, Russia possesses a substantial uranium endowment, including large volumes of reasonably assured and inferred resources at < US$ 130/kgU. This suggests that price increases would not pose an immediate threat to supply security, as sufficient reserves are already technically and economically viable at current market levels. However, the analysis also shows that only a small share of these reserves is actively being developed. As production is expected to decline after 2030, Russia's long-term self-sufficiency will depend not only on the size of its reserves but also on its ability to bring new capacity online. Without proactive development, even large resource bases can become bottlenecks—particularly if geopolitical uncertainty or global market tightening reduces access to foreign supply chains (58, 59).

It is also important to acknowledge that the presented figures and forecasts are based on conventional uranium fuel use and do not account for potential shifts toward MOX fuel or broader implementation of closed fuel cycle technologies. As these technologies continue to evolve, they may fundamentally alter the economics of resource use, extending the effective lifespan of current reserves and introducing new recovery pathways from previously uneconomic or unrecoverable sources (60). Similarly, improved processing, modular mining systems, or environmental permitting innovations could lead to a reclassification of existing reserves across price tiers (61-63). In this sense, today's marginal resources may become tomorrow's baseline supply (64).

Altogether, while the current outlook favors Russia in terms of scale and buffer capacity, and poses more urgent concerns for Iran, both countries face important challenges. For Iran, these relate to securing external partnerships or developing alternative technologies (65). For Russia, the challenge lies in converting its geological potential into timely, sustained production. Addressing these asymmetries will require long-term strategic planning that goes beyond price forecasting and into active resource policy and technology integration (66-68).

An additional consideration is the geological similarity of uranium reserves in both countries. As shown in the deposit-type analysis, metasomatite-hosted deposits dominate the resource base in Iran and also form a significant part of Russia's uranium portfolio. This commonality may offer technical advantages for bilateral cooperation, particularly in the areas of ore processing, exploration techniques, and mine planning. Given

Russia's greater experience in developing metasomatite resources, shared geological challenges could serve as a practical foundation for deeper collaboration—not only in fuel cycle agreements, but also in joint field development, personnel training, and technology transfer (69).

6. CONCLUSIONS

The results of this study provide a detailed response to the stated aim and objectives. The analysis of uranium resources in Russia and Iran by type (RAR and IR) and by economic cost categories (<40, <80, <130, and < US$ 260/kgU) reveals significant differences in scale and composition. Russia possesses one of the world's largest uranium resource bases, with a substantial share of reserves classified in categories below US$ 130/kgU. In absolute terms, Iran's endowment is much smaller, and although its reserves also fall within these cost categories, their limited volume constrains the country's long-term supply potential.

Since Russia reports recoverable resources while Iran reports in situ values, recovery factors specific to each mining method were applied to adjust the Iranian figures. This adjustment enabled a comparable assessment of the resource bases of both countries. Based on this, a structural analysis by deposit type shows that Iran's uranium resources are concentrated mainly in metasomatite deposits, whereas Russia's reserves are distributed across a broader spectrum, including metasomatite, sandstone, volcanic, and other deposits. Geological diversity strengthens Russia's resilience, while Iran's reliance on a single deposit type increases its vulnerability to domestic constraints.

The evaluation of production and consumption further underscores these differences. Russia has traditionally maintained substantial uranium output, but a decline is projected after 2030 if new projects are not developed. Forecasts to 2045 indicate that Russia can remain self-sufficient provided additional capacity is commissioned. In Iran, current uranium production remains limited, reflecting both the smaller scale of its resource base and institutional barriers to development. At the same time, the country is preparing feasibility studies for additional deposits such as Narigan and Khushumi, which may expand its production potential in the future. However, under both low and high consumption scenarios, the currently identified resources are insufficient to sustain demand beyond the 2030s.

The analysis also reveals complementarities between the two countries. Both possess significant metasomatite-hosted deposits, and Russia's broader experience in the exploration and development of such deposits could provide a basis for bilateral cooperation in exploration, mining, processing, and technology transfer. This creates a practical foundation for strengthening uranium security through joint initiatives, while also underscoring the broader strategic implications of geological similarities.

Overall, the findings demonstrate that Russia's extensive and geologically diverse uranium base can ensure long-term self-sufficiency with timely investment in new projects, whereas Iran's more limited endowment makes external cooperation and technological breakthroughs essential for sustaining its nuclear program. These results address the stated objectives by delivering a comparative geological and economic assessment, establishing data comparability through recovery factors, analyzing resource structures by deposit type and cost category, evaluating production and consumption balances, and estimating long-term self-sufficiency horizons together with opportunities for bilateral collaboration.

In contrast to earlier works, which primarily focused on the political, diplomatic, or institutional dimensions of Russian–Iranian cooperation or emphasized Rosatom's foreign policy role and global positioning, the present study provides a quantitative geological and economic perspective. It takes into account that Russian uranium resources are reported as recoverable while Iranian resources are presented in situ, and it applies appropriate recovery factors to bring them onto a comparable basis. By examining reserve volumes by deposit type and economic cost category and by estimating long-term self-sufficiency horizons, this study presents findings that have not been addressed in previous literature and thereby broadens the understanding of uranium supply security in both Russia and Iran.

Acknowledgements

The authors express their gratitude to colleagues and reviewers for their valuable comments and suggestions, which contributed to improving the quality of this manuscript.

Funding

This research received no external funding.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The authors declare no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors did not use generative artificial intelligence or AI-assisted technologies in the writing or preparation of the manuscript. The authors take full responsibility for the content of the published article.

Authors Biosketches

The authors are researchers in the field of sectoral economics, focusing on the analysis of industrial sectors and the assessment of their economic performance. Their research interests include quantitative methods, comparative analysis, and applied economic evaluation.

REFERENCES

1. Degueldre C. Uranium as a renewable for nuclear energy. Progress in Nuclear Energy. 2017;94:174-86. 10.1016/j.pnucene.2016.03.031.
2. Nguyen M. P., Ponomarenko T., Nguyen N. Energy Transition in Vietnam: A Strategic Analysis and Forecast. Sustainability. 2024;16(5). 10.3390/su16051969.
3. Babyr N. V. Topical Themes and New Trends in Mining Industry: Scientometric Analysis and Research Visualization. International Journal of Engineering Transactions B: Applications. 2024;37(2):439-51. 10.5829/ije.2024.37.02b.18.
4. Taylor R., Bodel W., Banford A., Butler G., Livens F. Sustainability of Nuclear Energy—A Critical Review from a UK Perspective. Sustainability. 2024;16(24). 10.3390/su162410952.
5. Marinina O. A., Ilyushin Y. V., Kildiushov E. V. Comprehensive Analysis and Forecasting of Indicators of Sustainable Development of Nuclear Industry Enterprises. International Journal of Engineering Transactions B: Applications 2025;38(11):2527-36. 10.5829/ije.2025.38.11b.05.
6. Lyócsa Š., Todorova N. What drives the uranium sector risk? The role of attention, economic and geopolitical uncertainty. Energy Economics. 2024;140:107980. 10.1016/j.eneco.2024.107980.
7. Ilyushin Y. V., Nosova V. A. Development of Mathematical Model for Forecasting the Production Rate. International Journal of Engineering Transactions B: Applications. 2025;38(8):1749-57. 10.5829/ije.2025.38.08b.02.
8. Monnet A., Gabriel S., Percebois J. Long-term availability of global uranium resources. Resources Policy. 2017;53:394-407. 10.1016/j.resourpol.2017.07.008.
9. Pashkevich N. V., Khloponina, V. S., Pozdnyakov, N. A., Avericheva, A. A. Analysing the problems of reproducing the mineral resource base of scarce strategic minerals. Journal of Mining Institute. 2024;270:1004-23. EDN HNTQBF.
10. Semenova T., Churrana N. Assessment of the Projects' Prospects in the Economic and Technological Development of the Oil and Gas Complex in the Republic of Mozambique. Resources. 2025;14(7). 10.3390/resources14070106.
11. Nuclear Energy Agency (NEA). Uranium 2024: Resources, Production and Demand. Paris: OECD Publishing; 2025.
12. Agarwal D. K., and Sreenivas B. An appraisal of uranium deposits of India and their style of deposition with reference to the Paleoproterozoic great oxidation event. International Geology Review. 2021;63(5):571-84. 10.1080/00206814.2020.1728583.
13. Hillacre S., Ansdell K., McEwan B. Geology, Structural Analysis, and Paragenesis of the Arrow Uranium Deposit, Western Athabasca Basin, Saskatchewan, Canada: Implications for the Development of the Patterson Lake Corridor. Economic Geology. 2021;116(2):285-321. 10.5382/econgeo.4797.
14. Groves D. I., Bierlein F. P., Meinert L. D., Hitzman M. W. Iron Oxide Copper-Gold (IOCG) Deposits through Earth History: Implications for Origin, Lithospheric Setting, and Distinction from Other Epigenetic Iron Oxide Deposits. Economic Geology. 2010;105(3):641-54. 10.2113/gsecongeo.105.3.641.
15. Stepanov V. A. Witwatersrand composite paleoplacers. Proceedings of higher educational establishments Geology and Exploration. 2021;63(4):8-22. 10.32454/0016-7762-2020-63-4-8-22.
16. Mishra P., Sati M., Krishnamurthi R. A Review on Uranium Mineralization Related to Na-Metasomatism: Indian and International Examples. Geosciences. 2024;14(11). 10.3390/geosciences14110304.
17. Lapido Loureiro F. E. D. V., Dos Santos R. C. The intra-intrusive uranium deposits of poços de caldas, Brazil. Ore Geology Reviews. 1988;3(1):227-40. 10.1016/0169-1368(88)90020-0.
18. Shanyengana S. H., Fan H., Xue C., Shanyengana E. S., Mapani B., Shilongo E., et al. An example of uraniferous leucogranites in the Rössing South-West deposit, Namibia. Journal of African Earth Sciences. 2020;162:103698. 10.1016/j.jafrearsci.2019.103698.
19. Steenfelt A., Kolb J., Thrane K. Metallogeny of South Greenland: A review of geological evolution, mineral occurrences and geochemical exploration data. Ore Geology Reviews. 2016;77:194-245. 10.1016/j.oregeorev.2016.02.005.
20. Pek A. A., Malkovsky V. I., Petrov V. A. Mineral System of the Streltsovka Caldera Uranium Deposits (East Transbaikalia). Geology of Ore Deposits. 2020;62(1):31-48. 10.1134/S1075701520010055.
21. Yu H., Deng J., Tang B., Egbert G., Chen H. Electrical image of magmatic system beneath the Xiangshan volcanogenic uranium deposit, southeast China: Linking magmatic evolution and uranium metallogenesis. Geology. 2023;51(9):870-4. 10.1130/G50783.1.
22. Bonnetti C., Liu X., Mercadier J., Cuney M., Deloule E., Villeneuve J., et al. The genesis of granite-related hydrothermal uranium deposits in the Xiazhuang and Zhuguang ore fields, North Guangdong Province, SE China: Insights from mineralogical, trace elements and U-Pb isotopes signatures of the U mineralisation. Ore Geology Reviews. 2018;92:588-612. 10.1016/j.oregeorev.2017.12.010.
23. Ballouard C., Poujol M., Boulvais P., Mercadier J., Tartèse R., Venneman T., et al. Magmatic and hydrothermal behavior of uranium in syntectonic leucogranites: The uranium mineralization associated with the Hercynian Guérande granite (Armorican Massif, France). Ore Geology Reviews. 2017;80:309-31. 10.1016/j.oregeorev.2016.06.034.
24. Cuney M. Evolution of Uranium Fractionation Processes through Time: Driving the Secular Variation of Uranium Deposit Types. Economic Geology. 2010;105(3):553-69. 10.2113/gsecongeo.105.3.553.

25. Vind J., Tamm K. Review of the extraction of key metallic values from black shales in relation to their geological and mineralogical properties. Minerals Engineering. 2021;174:107271. 10.1016/j.mineng.2021.107271.
26. Tillman F. D., Anderson J. R., Unema J. A., Chapin T. P. Assessing uranium and select trace elements associated with breccia pipe uranium deposits in the Colorado River and main tributaries in Grand Canyon, USA. PLOS ONE. 2020;15(11):e0241502. 10.1371/journal.pone.0241502.
27. Osmanlioglu A. E. Uranium Mining Techniques and Waste Management. European Journal of Sustainable Development Research. 2022;6(4):1-4. 10.21601/ejosdr/12273.
28. Li G., Yao J. A Review of In Situ Leaching (ISL) for Uranium Mining. Mining. 2024;4(1):120-48. 10.3390/mining4010009.
29. Alrawashdeh R., Altiti A., Alnawafleh H. Open Pit Mining. In: Soni AK, editor. Mining Techniques - Past, Present and Future. Rijeka: IntechOpen; 2021.
30. Tien D. L., Trung T. V., Anh S. D., Babyr N. V. Ground Pressure and Methods to Enhance Roof Stability in Mechanized Coal Mining. International Journal of Engineering Transactions A: Basics. 2026;39(4):862-9. 10.5829/ije.2026.39.04a.05.
31. Tulyakov T. F., Afanaseva O. V. Comparative Analysis of Image Segmentation Methods in Power Line Monitoring Systems. International Journal of Engineering Transactions A: Basics. 2026;39(1):1-11. 10.5829/ije.2026.39.01a.01.
32. Toews K. L., Takala J. M. Radiation protection strategies in high-grade underground uranium mines. Annals of the ICRP. 2024;53(1_suppl):296-300. 10.1177/01466453241283931ap.
33. Ilyushin Y., Nosova V., Krauze A. Application of Systems Analysis Methods to Construct a Virtual Model of the Field. Energies. 2025;18(4). 10.3390/en18041012.
34. Sidorenko S., Trushnikov V., Sidorenko A. Methane Emission Estimation Tools as a Basis for Sustainable Underground Mining of Gas-Bearing Coal Seams. Sustainability. 2024;16(8). 10.3390/su16083457.
35. Mudd G. M., Weng Z., Jowitt S. M., Turnbull I. D., Graedel T. E. Quantifying the recoverable resources of by-product metals: The case of cobalt. Ore Geology Reviews. 2013;55:87-98. 10.1016/j.oregeorev.2013.04.010.
36. Michael A., Varnava N., Ioannidis I., Pashalidis I. Uranium recovery from phosphate rocks/minerals – A comprehensive review. Sustainable Chemistry for the Environment. 2024;5:100055. 10.1016/j.scenv.2023.100055.
37. Shklyarskiy Y. E., Skamyin A. N., Jiménez Carrizosa M. Energy efficiency in the mineral resources and raw materials complex. Journal of Mining Institute. 2023;261:323-4.
38. Zhao J., Liu C. Unlocking the microscale secrets: A comprehensive study on uranium occurrence and eco-friendly processing techniques in Lanhe granite-type uranium deposits, Northern Guangdong Province, China. Minerals Engineering. 2024;215:108817. 10.1016/j.mineng.2024.108817.
39. Wang Z., Song H., Chen Y., Song J., Hou M., Li Q., et al. Uranium resource of Europe: Development status, metallogenic provinces and geodynamic setting. Energy Strategy Reviews. 2024;54:101467. 10.1016/j.esr.2024.101467.
40. Togizov K., Kenzhetaev Z., Temirkhanova R., Muzapparova A., Omirgali A., Altaibayev B. The Influence of the Physicochemical Characteristics of Ores on the Efficiency of Underground Well Leaching of Uranium Deposits in Kazakhstan. Minerals. 2024;14(4). 10.3390/min14040381.
41. Ye Y., Fan B., Qin Z., Tang X., Feng Y., Lv M., et al. Electrochemical removal and recovery of uranium: Effects of operation conditions, mechanisms, and implications. Journal of Hazardous Materials. 2022;432:128723. 10.1016/j.jhazmat.2022.128723.
42. Langanay J., Romary T., Freulon X., Langlais V., Petit G., Lagneau V. Uncertainty quantification for uranium production in mining exploitation by In Situ Recovery. Computational Geosciences. 2021;25(3):831-50. 10.1007/s10596-020-10018-x.
43. Bruneton P., Cuney M. 2 - Geology of uranium deposits. In: Hore-Lacy I, editor. Uranium for Nuclear Power: Woodhead Publishing; 2016. p. 11-52.
44. Wang Z., Xing W. Study on the Characteristics and Evolution Trends of Global Uranium Resource Trade from the Perspective of a Complex Network. Sustainability. 2022;14(22). 10.3390/su142215295.
45. Sadiq M., Shinwari R., Wen F., Usman M., Hassan S. T., Taghizadeh-Hesary F. Do globalization and nuclear energy intensify the environmental costs in top nuclear energy-consuming countries? Progress in Nuclear Energy. 2023;156:104533. 10.1016/j.pnucene.2022.104533.
46. Egorova I. Prospects for the natural uranium world market. Ores and Metals. 2023;1:6-16. 10.47765/0869-5997-2023-10001.
47. Bahl B. Iranian nuclear energy: history and advancements. Jindal Journal of International Affairs. 2022;10(1):77-89. 10.54945/jjia.v1i6.108.
48. El Khalfi M. A. Agreement on the Joint Comprehensive Plan of Action (JCPOA) Between Iran and The United States. Jurnal Pembaharuan Hukum. 2020;7(2):183. 10.26532/jph.v7i2.11296.
49. Kalbasi R. A. Prospects for Russian-Iranian relations in the nuclear Field. Post-Soviet Issues. 2019;6(2):195-202. 10.24975/2313-8920-2019-6-2-195-202.
50. Mirgorod D., Parubochaya E. Nuclear Energy as a Russian Foreign Policy Instrument in the Middle East. Vestnik Volgogradskogo gosudarstvennogo universiteta Serija 4 Istorija Regionovedenie Mezhdunarodnye otnoshenija. 2023;28(2):214-25. 10.15688/jvolsu4.2023.2.18.
51. Siddi M., Silvan K. Nuclear energy and international relations: the external strategy of Russia's Rosatom. International Politics. 2024. 10.1057/s41311-024-00618-0.
52. Kharitonova D. V. Prospects for Russian-Iranian Energy Cooperation. Geoeconomics of Energetics. 2024(4):77-100. 10.48137/26870703_2023_24_4_77.
53. Malozyomov B. V., Martyushev N. V., Babyr N. V., Pogrebnoy A. V., Efremenkov E. A., Valuev D. V., et al. Modelling of Reliability Indicators of a Mining Plant. Mathematics. 2024;12(18). 10.3390/math12182842.
54. Golovina E. I., Tselmeg B. Cost Estimate as a Tool for Managing Fresh Groundwater Resources in the Russian Federation. Geology and mineral resources of Siberia. 2023;a(4):81-91. 10.20403/2078-0575-2023-4a-81-91.
55. Galevskiy S., Qian H. A Binary Discounting Method for Economic Evaluation of Hydrogen Projects: Applicability Study Based on Levelized Cost of Hydrogen (LCOH). Energies. 2025;18(14). 10.3390/en18143839.
56. Mirkhusanov U. T., Semenova D. Y., Kharitonov V. V. Forecasting the Cost and Volume of Uranium Mining for different World Nuclear Energy Development Scenario. Izvestiya Wysshikh Uchebnykh Zawedeniy, Yadernaya Energetika. 2024(1):119-31. 10.26583/npe.2024.1.10.
57. Uranium price. Cameco Corporation. Available online: https://www.cameco.com/invest/markets/uranium-price (accessed on 06.04.2025).
58. Semenova T., Sokolov I. Theoretical Substantiation of Risk Assessment Directions in the Development of Fields with Hard-to-Recover Hydrocarbon Reserves. Resources. 2025;14(4). 10.3390/resources14040064.
59. Golovina E., Karennik K. Modern Trends in the Field of Solving Transboundary Problems in Groundwater Extraction. Resources. 2021;10(10). 10.3390/resources10100107.

60. Galevskiy S. G., Qian H. Developing and validating comprehensive indicators to evaluate the economic efficiency of hydrogen energy investments. ORESTA. 2024;7(3). 10.5281/zenodo.15093154.
61. Kukharova T. V., Utkin, V. A., Pershin, I. M. Modeling of a Decision Support System for a Psychiatrist Based on the Dynamics of Electrical Conductivity Parameters. 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus): IEEE; 2021. p. 975-8. 10.1109/ElConRus51938.2021.9396273
62. Martirosyan A. V., Martirosyan, K. V., Grudyaeva, E. K., Chernyshev, A. B. Calculation of the Temperature Maximum Value Access Time at the Observation Point. 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus); St. Petersburg, Moscow, Russia2021. p. 1014-8. 10.1109/ElConRus51938.2021.9396287
63. Martirosyan A. V., Martirosyan, K. V., Chernyshev, A. B. Calculation of the First Switch-on Time of Distributed Object's Control Action. 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus); St. Petersburg and Moscow, Russia2020. p. 750-4.
64. Kukharova T., Maltsev P., Novozhilov I. Development of a Control System for Pressure Distribution During Gas Production in a Structurally Complex Field. Applied System Innovation. 2025;8(2). 10.3390/asi8020051.
65. Semenova T., Martínez Santoyo J. Y. Determining Priority Areas for the Technological Development of Oil Companies in Mexico. Resources. 2025;14(1). 10.3390/resources14010018.
66. Farkhutdinov I. M., Khayrullin R. R., Soktoev B. R., Zlobina A. N., Chesalova E. I., Farkhutdinov A. M., et al. Uranium in man-made carbonates on the territory of Ufa. Journal of Mining Institute. 2023;260:226-37. 10.31897/PMI.2023.4.
67. Cherepovitsyn A. E., Dorozhkina I. P., Solov'eva V. M. Forecasts of Rare-earth Elements Consumption in Russia: Basic and Emerging Industries. Studies on Russian Economic Development. 2024;35(5):688-96. 10.1134/S1075700724700229.
68. Kongar-Syuryun C., Babyr N., Klyuev R., Khayrutdinov M., Zaalishvili V., Agafonov V. Model for Assessing Efficiency of Processing Geo-Resources, Providing Full Cycle for Development—Case Study in Russia. Resources. 2025;14(3). 10.3390/resources14030051.
69. Kotov D., Pervukhin D., Davardoost H., Afanaseva O. Prospects for the Use of Autonomous Underwater Vehicles (AUV) to Solve the Problems of the Mineral Resources Complex (MRC) of the Russian Federation. Journal of Maritime Research. 2024;21:309-17.

Persian Abstract

چکیده

این مطالعه ارزیابی مقایسه ای زمین شناسی و اقتصادی منابع اورانیوم در روسیه و ایران را با تمرکز بر توانایی آنها در حمایت از توسعه انرژی هسته ای داخلی انجام می دهد . برخلاف تحقیقات قبلی که عمدتا همکاری روسیه و ایران را از دیدگاه سیاسی یا نهادی بررسی کرده است ، این کار در دسترس بودن اورانیوم را با استفاده از طبقه بندی های زمین شناسی و دسته بندی های هزینه اقتصادی ذخایر ارزیابی می کند .این روش بین منابع قابل بازیابی گزارش شده برای روسیه و منابع داخلی گزارش شده برای ایران تمایز قائل است و این مورد دوم از طریق استفاده از عوامل مناسب بازیابی به یک پایه قابل مقایسه تبدیل شده است .علاوه بر تجزیه و تحلیل ساختار و پتانسیل پایگاه منابع ، این مطالعه تولید اورانیوم تاریخی و پیش بینی شده و همچنین سناریوهای مصرف برای هر دو کشور را در نظر می گیرد .بر این اساس ، افق های بلند مدت خودکفایی تا سال ۲۰۴۵ تخمین زده می شود .یافته ها نشان می دهد که روسیه دارای یک پایگاه اورانیوم گسترده و زمین شناسی متنوع است که قادر به تضمین عرضه بلند مدت است اگر پروژه های معدن جدید برای جبران کاهش تولید پس از سال ۲۰۳۰ توسعه یابد .در مقابل ، ایران ذخایر محدودتر و گرانتر دارد که عمدتا در ذخایر متاسوماتیت متمرکز شده است ، که دستیابی به خودکفایی پایدار را بدون همکاری خارجی یا پیشرفت های تکنولوژیکی بعید می کند .در عین حال ، شباهت های زمین شناسی بین دو کشور پایه ای برای همکاری بالقوه در اکتشاف و توسعه اورانیوم ایجاد می کند .با پرداختن به جنبه نسبتا کم کشف شده ادبیات ، این مطالعه ارزیابی ساختاری از پایداری عرضه اورانیوم را ارائه می دهد و دیدگاه مقایسه ای را نه تنها برای روسیه و ایران بلکه برای سایر کشورهای دارای برنامه های هسته ای در حال ظهور نیز معرفی می کند.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2434-2451

International Journal of Engineering

Journal Homepage: www.ije.ir

VidaFormer: A Hybrid Transformer for Image Steganography

V. Yousefi Ramandi, M. Fateh*, M. Rezvani

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

P A P E R I N F O

Paper history:
Received 23 July 2026
Received in revised form 15 December 2025
Accepted 04 January 2026

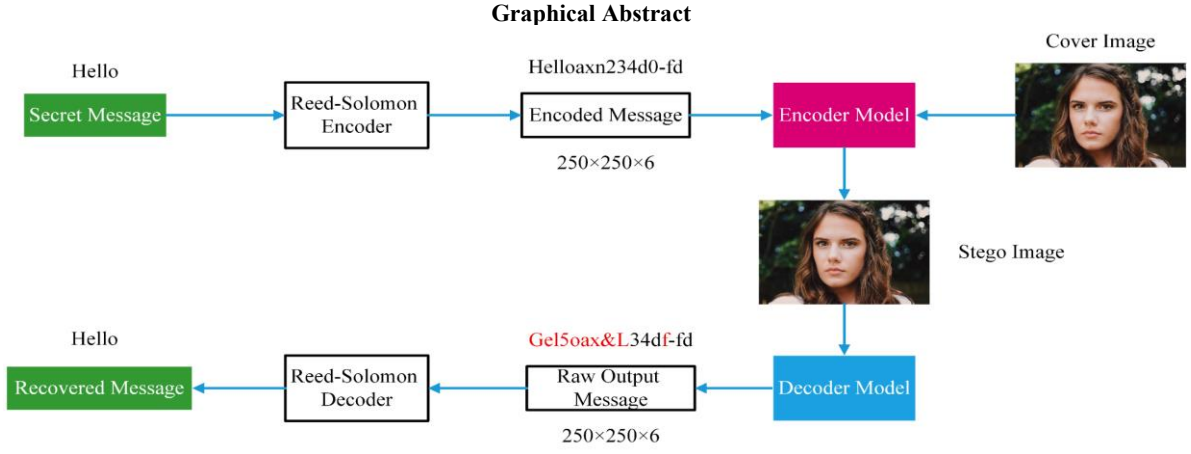
Keywords:
Image Steganography
Information Hiding
Deep Learning
Convolution
Transformers

A B S T R A C T

Image steganography has advanced significantly with the integration of deep learning, particularly through the use of convolutional neural networks (CNNs) for extracting complex features. While CNNs have long been a cornerstone of image analysis due to their efficiency in capturing local dependencies, the emergence of transformer models has revolutionized the field by achieving superior accuracy and effectively capturing global relationships within data. Transformers encounter issues with computational expense and memory consumption, especially when handling high-resolution images. To address these limitations, we propose a hybrid architecture that uses convolutional layers in high-resolution stages and transformer blocks in low-resolution stages. This design balances efficiency and performance by leveraging convolutional layers to capture short-range dependencies and transformers to model long-range relationships. Additionally, by replacing standard convolutions with CoordConv layers, we enhance the model's spatial awareness and feature localization capabilities. We introduce VidaFormer, an innovative steganography framework that integrates deep learning with these architectural enhancements to embed arbitrary binary data into images while maintaining high visual quality in the resulting stego images. VidaFormer outperforms state-of-the-art models, achieving an effective capacity of 4.89 bits-per-pixel on the DIV2K dataset, a 1.7% improvement over the previous best model. This demonstrates VidaFormer's superior performance and scalability for modern image steganography applications.

doi: *10.5829/ije.2026.39.10a.07*

Graphical Abstract



1. INTRODUCTION

Image steganography is the practice of embedding hidden information within digital images, making the concealed data imperceptible to the human eye. This technique is valuable for applications requiring secret data transmission (1) and digital watermarking (2, 3), as it provides a layer of secrecy without raising suspicion.

*Corresponding Author Email: mansoor_fateh@shahroodut.ac.ir (M. Fateh)

Please cite this article as: Yousefi Ramandi V, Fateh M, Rezvani M. VidaFormer: A Hybrid Transformer for Image Steganography. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2434-51.

Image steganography primarily involves two tasks: embedding and extraction. The embedding task, performed by the encoder (see Figure 1), conceals a message within a cover image to produce a stego image, which can then be transmitted through public channels to the receiver. The extraction task, executed by the decoder, retrieves the concealed message from the stego image, ensuring no part of the message is lost or corrupted. These tasks are inherently adversarial: making the stego image indistinguishable from the cover image may compromise the embedded message's integrity. Conversely, relaxing the similarity constraint between the cover and stego images can produce unrealistic stego images, potentially causing noticeable distortions.

Image steganography encompasses several key concepts. Embedding capacity, or simply capacity, measured in bits-per-pixel (BPP), determines how many bits can be hidden per pixel of an image. Recovery accuracy is defined as the fraction of correctly extracted message bits. To correct erroneous bits in the extracted message, error-correction algorithms are employed, which require adding extra data to the original message before embedding. Effective capacity, also measured in bits-per-pixel (BPP), represents the portion of the embedding capacity that corresponds to the original message, excluding additional error-correction data (4). It is determined by embedding capacity and recovery accuracy as discussed in the next the sections.

Transparency aims to make the concealed message harder for attackers to detect and recover. As demonstrated in the experiments, effective capacity and transparency objectives are adversarial: for a given model, increasing effective capacity reduces transparency, and vice versa. Robustness means being resilient to attacks such as noise or image compression (5). A steganographic attack on an image involves a third party degrading it in some way, resulting in a compromised image. A more robust algorithm experiences less reduction in capacity due to such an attack. However, robustness conflicts with capacity and transparency, so increasing robustness typically reduces the other two, and vice versa (6). Traditional methods, such as the least significant bit (LSB) approach, often rely on heuristic processing (7). These methods embed secret data by modifying the least significant bit of each pixel in an image, resulting in minimal visual distortion while covertly embedding the information. Although simple and effective, the LSB method is relatively vulnerable to detection.

Over the past few years, image steganography has transitioned from traditional embedding methods to more advanced deep learning-based approaches. These modern techniques focus on improving two key objectives: embedding information into images in a way that keeps the hidden data undetectable to unauthorized viewers and retrieving longer messages accurately without any loss of data.

Almost all recently proposed methods (8-10) leverage the power of deep learning to increase the capacity of data that an image can hold while maintaining its fidelity. These methods typically use convolutional layers in architectures designed for dense prediction, enabling the embedding and extraction of data at the pixel level. Common architectures employed in these approaches include U-Net (11), CSP-Net (12), and Dense-Net (13).

Although Convolutional Neural Networks (CNNs) have significantly propelled advancements in computer vision, their dependence on local receptive fields poses challenges, especially when it comes to effectively modeling long-range dependencies (14). These challenges can be effectively addressed by incorporating vision transformers, which leverage global attention mechanisms to enhance feature representation and spatial understanding.

Vision transformers (ViTs) (15) are a type of deep learning model that utilizes the self-attention mechanism instead of traditional convolutions to process images. These models divide images into patches and treat them as sequences, similar to natural language processing, allowing them to capture long-range dependencies and global context (16).

Newer transformers like Swin (17) and Swin V2 (18) introduced the concept of shifted windows to overcome the limitations of earlier models. Unlike ViT, which divides images into fixed-size patches and applies global self-attention across the entire image, Swin models use local windows to perform self-attention on smaller, non-overlapping regions. This local attention approach significantly reduces the computational complexity of vision transformer models, making Swin transformer more efficient and practical for various applications.

Although Swin transformers are more efficient than ViT models in terms of computational complexity and scalability, both architectures encounter significant challenges when dealing with high-resolution images. This limitation becomes particularly evident in tasks where preserving full-resolution data is critical, such as image steganography. Vision transformers encode each patch of pixels into an embedding vector, which makes restoring data to full resolution much harder (19, 20). We addressed this challenge in VidaFormer by using transformers only in the bottleneck stage.

To achieve superior results, we had to overcome the issue that standard convolutions lack an understanding of spatial positions (21). Augmenting the processing with positional encoding can improve the performance of convolutional layers. To achieve this, we used CoordConv (22) layers, which enhance traditional convolutional layers by explicitly incorporating spatial coordinates into the input, i.e. adding two additional channels to the input: One encoding the x-coordinates and the other encoding the y-coordinates of each pixel.

This enhancement allows the network to better capture spatial relationships, enabling it to handle tasks requiring precise localization more effectively.

This study presents VidaFormer, an innovative hybrid architecture for image steganography that combines convolutional methods with transformer-based approaches. VidaFormer sets a new benchmark by excelling in hiding capacity and recovery accuracy, all while maintaining the stego image's high fidelity. This approach is founded on a U-Net-inspired architecture, enhanced with CoordConv blocks and Swin V2 transformers, to create a robust and efficient framework for securely embedding and recovering hidden messages in images. To further enhance reliability, Reed-Solomon error correction (23) is utilized, enabling recovery even if portions of the message are corrupted. This combination of techniques addresses the dual challenges of fidelity and data accuracy in steganographic applications.

The proposed architecture marks a notable breakthrough in image steganography, achieving a balance between hiding capability and exceptional performance across various scenarios. By leveraging the strengths of U-Net, CoordConv, and Swin transformer, VidaFormer establishes a powerful framework for secure, high-capacity data hiding, setting a new benchmark for practical and resilient steganographic systems.

Our approach shows a remarkable enhancement in recovery accuracy, driven by the adoption of vision transformers for capturing long-range dependencies and the incorporation of CoordConv, which improves the model's spatial awareness. VidaFormer attains over 90% accuracy at 6 BPP (bits-per-pixel) capacity and achieves effective capacity of 4.89 BPP, substantially surpassing the performance of the current state-of-the-art models. The key contributions of this work are as follows:

(1) Developed a novel architecture integrating two hybrid models, each combining convolutional and transformer stages.
(2) Pioneered the use of CoordConv to provide positional awareness in the image steganography task.
(3) Achieved state-of-the-art accuracy, PSNR, and Reed-Solomon Bit Per Pixel (RS-BPP) on the DIV2K dataset.
(4) Evaluated the transparency of our method using a steganalysis tool to assess its effectiveness in concealing secret messages.
(5) Analyzed the trade-off between capacity and transparency by testing various transparency targets.

This paper is organized into five sections, each addressing key aspects of our research. The first section provides an overview, discussing the principles, challenges, and foundational concepts of our proposed method, while highlighting our main contributions. The second section reviews related works, offering a detailed analysis of prior methods, their strengths, limitations, and relevance to our approach. The third section describes the proposed method in depth, outlining its architecture, key innovations, and the rationale for our design choices. The fourth section presents a comprehensive evaluation of our model through experiments and ablation studies, demonstrating its effectiveness. Finally, the fifth section concludes the paper by summarizing findings and proposing future research directions.

2. RELATED WORK

In this section, different methods of image steganography are introduced. The discussion begins with statistical steganography, encompassing traditional techniques for hiding information within images. Modern approaches are then explored, including CNN-based methods that process images through convolutional layers, GAN-based techniques that generate realistic cover images using adversarial training, and finally vision transformers which use self-attention mechanism to capture long range dependencies inside an image.

2. 1. Statistical Steganography The least significant bit (LSB) method (7, 24-26) is the most commonly utilized technique in this category. This approach identifies bits within an image that can be altered without compromising its visual integrity and substitutes them with the binary representation of the secret message. Numerous variations of LSB-based methods have been proposed. For instance, Singh and Singh (27) embedded the secret message across the R, G, and B planes, while Das and Tuithung (28) utilized Huffman encoding to generate a binary form of the secret message, which is then embedded using LSB. A novel Reversible Data Hiding (RDH) method was proposed by Arham and Nugroho (29), combining difference expansion and modulus function to embed 3-bit data into 2-bit LSB differences of pixel pairs in rectangular blocks. It achieves high payload capacities with excellent visual quality. A novel steganography scheme was proposed by Chhajed and Garg (30) which uses a cover pattern histogram-based decision tree to hide and extract data in binary images, leveraging high-frequency 3x3 pixel block patterns and a secret key for encryption. It achieves high embedding capacity, minimal visual distortion (50-80% of bits embedded without pixel flipping), and enhanced security, outperforming similar methods for steganography and watermarking. A secure spatial steganography algorithm was introduced by Alexan et al. (31) that encrypts a secret message using tan logistic map, transforms it into QR codes, applies DNA coding with a Mersenne Twister key, and embeds

the resulting bit-streams into the LSB channel. Additionally, such as Qu et al. (32) and Wang et al. (33) explored LSB techniques within quantum images rather than the spatial domain.

Alternative methods utilize statistical analysis, including Pixel Value Differencing (PVD), a well-known technique that determines the difference between consecutive pixels to locate appropriate spots for embedding the secret message. Swain (34) introduced a hybrid approach combines LSB and PVD, applying LSB to the first two bits of every 8-bit segment of the secret message and PVD to the remaining six bits. Furthermore, some studies propose coverless solutions for image steganography, where a secret message is used to generate an image inherently designed to conceal the message (35). An errorless robust JPEG steganography method proposed by Qiu et al. (36) ensures message recovery after JPEG recompression by creating a robust set of DCT coefficients through lattice-based embedding. The approach guarantees error-free recovery while maintaining security and robustness against various JPEG compressors and quality factors. A steganography technique for JPEG2000 compressed images proposed by Butora et al. (37) embeds secret data into singular values obtained via singular value decomposition (SVD) in the discrete wavelet transform (DWT) domain, optimized by a genetic algorithm (GA) for scaling factor. It achieves up to 25% embedding capacity with higher PSNR than existing methods.

2. 2. Steganography by CNN These methods often rely on models designed to produce dense prediction outputs, typically using an encoder-decoder architecture. Within this framework, the secret message is concatenated with the cover image and processed to create the stego image. For instance, several researchers (10, 38, 39) employed U-Net as the basis for their encoder-decoder architectures. Rahim and Nadeem (40) presented a dual-path approach where two architectures process the cover image and secret message independently and simultaneously. A different approach was adopted by Wang et al. (41) which uses VGG (42) as the baseline to create a stego image styled after another input image. Alternative paradigms have also been explored; for example, Baluja (43) introduced three networks: one for data preparation, one for embedding the secret message, and one for extracting it. A deep learning-based steganography model was presented by Duan et al. (8) that can embed two secret images within a single carrier image and reconstruct them successfully. A Secret-to-Image Reversible Transformation (SIRT) framework for generative steganography was introduced by Zhou et al. (44), which encodes secret data into synthetic images with high embedding capacity, ensuring both imperceptibility and lossless data recovery while maintaining the fidelity and security of the generated images. A black-box generative steganography method was introduced by Zhang et al. (45) that leverages model volatility to mask steganographic modifications by modeling pixel distributions. The approach achieves high transparency and robustness against feature-based and CNN-based StegAnalyzer.

2. 3. Steganography by GAN These methods are built on Generative Adversarial Networks (GANs) (46), where the adversarial framework encourages the generator to produce high-quality, realistic images. Generative Steganographic Models (47-50) typically add a StegAnalyzer network that is responsible for detecting the presence of the secret message. An innovative approach by Bi et al. (51) integrates image style transfer into steganography, concealing secret information within stylized images while fusing secret and style features at multiple scales for improved concealment. A generic generative steganography framework was achieved by Su et al. (52) by enabling a distribution-preserving secret data modulator and a versatile secret data extractor. A novel approach by Ramandi et al. (49) introduced balancing transparency and capacity during training by dynamically adjusting the weights of the respective losses. A reversible generative steganography method by Tang et al. (53) proposed a distribution-preserving message mapping strategy and a reversible Glow model to encode secret information into latent vectors, achieving high security and accurate extraction.

2. 4. Steganography by Transformers Vision transformers (ViTs) (15) have transformed the field of computer vision by utilizing self-attention (16) mechanisms to effectively capture both local and long-range dependencies in image data, surpassing convolutional networks in tasks such as image classification. The Swin transformer family (17-19) employs a hierarchical structure with shifted windows, allowing for improved scalability and efficiency in dense prediction tasks.

On the steganography front, a transformer-based steganography scheme for enhanced feature extraction was introduced by Wang et al. (54). It also incorporates a recursive permutation image encryption algorithm to strengthen the security of secret images. Zhou et al. (55) used a channel-wise attention transformer model for image steganography that constructs long-range dependencies and optimized data embedding by considering global and local features and a combination of channel self-attention, non-linear enhancement, cross-attention, and global-local aggregation modules to improve data concealment. A transformer-based adversarial network steganography framework proposed by Xiao et al. (56) used attention mechanisms, maintains spatial relationships, and employs a WGAN discriminator to improve imperceptibility.

3. PROPOSED METHOD

3. 1. Overview In this research, we propose the Versatile Dynamic Transformer (VidaFormer) architecture, which achieves high hiding capacity while maintaining high image fidelity, as measured by PSNR (57) and SSIM (58), two primary metrics for image quality assessment (see Section [Evaluation metrics]). It employs MSE targeting, introduced by Ramandi et al. (49), to dynamically adjust the balance between effective capacity and transparency based on a predetermined transparency target.

The overall architecture is presented in Figure 1. For the embedding task, the encoder E takes the secret message M and the cover image C to generate the stego image S, as shown in Equation 1.

$$S = E(C, M) \quad (1)$$

For the extraction task, the decoder D takes the stego image and extracts the recovered message R, as in Equation 2.

$$R = D(S) \quad (2)$$

To enforce , it is crucial that the stego image closely resembles the cover image, making it hard to distinguish between the two.

The encoder has two distinct and competing objectives. On the one hand, it must produce a stego image that closely resembles the cover image, making it difficult for an attacker to detect the presence of a secret message, thereby achieving transparency. Conversely, the encoder must embed the secret message into the stego image for later extraction. To ensure effective recoverability and capacity, the encoder needs to alter pixel values to some degree, which can distort the original cover image and compromise transparency.

The decoder retrieves the original message by decoding the stego image and identifying patterns of hidden content embedded by the encoder. The accuracy of message recovery depends on the level of distortion, which is inversely related to the stego image's transparency. The algorithm's effective capacity is proportional to recovery accuracy; thus, increasing effective capacity reduces transparency and vice versa. Transparency is enforced through a loss function that penalizes pixel value differences between the cover and stego images. Another loss function, based on the difference between the recovered and secret messages, ensures recoverability.

3. 2. Encoder Architecture The encoder, depicted in Figure 2, embeds a binary secret message into a cover image, resulting in the creation of the stego image. Its architecture is inspired by U-Net (11), comprising a feature extraction path (left blocks), a bottleneck (central orange blocks), and a reconstruction path (right blocks). Data undergoes several stages of processing before being reconstructed to generate the desired output.

Initially, the cover image, with three RGB channels, and the secret message, with six channels each representing one bit per pixel, are concatenated to form a single nine-channel feature map. In this configuration, 6 bits are embedded per pixel, resulting in a 6 bits-per-pixel (6-BPP) embedding capacity setting.

In the first stage, three 48-channel convolutional blocks operate at full resolution. Each block, represented by a blue box, consists of three convolutional layers, resulting in a total of nine convolutional layers in this stage. The first convolution in the initial block takes nine input channels and produces 48 output channels, accommodating the channel count difference. The model benefits from extensive processing at full resolution, ensuring comprehensive pixel analysis and effective capture of local relationships between neighboring pixels.

At the beginning of the second stage, a max-pooling layer reduces the feature map's resolution by half. This operation expands the receptive field, enabling the model to capture broader contextual information while lowering computational costs by progressively reducing feature map resolution. The first convolution in the initial block doubles the channel count to 96, enhancing the model's ability to learn high-level

features. Consequently, with W and H representing the input image's width and height, the feature map's shape transitions from (W, H, 48) in the first stage to (W/2, H/2, 96) at the beginning of the second stage. The second stage comprises six convolutional blocks. Feature maps at this stage are more complex, necessitating increased processing, which justifies the greater number of blocks compared to the first stage. The expanded receptive field further enables the model to capture long-range pixel relationships.

At the start of the bottleneck stage, the feature map is reshaped from (W/2, H/2, 96) to (W/4, H/4, 384). As standard in vision transformers (18) (15), this reshaping before the transformer blocks incurs no data loss. Using a 2×2 patch size, the resolution is halved, and the channel count is quadrupled. Positioned between the feature extraction and reconstruction paths, the bottleneck stage (center orange blocks) employs three Swin transformer V2 blocks (18), optimized to capture long-range dependencies and enhance the model's ability to process global contextual information. By incorporating these blocks, the model leverages self-attention mechanisms to detect and encode complex patterns across the cover image, ensuring secure information embedding while preserving visual integrity.

The encoder has completed capturing all necessary information.

A two-stage reconstruction path (right blue blocks) uses this information to generate the embedded stego

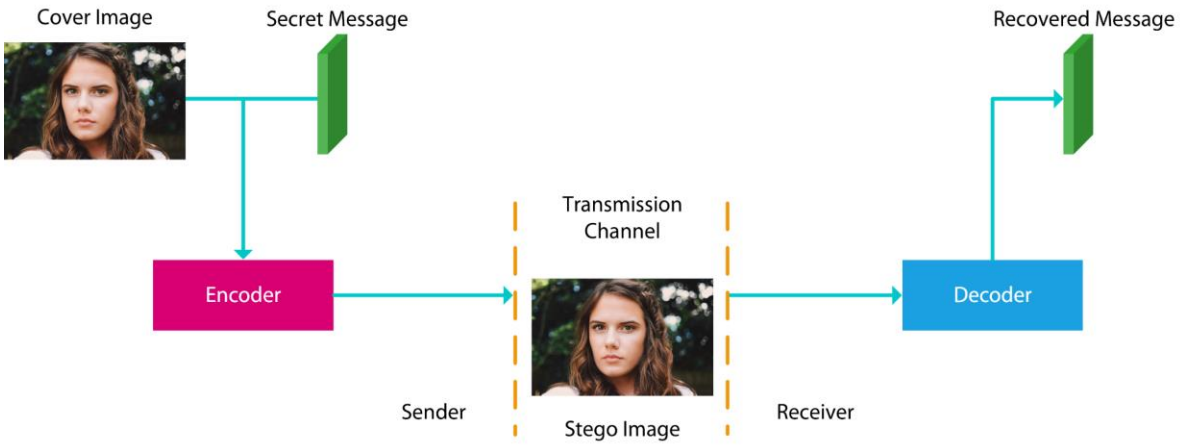


Figure 1. The overall process of image steganography

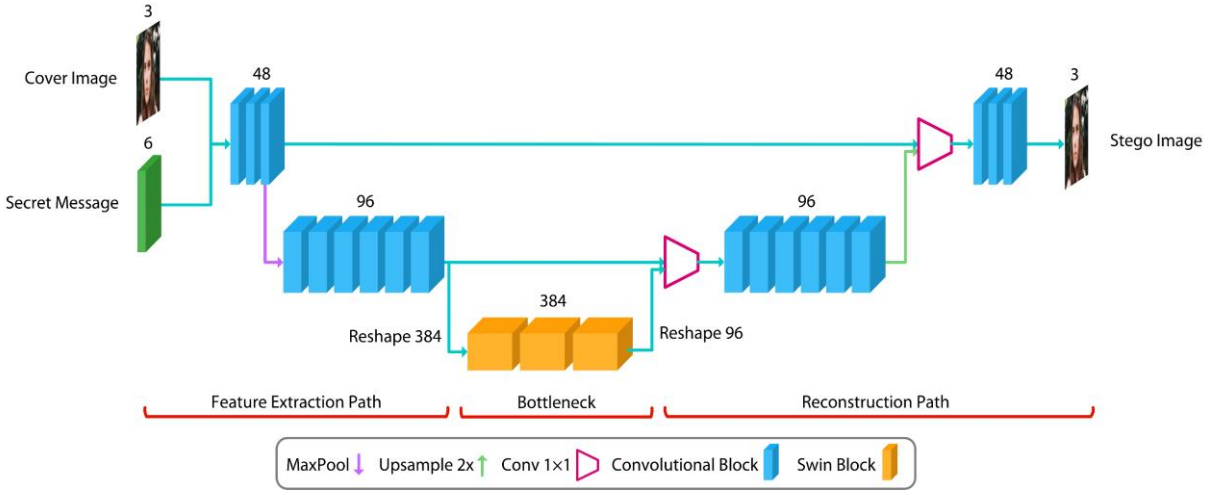


Figure 2. The architecture of the encoder

image. First, the bottleneck stage output is reshaped from (W/4, H/4, 384) to (W/2, H/2, 96) to align with the corresponding feature extraction path output and concatenated with it. The resulting feature map combines high-level information from the bottleneck with low-level information from the feature extraction path. A 1×1 convolution (first magenta trapezoid) projects the feature map to 96 channels. This feature map is then processed through six convolutional blocks, which further refine the integrated low- and high-level features.

The feature map is then unsampled (green arrow) to match the resolution of the first stage of the feature extraction path and concatenated with its output. This concatenated feature map is processed through a 1×1 convolution (second magenta trapezoid) and three additional convolutional blocks to complete processing at full resolution. A final convolution reduces the channel count to three, producing the stego image.

This symmetrical design facilitates the accurate reconstruction of the stego image. The overall architecture is designed to minimize pixel-level distortion in the cover image, ensuring that the resulting stego image closely resembles the original while securely embedding the secret message.

3. 3. Convolutional Blocks Each convolutional block in Figure 3 processes a feature map through a series of convolutional and normalization steps. Initially,

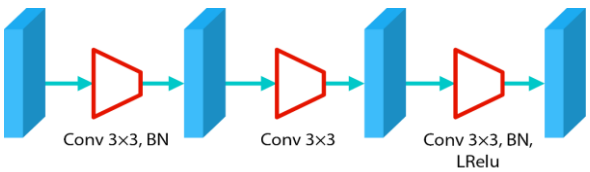


Figure 3. Convolutional block used in VidaFormer encoder and decoder

the input feature map passes through a convolution layer with a 3×3 kernel, followed by batch normalization to stabilize and accelerate convergence. This process is repeated with two additional 3×3 convolutional layers. Finally, batch normalization and a Leaky ReLU activation function are applied to introduce non-linearity while avoiding dead neurons by maintaining small gradients even for negative inputs.

Replacing standard convolutions with CoordConv (22) layers enhances the model's ability to retain spatial localization. CoordConv achieves this by adding explicit coordinate channels (x, y coordinates) to the input of the convolutional layers. This allows the network to retain spatial information and learn spatial dependencies more effectively than traditional convolutions. The added positional awareness enables the network to better understand and preserve the spatial structure of the input data. In steganography, where precise positioning is essential to embed data subtly without disrupting the image's visual integrity, CoordConv proves ideal for enhancing spatial control during the embedding process.

3. 4. Swin Transformer Stage The bottleneck stage in the encoder architecture consists of three Swin V2 (18) transformer blocks with a patch size of 2. Positioned at the network's bottleneck, it processes the feature map with the highest channel count and the lowest spatial resolution. This configuration allows the model to capture intricate, multi-scale dependencies across channels while preserving computational efficiency. It leverages Swin V2's local-global attention capabilities at the most condensed spatial representation.

Each block in this bottleneck stage, inspired by Swin V2, includes two multi-head self-attention modules. The first module applies attention within standard windows, while the second employs a shifted window configuration. This dual approach enhances performance by providing broader contextual awareness while keeping computational complexity linear rather than quadratic to the input size.

The query (Q), key (K), and value (V) matrices are generated through linear projections, transforming the input features as described in the below equations. These matrices are then used to compute multi-head self-attention, as detailed in Equation 6 (16).

$$Q = XW_Q \tag{3}$$

$$K = XW_K \tag{4}$$

$$V = XW_V \tag{5}$$

$$Attention = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{6}$$

In the above equations, Q, K, and V are query, key, and value tokens, B is the relative positional bias, and d is the query dimension. This approach is especially effective in the bottleneck, where spatial dimensions are reduced, as it leverages shifted window attention.

3. 5. Decoder Architecture Decoder architecture, illustrated in Figure 4, closely mirrors the encoder, with the primary difference being its input and output. Instead of embedding data, the decoder receives the stego image as input, processes it, and extracts the hidden information, ultimately recovering the binary secret message to align with the original.

3. 6. Network Loss The framework employs two primary loss functions: one for the encoder and one for the decoder as shown in Figure 5. For the encoder, the Mean Squared Error (MSE) loss function, as depicted in Equation 7, minimizes the difference between the cover image and the stego image, thereby ensuring the integrity and visual realism of the stego image. For the decoder, Equation 8 introduces a binary cross-entropy loss, which constrains the recovered message to match the original secret message with high accuracy.

$$L_E = \frac{1}{3\times W\times H}\sum(C - S)^2 \tag{7}$$

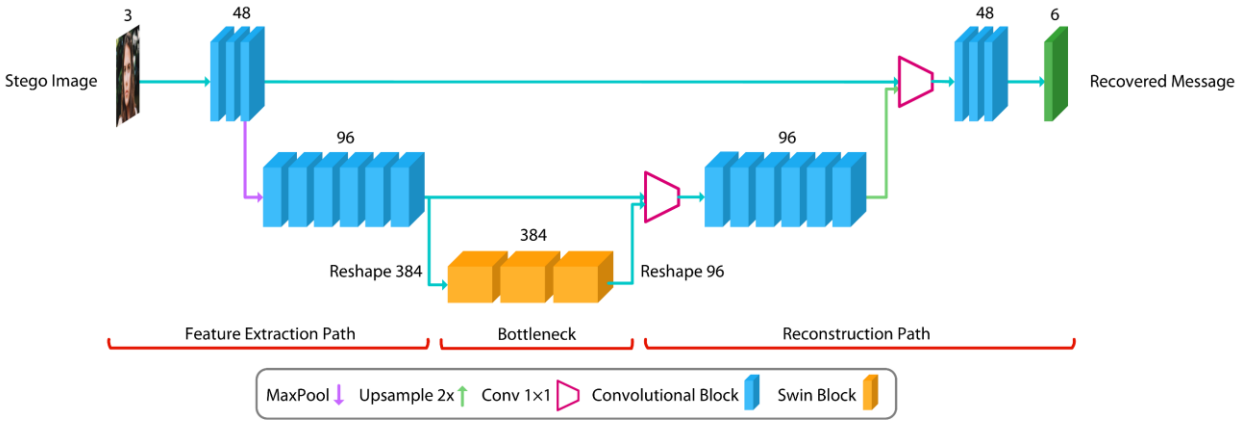


Figure 4. The architecture of the decoder which differs from the encoder in input and output

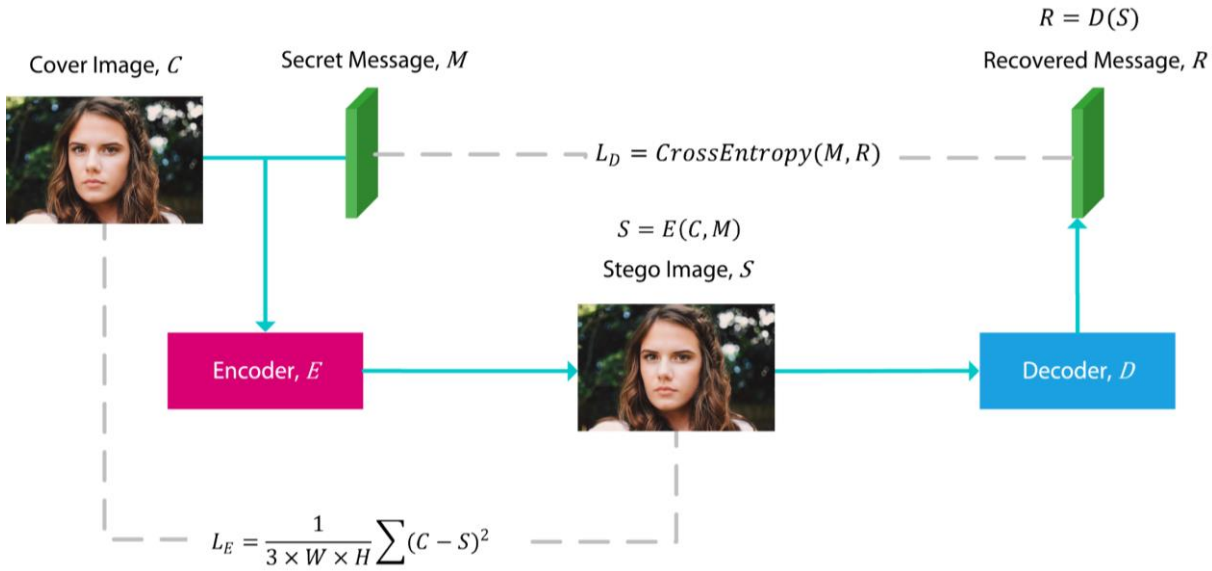


Figure 5. Network loss functions

$$L_D = CrossEntropy(M, R) \tag{8}$$

$$L_{total} = L_E + \theta L_D \tag{9}$$

The overall loss is computed using Equation 9, where θ is adjusted through MSE targeting, as proposed by Ramandi et al. (49). MSE targeting ensures that the stego image maintains sufficient integrity and realism based on the MSE loss while dynamically adjusting θ for each training iteration. This iterative adjustment balances the stego image's integrity and effective capacity.

3. 7. Error Correction Steganographic algorithms aim to extract the original message from the stego image, but the extracted message may contain errors, with some bits differing from the original secret message. However, the algorithm must fully and accurately extract the secret message without compromising its content. To achieve this, we incorporate Reed-Solomon error correction (23). This technique oversamples a polynomial constructed from the secret message, generating extended message that includes error-checking bits. It is particularly effective in steganography, where extracted data may contain errors.

The process is as follows: 1) The original message is processed by the Reed-Solomon algorithm, which adds correction bits to create an extended message. 2) The extended message is input to the encoder to generate the stego image. 3) The extended message, which may contain errors, is extracted from the stego image. 4) The Reed-Solomon algorithm corrects these errors, recovering the original message without any errors.

If the percentage of erroneous bits is below a threshold, the additional bits enable Reed-Solomon error correction to reconstruct the message accurately. Equation 10 specifies the minimum number of extra bits required for error correction, where p is the probability of error per bit, k is the length of the original bit array, and n is the extended length including error correction. Given p and k, we can calculate n to fully recover the message.

$$p.n \leq \frac{n-k}{2} \tag{10}$$

Maximizing k will give effective capacity which is the portion of the embedding capacity that corresponds to the original message, excluding additional error-correction data. We find k_{max} by turning Equation 10 to equality and rearrange as in Equation 11.

$$k_{max} = n(1 - 2p) \tag{11}$$

We denote recovery accuracy as a, defined as the probability of correctly extracting a bit, as opposed to erroneous extraction (p):

$$a = 1 - p \tag{12}$$

By definition, n is the embedding capacity. Combining the above equations will lead to Equation 13.

$$EC = Cap(2a - 1) \tag{13}$$

where EC is effective capacity and Cap is embedding capacity. Therefore, effective capacity can be easily calculated by embedding capacity and recovery accuracy.

4. EXPERIMENTS

The VidaFormer network is implemented using the PyTorch framework. Training and evaluation are

conducted on an NVIDIA RTX 3060 GPU, using images from the DIV2K dataset (59-61). To optimize performance, the embedding capacity is tested across configurations ranging from 1 to 6 bits, with the 6-bit configuration yielding the most effective results. Embedded capacity refers to the number of message bits embedded per pixel of the image (number above the green tensor in Figure 2). The list of hyperparameters is presented in Table 1. Data augmentation techniques, such as horizontal flipping and random cropping, are applied to enhance model generalization. The encoder and decoder are trained using the loss functions described in Network Loss.

4. 1. Evaluation Metrics We use three evaluation metrics to evaluate the quality of the stego image. The first metric is the Mean Squared Error (MSE), defined in Equation 14, which quantifies the difference between the stego image and the cover image.

$$MSE = \frac{1}{W \times H}(y - \hat{y})^2 \quad (14)$$

The second metric is the Peak Signal-to-Noise Ratio (PSNR), described in Equation 15, which evaluates image quality by quantifying pixel noise intensity.

$$PSNR = 10 \times \log_{10} \frac{s^2}{MSE} \quad (15)$$

where s is the maximum possible pixel value of the image.

The third metric is the Structural Similarity Index (SSIM) (57), as shown in Equation 16. SSIM evaluates the perceptual quality of the stego image by comparing its structural features, brightness, and contrast with those of the cover image.

$$SSIM = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)} \quad (16)$$

where $C_1 = 0.0001$ and $C_2 = 0.0009$, and the output is between -1 and 1.

These metrics enable a comprehensive quantitative comparison between the cover and stego images, ensuring the stego image's fidelity and quality.

4. 2. Main Results Table 2 presents the performance of our steganographic method across various embedding capacity, evaluating recovery accuracy, effective capacity, PSNR, and SSIM. As embedding capacity increases, recovery accuracy decreases, from perfect recovery at 1 to 0.907 at 6. This trend underscores the challenge of accurately recovering larger embedded data volumes. Nevertheless, effective capacity consistently improves, reaching 4.89 BPP at a capacity of 6, demonstrating the method's ability to handle higher data volumes.

PSNR values remain stable at lower capacities, peaking at 40.41 for a capacity of 2 but declining to 37.59 at a capacity of 6. SSIM values consistently surpass 0.90 at all capacities, indicating that the structural features of the images are effectively preserved. The highest SSIM of 0.96 is achieved at a capacity of 2. At an embedding capacity of 6, the model performs optimally when prioritizing capacity, achieving effective capacity of 4.89 BPP, 90.7% recovery accuracy, and acceptable image quality (PSNR of 37.59 and SSIM of 0.91). This demonstrates the method's strength in balancing high effective capacity with maintaining good overall performance.

Table 3 compares the performance of the proposed method with other advanced steganography techniques, including VidaGAN, SteganoGAN, AFHS-GAN, and FNNS variants, using recovery accuracy, PSNR, and SSIM as evaluation metrics. The proposed method achieves the highest recovery accuracy of 0.987,

TABLE 1. Hyperparameters used in the training of VidaFormer

Hyperparameter	Value
Epochs	80
Embedding Capacity	1-6
Resolution	256×256
Learning Rate	10^{-4}
Weight Decay	5×10^{-6}

TABLE 2. VidaFormer output for different capacities

Embedding Capacity (BPP)	Recovery Accuracy	Effective Capacity (BPP)	PSNR	SSIM
1	**1.00**	1.00	39.57	0.95
2	0.999	1.99	**40.41**	**0.96**
3	0.998	2.99	40.00	0.92
4	0.987	3.89	39.31	0.92
5	0.918	4.18	39.71	0.92
6	0.907	**4.89**	37.59	0.91

TABLE 3. Comparison of VidaFormer with other state-of-the-art methods applied to the DIV2K dataset (4-BPP)

Method	Recovery Accuracy	PSNR	SSIM
SteganoGAN (50)	0.820	37.49	0.88
FNNS-R (62)	0.891	28.60	0.76
FNNS-D (62)	0.945	25.74	0.65
Secure FNNS (5)	0.912	25.79	0.77
VidaGAN (49)	0.970	37.51	0.90
AFHS-GAN (63)	0.970	**40.14**	**0.94**
Ours	**0.987**	39.31	0.92

surpassing AFHS-GAN and VidaGAN, which both rank second at 0.970. This highlights the method's robust capability to reliably extract hidden data, even outperforming Secure FNNS. For PSNR, a measure of image quality, AFHS-GAN scores 40.14, the highest among all techniques, with the proposed method and VidaGAN following at 39.31 and 37.51, respectively, while FNNS-based methods exhibit significantly lower values. In terms of SSIM, which assesses the preservation of structural details in the image, AFHS-GAN achieves the highest score of 0.94, surpassing the proposed method (0.92) and VidaGAN (0.90). FNNS-based methods perform poorly in this regard, with FNNS-D recording the lowest SSIM of 0.65. Overall, the proposed method competes toe-to-toe with AFHS-GAN, surpassing in effective capacity but falling behind slightly in terms of transparency.

4. 3. Statistical Variance To provide a comprehensive evaluation of VidaFormer's performance and ensure the robustness of the reported metrics, we analyze the statistical variance of key evaluation metrics: recovery accuracy, effective capacity (BPP), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) across multiple runs on the DIV2K dataset. This analysis quantifies the consistency of the model's performance under varying training conditions and highlights its stability for image steganography tasks.

The VidaFormer model was trained and evaluated over 5 independent runs on the DIV2K dataset, using the same experimental setup described in experiment section. For each run, we computed the metrics at embedding capacities of 6 bits-per-pixel (BPP), which maximizes effective capacity. We report the mean and standard deviation of recovery accuracy, effective capacity, PSNR, and SSIM to capture the variability across runs. Table 4 presents the mean and standard deviation of the metrics for the 6-BPP configuration, which is the most challenging due to its high embedding capacity.

The mean recovery accuracy of 0.91 indicates reliable message extraction, with a low standard deviation, suggesting consistent performance across runs. The tight reflects the model's ability to maintain high accuracy despite the high embedding load. The effective capacity averages 4.89 BPP, with a standard deviation of 0.1150 BPP. This minor variability underscores the model's stability in embedding large data volumes while preserving message integrity. Variations are primarily due to differences in initialization and data augmentation effects. The mean PSNR of 37.59 dB, with a standard deviation of 0.4219 dB, indicates stable image quality across runs. The relatively low variance suggests that the model consistently produces stego images with minimal distortion, even under different training conditions. The SSIM mean of 0.91, with a standard deviation of 0.005094, confirms consistent preservation of structural features in stego images. The low variance indicates that VidaFormer reliably maintains perceptual quality, with slight fluctuations attributable to the stochastic nature of training.

TABLE 4. Statistical Variance of Metrics for VidaFormer (6-BPP)

Metric	Mean	Standard Deviation
Recovery Accuracy	0.9071	0.009601
Effective Capacity (BPP)	4.887	0.1150
PSNR (dB)	37.59	0.4219
SSIM	0.9073	0.005094

4. 4. Ablation Studies The ablation studies, presented in Table 5, examine how modifications to the baseline network affect steganographic performance.

Performance is evaluated using recovery accuracy, effective capacity (BPP), PSNR, and SSIM. The baseline model achieves a recovery accuracy of 0.822, an effective capacity of 3.86 BPP, a PSNR of 38.29, and an SSIM of 0.92, serving as a reference for further improvements.

Incorporating CoordConv, which enhances spatial processing, significantly improves performance. Recovery accuracy rises to 0.859, effective capacity increases to 4.31 BPP, PSNR improves to 39.19, and SSIM reaches 0.93, indicating enhanced data handling and image quality.

Using transformer layers in the bottleneck further improves recovery accuracy to 0.907 and effective capacity to 4.89 BPP. However, this entails a slight trade-off, with PSNR decreasing to 37.59 and SSIM to 0.89, reflecting a balance between capacity and image quality.

Compared to AFHS-GAN, which at 6 BPP capacity, achieves an effective capacity of 4.86 BPP, a PSNR of 37.69, and an SSIM of 0.89, the two models offer similar results. This demonstrate that architectural improvements, particularly CoordConv and transformer layers, significantly enhance network performance while maintaining image quality.

TABLE 5. Ablation studies with each main part of the architecture in comparison with VidaGAN (6-BPP)

Network	Recovery Accuracy	Effective Capacity (BPP)	PSNR	SSIM
VidaGAN (49)	0.825	3.9	38.56	0.88
Baseline (conv network)	0.822	3.86	38.29	0.92
+ CoordConv	0.859	4.31	**39.19**	**0.93**
+ transformer	**0.907**	**4.89**	37.59	0.89

4. 5. MSE Adjustment MSE targeting, introduced by VidaGAN (49), balances the fidelity of the cover and stego images by setting a target Mean Squared Error (MSE) value. Figure 6 illustrates how adjusting this MSE value affects recovery accuracy, PSNR, and SSIM when embedding 6 bits-per-pixel (6 BPP).

Increasing the MSE threshold improves recovery accuracy, rising from 0.835 at the lowest threshold (0.00025) to 0.963 at the highest (0.000875). This indicates that higher MSE values enhance the method's ability to reliably recover data. However, this improvement comes at the cost of image quality, as PSNR decreases from 40.038 to 33.142, reflecting increased distortion in the stego images.

SSIM, which measures structural similarity, decreases as the MSE threshold increases, falling from 0.945 to 0.789. Higher thresholds improve data recovery but compromise both structural and visual image quality. Figure 7 illustrates that enhanced recovery accuracy at higher thresholds increases embedding capacity but requires a trade-off in image fidelity.

These findings highlight the flexibility of MSE targeting. Lower thresholds prioritize image quality, yielding high PSNR and SSIM, while higher thresholds emphasize improved recovery accuracy at the cost of visual clarity. This adaptability enables the method to meet diverse application requirements.

4. 6. Computational Cost The model's computational demands are primarily driven by the transformer blocks, which employ self-attention mechanisms that scale quadratically with sequence length, although mitigated here by the shifted window approach in Swin V2 and their placement only in the bottleneck (where spatial dimensions are reduced to W/4× H/4). During training, VidaFormer requires approximately 80 epochs on an NVIDIA RTX 3060 GPU, with each epoch processing the DIV2K dataset taking less than 5 minutes. This results in a total training time of about 6.5 hours for the full model, including data augmentation steps like random cropping and flipping.

Inference times, as detailed in Table 6, are more practical for individual images: embedding completes in 0.2 seconds, benefiting from the efficient convolutional paths, while extraction takes 1.45 seconds, largely due to the Reed-Solomon error correction post-processing. When benchmarked against state-of-the-art methods (Table 7), VidaFormer's times are moderate. It is faster than FNNS (which exceeds 44 seconds) but slower than purely convolutional models like LISO (0.33 seconds) (64). The GPU memory footprint peaks at around 4-6 GB during inference, making it feasible on mid-range hardware but potentially challenging on edge devices without optimization.

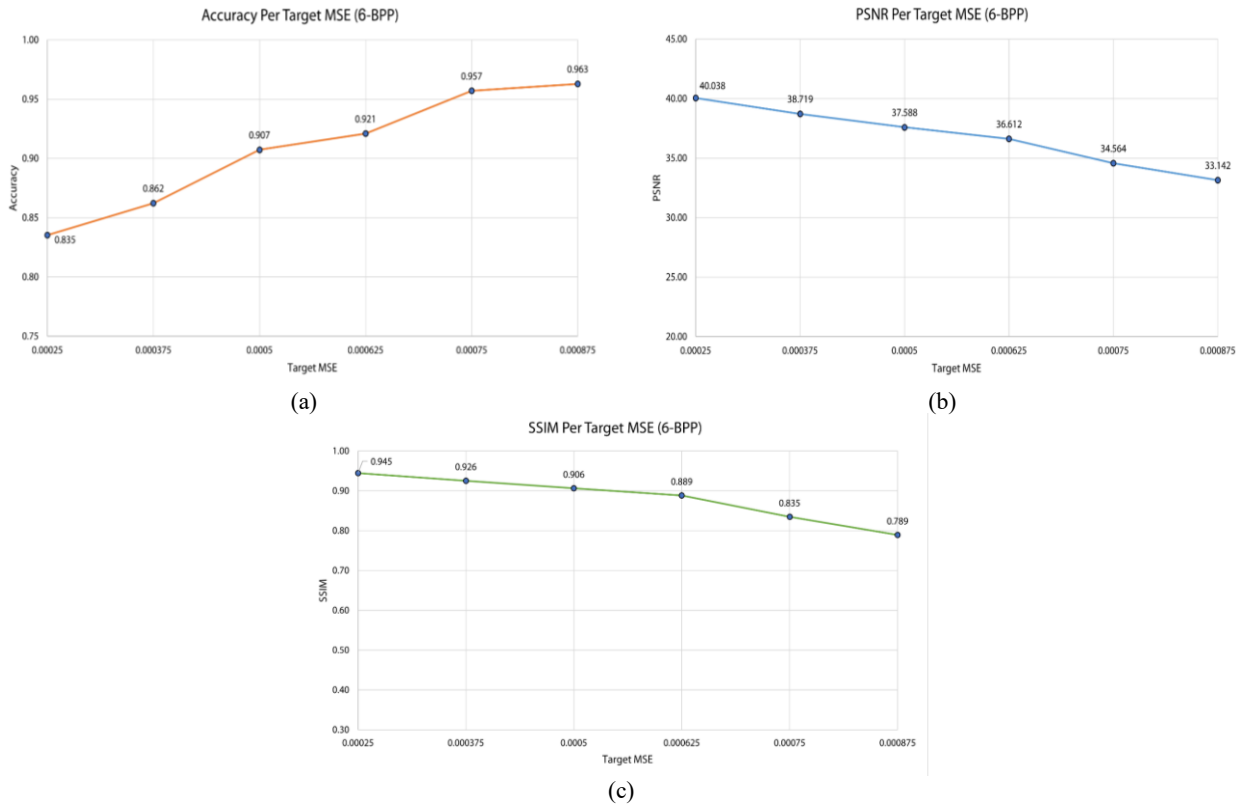
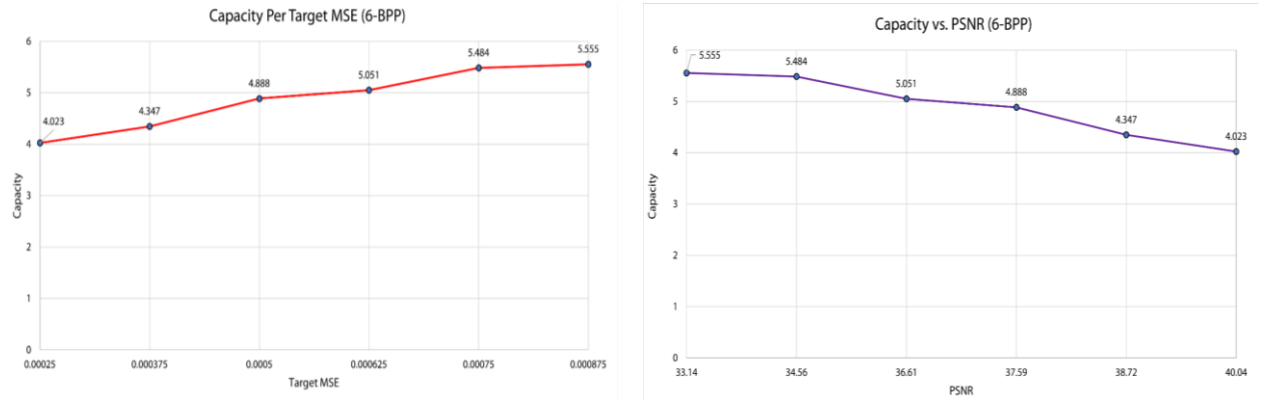


Figure 6. This Figure illustrates the impact of tweaking MSE targeting on three important metrics: Accuracy (a), PSNR (b), and SSIM (c)



(a) (b)

Figure 7. Capacity vs (a) MSE, (b) PSNR

TABLE 6. Computation time for images (4-BPP)

Task	Time Taken (S)
Embedding	0.20
Extraction	1.45
Embedding + Extraction	1.65

TABLE 7. Computation time for different methods (4-BPP)

Method	Time Taken (S)
SteganoGAN (50)	0.9
FNNS-D (62)	44.29
LISO (64)	0.33
VidaGAN (49)	1.28
Ours	1.65

VidaFormer demonstrates good scalability to higher resolutions due to its hybrid design: convolutional layers handle early high-resolution stages with linear complexity, while transformers are confined to down sampled features, avoiding the full quadratic scaling issues of pure ViT models.

Overall, while VidaFormer is computationally more intensive than traditional CNN-based steganography, its design choices enhance scalability compared to full transformer models. Future enhancements, as outlined in future work, such as adopting lighter transformers could further improve its viability for real-time and large-scale deployments, making it a scalable solution for secure data hiding in diverse scenarios.

4. 7. Steganalysis Steganography focuses on concealing messages to evade detection, while steganalysis seeks to uncover hidden content in digital media. GBRAS-Net (65), a recent CNN-based steganalysis tool, performs spatial image steganalysis by incorporating filter banks, depth-wise and separable convolutional layers, and skip connections. It outperforms recent works in detecting steganographic images across multiple adaptive algorithms using BOSSbase 1.01 and BOWS 2 datasets. The approach significantly improves classification accuracy for various steganographic methods and should be able to challenge steganographic algorithms.

In this study, GBRAS-Net was evaluated on 100 masked and hidden images to assess its effectiveness in detecting embedded messages. The results yielded an area under the Receiver Operating Characteristic curve (AUROC) of 0.510 for the secret message detection task, a s shown in Figure 8. This suggests that GBRAS- Net was unable to distinguish between cover and stego images, with predictions only slightly better than random guessing. The classification report and confusion matrix for GBRAS-Net are presented in Tables 8 and 9, respectively, indicating that the steganalysis tool is heavily biased toward classifying images as stego.

To visually assess the effectiveness and embedding characteristics of various steganographic methods, we compare the residuals, the difference between the stego

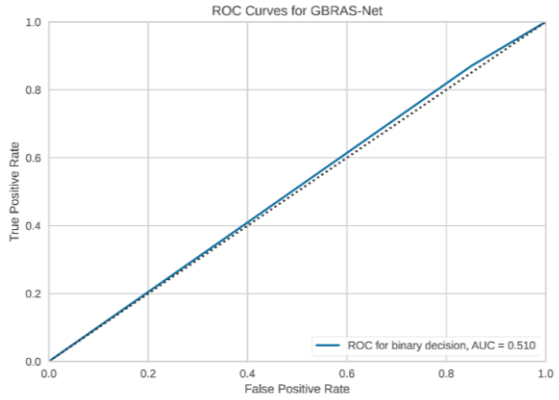


Figure 8. ROC curve for classification between cover and stego images by GBRAS-Net

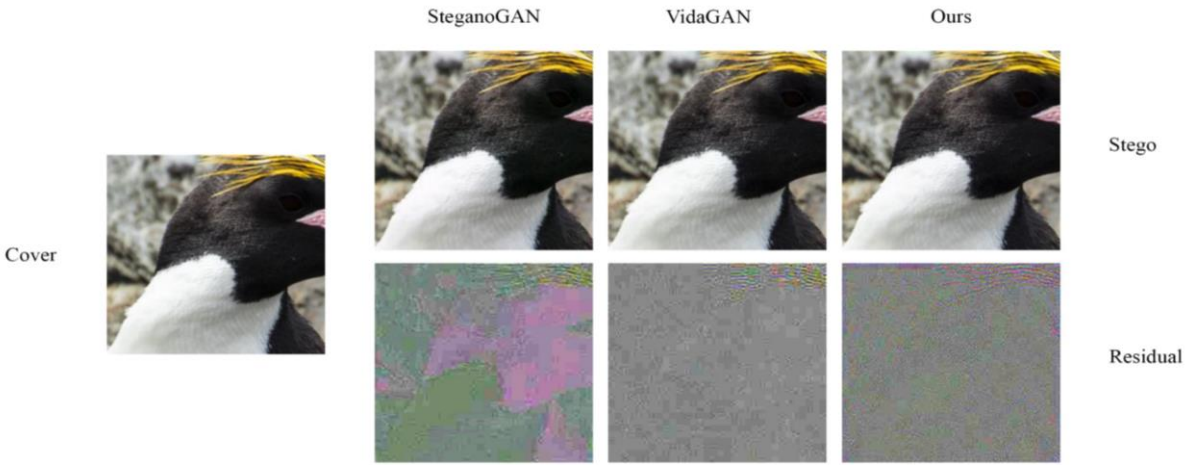


Figure 9. Qualitative results for 3 BPP stegos by different methods. For each method, the residual image is the difference between stego and cover image multiplied by 10.

and cover images, for SteganoGAN (50), VidaGAN (49), and our proposed method. As shown in Figure 9, the residuals are magnified by a factor of 10 to enhance perceptibility and highlight subtle embedding patterns that may not be visible in the original stego images.

In the SteganoGAN residual, we observe large spatial regions where the patterns closely resemble textures from the cover image. This suggests that the data embedding is not uniformly distributed and is influenced by the content of the cover image. Specifically, areas with higher color variation or certain textures seem to carry more embedded information, implying a channel-dependent and image-aware encoding strategy.

In the SteganoGAN residual, we observe large spatial regions where the patterns closely resemble textures from the cover image. This suggests that the data embedding is not uniformly distributed and is influenced by the content of the cover image. Specifically, areas with higher color variation or certain textures seem to carry more embedded information, implying a channel-dependent and image-aware encoding strategy.

The VidaGAN residual, on the other hand, appears predominantly grayscale with less structural variation. This indicates that the embedding is more uniformly spread across the RGB channels, with less dependence on the original content of the image. The uniformity hints at a more balanced, channel-agnostic embedding process.

In contrast, the residual generated by our method displays higher frequency patterns and more diverse coloration, suggesting a finer-grained and denser embedding. These characteristics are consistent with the observed improvement in decoding accuracy. The presence of high-frequency residuals implies that our method is capable of encoding more bits without significantly altering the perceptual quality of the stego image.

TABLE 8. Classification Report for GBRAS-Net

Class	Precision	Recall	F1-Score
Cover	0.536	0.150	0.234
Stego	0.506	0.870	0.640

TABLE 9. Confusion matrix for GBRAS-Net

True Class	Predicted Class	
	Cover	Stego
Cover	15%	85%
Stego	13%	87%

Overall, the qualitative comparison highlights that our method achieves more efficient and content-independent embedding, which translates into higher capacity compared to SteganoGAN and VidaGAN.

4. 8. Lightweight Variant To address the computational complexity of the Swin V2 transformer blocks used in the bottleneck stage of VidaFormer, we explore a lightweight transformer variant, MobileViT (61), as an alternative to demonstrate the trade-off between efficiency and steganographic performance. MobileViT combines convolutional layers with transformer-like blocks designed for resource-constrained environments, offering reduced computational overhead while maintaining competitive feature extraction capabilities. This makes it a suitable candidate for optimizing VidaFormer for scenarios requiring lower latency or deployment on edge devices.

We replaced the three Swin V2 transformer blocks in the bottleneck stage of the VidaFormer encoder (as described in Section 3.4) with MobileViT blocks,

specifically the MobileViT-XS configuration, which balances efficiency and performance. The MobileViT blocks process the feature map at the same resolution (W/4 × H/4, 384 channels) as the original Swin V2 blocks, ensuring compatibility with the existing U-Net-inspired architecture. The MobileViT blocks utilize a combination of depthwise separable convolutions and transformer-like self-attention mechanisms, reducing the computational complexity from the quadratic scaling of Swin V2's shifted window attention to a more efficient linear scaling. The rest of the architecture, including the CoordConv layers in the convolutional blocks and the Reed-Solomon error correction mechanism, remains unchanged.

As presented in Table 10, the MobileViT-based VidaFormer achieves a recovery accuracy of 0.836, lower than the Swin V2-based VidaFormer (0.907), resulting in an effective capacity of 4.03 BPP compared to 4.89 BPP. This indicates a trade-off in steganographic performance, as the lightweight MobileViT blocks are less effective at capturing long-range dependencies compared to Swin V2's shifted window attention. However, the PSNR (37.82 dB) is slightly improved.

In terms of efficiency, the MobileViT-based model significantly reduces computational demands. The training time per epoch drops to 3.2 minutes from 4.8 minutes and the total training time is approximately 4.5 hours compared to 6.5 hours for the Swin V2-based model. Inference times are also improved, with embedding completed in 0.15 seconds and extraction in 1.39 seconds. The GPU memory footprint is reduced to 3.9 GB from 5.8 GB, making the MobileViT-based VidaFormer more suitable for deployment on resource-constrained devices. The integration of MobileViT demonstrates a clear trade-off between efficiency and steganographic performance. The reduction in recovery accuracy and effective capacity (from 4.89 BPP to 4.03 BPP) reflects the limitations of MobileViT's lightweight self-attention mechanism in modeling complex global dependencies compared to Swin V2.

TABLE 10. Performance Comparison of VidaFormer with lightweight variant (6-BPP)

Metric	VidaFormer	VidaFormer (MobileViT-XS)
Recovery Accuracy	0.907	0.836
Effective Capacity (BPP)	4.89	4.03
PSNR (dB)	37.59	37.82
SSIM	0.91	0.91
Epoch Training Time (min)	4.8	3.2
Embedding Time (s)	0.2	0.15
Extraction time (s)	1.45	1.39
GPU Memory (GB)	5.8	3.9

5. CONCLUSION

This study introduces VidaFormer, a new type of hybrid transformer model that sets a high standard for image steganography by balancing capacity, recovery accuracy, and image quality. It combines convolutional layers with CoordConv for better spatial understanding and Swin V2 transformer blocks to capture long-distance connections in images. On the DIV2K dataset, VidaFormer achieves an effective capacity of 4.89 bits-per-pixel (BPP), which is 1.7% better than the previous best model. It uses Reed-Solomon error correction to ensure reliable message recovery and a U-Net-inspired design to keep images clear, with a PSNR of 37.59 and SSIM of 0.91 at 6 BPP. Testing with GBRAS-Net shows it hides messages well, with an AUROC of 0.510, meaning detection is almost like random guessing.

VidaFormer's design tackles the challenges of steganography, such as balancing capacity, transparency, and processing speed, through a well-planned structure. CoordConv helps accurately place hidden data, and transformer layers improve feature detection, overcoming issues with traditional convolutional networks. Tests confirm these parts work well, and MSE targeting allows the model to adapt for different needs, like focusing on image quality or data recovery.

While VidaFormer is a big step forward, it has some issues, like high computational demands and limited dataset variety. Future work will aim to improve efficiency, adaptability, and explore new uses. VidaFormer is a strong foundation for secure, high-capacity data hiding and shows how hybrid deep learning can improve steganography, offering a reliable solution for secure communication in a digital world.

5. 1. Limitations Despite VidaFormer's state-of-the-art performance in image steganography, particularly in achieving high effective capacity and recovery accuracy, several limitations must be acknowledged. First, the integration of transformer blocks in the bottleneck stage, while optimized through a hybrid architecture, introduces computational complexity compared to purely convolutional models. The extraction process, for example, takes approximately 1.45 seconds per image on an NVIDIA RTX 3060 GPU, primarily due to the Reed-Solomon error correction mechanism. This latency may limit its suitability for real-time applications, especially when processing high-resolution images or large datasets.

Second, the evaluation is conducted primarily on the DIV2K dataset, which, while high-quality, is relatively small, with only 100 validation images commonly used in benchmarks. This raises concerns about the model's generalizability to more diverse image sets, such as low-resolution web images or those from real-world scenarios like social media platforms. Furthermore,

while transparency is evaluated using GBRAS-Net steganalysis, achieving an AUROC of 0.510, the tool's bias toward classifying images as stego suggests that additional testing against advanced steganalysis methods (e.g., ensemble detectors or deeper CNN-based analyzers) is needed to fully confirm the model's undetectability.

Finally, the fixed embedding capacity of up to 6 bits-per-pixel (BPP), with an effective capacity of 4.89 BPP, highlights an inherent trade-off between capacity, transparency, and robustness. Prioritizing one of these objectives often compromises the others, and further optimization is needed to achieve a better balance.

5. 2. Future Works To address the identified limitations and further enhance VidaFormer's capabilities, several research directions are proposed. First, improving computational efficiency is a key priority. Exploring more lightweight transformer architectures, such as EfficientFormer, could reduce latency in both embedding and extraction phases, enabling real-time applications on resource-constrained devices like mobile phones or for processing video streams.

Second, expanding the evaluation to include larger and more diverse datasets, such as COCO or ImageNet subsets, will help validate the model's generalizability and reduce the risk of overfitting to the DIV2K dataset.

Third, investigating alternative positional encoding mechanisms beyond CoordConv, such as learnable embeddings or hybrid CNN-transformer integrations at multiple stages of the architecture, could further improve spatial awareness without increasing input channel complexity. Additionally, developing dynamic embedding strategies that allocate capacity based on image content such as embedding more bits in textured regions versus smooth areas could optimize the trade-off between capacity and transparency.

Finally, extending VidaFormer to other domains, such as video steganography, audio data hiding, or multi-modal concealment, presents exciting opportunities. Integrating the model with generative approaches for coverless steganography could enhance security in applications like secure communication over online social networks. These advancements would position VidaFormer as a versatile framework for next-generation steganographic systems.

Acknowledgements

The author appreciates Kharazmi Campus and Faculty of Computer Engineering in Shahrood University of Technology for providing necessary facilities to conduct present research.

Funding

This work has not received any financial support.

Declarations Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable. Competing interests the author declares no financial or organizational conflicts of interest.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.
The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Authors Biosketch

Vida Yousefi Ramandi holds a Ph.D. in Artificial Intelligence from Shahrood University of Technology. Her research focuses on image steganography using deep learning techniques, and she has published several papers in this field. She completed her Ph.D. program with an A+ academic standing and defended her dissertation with distinction.

Mansoor Fateh received the M.S. degree in biomedical engineering and the Ph.D. degree from Tarbiat Modares University, Tehran, Iran. He is currently a Faculty Member with the Faculty of Computer Engineering, Shahrood University of Technology, Iran. His research interests include deep learning and image processing.

Mohsen Rezvani received the PhD degree in computer science from the University of New South Wales, Australia. He is a faculty member in the Faculty of Computer Engineering, Shahrood University of Technology, Iran. His research focuses on computer security, privacy, trust and reputation systems.

REFERENCES

1. You Z, Ying Q, Li S, Qian Z, Zhang X, editors. Image generation network for covert transmission in online social network. Proceedings of the 30th ACM International Conference on Multimedia; 2022. DOI: 10.1145/3503161.3548139.

2. Evsutin O, Melman A, Meshcheryakov R. Digital Steganography and Watermarking for Digital Images: A Review of Current Research Directions. IEEE Access. 2020;8:166589-611. DOI: 10.1109/ACCESS.2020.3022779.
3. Saadati M, Vahidi J, Seydi V, Sheikholharam Mashhadi P. Proposing a New Image Watermarking Method Using Shearlet Transform and Whale Optimization Algorithm. International Journal of Engineering Transactions A: Basics. 2021;34(4):843-53. DOI: 10.5829/ije.2021.34.04a.10.
4. Qi Y, Chen K, Zhao N, Yang Z, Zhang W. Provably Secure Robust Image Steganography via Cross-Modal Error Correction. arXiv e-prints. 2024:arXiv: 2412.12206. DOI: 10.48550/arXiv.2412.12206.
5. Luo Z, Li S, Li G, Qian Z, Zhang X, editors. Securing Fixed Neural Network Steganography. Proceedings of the 31st ACM International Conference on Multimedia; 2023. DOI: 10.1145/3581783.3611920.
6. Fateh M, Mohsen R, Yasser I. A new method of coding for steganography based on LSB matching revisited. Security and Communication Networks. 2021:1-15. DOI: 10.1155/2021/6610678.
7. Mielikainen J. LSB matching revisited. IEEE signal processing letters. 2006;13(5):285-7. DOI: 10.1109/LSP.2006.870357.
8. Duan X, Liu N, Gou M, Wang W, Qin C. SteganoCNN: Image steganography with generalization ability based on convolutional neural network. Entropy. 2020;22(10):1140. DOI: 10.3390/e22101140.
9. Kumar V, Laddha S, Aniket ND. Steganography techniques using convolutional neural networks. J Homepage. 2020;7:66-73. DOI: 10.18280/rces.070304.
10. Wu P, Yang Y, Li X, editors. Image-into-image steganography using deep convolutional network. Advances in Multimedia Information Processing–PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, September 21-22, 2018, Proceedings, Part II 19; 2018: Springer. DOI: 10.1007/978-3-030-00767-6_73.
11. Ronneberger O, Fischer P, Brox T, editors. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical image computing and computer-assisted intervention; 2015. DOI: 10.1007/978-3-319-24574-4_28.
12. Wang C-Y, Liao H-YM, Wu Y-H, Chen P-Y, Hsieh J-W, Yeh I-H, editors. CSPNet: A new backbone that can enhance learning capability of CNN. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops; 2020. DOI: 10.1109/CVPRW50498.2020.00203.
13. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ, editors. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. DOI: 10.1016/j.patcog.2020.107610
14. Luo W, Li Y, Urtasun R, Zemel R. Understanding the effective receptive field in deep convolutional neural networks. Advances in neural information processing systems. 2016;29. DOI: 10.5555/3157382.3157645.
15. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al., editors. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations; 2020. DOI: 10.48550/arXiv.2010.11929.
16. Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017. DOI: 10.5555/3295222.3295349.
17. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al., editors. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision; 2021. DOI: 10.1109/ICCV48922.2021.00986.
18. Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, et al., editors. Swin transformer v2: Scaling up capacity and resolution. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. DOI: 10.1109/CVPR52688.2022.01170.
19. Dong X, Bao J, Chen D, Zhang W, Yu N, Yuan L, et al., editors. Cswin transformer: A general vision transformer backbone with cross-shaped windows. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. DOI: 10.1109/CVPR52688.2022.01181.
20. Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, et al. A survey on vision transformer. EEE transactions on pattern analysis and machine intelligence. 2022;45(1):87-110. DOI: 10.1109/TPAMI.2022.3152247.
21. Kayhan OS, Gemert JCv, editors. On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. DOI: 10.1109/CVPR42600.2020.01428.
22. Liu R, Lehman J, Molino P, Petroski Such F, Frank E, Sergeev A, et al. An intriguing failing of convolutional neural networks and the coordconv solution. Advances in neural information processing systems. 2018;31. DOI: 10.5555/3327546.3327630.
23. Reed IS, Solomon G. Polynomial codes over certain finite fields. Journal of the society for industrial and applied mathematics. 1960;8(2):300-4. DOI: 10.1137/0108018.
24. Fateh M, Rezvani M, Irani Y. A New Method of Coding for Steganography Based on LSB Matching Revisited. Security and Communication Networks. 2021;2021(1):6610678. 10.1155/2021/6610678.
25. Karim SM, Rahman MS, Hossain MI, editors. A new approach for LSB based image steganography using secret key. 14th international conference on computer and information technology (ICCIT 2011); 2011: IEEE. DOI: 10.1109/ICCITechn.2011.6164800.
26. mansoor fateh sr, elahe alipour. Review of Image Steganography Based on LSB and Pixel Classification and Providing a New Method in This Area. Biannual Journal Monadi for Cyberspace Security (AFTA). 2017;5(2):9. http://monadi.isc.org.ir/article-1-82-en.html.
27. Singh A, Singh H, editors. An improved LSB based image steganography technique for RGB images. 2015 IEEE International Conference on electrical, computer and communication technologies (ICECCT); 2015: IEEE. DOI: 10.1109/ICECCT.2015.7226122.
28. Das R, Tuithung T, editors. A novel steganography method for image based on Huffman Encoding. 2012 3rd National Conference on Emerging Trends and Applications in Computer Science; 2012: IEEE. DOI: 10.1109/NCETACS.2012.6203290.
29. Arham A, Nugroho HA. Enhanced reversible data hiding using difference expansion and modulus function with selective bit blocks in images. Cybersecurity. 2024;7(1):61. DOI: 10.1186/s42400-024-00251-7.
30. Chhajed G, Garg B. Novel Scheme for Data Hiding in Binary Images using Cover Pattern Histogram. International Journal of Engineering Transactions B: Applications. 2023;36(11):2124-36. DOI: 10.5829/ije.2023.36.11b.16.
31. Alexan W, Mamdouh E, Aboshousha A, Alsahafi YS, Gabr M, Hosny KM. Stegocrypt: A robust tri-stage spatial steganography algorithm using TLM encryption and DNA coding for securing digital images. IET Image Processing. 2024;18(13):4189-206. DOI: 10.1049/ipr2.13242.
32. Qu Z, Cheng Z, Liu W, Wang X. A novel quantum image steganography algorithm based on exploiting modification direction. Multimedia Tools and Applications. 2019;78:7981-8001. DOI: 10.1007/s11042-018-6476-5.

33. Wang S, Sang J, Song X, Niu X. Least significant qubit (LSQb) information hiding algorithm for quantum image. Measurement. 2015;73:352-9. DOI: 10.1016/j.measurement.2015.05.038.
34. Swain G. Very high capacity image steganography technique using quotient value differencing and LSB substitution. Arabian Journal for Science and Engineering. 2019;44(4):2995-3004. DOI: 10.1007/s13369-018-3372-2.
35. Qiu A, Chen X, Sun X, Wang S, Guo W. Coverless image steganography method based on feature selection. Journal of Information Hiding and Privacy Protection. 2019;1(2):49. DOI: 10.32604/jihpp.2019.05881.
36. Butora J, Puteaux P, Bas P. Errorless robust JPEG steganography using outputs of JPEG coders. IEEE Transactions on Dependable and Secure Computing. 2023. DOI: 10.1109/TDSC.2023.3306379.
37. Bhatia SS, Singh K, Kasana G. Singular Value Decomposition based Steganography Technique for JPEG2000 Compressed Images. International Journal of Engineering Transactions C: Aspects. 2015;28(12):1720-7. DOI: 10.5829/idosi.ije.2015.28.12c.04.
38. Wu P, Yang Y, Li X. Stegnet: Mega image steganography capacity with deep convolutional network. Future Internet. 2018;10(6):54. DOI: 10.3390/fi10060054.
39. Duan X, Jia K, Li B, Guo D, Zhang E, Qin C. Reversible image steganography scheme based on a U-Net structure. IEEE Access. 2019;7:9314-23. DOI: 10.1109/ACCESS.2019.2891247.
40. Rahim R, Nadeem S, editors. End-to-end trained CNN encoder-decoder networks for image steganography. Proceedings of the European conference on computer vision (ECCV) workshops; 2018. DOI: 10.1007/978-3-030-11018-5_64.
41. Wang Z, Gao N, Wang X, Xiang J, Liu G, editors. STNet: A style transformation network for deep image steganography. Neural Information Processing: 26th International Conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, Proceedings, Part II 26; 2019: Springer. DOI: 10.1007/978-3-030-36711-4_1.
42. Simonyan K, Zisserman A, editors. Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations (ICLR 2015); 2015: Computational and Biological Learning Society.
43. Baluja S. Hiding images in plain sight: Deep steganography. Advances in neural information processing systems. 2017;30. DOI: 10.5555/3294771.3294968.
44. Zhou Z, Su Y, Li J, Yu K, Wu QJ, Fu Z, et al. Secret-to-image reversible transformation for generative steganography. IEEE Transactions on Dependable and Secure Computing. 2022;20(5):4118-34. DOI: 10.1109/TDSC.2022.3217661.
45. Zhang J, Chen K, Li W, Zhang W, Yu N. Steganography with Generated Images: Leveraging Volatility to Enhance Security. IEEE Transactions on Dependable and Secure Computing. 2023. DOI: 10.1109/TDSC.2023.3341427.
46. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27. DOI: 10.5555/2969033.2969125.
47. Volkhonskiy D, Borisenko B, Burnaev E. Generative adversarial networks for image steganography. 2016.
48. Volkhonskiy D, Nazarov I, Burnaev E, editors. Steganographic generative adversarial networks. Twelfth international conference on machine vision (ICMV 2019); 2020: SPIE. DOI: 10.1117/12.2559429.
49. Ramandi VY, Fateh M, Rezvani M. VidaGAN: Adaptive GAN for image steganography. IET Image Processing. 2024;18(12):3329-42. DOI: 10.1049/ipr2.13177.
50. Zhang KA, Cuesta-Infante A, Xu L, Veeramachaneni K. SteganoGAN: High Capacity Image Steganography with GANs. ArXiv. 2019;abs/1901.03892. DOI: 10.48550/arXiv.1901.03892.
51. Bi X, Yang X, Wang C, Liu J. High-Capacity Image Steganography Algorithm Based on Image Style Transfer. Security and Communication Networks. 2021;2021(1):4179340. DOI: 10.1155/2021/4179340.
52. Su W, Ni J, Sun Y. StegaStyleGAN: towards generic and practical generative image steganography. Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence: AAAI Press; 2024. p. Article 28. DOI: 10.1609/aaai.v38i1.27776
53. Tang W, Rao Y, Yang Z, Peng F, Cui X, Huang J, et al. Reversible generative steganography with distribution-preserving. Cybersecurity. 2025;8(1):18. DOI: 10.1186/s42400-024-00317-6.
54. Wang Z, Zhou M, Liu B, Li T. Deep Image Steganography Using Transformer and Recursive Permutation. Entropy. 2022;24(7):878. DOI: 10.3390/e24070878.
55. Zhou Y, Luo T, He Z, Jiang G, Xu H, Chang C-C. CAISFormer: Channel-wise attention transformer for image steganography. Neurocomputing. 2024;603:128295. DOI: 10.1016/j.neucom.2024.128295.
56. Xiao C, Peng S, Zhang L, Wang J, Ding D, Zhang J. A transformer-based adversarial network framework for steganography. Expert Systems with Applications. 2025;269:126391. DOI: 10.1016/j.eswa.2025.126391.
57. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing. 2004;13(2):600-12. DOI: 10.1109/TIP.2003.819861.
58. Setiadi DRIM. PSNR vs SSIM: imperceptibility quality assessment for image steganography. Multimedia Tools and Applications. 2021;80(6):8423-44. DOI: 10.1007/s11042-020-10035-z.
59. Agustsson E, Timofte R. NTIRE 2017 Challenge on Single Image Super-Resolution: Dataset and Study. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW): IEEE Computer Society; 2017. p. 1122-31. DOI: 10.1109/cvprw.2017.150.
60. Kinga D, Adam JB, editors. A method for stochastic optimization. International conference on learning representations (ICLR); 2015: California;.
61. Mehta, S., & Rastegari, M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. ArXiv. 2021; DOI: 10.48550/arXiv.2110.02178.
62. Kishore V, Chen X, Wang Y, Li B, Weinberger KQ, editors. Fixed neural network steganography: Train the images, not the network. International Conference on Learning Representations; 2021.
63. Zhang S, Li H, Li L, Lu J, Zuo Z. A high-capacity steganography algorithm based on adaptive frequency channel attention networks. Sensors. 2022;22(20):7844. DOI: 10.3390/s22207844.
64. Chen X, Kishore V, Weinberger KQ, editors. Learning iterative neural optimizers for image steganography. The Eleventh International Conference on Learning Representations; 2023. https://openreview.net/forum?id=gLPkzWjdhBN.
65. Reinel T-S, Brayan A-AH, Alejandro B-OM, Alejandro M-R, Daniel A-G, Alejandro A-GJ, et al. GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis. IEEE Access. 2021;9:14340-50. DOI: 10.1109/ACCESS.2021.3052494.

Persian Abstract

چکیده

نهان‌نگاری تصویر با به‌کارگیری یادگیری عمیق و به‌ویژه شبکه‌های عصبی کانولوشنی برای استخراج ویژگی‌های پیچیده، پیشرفت چشمگیری داشته است. در حالی که شبکه‌های کانولوشنی به دلیل کارایی بالا در شناسایی وابستگی‌های محلی، مدت‌هاست به عنوان ابزاری قدرتمند در پردازش تصویر شناخته می‌شوند، ظهور مدل‌های ترنسفورمر تحول بزرگی در این حوزه ایجاد کرده است زیرا این مدل‌ها با دقت بالاتر و توانایی درک روابط سراسری در داده‌ها، عملکرد بهتری ارائه می‌دهند. با این حال، ترنسفورمرها در پردازش تصاویر با وضوح بالا با چالش‌هایی مانند هزینه محاسباتی و مصرف حافظه مواجه هستند. برای رفع این محدودیت‌ها، ما معماری هیبریدی‌ای پیشنهاد می‌دهیم که در مراحل با وضوح بالا از لایه‌های کانولوشنی و در مراحل با وضوح پایین از بلوک‌های ترنسفورمر استفاده می‌کند. این طراحی با بهره‌گیری از لایه‌های کانولوشنی برای استخراج وابستگی‌های کوتاه‌برد و ترنسفورمرها برای مدل‌سازی روابط بلندبرد، تعادلی میان کارایی و عملکرد ایجاد می‌کند. همچنین با جایگزینی کانولوشن‌های استاندارد با کانولوشن مختصات، آگاهی فضایی و قابلیت مکان‌یابی ویژگی‌ها در مدل را بهبود می‌بخشیم. ما چارچوبی نوآورانه برای نهان‌نگاری تصویر به نام VidaFormer معرفی می‌کنیم که یادگیری عمیق را با این بهبودهای معماری ترکیب می‌کند تا داده‌های دودویی دلخواه را در تصاویر جاسازی کند و در عین حال کیفیت بصری بالای تصاویر نهان را حفظ نماید. روش پیشنهادی با دستیابی به ظرفیت مؤثر ٤/٨٩ بیت بر پیکسل روی دادگان DIV2K که ٪۱/۷ بهتر از بهترین مدل قبلی است، عملکرد بهتری نسبت به مدل‌های پیشرفته فعلی دارد. این نتایج نشان‌دهنده عملکرد و مقیاس‌پذیری برتر VidaFormer برای کاربردهای مدرن نهان‌نگاری تصویر است. کد منبع در نشانی https://github.com/vidayousefi/vidaformer-ije در دسترس است.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2452-2460

International Journal of Engineering

Journal Homepage: www.ije.ir

Evaluation of Aluminum-Crosslinked Gel Compositions for Controlling Water Cut and Improving Flow Conformance in Oil-Bearing Formations

D. G. Podoprigora, D. Markushina*

Petroleum Engineering Department, Empress Catherine II Saint Petersburg Mining University, Saint Petersburg, Russian Federation

PAPER INFO

Paper history:
Received 15 September 2025
Received in revised form 16 October 2025
Accepted 14 December 2025

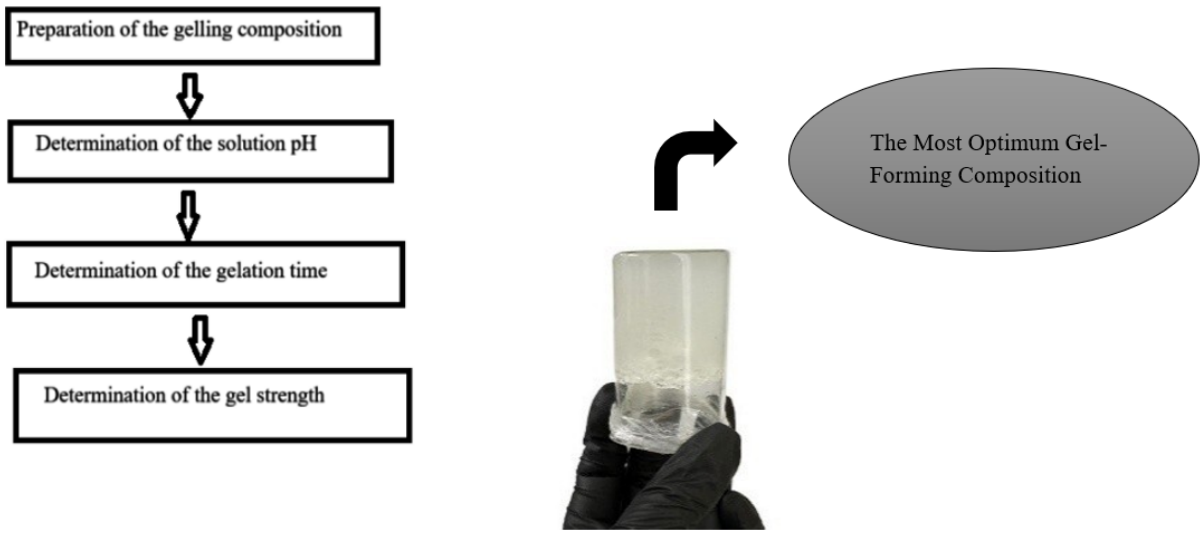
Keywords:
Aluminum
Conformance Control
Crosslinked Gels
Enhanced Oil Recovery
Gelation Time
Gel Strength
Polyacrylamide
pH

ABSTRACT

High water cut in produced fluids is a common challenge across most hydrocarbon reservoirs. In this regard, conformance control technologies based on gel-forming compositions are considered one of the most effective methods of enhancing oil recovery. This study investigated a gel-forming composition based on an acrylamide polymer and an aluminum-containing crosslinker. Despite the extensive range of developed crosslinked polymer systems, achieving precise control over the gelation rate without introducing auxiliary retarders remains a significant challenge, particularly when inorganic crosslinkers such as aluminum-based compounds are employed. The aim of the research was to determine the optimal crosslinker-to-polymer ratio required to obtain a gel-forming system with high mechanical strength and adjustable gelation time that meets the technological requirements for reservoir injection. Experimental studies were carried out to determine the gelation time, solution pH, and mechanical strength of the resulting gel. The gelation time reached up to 197 minutes at a polymer concentration of 2 wt.% under standard temperature conditions (20 °C). The measured storage modulus of the gel reached 1350 Pa at a pH value of 4.66. The efficiency of the investigated aluminum-containing crosslinker is compared with existing analogues currently used in the oil and gas industry for water shutoff and conformance control applications.

***doi**: 10.5829/ije.2026.39.10a.08*

Graphical Abstract



*Corresponding Author Email: s233002@stud.spmi.ru (D. Markushina)

Please cite this article as: Podoprigora DG, Markushina D. Evaluation of Aluminum-Crosslinked Gel Compositions for Controlling Water Cut and Improving Flow Conformance in Oil-Bearing Formations. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2452-60.

1. INTRODUCTION

Global practice demonstrates that hydrocarbon production is often accompanied by high water cut (1-3). This problem is particularly pronounced in mature oilfields, where the ratio of produced water to hydrocarbons steadily increases over time as the reservoir becomes depleted. High water cut not only reduces the efficiency of hydrocarbon recovery but also leads to an increase in operating costs, as the handling, separation, and disposal of produced water represent a significant burden for the oil industry. A promising approach to improving the oil recovery factor in mature reservoirs is the implementation of conformance control technologies using gel-forming compositions, making their development an urgent task for the petroleum industry (4-6). These technologies are designed to modify reservoir permeability profiles, block high-permeability channels, and restrict the unwanted movement of injected fluids, thereby improving sweep efficiency and extending the productive life of oilfields. Approaches to mineral resource management are outlined by Litvinenko et al. (7), which are also relevant to these technologies. In practice, the application of such methods is considered a cost-effective alternative to large-scale redevelopment or drilling of infill wells, highlighting the strategic importance of conformance control solutions.

Among the wide range of gel-forming systems, chemical methods based on polymers (8, 9) are most widely applied due to their effective mechanisms of action (10, 11). These systems are capable of forming strong three-dimensional networks that selectively reduce water mobility while leaving oil permeability largely unaffected. Polyacrylamide (PAM) and its derivatives are the most commonly used reagents for the preparation of polymer gels employed in oil reservoirs (12-14). Their advantages include high solubility in water, tunable molecular weight, and the ability to undergo controlled gelation under reservoir conditions. The versatility of PAM-based systems has led to their adoption across different types of reservoirs, ranging from sandstone formations to fractured carbonates. Some polymer systems have been described for water shutoff operations (15, 16), where they are injected to seal water-bearing layers or fractures, while others have been applied for well-killing purposes (17, 18), ensuring safe temporary plugging during workover and intervention activities.

Polymer gels may utilize organic crosslinkers (polyphenol, phenol-formaldehyde, hydroquinone, etc.) (19, 20) or inorganic ones (aluminum, chromium, iron compounds, etc.) (21-23). It is well established that polymer compositions with inorganic crosslinkers are more frequently used for in-situ reservoir isolation due to their selectivity (24, 25) and adjustable gelation time compared with those containing organic crosslinkers (22, 26).

In this study, an aluminum-based crosslinker was selected owing to its ability to form stable gel networks through interactions with amide and hydroxyl groups of polymers and its non-toxic properties (27, 28).

The mechanism of polymer crosslinking with the crosslinker under consideration is shown in Figure 1.

It is well established that aluminum ions are capable of binding several polymer chains simultaneously (29). Nevertheless, the main challenge associated with aluminum-based gel systems lies in controlling the rate of gelation. Therefore, one of the key scientific tasks in such systems is to extend the gelation time without compromising the mechanical strength of the gel and without using toxic or corrosion-active retarders.

Although aluminum as a crosslinking agent has certain limitations, it has been extensively studied in several works, where it has been demonstrated that aluminum exhibits high stability in a mildly acidic medium with a pH range of 4 to 6 and the ability to maintain gel integrity at elevated temperatures up to 200 °C (23, 27). However, as reported in other studies (30,31), at hydrogen ion concentrations below pH 3.5 and above pH 9.0, gel degradation and precipitation occur. The interaction of aluminum ions with the functional groups of polyacrylamide provides the formation of a stable three-dimensional gel network that retains its structural integrity even under high salinity conditions, with sodium chloride concentrations up to 50,000 ppm (31).

Several studies have shown that the structural and chemical stability of inorganic compounds significantly affects their functional performance. Aziz et al. (32) developed Ni/ZnO nanocomposites exhibiting enhanced chemical resistance and catalytic activity due to the precise control of interactions between inorganic particles. A similar concept can be applied to aluminum-

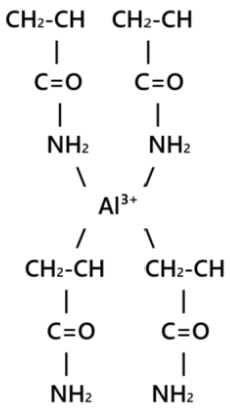


Figure 1. Mechanism of cross-linking of acrylic polymer (PAM) with trivalent aluminum ions (Al^{3+})

based crosslinked gel systems, where the gel strength and stability depend strongly on the bonding interactions between aluminum ions and the amide groups of polyacrylamides. Furthermore, the development of flow-diverting technologies in the oil and gas industry is closely associated with global digitalization and the implementation of innovative approaches. Kongdachudomkul et al. (33) proposed a model for assessing the digital maturity of oil and gas companies, emphasizing the importance of integrating advanced materials and modern process management methods. In the present study, experimental investigations were carried out to analyze the behavior of an aluminum-containing crosslinker and to establish the relationships between its concentration, gelation time, pH and the mechanical properties of the resulting gel-forming composition.

2. METHODS AND MATERIALS

2. 1. Gel Preparation Process To obtain the gel-forming composition, the procedure shown in Figure 3 was followed. A magnetic stirrer IKA - RT 5 (Figure 2) was used to mix the solution, providing up to 1000 rpm at room temperature.

The polymer concentration in all experiments was fixed at 4%, while the crosslinker concentration varied from 0.1 to 1 wt.%.

Figure 2. Illustration of the mixing process on a magnetic stirrer IKA - RT 5

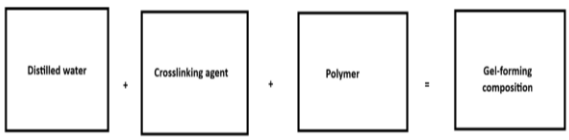


Figure 3. Procedure for preparing the gel-forming composition

2. 2. Determination of Gelation Time Gelation time is a fundamental parameter when selecting the optimal concentration to form an effective crosslinked system (34, 35). As in other studies (19, 26, 36), this work employed the bottle test, a rapid and cost-effective method based on visual observation. Additionally, the Sydansk gel code (A–J) was used to classify gel strength (10, 37, 38). For example: Code A – No detectable gel formed; B - highly flowing gel; code C - flowing gel; code D - moderately flowing gel; Code E – barely flowing gel; F- highly deformable nonflowing gel; G - moderately deformable nonflowing gel; code H - Slightly deformable nonflowing gel; Code I – rigid gel; Code J – ringing rigid gel.

2. 3. Determination of pH The hydrogen ion concentration is a key parameter influencing gelation in crosslinked polymer systems (16, 39). It is important to identify the pH interval where strong and stable gels form. A Testo-206 pH meter (Figure 4) was used.

2. 4. Determination of Gel Strength The efficiency of gel-forming systems is largely determined by their ability to resist water breakthrough, which depends on gel strength. Strength tests were conducted 24 and 48 hours after preparation. The Rebinder method was initially applied, measuring cone penetration under constant load for 10 minutes. (Figure 5) (40, 41).

Figure 4. Testo-206 pH meter

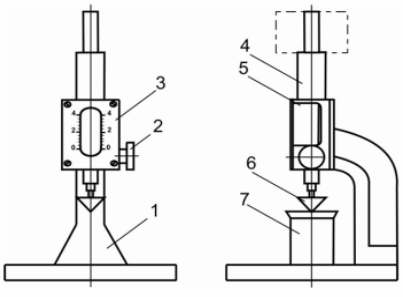


Figure 5. Scheme of a conical plastometer: 1—frame; 2—clamp; 3—measuring scale; 4—rod; 5—vertical clip; 6—cone; 7—glass beaker with test gel

However, due to the softness of early-stage gels, more accurate rheological testing with an Anton Paar MCR 302 rheometer was employed. The oscillatory test determined the storage modulus (G′), a key indicator of gel strength.

After 24 hours of preparation of the gel-forming composition, this test was performed in a rheometer. A plate-plate measuring system with a diameter of 25 mm and a gap of 2 mm was used. The test is called oscillatory and is performed on the basis of an amplitude ramp, conducted in the range from 1 to 1000% at a constant frequency of 1 Hz (within the linear viscoelastic region). The experiments were conducted at a temperature of 25 °C.

3. RESULTS AND DISCUSSION

3. 1. Gelation Time The authors conducted experimental studies to determine the gelation time at various concentrations of polyacrylamide (PAA) in the range of 1–5 wt.% and an aluminum-based crosslinker within 0.1–1.0 wt.%. The selected concentration range for the crosslinker was based on preliminary tests, which showed that gel formation does not occur at aluminum reagent concentrations below 0.1 wt.% due to an insufficient number of coordination centers for forming a three-dimensional polymer network. Furthermore, at crosslinker concentrations above 0.8 wt.%, precipitate formation and system destabilization were observed, attributed to excessive hydrolysis of aluminum ions and the formation of insoluble hydroxide phases (29). Figure 6 shows the dependence of gelation time on crosslinker concentration. The data indicate that the gelation time increases up to 197 minutes with an increase in the aluminum-based component concentration to 0.7 wt.% and a decrease in polymer concentration to 2 wt.%. This behavior can be explained by the increased formation of oligomeric aluminum species at higher crosslinker concentrations, which are less reactive than monomeric Al^{3+} ions. This slows down the formation of a uniform network. At the same time, a decrease in polymer concentration reduces the density of active points involved in crosslinking, further contributing to the increase in gelation time (42). The optimal curing time of gel-forming compositions is generally considered to be 30 minutes or longer, depending on specific parameters (temperature, pH) and the particular objectives of each researcher (19, 43). At a low polymer concentration - 1 wt.%, no gel formation was observed during the bottle test, as indicated by meniscus deviation, confirming the system's fluid state. This is probably due to the insufficient number of polymer chains available for interchain coordination with aluminum ions. Several studies (23, 31, 44) report similar gelation time ranges for aluminum-based crosslinkers without additives. According to the Sydansk method, gels obtained at concentrations from 0.1 to 0.7% correspond to Code I (Figure 7).

3. 2. Effect of pH The optimal pH range is determined by the polymer nature and crosslinker type, as the acid-base environment affects both the complexation kinetics and bond stability. For chromium-containing systems, studies (16, 45) indicate optimal crosslinking near neutral pH, where chromium ions maintain coordination activity without premature hydrolysis. In contrast, aluminum-based systems exhibit different behavior. Research conducted by Shamlooh et al. (31) and Altunina et al. (46) demonstrates that stable aluminum-gel compositions form across a broader pH range (0.5–14), depending on the aluminum compound and complexing additives. Specifically, gelation rate decreases with increasing pH, suggesting that mildly acidic conditions (4–5) promote active coordination and stable bond formation, leading to robust polymer networks (31). As noted in work (29), higher pH values induce aluminum ion hydrolysis, forming larger polymeric structures and altering crosslinking mechanisms. pH is a key driver of aluminum speciation in all environments. In this study, the pH range of the systems was 3.54–5.6. Figure 8 illustrates the pH dependence on aluminum crosslinker concentration: increasing crosslinker content (0.1–1.0 wt.%) and decreasing polymer concentration (5–1 wt.%) reduce pH due to enhanced aluminum ion hydrolysis and H^+ release.

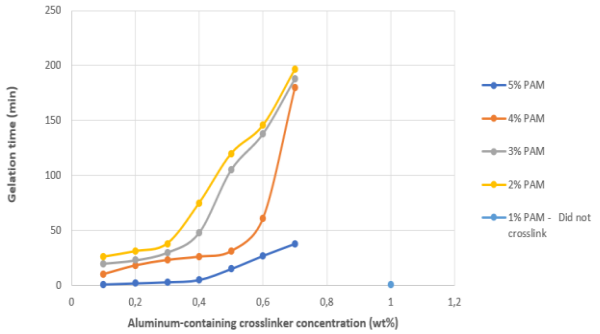


Figure 6. Determination of gelation time

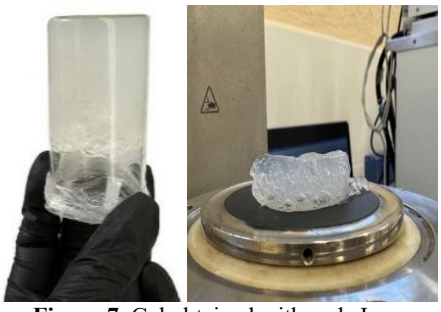


Figure 7. Gel obtained with code I

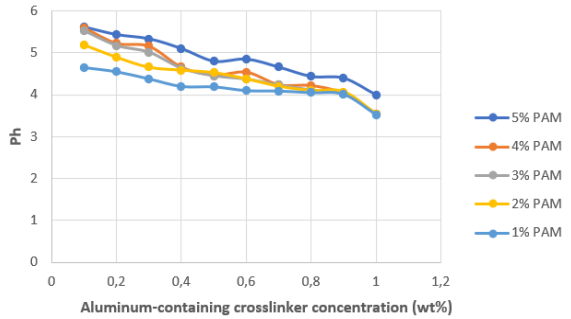


Figure 8. pH of gel-forming composition

In research by Berger et al. (42) further confirm that monomeric aluminum hydroxy species dominate at low pH, transitioning to oligomeric/polymeric forms as pH rises, reducing free H^+ concentration. Thus, literature and experimental data confirm that the optimal pH range for aluminum-gel systems is approximately 4–5. This interval balances reaction rate, aluminum hydrolysis, and coordination complex stability, ensuring the formation of strong and durable gel structures.

3. 3. Effect of Crosslinker Concentration on Gel Strength At the initial stage, the strength of the gel-forming compositions was assessed using the Rebinder method. However, the experiments revealed that the gels exhibited low mechanical rigidity during the early stages of gel network development, resulting in strength values below the detection threshold of the method. The cone penetrated completely into the sample under its own weight within 24 hours of preparation, making it impossible to obtain reliable data. This indicates that the Rebinder method has limited applicability for the characterization of soft gel-forming systems, particularly during the first day after preparation. Therefore, for subsequent evaluations, a more precise, sensitive, and reliable approach was adopted—rheological testing using an Anton Paar MCR 302 rheometer (31, 47) (Figure 9). As shown in Figure 10, the storage modulus (G′) increases with higher concentrations of both the aluminum-based crosslinker and polyacrylamide. This is attributed to increased crosslink density between macromolecules, forming a more robust three-dimensional network. Increasing the crosslinker concentration from 0.1 to 0.7 wt.% raised G′ to 1350 Pa, indicating the development of a robust three-dimensional polymer network capable of withstanding substantial mechanical stresses during water shutoff operations in the reservoir. At PAM concentrations of 4-5 wt%, maximum G′ values of 1100-1350 Pa were achieved after 24 hours (Figure 10). Additional measurements after 48 hours showed further increase to 1990-2200 Pa, confirming strength development through slow bond reorganization and ongoing aluminum chain hydrolysis. These results align with literature values for unmodified aluminum-complexed systems, where G′ ranges from 125 Pa to 2000 Pa were reported in some works (29,31). As noted by Raupov et al. (16) optimal high-strength gels with low initial viscosity and delayed gelation can be achieved using chemical modifiers such as gel-time regulators, retarders and radiation-treated polymers. Figure 11 shows a graph showing the dependence of elastic modulus (G′) on strain for a crosslinker concentration of 0.7% wt.% at a polymer ratio - 4 wt.% obtained by rheometer.

Table 1 shows the comparison of the investigated gel-forming composition with analogs.

Figure 9. The process of investigating the strength of gel-forming composition using Anton Paar MCR 302 rheometer

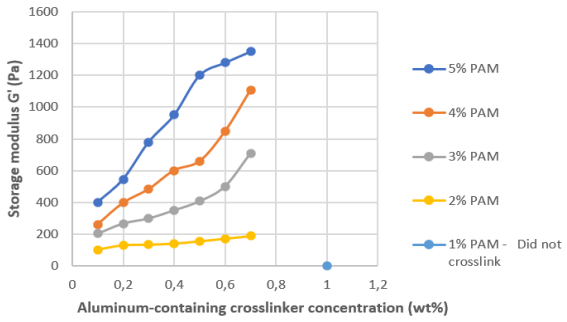


Figure 10. Storage modulus of the obtained gel as a function of the crosslinker concentration

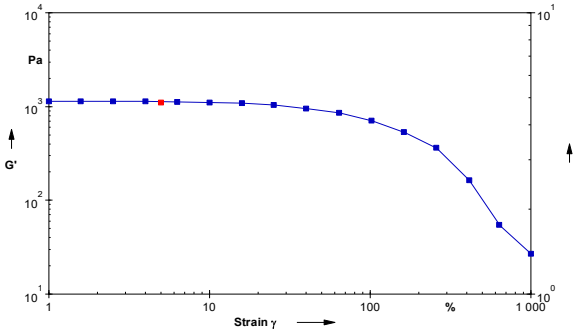


Figure 11. Obtained function of elastic modulus (G′) on strain by Anton Paar MCR 302 rheometer

TABLE 1. Comparison of the investigated gel-forming composition with analogs

Gel-forming composition	Gelation time (min)	pH range with gel formation	Storage modulus G′ (Pa)	Specific characteristics	References
Investigated gel-forming composition based on polyacrylamide and aluminum-containing crosslinker	<197	3.54-5.6	1350	Without additives or retarders; high strength under natural pH conditions	The authors' composition under study
Polyacrylamide + tris(2-aminoacetate) aluminum	Not specified	0.5-10	<125	Requires pH control; reaction proceeds more slowly	(31)
Polyacrylamide + aluminum/acetate	50	0.5-8.5	2000	Requires pH control	(31)
Polyacrylamide + aluminum/zirconium	40-60	4.0-6.5	<100	Complex structure with zirconium ions; suitable for high reservoir temperatures up to 200 °C	(23)
Polyacrylamide + aluminum/nitrate	Not specified	2.5-14	2000	Narrow pH range	(31)
Polyacrylamide + aluminum	150	Not specified	Not specified	Various salinity ranges	(44)

Table 1 highlights the advantages of the investigated gel system, confirming its practical potential.

4. Future Research Directions and Recommendations

Future work will focus on expanding the study of the developed gel-forming composition through detailed characterization of the chemical composition and crosslinking mechanism of the aluminum-based agent. A key objective is the selection of a gelation kinetics modifier to increase the gelation time, which is essential for large-volume treatments in conformance control applications. To enhance data accuracy, rheological tests will be conducted to determine the gelation induction period at specific shear rates, simulating the mechanical stresses experienced during field injection. Furthermore, the adhesive properties of the gel towards oil- and water-saturated rocks of varying lithology will be investigated. This will assess the gel's ability to adhere to and remain stable on pore walls, a critical factor for long-term isolation performance. Additional studies will examine gel performance across different temperature ranges and core model test to evaluate the composition's effectiveness in isolating high-permeability channels.

5. CONCLUSIONS

The study presents a set of laboratory investigations aimed at developing gel-forming systems, including determination of gelation time by visual inspection, measurement of pH using a Testo-206 pH meter, and evaluation of gel strength by both the Rebinder method and rheological analysis with an Anton Paar MCR 302 rheometer. The following conclusions can be drawn from the research:

- The maximum gelation time of the studied composition based on polyacrylamide and an aluminum-based crosslinker, determined by the bottle test, is 197 minutes at a polymer concentration of 2 wt.% and standard temperature.
- Gel formation for the studied composition occurs within a pH range of 3.54 to 5.6.
- At the initial stage, the Rebinder method was used to determine strength characteristics; however, its limitations were identified. Consequently, a methodology based on an oscillatory test using an Anton Paar MCR 302 rheometer was employed.
- An increase in the crosslinker concentration to 0.7 wt.% leads to a significant rise in the storage modulus (G′) to 1350 Pa, measured 24 hours after preparation at a polymer concentration of 5 wt.%.
- The experimental results and a comparison of the studied composition with similar systems from the literature confirm its competitiveness.

Acknowledgements

The experimental investigations reported in this study were conducted at the laboratories of Saint Petersburg Mining University of Empress Catherine II, Scientific Center “Arctic”. The authors gratefully acknowledge the Scientific Center “Arctic” for providing the facilities, materials, chemical reagents, and equipment required to carry out this research.

Funding

This research received no external funding.

Ethics Approval and Consent to Participate

This study did not involve human participants or animals. Therefore, ethical approval and informed consent were not required.

Competing Interests

The authors declare that they have no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors used artificial intelligence–based tools to assist with the accurate scientific translation of selected phrases from Russian into English. All translated content was carefully reviewed, edited, and validated by the authors, who take full responsibility for the integrity, originality, and accuracy of the final manuscript.

Author Contributions

D. G. Podoprigora, Ph.D. in Petroleum Fngineering: Conceptualization, problem definition, article writing, and execution of experimental studies.
F. D. Markushina, Ph.D. candidate: Literature review, experimental studies, data systematization, and article writing.

REFERENCES

1. Duryagin V, Nguyen Van T, Onegov N, Shamsutdinova G. Investigation of the selectivity of the water shutoff technology. Energies. 2022;16(1):366. https://doi.org/10.3390/en16010366
2. Fan G, Li M, Chen X, Palyanitsina A, Timoshin A. Polymer-nanosilica-assisted to evaluate oil recovery performances in sandstone reservoirs. Energy Reports. 2021;7:2588-2593. https://doi.org/10.1016/j.egyr.2021.04.047
3. Podoprigora D, Byazrov R, Sytnik J. The comprehensive overview of large-volume surfactant slugs injection for enhancing oil recovery: status and outlook. Energies. 2022;15(21):8300. 10.3390/EN15218300
4. Dvoynikov MV, Kuchin VN, Mintzaev MS. Development of viscoelastic systems and technologies for isolating water-bearing horizons with abnormal formation pressures during oil and gas wells drilling. Journal of Mining Institute. 2021;247:57-65. 10.31897/PMI.2021.1.7
5. Raupov IR, Kone A, Podjapolski EV, Milic J. Justification of conformance control in mature oil fields. International Journal of Engineering Transactions A: Basics. 2025;38(1):12-20. 10.5829/ije.2025.38.01a.02
6. Zhou M, Bu J, Wang J, Guo X, Huang J, Huang M. Study on three phase foam for enhanced oil recovery in extra-low permeability reservoirs. Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles. 2018;73:55. https://dx.doi.org/10.2516/ogst/2018059
7. Litvinenko VS, Petrov EI, Vasilevskaya DV, Yakovenko AV, Naumov IA, Ratnikov MA. Assessment of the role of the state in the management of mineral resources. Journal of Mining Institute. 2023;259:95-111. 10.31897/PMI.2022.100
8. Petrakov D, Loseva A, Alikhanov N, Jafarpour H. Standards for selection of surfactant compositions used in completion and stimulation fluids. International Journal of Engineering Transactions C: Aspects. 2023;36(9):1605-1610. 10.5829/ije.2023.36.09c.03
9. Korobov GY, Parfenov DV, Nguyen VT. Long-term inhibition of paraffin deposits using porous ceramic proppant containing solid ethylene-vinyl acetate. International Journal of Engineering Transactions B: Applications. 2025;38(8):1887-1897. 10.5829/ije.2025.38.08b.13
10. Sydansk RD, Al-Dhafeeri AM, Xiong Y, Seright RS. Polymer gels formulated with a combination of high- and low-molecular-weight polymers provide improved performance for water-shutoff treatments of fractured production wells. SPE Production & Facilities. 2004;19(4):229-236. https://doi.org/10.2118/89402-PA
11. Dvoynikov M, Sidorov D, Kambulov E, Rose F, Ahiyarov R. Salt deposits and brine blowout: development of a cross-linking composition for blocking formations and methodology for its testing. Energies. 2022;15(19):7415. https://doi.org/10.3390/en15197415
12. Aguiar KLNP, De Oliveira PF, Mansur CRE. A comprehensive review of in situ polymer hydrogels for conformance control of oil reservoirs. Oil & Gas Science and Technology – Revue d'IFP Energies nouvelles. 2020;75(8):1-17. https://doi.org/10.2516/ogst/2019067
13. Ketova YA, Bai B, Khizhnyak GP, Gladkikh EA, Galkin SV. Testing of preformed particle polymer gel technology on core filtration models to limit water inflows. Journal of Mining Institute. 2020;241:91-96. 10.31897/PMI.2020.1.91
14. Seright RS, Lane RH, Sydansk RD. A strategy for attacking excess water production. SPE Permian Basin Oil and Gas Recovery Conference; 2001 May 15–17; Midland, Texas. SPE Paper No. 70067. https://doi.org/10.2118/70067-MS
15. Palyanitsina A, Safiullina E, Byazrov R, Podoprigora D, Alekseenko A. Environmentally safe technology to increase efficiency of high-viscosity oil production for the objects with advanced water cut. Energies. 2022;15(3):753. https://doi.org/10.3390/en15030753
16. Raupov I, Rogachev M, Sytnik J. Design of a polymer composition for the conformance control in heterogeneous reservoirs. Energies. 2023;16(1):515. https://doi.org/10.3390/en16010515
17. Mardashov DV, Limanov MN. Improving the efficiency of oil well killing at the fields of the Volga-Ural oil and gas province with abnormally low reservoir pressure. Bulletin of the Tomsk Polytechnic University Geo Assets Engineering. 2022;333(7):185-194. 10.18799/24131830/2024/6/436
18. Islamov SR, Bondarenko AV, Gabibov AF, Mardashov DV. Polymer compositions for well killing operation in fractured reservoirs. In: Advances in Raw Material Industries for Sustainable Development Goals. London: Taylor & Francis; 2021. p. 343-351. 10.1201/9781003164395
19. Zhu D, Hou J, Meng X, Zheng Z, Wei Q, Chen Y, Bai B. Effect of different phenolic compounds on performance of organically cross-linked terpolymer gel systems at extremely high temperatures. Energy & Fuels. 2017;31(8):8120-8130. https://doi.org/10.1021/acs.energyfuels.7b01386
20. Liu Y, Dai C, Wang K, Zhao M, Zhao G, Yang S, You Q. New insights into the hydroquinone (HQ)–hexamethylenetetramine (HMTA) gel system for water shut-off treatment in high temperature reservoirs. Journal of Industrial and Engineering Chemistry. 2016;35:20-28. 10.1016/j.jiec.2015.09.032

21. Lu S, Bo Q, Zhao G, Shaikh A, Dai C. Recent advances in enhanced polymer gels for profile control and water shutoff: a review. Frontiers in Chemistry. 2023;11:1067094. https://doi.org/10.3389/fchem.2023.1067094
22. Pu J, Geng J, Bai B. Effect of chromium ion diffusion in polyacrylamide/chromium acetate gelation system with surrounding water on gelation behavior. Journal of Petroleum Science and Engineering. 2018;171:1067-1076. https://doi.org/10.1016/j.petrol.2018.08.014
23. Shi S, Sun J, Mu S, Lv K, Liu J, Bai Y, Li J. Effect of chemical composition of metal–organic crosslinker on the properties of fracturing fluid in high-temperature reservoir. Molecules. 2024;29(12):2798. https://doi.org/10.3390/molecules29122798
24. Shagiakhmetov A, Yushchenko S. Substantiation of in situ water shut-off technology in carbonate oil reservoirs. Energies. 2022;15(14):5059. https://doi.org/10.3390/en15145059
25. Sharma P, Kudapa VK. Study on the effect of cross-linked gel polymer on water shutoff in oil wellbores. Materials Today: Proceedings. 2022;48:1103-1106. https://doi.org/10.1016/j.matpr.2021.07.506
26. Zhang H, Yang H, Zhou B, Li X, Zhao H, Wang F, Gabdullin M. Effects of cyclodextrin polymer on the gelation of amphiphilic polymer in inclusion complex. Journal of Molecular Liquids. 2020;305:112850. https://doi.org/10.1016/j.molliq.2020.112850
27. Mao J, Wang D, Yang X, Zhang Z, Yang B, Zhang C, Zhao J. Experimental study on high temperature resistance aluminum-crosslinked non-aqueous fracturing fluids. Journal of Molecular Liquids. 2018;258:202-210. https://doi.org/10.1016/j.molliq.2018.03.032
28. Al-Assi AA, Willhite GP, Green DW, McCool CS. Formation and propagation of gel aggregates using partially hydrolyzed polyacrylamide and aluminum citrate. SPE Journal. 2009;14(3):450-461. https://doi.org/10.2118/100049-PA
29. Bi S, Wang C, Cao Q, Zhang C. Studies on the mechanism of hydrolysis and polymerization of aluminum salts in aqueous solution: correlations between the “Core-links” model and “Cage-like” Keggin-Al_{13} model. Coordination Chemistry Reviews. 2004;248(5–6):441-455. 10.1016/j.ccr.2003.11.001
30. Hamza A, Shamlooh M, Hussein IA, Nasser MS, Onawole AT, Magzoub M, Salehi S. Impact of aluminium acetate particle size on the gelation kinetics of polyacrylamide-based gels: rheological and molecular simulation study. Canadian Journal of Chemical Engineering. 2022;100(6):1169-1177. https://doi.org/10.1002/cjce.24152
31. Shamlooh M, Hussein IA, Nasser MS, Magzoub M, Salehi S. Development of pH-controlled aluminum-based polymeric gel for conformance control in sour gas reservoirs. ACS Omega. 2020;5(38):24504-24512. https://doi.org/10.1021/acsomega.0c02967
32. Aziz S, Rehman NU, Soonmin H, Ahmad N. Novel Ni/ZnO nanocomposites for the effective photocatalytic degradation of malachite green dye. Civil Engineering Journal. 2024;10(8):2601-2614. http://dx.doi.org/10.28991/CEJ-2024-010-08-011
33. Kongdachudomkul C, Puriwat W, Hoonsopon D. Digital Transformation Maturity Measurement (DTMM) for the Oil and Gas Industry. *HighTech and Innovation Journal*. 2025;6(2). 2723-9535 http://dx.doi.org/10.28991/HIJ-2025-06-02-018
34. Henaut I, Pasquier D, Rovinetti S, Espagne B. HP-HT drilling mud based on environmentally-friendly fluorinated chemicals. Oil & Gas Science and Technology – Revue d’IFP Energies nouvelles. 2015;70(6):917-930. http://dx.doi.org/10.2516/ogst/2014047
35. Pham LT, Hatzignatiou DG. Rheological evaluation of a sodium silicate gel system for water management in mature, naturally-fractured oilfields. Journal of Petroleum Science and Engineering. 2016;138:218-233. https://doi.org/10.1016/j.petrol.2015.11.039
36. Sydansk RD. A newly developed chromium (III) gel technology. SPE Reservoir Engineering. 1990;5(3):346-352. https://doi.org/10.2118/19308-PA
37. Sydansk RD, Argabright PA. Conformance improvement in a subterranean hydrocarbon-bearing formation using a polymer gel. US Patent 4683949. 1987.
38. Tessarolli FG, Souza ST, Gomes AS, Mansur CR. Influence of polymer structure on the gelation kinetics and gel strength of acrylamide-based copolymers, bentonite and polyethylenimine systems for conformance control of oil reservoirs. Journal of Applied Polymer Science. 2019;136(22):47556. https://doi.org/10.1002/app.47556
39. Xiao K, Mu L, Wu X, Yang L, Xu F, Zhu G, Zhu Y. Comprehensive study of polymer gel profile control for WAG process in fractured reservoir: using experimental and numerical simulation. SPE EOR Conference at Oil and Gas West Asia; 2016. 10.2118/179860-MS
40. Mardashov D, Duryagin V, Islamov S. Technology for improving the efficiency of fractured reservoir development using gel-forming compositions. Energies. 2021;14(24):8254. https://doi.org/10.3390/en14248254
41. Nikitin MN, Petukhov AV. Gelling composition based on sodium silicate for limiting water influx in complex fractured reservoirs. Electronic Scientific Journal Oil and Gas Business. 2011;5:143-153. EDN: PDWISD
42. Berger S, Nolde J, Yüksel T, Tremel W, Mondeshki M. 27Al NMR study of the pH-dependent hydrolysis products of $Al_2(SO_4)_3$ in different physiological media. Molecules. 2018;23(4):808. https://doi.org/10.3390/molecules23040808
43. Galkin SV, Rozhkova YA. Analysis of experience in the use of preformed particle polymer gels in the development of high-water-cut production facilities in low-temperature oil reservoirs. Journal of Mining Institute. 2024;265:55-64. EDN: CNCFIW
44. Zhang L, Khan N, Pu C. Influence of salinity on the properties of the different HPAM/Al3+ systems. Oil & Gas Science and Technology – Revue d’IFP Energies nouvelles. 2019;74:37. https://doi.org/10.2516/ogst/2019011
45. Sun F, Lin M, Dong Z, Zhu D, Wang SL, Yang J. Effect of composition of HPAM/chromium(III) acetate gels on delayed gelation time. Journal of Dispersion Science and Technology. 2016;37(6):753-759. https://doi.org/10.1080/01932691.2015.1041034
46. Altunina LK, Kuvshinov VA, Stasieva LA, Kuvshinov IV. Enhanced oil recovery for deposits of high-viscosity oils using multifunctional systems based on surfactants. Journal of Physics: Conference Series. 2020;1611(1):012033. 10.1088/1742-6596/1611/1/012033
47. Shamlooh M, Hamza A, Hussein IA, Nasser MS, Magzoub M, Salehi S. Investigation of the rheological properties of nanosilica-reinforced polyacrylamide/polyethyleneimine gels for wellbore strengthening at high reservoir temperatures. Energy & Fuels. 2019;33(7):6829-6836. 10.1021/acs.energyfuels.9b00974

Persian Abstract

چکیده

درصد بالای آب در سیالات تولیدی یکی از چالش‌های متداول در اغلب مخازن هیدروکربنی به شمار می‌رود. در این زمینه، فناوری‌های کنترل انطباق (Conformance Control) مبتنی بر ترکیبات ژل‌زا به‌عنوان یکی از مؤثرترین روش‌های ازدیاد برداشت نفت شناخته می‌شوند. در این پژوهش، ترکیب ژل‌زایی بر پایه‌ی پلیمر آکریل‌آمید و عامل شبکه‌ساز حاوی آلومینیوم مورد بررسی قرار گرفت. با وجود گستره‌ی وسیعی از سامانه‌های پلیمری شبکه‌ای توسعه‌یافته، کنترل دقیق نرخ ژلاسیون بدون استفاده از مواد بازدارنده‌ی کمکی همچنان یک چالش اساسی محسوب می‌شود، به‌ویژه هنگامی که از عامل‌های شبکه‌ساز معدنی نظیر ترکیبات پایه آلومینیوم استفاده گردد. هدف از این تحقیق تعیین نسبت بهینه‌ی عامل شبکه‌ساز به پلیمر برای دستیابی به سامانه‌ای ژل‌زا با استحکام مکانیکی بالا و زمان ژلاسیون قابل تنظیم بود، به‌گونه‌ای که نیازهای فناوری تزریق در مخزن را برآورده سازد. آزمایش‌های تجربی به‌منظور تعیین زمان ژلاسیون، مقدار pH محلول و استحکام مکانیکی ژل حاصل انجام گرفت. زمان ژلاسیون در غلظت پلیمری ۲ درصد وزنی و در شرایط دمای استاندارد (۲۰ درجه سانتی‌گراد) تا ۱۹۷ دقیقه اندازه‌گیری شد. مدول ذخیره‌ی ژل در مقدار pH برابر ٤٠٦٦ به ۱۳۵۰ پاسکال رسید. بازده عامل شبکه‌ساز حاوی آلومینیوم بررسی‌شده با نمونه‌های مشابه موجود که در صنعت نفت و گاز برای عملیات کاهش آب‌دهی و کنترل انطباق به‌کار می‌روند مقایسه گردید.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2461-2471

International Journal of Engineering

Journal Homepage: www.ije.ir

Improving Person Re-Identification Performance Using ESRGAN Image Enhancement

F. Ilbeigi, H. Farsi*, S. Mohamadzadeh, A. Sezavar

Department of Electrical and Computer Engineering, University of Birjand, Birjand, Iran

P A P E R I N F O

Paper history:
Received 13 September 2025
Received in revised form 20 October 2025
Accepted 03 Januray 2026

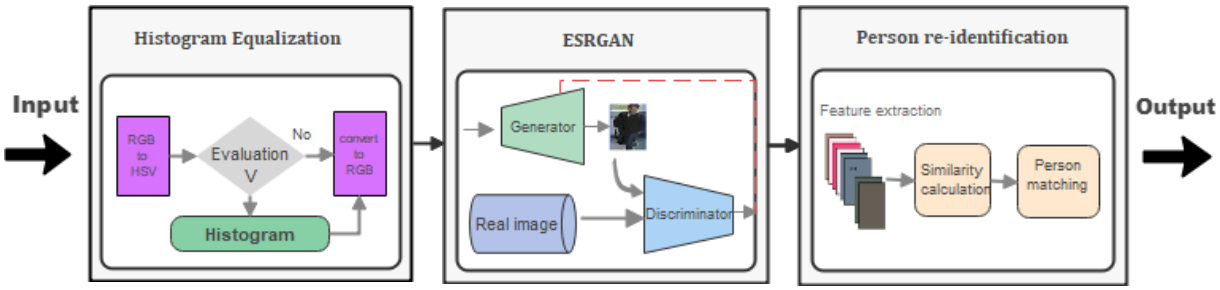
Keywords:
Histogram
Images
Super-Resolution
Person Re-Identification

A B S T R A C T

Person re-identification is one of the most important challenges in the field of image processing and computer vision, which has attracted significant attention in recent years. One of the main challenges in this domain is the low quality of images captured by different cameras, which can result in the loss of important details and consequently a reduction in the accuracy of re-identification algorithms. This paper presents a framework that enhances network accuracy and capability in determining individuals' identities by improving the brightness of low-light images using histogram equalization techniques and increasing their resolution using the ESRGAN network. The experimental results show that the proposed method, by improving the brightness and resolution of images which plays a key role in revealing important details, enhances the quality of the re-identification network's input. Also, compared to existing methods, our model increases the accuracy of person re-identification, such that it achieved Rank-1 accuracy of 93.6% for the CUHK01 dataset and 94.5% for the CUHK03 dataset.

***doi**: 10.5829/ije.2026.39.10a.09*

Graphical Abstract



1. INTRODUCTION

Person re-identification is the process of matching and identifying an individual in photos captured by different cameras to retrieve the top-k most similar images to the probe image of a person within a large dataset (gallery) recorded by different cameras (1). In recent years, with the increasing proliferation of surveillance cameras in urban environments, the importance of person re-identification has become more evident than ever. This technique has wide-ranging applications in border control, searching for missing people, video surveillance, traffic management, and urban security. The ability to reliably identify individuals across vast networks of surveillance cameras can play an important role in enhancing security. One of the main challenges of re-identification systems is the low resolution of images, which is particularly exacerbated when cameras capture images from long distances, unfavorable angles, or under adverse weather conditions, turning person re-identification into a complex problem (2). The reduction in image quality makes it difficult to extract accurate and distinctive features that are essential for re-identifying individuals, leading to decreased accuracy and an

*Corresponding Author Email: hfarsi@birjand.ac.ir (H. Farsi)

Please cite this article as: Ilbeigi F, Farsi H, Mohamadzadeh S, Sezavar A. Improving Person Re-Identification Performance Using ESRGAN Image Enhancement. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2461-71.

increased error rate (3). So far, a lot of studies have been done to deal with these problems (4–9). Some methods (10, 11), instead of reconstructing specific appearance features lost due to quality degradation, focus on mapping representations in a defined feature space across different resolutions. This method does not solve the problem of unequal access to information because it is impossible to get high-resolution pixel-level supervision for pictures of people taken from different angles. Some researchers (12, 13) only use handcrafted visual features and did not use neural networks to learn and improve discriminative attributes automatically. Because of this dependence, they have lower efficiency at dealing with complex changes in data, such as changes in angles, lighting, or crowded backgrounds, while artificial neural networks (ANNs) are capable of dealing with complex patterns (14).

This paper proposes a three-step process for improving the accuracy of person re-identification with low-resolution images. The primary focus of this study is on Low-Resolution Person Re-Identification (LR Re-ID). Images with low resolution usually contain scattered information and are sensitive to minor changes (15). These images have low contrast, and their details and edges are blurred and difficult to see, which makes it harder to get basic features like color, pixel intensity, pixel gradients, orientation, and so on (16). These basic features are what allow us to get advanced and semantic attributes, like facial structures or clothing patterns, which are very important for re-identification. The first step in the suggested model is to fix images with brightness levels below a certain threshold (which is based on a certain average pixel intensity) using histogram equalization methods. This first step makes the images bright enough to make hidden details in darker areas easier to see, but the images may still not be very sharp or have a lot of fine details. After that, a deep learning-based super-resolution network processes the images. We used an ESRGAN to improve the network's ability to extract features. It does this by creating a high-resolution image from a low-resolution input, which greatly improves the quality of the images. This enables rapid extraction of distinctive appearance features from the images. Finally, the person re-identification network (17) is used for feature extraction.

2. RELATED WORKS

In this section, the studies conducted in the areas of image super-resolution, person re-identification, and the application of super-resolution for person re-identification are reviewed separately.

2. 1. Image Super Resolution Image Super-Resolution (SR) is a prevalent method to make images clearer by making a high-resolution (HR) version from a low-resolution (LR) image (18). The widespread use of deep learning has sped up the development of super-resolution methods. Zhao et al. (19) were the first to succeed in this area by creating a deep Convolutional Neural Network (CNN) that linked low-resolution (LR) and high-resolution (HR) images from start to finish by lowering reconstruction error. This study laid the groundwork for several other studies in the development of CNN-based super-resolution models. Later, Wang et al. (20) came up with a hybrid model that used clustering of different blur kernels to get a better Peak Signal-to-Noise Ratio (PSNR) than basic models. In addition, image reconstruction was more effective using aggregation-based models like MAP (21) and ESCN (22). However, the majority of these models focus solely on pixel-wise similarity, resulting in output images that lack visual appeal.

Recently Generative Adversarial Networks (GAN) are widely used in image processing. Ledig (23) used a Generative Adversarial Network (24) and came up with a perceptual loss function to make textures that look and feel more real. Later, Zhao et al. (25) and Wang et al. (26) improved the GAN architecture and learning method to provide superior high-resolution outputs. Shamsolmoali et al. (27) used the least squares loss as a discriminator loss function to verify that training was stable. They also came up with a progressive GAN that could recreate more details in images. Subsequently, models such as SRGAN frequently produced outputs with overly smooth textures. The ESRGAN model was introduced to resolve this issue (26). Residual-in-Residual Dense Blocks are used to enhance image restoration and give it a more realistic quality. This study utilizes the ESRGAN model's superior ability to replicate image structures and intricate details. Compared to older models like SRGAN, this model is more successful in enhancing the ability to see and understand images.

2. 2. Person Re-Identification The use of deep networks in person reidentification has recently garnered considerable interest from researchers (28). Early CNN architectures, through the utilization of sequential convolutional layers, were able to significantly enhance the process of feature extraction and learning (29). Sezavar et al. (30) used a CNN-based feature extractor to obtain initial features for person re-identification, followed by a sparse representation for matching individuals across diverse viewpoints that this combination has facilitated the identification of individuals possessing diverse characteristics that integrate well within the group. A combination of an autoencoder and a deep convolutional neural network is used for person re-identification (17). The CNN initiates by identifying visual features. The autoencoder reduces their size and enhances their uniqueness through

compression and reconstruction. This compression reduces noise, enhancing the model's utility across diverse scenarios.

Chen et al. (31) examined a novel deep network known as the Deep Quadruplet Network. This network uses a quadruplet loss function. Instead of the standard triplet loss, this new loss function helps the model handle the distance between positive and negative samples. It also helps create a more structured and discriminative feature space. Zhao et al. (32) introduced a network called Spindle Net that analyzes features based on different human body regions for person reidentification. First, the network divides the person's image into separate body parts. Then, it extracts features from each segment on its own and finally combines these features. Liu et al. (33) introduced a network termed the Comparative Attention Network (CAN), which is trained end-to-end for person reidentification. The network adaptively focuses on critical regions of the images and extracts significant features in a comparative manner to strengthen discrimination between individuals. The use of an attention mechanism enables the model to concentrate on the more relevant parts of the body while ignoring less useful regions. This technique makes re-identification more accurate, especially in complex scenarios with diverse backgrounds.

2. 3. Low-Resolution Person Re-Identification In recent years, a lot of work has gone into solving the problem of person re-identification in low-resolution images. Early approaches primarily relied on dictionary learning and metric learning, but their performance was limited due to the lack of sufficient fine-grained details in low-resolution images. Many SR-based techniques have been proposed that greatly enhance matching accuracy thanks to progress in neural networks and super-resolution technology. Jiao et al. (34) were the first to put the SRCNN network and a re-identification network together into a single framework and suggest a way to train both networks at the same time. Wang et al. (35) used SR-GAN in a cascading structure by putting a number of SR-GAN modules in a row. This way, each low-resolution picture is sent to a different generator, which makes sure that all of the pictures are the same size. But neither of these methods trains SR modules together, so they don't take into account the extra features that each one finds. This makes it easy to improve them all in one place. They also did not consider generative artifacts, which hurt recognition performance, especially when the scaling factor gets bigger. Cheng et al. (36) improved the SR-ReID framework by using a better training method, which led to better results. Their method makes the two sub-networks work better together by using the natural relationship between super-resolution and re-identification. Mao et al. (10) added two parts: the first is a foreground-centric super-resolution model (FFSR) that fixes resolution loss in low-resolution input images, and the second is a resolution-invariant person re-identification module that helps with feature extraction. Even though this method has shown promising results, it requires accurate labeling of the foreground region to guide the image recovery learning process. Chen et al. (37) came up with the Resolution-Adaptive and Identity-Preserving Network (RAIN), which uses reconstruction loss to bring the feature distributions of high-resolution and low-resolution images closer together. This helps reduce the disparity between the deep features of images at different resolutions. Li et al. (38) developed a method that integrates resolution-invariant representation learning with image super-resolution to make cross-resolution person re-identification (ReID) work better.

3. PROPOSED METHOD

Figure 1 shows a general image of the proposed network. The model consists of three fundamental components. The first part enhances the image's brightness by equalizing the histogram in the HSV color space. In the second part, the ESRGAN network is used to improve the image resolution. Finally, the reconstructed images are fed into a re-identification network, which pulls out features that are related to a person's identity. The next sections will explain the details of each phase. The proposed method is represented in Algorithm 1.

3. 1. Histogram Equalization Many surveillance images, particularly those captured at night or in poor weather conditions, suffer from low brightness. This diminishes the contrast, complicating the visibility of essential detail such as edges and facial or bodily features. To alleviate these impacts, the initial stage is the application of a histogram equalization approach within the HSV color space. This color space displays the parts of Hue, Saturation, and Value, making it easy to change the brightness of an image directly and on its own without changing its color. In the proposed method color images are first converted from the RGB color space to the HSV color space. Next, the average brightness of the V channel is found. If the average brightness of the V dark regions. After that, the enhanced images are converted back to the RGB color space, and it is ready Channel is equal to or higher than a set threshold, the original images are sent to the next stage without any changes. Otherwise the images are considered dark, and histogram equalization is applied to enhance details in the for the next step, which is the super-resolution network. Figure 2 shows that the overall brightness of a very dark image can be improved without changing its colors. Histogram equalization alone, on the other hand, can not make images sharper or bring out fine details; it just makes the input better for the next steps in processing.

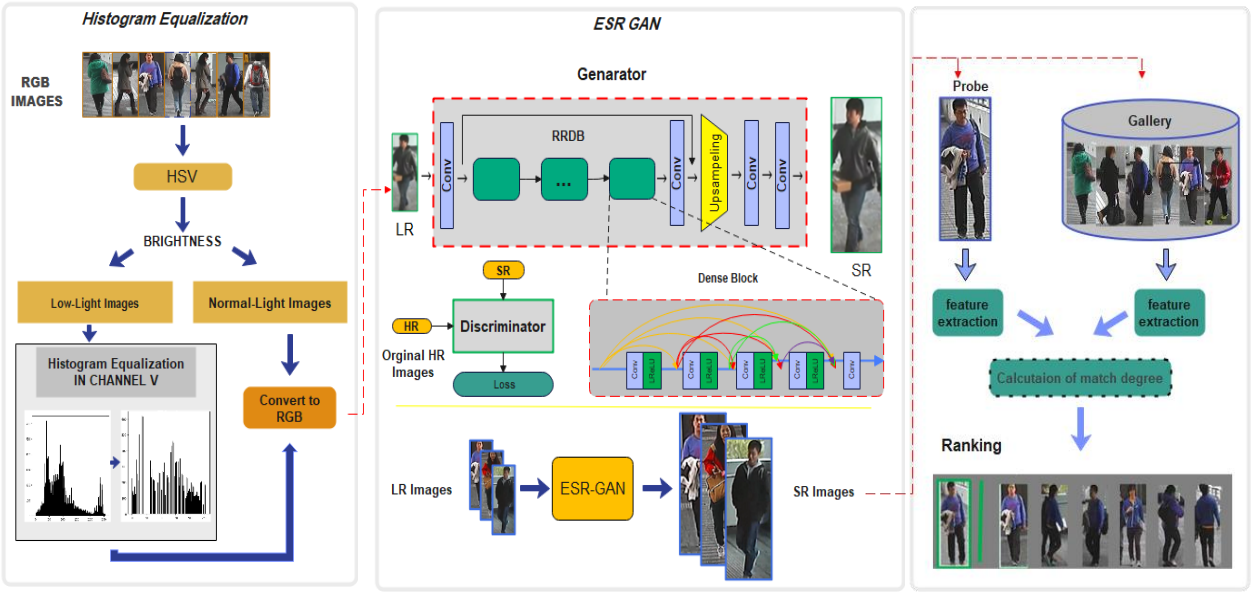


Figure 1. An illustration of the proposed network is presented, which consists of histogram equalization, a super-resolution network (ESRGAN), and a person re-identification network (ReID Network)

Algorithm 1. Algorithm of the proposed method

Step 1 – Histogram Equalization
Input: RGB image
Enhanced image

1. 1.Convert input images from RGB → HSV.
2. Compute the average brightness $\bar{V}$ from the V channel.
3. If $\bar{V}$ is small
   Apply histogram equalization on the V channel
   Reconstruct HSV (H, S, V′)
   Convert HSV back to RGB
   Else:
   Convert HSV back to RGB (without changes)

Step 2 – Super-Resolution Enhancement (ESRGAN):
Input: Enhanced image
Output: Super-resolved image

1. Training procedure (for each iteration):
2. Generating an SR image from an LR image
3. Compute perceptual loss
4. Compute L1 loss (SR vs HR)
5. Compute discriminator output on HR and SR (relativistic GAN)
6. Updating Discriminator
7. 1) Compute discriminator loss L_{Dis}^{Ra}
   2) Update parameters
8. Updating Generator
   1) Compute generator loss
   $L_{Gen} = \gamma_1 L_1 + \gamma_2 L_{perceptual} + \gamma_3 L_{Gen}^{Ra}$
   2) Update parameters
9. After training, the generator produces a high-resolution image from the enhanced input.

Step 3 –Re-Identification
Input: Super-resolved image
Output: Recognition result

1. Feed the SR image into the Re-ID network to extract a feature vector from the image
2. Compute similarity score
3. Rank gallery images based on similarity scores
4. Re-Identification result

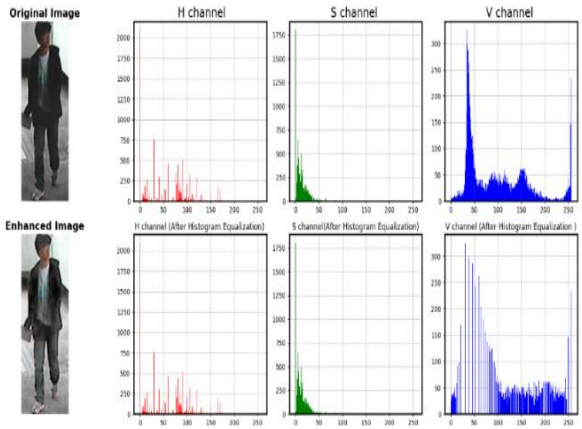


Figure 2. The original RGB image (top image) and the enhanced version after applying histogram equalization on the Value channel in the HSV color space (bottom image)

3. 2. Super Resolution Network The next step after enhancing the image brightness is to use ESRGAN, a deep super-resolution network based on GAN to make the images clearer. ESRGAN is one of the best ways to super-resolve pictures. The structure of GAN is what this model is based on. The three main parts of ESRGAN are the generator, the discriminator, and the loss function. The generator produces high-resolution images with fine details, and the discriminator tells the difference between the original high-resolution pictures and the ones made by the generator. The loss function is also composed of three parts: pixel-wise loss, perceptual loss, and relativistic GAN loss.

ESRGAN utilizes a relativistic version of the GAN called RaGAN (Relativistic Average GAN) (35). This method uses relative discrimination to tell the difference between real and fake pictures. The discriminator doesn't make a clear decision about whether or not an image is real. Instead, it looks at how real the real image looks compared to the fake one. This function makes the training process more stable and the results better. The perceptual loss is calculated using the feature map of the image that the generator made and that was taken out before the final activation function. This loss makes the final output look more like real photos in terms of how they look and how many small details there are (39). After that, the pixel-wise error is calculated, and the resulting image is sent to the discriminator network. During training, the discriminator is also given the original image, which helps it tell the difference between real and fake images. The generator then changes its adversarial loss and makes a new version of the same image. The process goes on until the discriminator can no longer tell the difference between real and fake images. The total loss of the network is defined as follows:

$$L_{totall} = L_{Gen} + L_{Dis}^{Ra} \quad (1)$$

where L_{Gen} is the loss of the generator, L_{Dis}^{Ra} is the loss of the relativistic discriminator, and the generator loss is the sum of the next three loss functions:

$$L_{Gen} = \gamma_1 L_1 + \gamma_2 L_{perceptual} + \gamma_3 L_{Gen}^{Ra} \quad (2)$$

γ_1 , γ_2 and γ_3 are weights for different loss types, which control the influence of each loss and are chosen approximately. $L_{perceptual}$ is the perceptual loss, L_{Gen}^{Ra} is the relativistic generator loss, and L_1 is the pixel-wise loss.

3. 2. 1. Pixel-Wise Loss The network uses a pixel-wise loss function to enhance the pixel-level accuracy of the generated image. This helps ensure that the high-resolution image I_{SR} closely resembles the real image I_{HR} in terms of pixel values. To improve performance and convergence the loss function L_1 is used, as the standard loss function L_2 often leads to overly smooth results. L_1 loss is is computed as follows:

$$L_1 = \sum_{x}^{W} \sum_{y}^{H} \| I_{SR}(x, y) - I_{HR}(x, y) \| \quad (3)$$

where W and H are the image's width and height, respectively. This equation finds the L1 distance between the pixels of the high-resolution image that was made and the real image.

3. 2. 2. Perceptual Loss Perceptual loss functions by contrasting the feature representations of the original and generated images, hence enhancing the preservation of image quality and fine details. Feature maps are derived from the generator's output prior to the final activation function, retaining high-level semantic information for the calculation of perceptual loss. This method enhances the correlation between the features of the input and the reconstructed images, leading to enhanced reconstruction quality and increased overall model performance (26). The feature maps of the high-resolution original image I_{HR} and the image generated I_{SR} by the generator are compared utilizing the Euclidean distance, and the perceptual loss is calculated based on the following equation:

$$L_{perceptual} = \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I_{HR})_{xy} - \phi_{i,j}(I_{SR})_{xy})^2 \quad (4)$$

The feature map $\phi_{i,j}$ comes from the j-th convolutional layer and the i before the i-th max-pooling layer.

3. 2 .3. Relativistic Gan Loss The relativistic adversarial generator network estimates the probability of distinguishing between real data and generated data by calculating the following distance:

$$D_{Ra}(I_{HR}, I_{SR}) \quad (5)$$

The Relativistic Generative Adversarial Network (RGAN) generates images of individuals with enhanced sharpness and superior visual and frequency details relative to the conventional Generative Adversarial Network. This principle is elucidated in Equations 6 and 7:

$$D_{Ra}(\mathrm{Real}, Fake) = C(\mathrm{Real}) - E(C(Fake)) \rightarrow 1 \quad (6)$$

$$D_{Ra}(Fake, \mathrm{Real}) = C(Fake) - E(C(\mathrm{Real})) \rightarrow 0 \quad (7)$$

These two equations show how much more realistic the real image is compared to the fake image and how much less realistic the fake image is compared to the real image. In this case, $E(\cdot)$ is the average of real or fake data in a batch and is the output of the discriminator. This small change makes the model work better than the old discriminator network. Equations 8 and 9 show, respectively, the loss function for the discriminator network and the adversarial loss for the relativistic generative network:

$$L_{Dis}^{Ra} = -E_{I_{HR}}[\log(D_{Ra}(I_{HR}, I_{SR}))] - E_{SR}[\log(1 - D_{Ra}(I_{SR}, I_{HR}))] \quad (8)$$

$$L_{Gen}^{Ra} = -E_{I_{HR}}[\log(D_{Ra}(I_{HR}, I_{SR})] - E_{SR}[\log(D_{Ra}(I_{SR}, I_{HR}))] \quad (9)$$

The relativistic generative adversarial network is a big step up from the standard GAN because it uses a relativistic loss function. This method lets the discriminator compare real and fake images instead of classifying them. In contrast to the conventional GAN, which ceases to learn from real data upon reaching an optimal state, RGAN persists in learning from both real

and synthetic data, as its gradients are influenced by both categories. Furthermore, RGAN produces images characterized by sharper edges, enhanced visual details, and more detailed frequency distribution. Utilizing relative comparison rather than absolute classification enhances the stability of the training process and mitigates problems such as mode collapse.

3. 3. Super Resolution Network This study uses a pre-trained CNN-based network to re-identify persons. Using CNN as an appropriate deep learning algorithm enables the extraction of meaningful and high-level image features (40). The preprocessed high-resolution images are fed into a CNN utilizing the VGGNet architecture to extract unique feature vectors. These vectors are compared to each other using the Euclidean distance metric. If the distance between two vectors is minimal, the images are presumed to depict the same individual. Otherwise, each image is perceived as depicting a distinct individual. In the ranking stage, the feature vector of the probing image is compared with all gallery images, and individuals are scored according to their similarity. The nearest match is the one with the least distance. The quality of the extracted feature vectors and the improvements made to the preprocessing stages will determine how accurate this stage is. Using a pre-trained network not only saves time on the computer, but it also makes it easier to get to the immense amounts of training data that were used to train the model. Because of this, it works well in the real world (17).

4. EXPERIMENTAL RESULTS

The experiments and evaluations in this study were conducted on two datasets: CUHK01 (41) and CUHK03 (42). The CUHK01 dataset is one of the well-known and widely ued datasets in the field of Person Re-Identification. In this dataset, photos of people are taken from many angles and in different situations. Two different cameras took the pictures, which show 971 different people. Each person is assigned four images, with each image having a resolution of 60 × 160 pixels.

There are 14,097 photos of 1,467 people in the CUHK03 dataset. These were taken with six different cameras to record individuals from various angles and under different conditions. Figure 3 illustrates some sample frames of these datasets.

To assess the effectiveness of the proposed method, its performance has been compared with several well-known person re-identification methods. The two main ways for evaluating the accuracy of a person re-identification system are the rank-based measure and the Cumulative Matching Characteristic (CMC) curve (43).

The Rank-r measure tells us how likely it is that the correct image that matches the query is one of the top r results that the algorithm gives you (44). It is very important to know the Rank-1 accuracy because it shows how well the system can find the right person as the first result. When matching is more flexible, higher rank values like Rank-5 and Rank-10 are used to rate performance.

We have comprehensively evaluated the re-identification accuracy for each dataset by showing results at four different rank levels: Rank-1, Rank-5, Rank-10, and Rank-20. Tables 1 and 2 show the results of the experiments. Although the proposed method achieved lower accuracy in the Rank-5 metric compared to some existing methods, it achieves higher accuracy in Rank-1, which holds greater importance in practical applications. For example, the Rank-1 accuracy on the CUHK01 dataset was 93.6%, and on the CUHK03 dataset, it was 95.1%.

The CMC curve indicates how well the system performs at different levels of rank. The horizontal axis of this chart shows the rank, and the vertical axis shows

Figure 3. Sample Images of re-id datasets: CUHK01 and CUHK03

TABLE 1. Numerical results and comparison on CUHK01

Methods	Rank1	Rank5	Rank10	Rank20
DCAE (17)	92.1%	-	97.5%	99.3%
PLCA (45)	46.8%	71.8%	80.5%	88.2%
EDPR (p =100) (46)	88.2%	98.20%	99.35%	-
DSA (47)	90.4%	97.8%	-	-
PersonNet (48)	71.14%	90.07%	95.00%	98.06%
DFSN (49)	83.95%	98.15%	98.97%	-
PN-GAN (50)	67.65%	86.64%	91.82%	-
PSD (51)	83.2%	-	97.1%	98.8%
EFS-Net (52)	70.1%	95.2%	99.1%	99.2%
Improved Quartet (53)	93.4%	98.8%	99.2%	-
The proposed model	93.6%	97.2%	99.4%	99.7%

TABLE 2. Numerical results and comparison on CUHK03

Methods	Rank1	Rank5	Rank10	Rank20
AACN (51)	91.39%	98.89%	99.48%	99.75%
PN-GAN (47)	79.76%	96.24%	98.56%	88.2%
PSD (48)	91.8%	-	99.1%	99.6%
Improved Quartet (53)	89.8%	99.4%	99.8%	-
DCAE(16)	94.4%	-	99.1%	99.65%
DFSN (46)	85.5%	98.74%	99.8%	-
SSM (54)	76.63%	94.59%	97.95%	-
BTL (28)	75.53%	95.15%	99.16%	-
Spindle Net (32)	88.5%	97.8%	98.6%	99.2%
CAN (33)	72.3%	93.8%	98.4%	99.2%
The proposed model	94.5%	96.3%	99.19%	99.75%

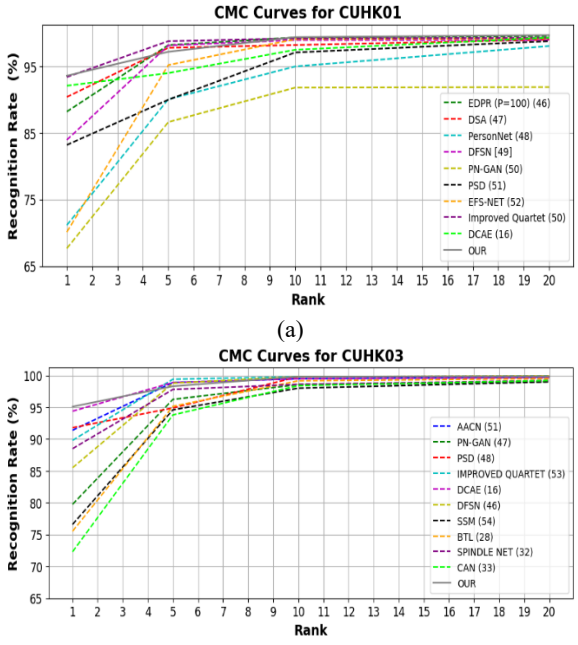
how well the system was able to find the right match up to that rank. In other words, this curve represents how likely it is that the correct identity is among the top r results that were calculated as follows:

$$CMC(r) = \frac{1}{N}\sum_{n=1}^{N} E(p_n \le r) \quad (10)$$

$$E(P_n \le r) = \begin{cases} 1 & P_n \le R \\ 0 & otherwise \end{cases} \quad (11)$$

In this equation, P_n stands for the position of the nth person on the ranked list, and N stands for the total number of probe IDs in the dataset. The CMC curve is a well-known and often-used statistic in many studies on person re-identification (46, 55, 56). This curve shows how well a model works overall by showing how accurate it is at all rank positions, not just one. Figure 4 shows the CMC curves for the CUHK01 and CUHK03 datasets.

Our experimental results indicate that combining the histogram method with ESRGAN improves person re-identification accuracy on these two datasets. Histogram equalization enhances brightness and contrast, revealing details hidden in dark areas and preparing the input for ESRGAN, allowing the network to concentrate on the real and clear information instead of on dark or low-light areas. Unlike many traditional super-resolution methods that add noise or edges that don't look natural, ESRGAN, which uses a strong discriminator and an improved perceptual loss function, preserves fine details and edges while generating more realistic and higher-quality images (26). The better quality of reconstruction helps get descriptive features out, which boosts the accuracy of person re-identification systems. Figure 5 shows examples of the initial low-resolution images and their corresponding high-resolution reconstructed versions.



(a)

(b)

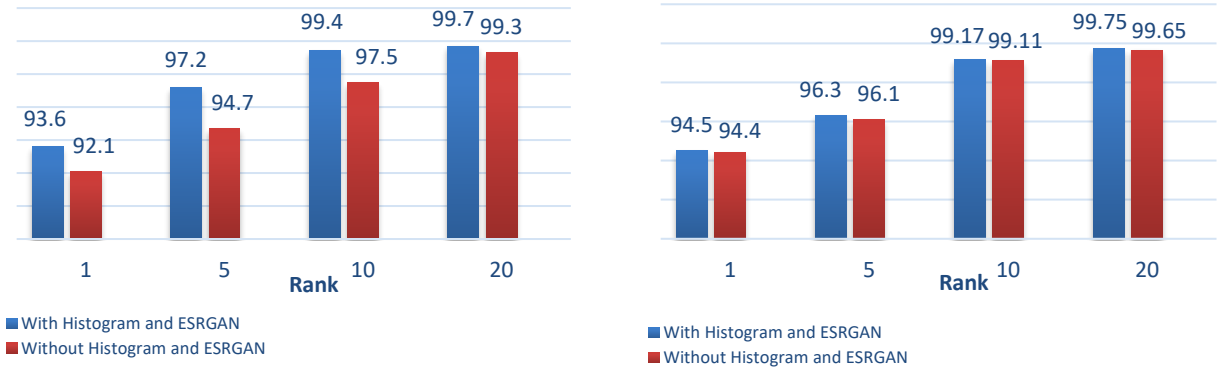
Figure 4. CMC Curve for CUHK01 and CUHK03

(a)

(b)

Figure 5. Examples of images with (a) low resolution that have been converted to (b) high-resolution images

Additionally, Figure 6 shows the system's performance in two situations: without and with the use of histogram equalization and ESRGAN. These charts clearly demonstrate that applying the proposed method, particularly at lower ranks such as Rank-1 and Rank-5, improves person re-identification accuracy. These results highlight the importance of enhancing image resolution, extracting more precise descriptive features, and improving system performance in real-world conditions.



(a) (b)

Figure 6. Performance comparison under two conditions, with and without the application of histogram equalization and ESRGAN for (a) CUHK01 and (b) CUHK03 datasets

5. CONCLUSION

In this paper, a framework for person re-identification with the objective of solving the problem of person recognition with low resolution is proposed. Initially, histogram equalization techniques based on the Human Visual System (HVS) have been used to enhance image brightness and contrast. This method, by considering the human eye's sensitivity to intensity variations and details, enhances image perceptibility and improves the clarity of edges and important structures. After this step, the images were input into the ESRGAN network, an advanced GAN-based network that demonstrates remarkable performance in enhancing image resolutions.

And with its advanced architecture, it effectively reconstructs textures and precise structural components, especially edges. As a result, the re-identification network utilized in the final stage is able to successfully extract superior and more distinctive visual features. Comprehensive studies performed using prevalent datasets indicate that the proposed method significantly improves Rank-1 recognition accuracy and demonstrates superior performance compared to other existing approaches.

Acknowledgements

The author appreciates University of Birjand for providing necessary facilities to conduct present research.

Funding

This work has no financial funding.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable. Competing interests the author declares no financial or organizational conflicts of interest.

Data Availability

The data of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors declare that they did not use any AI-assistant technology in writing process.

Author Biosketch

Fatemeh Ilbeigi is currently a Ph.D. student at the University of Birjand, Birjand, Iran. Her research interests include digital image and audio signal processing, deep learning, and artificial intelligence. Her Email is: fatemeh.ilbeigi@birjand.ac.ir

Hassan Farsi is a Professor of Communication Engineering in the Department of Electrical and Computer Engineering at the university of Birjand, Birjand, Iran. He received his Ph.D. from the Centre for Communication Systems Research (CCSR), University of Surrey, Guildford, UK. His research interests include speech, image, and video processing in wireless communications. His Emil is: hfarsi@birjand.ac.ir

Sajad Mohamadzadeh received his Ph.D. in Telecommunication Engineering from the University of Birjand, Birjand, Iran. He is currently an Associate Professor in Faculty of Technical and Engineering of Ferdows, university of Birjand, Birjand, Iran. His research interests include image processing and retrieval, pattern recognition, digital signal processing, and sparse representation. His email is: s.mohamadzadeh@birjand.ac.ir

Amir Sezavar received his Ph.D. degree in Telecommunication Engineering from the University of Birjand, Birjand, Iran. He is currently an Assistant Professor at the Faculty of Technical and Engineering of Ferdows, University of Birjand, Birjand, Iran. His research interests include digital image processing and retrieval, visual signal processing, deep learning, and artificial intelligence. His Email is: a.sezavar@birjand.ac.ir

REFERENCES

1. Zhao K, Cheng D, Kou Q, Li J, Liu R. Sequences consistency feature learning for video-based person re-identification. Electronics Letters. 2022;58(4):142-44. https://doi.org/10.1049/ell2.12382
2. Fu H, Zhang K, Li H, Wang J. MRRNet: Learning multiple region representation for video person re-identification. Engineering Applications of Artificial Intelligence. 2022;114:105108. https://doi.org/10.1016/j.engappai.2022.105108
3. Shahbakhsh MB, Hassanpour H. Empowering Face Recognition Methods using a GAN-based Single Image Super-Resolution Network. International Journal of Engineering, Transactions A: Basics. 2022;35(10):1858-66. https://doi.org/10.5829/ije.2022.35.10a.05
4. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Wang S, Sun J. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019:393-402. https://doi.org/10.1109/CVPR.2019.00048
5. Wang Z, Wang Z, Zheng Y, Wu Y, Zeng W, Satoh S. Beyond intra-modality: A survey of heterogeneous person re-identification. IJCAI International Joint Conference on Artificial Intelligence. 2020;2021:4973-80. https://doi.org/10.24963/ijcai.2020/692
6. Yang Q, Yu HX, Wu A, Zheng WS. Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019:3628-37. https://doi.org/10.1109/CVPR.2019.00375
7. Wang H, Du H, Zhao Y, Yan J. A Comprehensive Overview of Person Re-Identification Approaches. IEEE Access. 2020;8:45556-83. https://doi.org/10.1109/ACCESS.2020.2978344
8. Munjal B, Amin S, Tombari F, Galasso F. Query-guided end-to-end person search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:811-20. https://doi.org/10.48550/arXiv.1905.01203
9. Deng W, Zheng L, Ye Q, Kang G, Yang Y, Jiao J. Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018:994-1003. https://doi.org/10.1109/CVPR.2018.00110
10. Mao S, Zhang S, Yang M. Resolution-invariant person re-identification. arXiv preprint arXiv:1906.09748. 2019. https://doi.org/10.48550/arXiv.1906.09748
11. Chen YC, Huang PH, Yu LY, Huang JB, Yang MH, Lin YY. Deep Semantic Matching with Foreground Detection and Cycle-Consistency. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019;11363:347-62. https://doi.org/10.1007/978-3-030-20893-6_22
12. Jing XY, Zhu X, Wu F, Hu R, You X, Wang Y, Feng H, Yang JY. Super-Resolution Person Re-Identification with Semi-Coupled Low-Rank Discriminant Dictionary Learning. IEEE Transactions on Image Processing. 2017;26:1363-78. https://doi.org/10.1109/TIP.2017.2651364
13. Li X, Zheng WS, Wang X, Xiang T, Gong S. Multi-scale learning for low-resolution person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. 2015:3765-73. https://doi.org/10.1109/ICCV.2015.429
14. Imani M. Deep Learning Based Electricity Demand Forecasting in Different Domains. Iranian Journal of Energy and Environment. 2020;11(1):33-39. https://doi.org/10.5829/ijee.2020.11.01.06
15. Khodadadi Z, Owlia MS, Amiri A. Enhancing Fault Detection in Image Analysis: A Combined Wavelet-Fourier Technique for Advancing Manufacturing Quality Control. International Journal of Engineering, Transactions A: Basics. 2024;37(2):387-401. https://doi.org/10.5829/ije.2024.37.02b.14
16. Mortezaie Z, Amiri Z, Mortezaee M. Adaptive Fractional-order Differentiation for Enhanced Image Contrast Utilizing Caputo Masks. International Journal of Engineering, Transactions B: Applications. 2026;39(5):1275-92. https://doi.org/10.5829/ije.2026.39.05b.19
17. Sezavar A, Farsi H, Mohamadzadeh S. A New Model for Person Reidentification Using Deep CNN and Autoencoders. Iranian Journal of Energy and Environment. 2023;14(4):314-20. https://doi.org/10.5829/ijee.2023.14.04.01
18. Al-Kaabi HAH, Aghagolzadeh A, Mikaeeli E. Super-resolution Augmentation of Large Images with Architecture Treatment of Deeply Recursive Convolutional Network. International Journal of Engineering, Transactions B: Applications. 2025;38(5):1042-55. https://doi.org/10.5829/ije.2025.38.05b.08
19. Zhao J, Xiong R, Xu J, Wu F, Huang T. Learning a deep convolutional network for subband image denoising. In: Proceedings - IEEE International Conference on Multimedia and Expo. 2019:1420-25. https://doi.org/10.1109/ICME.2019.00246
20. Wang Y, Wang L, Wang H, Li P, Lu H. Blind single image super-resolution with a mixture of deep networks. Pattern Recognition. 2020;102:107169. https://doi.org/10.1016/j.patcog.2019.107169
21. Jiang J, Yu Y, Wang Z, Tang S, Hu R, Ma J. Ensemble Super-Resolution with a Reference Dataset. IEEE Transactions on

Cybernetics. 2020;50(11):4694-708. https://doi.org/10.1109/TCYB.2018.2890149

22. Wang L, Huang Z, Gong Y, Pan C. Ensemble based deep networks for image super-resolution. Pattern Recognition. 2017;68:191-98. https://doi.org/10.1016/j.patcog.2017.02.027
23. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, Shi W. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017:105-14. https://doi.org/10.1109/CVPR.2017.19
24. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Advances in neural information processing systems. Curran Associates, Inc. 2014;36:2672-80. https://doi.org/10.1016/s0925-2312(97)90018-7
25. Zhao L, Bai H, Liang J, Zeng B, Wang A, Zhao Y. Simultaneous color-depth super-resolution with conditional generative adversarial networks. Pattern Recognition. 2019;88:356-69. https://doi.org/10.1016/j.patcog.2018.11.028
26. Wang X, Yu K, Wu S, Gu J, Liu Y, Dong C, Qiao Y, Change Loy C. ESRGAN: Enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. 2018. https://doi.org/10.48550/arXiv.1809.00219
27. Shamsolmoali P, Zareapoor M, Wang R, Jain DK, Yang J. G-GANISR: Gradual generative adversarial network for image super resolution. Neurocomputing. 2019;366:140-53. https://doi.org/10.1016/j.neucom.2019.07.094
28. Dorrani Z, Farsi H, Mohamadzadeh S. Shadow Removal in Vehicle Detection Using ResUNet-a. Iranian Journal of Energy and Environment. 2023;14(1):87-95. https://doi.org/10.5829/ijee.2023.14.01.11
29. Sahragard E, Farsi H, Mohamadzadeh S. A Deep Learning-based Approach for Accurate Semantic Segmentation with Attention Modules. Iranica Journal of Energy and Environment. 2025;16(4):692-705. https://doi.org/10.5829/ijee.2025.16.04.12
30. Sezavar A, Farsi H, Mohamadzadeh S, Radeva P. A new person re-identification method by defining CNN-based feature extractor and sparse representation. Multimedia Tools and Applications. 2024;83(4):11043-59. https://doi.org/10.1007/s11042-023-15718-x
31. Chen W, Chen X, Zhang J, Huang K. Beyond triplet loss: A deep quadruplet network for person re-identification. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017:1320-29. https://doi.org/10.1109/CVPR.2017.145
32. Zhao H, Tian M, Sun S, Shao J, Yan J, Yi S, Wang X, Tang X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017:907-15. https://doi.org/10.1109/CVPR.2017.103
33. Liu H, Feng J, Qi M, Jiang J, Yan S. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing. 2017;26(7):3492-3506. https://doi.org/10.1109/TIP.2017.2700762
34. Jiao J, Zheng WS, Wu A, Zhu X, Gong S. Deep Low-Resolution Person Re-Identification. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1):6967-74. https://doi.org/10.1609/aaai.v32i1.12284
35. Wang Z, Ye M, Yang F, Bai X, Satoh S. Cascaded SR-GAN for scale-adaptive low resolution person re-identification. In: IJCAI International Joint Conference on Artificial Intelligence. 2018:3891-97. https://doi.org/10.24963/ijcai.2018/541
36. Cheng Z, Dong Q, Gong S, Zhu X. Inter-Task Association Critic for Cross-Resolution Person Re-Identification. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:2602-12. https://doi.org/10.1109/CVPR42600.2020.00268
37. Chen YC, Li YJ, Du X, Wang YCF. Learning resolution-invariant deep representations for person re-identification. In: 33rd AAAI Conference on Artificial Intelligence, AAAI 2019. 2019;33:8215-22. https://doi.org/10.1609/aaai.v33i01.33018215
38. Li YJ, Chen YC, Lin YY, Du X, Wang YCF. Recover and identify: A generative dual model for cross-resolution person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. 2019:8089-98. https://doi.org/10.1109/ICCV.2019.00818
39. Adil M, Mamoon S, Zakir A, Manzoor MA, Lian Z. Multi scale-adaptive super-resolution person re-identification using GAN. IEEE Access. 2020;8:177351-62. https://doi.org/10.1109/ACCESS.2020.3023594
40. Sezavar A, Farsi H, Mohamadzadeh S. A modified grasshopper optimization algorithm combined with CNN for content based image retrieval. International Journal of Engineering, Transactions A: Basics. 2019;32(7):924-30. https://doi.org/10.5829/ije.2019.32.07a.04
41. Li W, Zhao R, Wang X. Human Reidentification with Transferred Metric Learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2013;7724:31-44. https://doi.org/10.1007/978-3-642-37331-2_3
42. Li W, Zhao R, Xiao T, Wang X. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. 2014:152-59. https://doi.org/10.1109/CVPR.2014.27
43. Dong H, Lu P, Zhong S, Liu C, Ji Y, Gong S. Person re-identification by enhanced local maximal occurrence representation and generalized similarity metric learning. Neurocomputing. 2018;307:25-37. https://doi.org/10.1016/j.neucom.2018.04.013
44. Hassanpour H, Mortezaie Z, Beghdadi A. Sensing Image Regions for Enhancing Accuracy in People Re-identification. Iranian Journal of Energy and Environment. 2022;13(3):295-304. https://doi.org/10.5829/ijee.2022.13.03.09
45. Matsukawa T, Suzuki E. Person re-identification using CNN features learned from combination of attributes. In: 2016 23rd International Conference on Pattern Recognition (ICPR). 2016:2428-33. https://doi.org/10.1109/ICPR.2016.7900000
46. Guo Y, Cheung NM. Efficient and Deep Person Re-identification Using Multi-level Similarity. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018:2335-44. https://doi.org/10.1109/CVPR.2018.00248
47. Zhang Z, Lan C, Zeng W, Chen Z. Densely semantically aligned person re-identification. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019:667-76. https://doi.org/10.1109/CVPR.2019.00076
48. Wu L, Shen C, van den Hengel A. Personnet: Person re-identification with deep convolutional neural networks. arXiv preprint arXiv:1601.07255. 2016. https://doi.org/10.48550/arXiv.1601.07255
49. Khatun A, Denman S, Sridharan S, Fookes C. A Deep Four-Stream Siamese Convolutional Neural Network with Joint Verification and Identification Loss for Person Re-Detection. In: Proceedings - 2018 IEEE Winter Conference on Applications of

Computer Vision, WACV 2018. 2018:1292-301. https://doi.org/10.1109/WACV.2018.00146

50. Qian X, Fu Y, Xiang T, Wang W, Qiu J, Wu Y, Jiang YG, Xue X. Pose-normalized image generation for person re-identification. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019;11213:661-78. https://doi.org/10.1007/978-3-030-01240-3_40
51. Liu Y, Zhang Y, Coleman S, Bhanu B, Liu S. A new patch selection method based on parsing and saliency detection for person re-identification. Neurocomputing. 2020;374:86-99. https://doi.org/10.1016/j.neucom.2019.09.073
52. Sheshkal SA, Fouladi-Ghaleh K, Aghababa H. An Improved Person Re-identification Method by light-weight convolutional neural network. In: 2020 10th International Conference on Computer and Knowledge Engineering, ICCKE 2020. 2020:463-68. https://doi.org/10.1109/ICCKE50421.2020.9303722
53. Khatun A, Denman S, Sridharan S, Fookes C. Joint identification–verification for person re-identification: A four stream deep learning approach with improved quartet loss function. Computer Vision and Image Understanding. 2020;197-8:102989. https://doi.org/10.1016/j.cviu.2020.102989
54. Bai S, Bai X, Tian Q. Scalable person re-identification on supervised smoothed manifold. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. 2017:3356-65. https://doi.org/10.1109/CVPR.2017.358
55. Al-Kaabi, H. A. H., Aghagolzadeh, A., Mikaeeli, E. Super-resolution Augmentation of Large Images with Architecture Treatment of Deeply Recursive Convolutional Network. International Journal of Engineering, Transactions B: Applications. 2025;38(5):1042-55. https://doi.org/10.5829/ije.2025.38.05b.08
56. Dey, A., Biswas, S., Abualigah, L. Efficient Violence Recognition in Video Streams using ResDLCNN-GRU Attention Network. ECTI Transactions on Computer and Information Technology. 2024;18(3):329-41. https://doi.org/10.37936/ecti-cit.2024183.255679

Persian Abstract

چکیده

بازشناسی افراد یکی از مسائل مهم در زمینه‌پردازش تصویر و بینایی ماشین است که در سال‌های اخیر توجه زیادی را به خود جلب کرده است. یکی از چالش‌های اصلی در این حوزه، کیفیت پایین تصاویر ثبت‌شده توسط دوربین‌های مختلف است که می‌تواند منجر به از دست‌رفتن جزئیات مهم و در نتیجه کاهش دقت الگوریتم‌های بازشناسی می گردد. در این مقاله، چارچوبی برای بهبود دقت بازشناسی افراد پیشنهاد شده است که با بهبود روشنایی تصاویر کم‌نور از طریق تکنیک‌های همسان‌سازی هیستوگرام و افزایش وضوح آن‌ها با استفاده از شبکه‌ی ESRGAN ، دقت و توانایی شبکه را در بازشناسی افراد ارتقا می‌دهد. نتایج تجربی نشان می‌دهد که روش پیشنهادی با بهبود روشنایی و وضوح تصاویر که نقشی کلیدی در آشکارسازی جزئیات مهم دارند، کیفیت ورودی شبکه‌ی بازشناسی را افزایش داده و در مقایسه با روش‌های موجود، دقت بازشناسی افراد را ارتقا می‌دهد، به‌طوری‌که به‌دقت Rank-1 معادل ٪۹۳.٦ بر روی پایگاه‌داده CUHK01 و ٪۹٤.٥ بر روی پایگاه‌داده CUHK03 دست‌یافته است.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2472-2484

International Journal of Engineering

Journal Homepage: www.ije.ir

Justification of the Power-performance Characteristics of a Two-section Positive Displacement Motor with a Bi-rotary Mechanism

M. V. Dvoynikov, D. O. Morozov*

Well Drilling Department, Oil and Gas Faculty, Saint Petersburg Mining University, Saint Petersburg, Russia

PAPER INFO

Paper history:
Received 08 September 2025
Received in revised form 16 October 2025
Accepted 26 December 2025

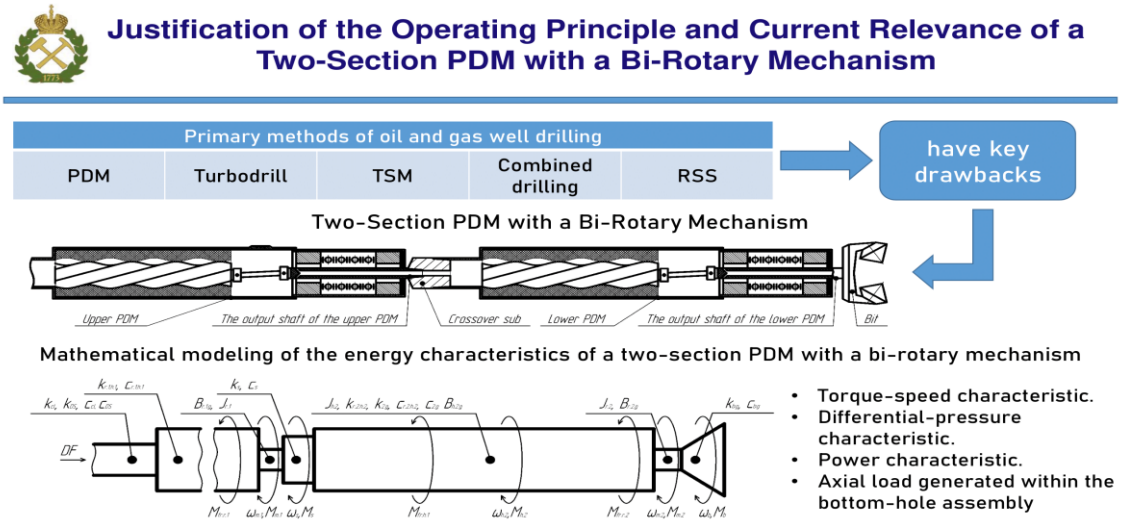
Keywords:
Drilling
Directional Drilling
Positive Displacement Motor
Bi-rotary Mechanism
Power-performance Characteristics
Working Elements

ABSTRACT

For drilling oil and gas wells, screw downhole motors – also known as positive displacement motors (PDMs) – are widely used both in Russia and abroad, accounting for over 80% of applications. However, PDMs exhibit a number of drawbacks affecting the efficiency of directional and overall drilling. First, harmful dynamic phenomena are generated in the motor's power section, which adversely impact drilling performance. Second, the low output shaft speed prevents the use of certain diamond/PDC bits. Third, the ratio of horizontal reach to true vertical depth almost never exceeds 4.0, whereas with rotary steerable systems (RSS) it can reach 6.0. At present, both motorized and non-motorized RSS are being used with increasing frequency to drill wells with greater departure from vertical and precisely prescribed trajectory profiles at maximum step-out. It is known that each of these and other solutions has its own advantages and limitations that are difficult, and sometimes impossible, to eliminate during drilling operations. A drilling technology employing a two-section PDM with a bi-rotary mechanism is proposed to increase the efficiency of directional drilling. This work presents prototypes of the developed motor, identifies its characteristic similarities and differences relative to earlier tools and technologies, and determines the power-performance characteristics of the two-section PDM – namely torque, rotational speed, power at the bit, and the axial load generated within the bottomhole assembly. The feasibility of the regulating motor operation by adjusting the drilling fluid (DF) flow rate to the power section is substantiated. Potential topics for future research are outlined.

doi: *10.5829/ije.2026.39.10a.10*

Graphical Abstract



*Corresponding Author Email: dmitrij_morozov_01@mail.ru (D. O. Morozov)

Please cite this article as: Dvoynikov MV, Morozov DO. Justification of the Power-performance Characteristics of a Two-section Positive Displacement Motor with a Bi-rotary Mechanism. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2472-84.

NOMENCLATURE

k	Torsional stiffness (N·m/rad)	v	Linear velocity (m/s)
z	Number of lobes of the WE	e	Eccentricity (m)
M	Torque characteristic (N·m)	L	Length of the PDM element (m)
P	Pressure drop (differential pressure) (Pa)	m	Rotor mass (kg)
Q	DF flow rate (m³/s)	R	Radius of the PDM element (m)
V	Working volume (m³/s)	N	Power characteristic (W)
E	Energy on the PDM element (J)	Rel	Rayleigh dissipation (N·m)
h	Moment arm of the forces generating the resistance torque (m)	**Greek Symbols**	
T	Stator helical surface pitch (m)	μ	Dynamic viscosity of the drilling fluid (Pa·s)
n	Rotational speed (rev/s)	τ	Output torque of the two-section PDM (N·m)
F	Force characteristic (N)	θ	Twist angle (rad)
c	Relative viscous loss coefficient on a PDM element (N·m·s/rad)	η	Efficiency (%)
B	Viscous loss coefficient on a PDM element (N·m·s/rad)	Λ	Number of contact lines of the working elements
J'	Polar moment of area (m⁴)	ω	Angular velocity (s⁻¹)
J	Moment of inertia (kg·m²)	$\dot{\omega}$	Angular acceleration (m/s²)

1. INTRODUCTION

The use of a PDM to drill long directional intervals while achieving the required curvature is constrained by the inability to deliver sufficient weight on bit (WOB). The operational capability of PDM-based directional drilling is reduced by friction forces acting between the drilling assembly and the formation. To transmit WOB to the bit, it is recommended to employ special tools installed in the drill string (DS) or to add lubricants to the DF. These measures typically increase the achievable length of oriented (sliding) directional drilling by no more than 500-1,000 m (1-3).

A further increase in the horizontal reach is possible only with additional rotation of the DS. However, rotation of the DS – and thus of the bent-housing PDM – precludes intentional trajectory deflection via oriented drilling (4). To improve the efficiency of directional intervals, in addition to the friction-reduction methods outlined above, the authors propose using a dynamic module implemented as the rotating housing of the lower motor in a two-section PDM based on a bi-rotary mechanism. The foregoing underscores the relevance of oriented directional drilling for complex well profiles – a challenge that can be addressed by developing a drilling methodology employing a two-section PDM with a co-rotating rotor-housing lower stage (two-section bi-rotor PDM).

Multi-section motors with rotors connected in series are known (1, 5). Their operating principle differs from that of a bi-rotary mechanism, which introduces specific operational considerations for the proposed motor. The most critical of these is selecting compatible motors so that the assembled two-section PDM not only performs efficiently but also avoids off-design and failure modes – since, as will be shown below, there is a high likelihood of one of the units entering a braking (stall) mode. The efficiency of drilling with a sectional PDM increases due to higher power at the rock-cutting tool (bit), yet the magnitude – and, in some cases, the very presence – of any gain in drilling performance (rate of penetration (ROP)) depends on the prevailing geologic and technical conditions. A known peculiarity of rock failure is that raising rotational speed (revolutions per minute (RPM)) at constant torque characteristics does not necessarily improve ROP. The feasibility of tuning PDM performance by varying DF flow rate through the bit is substantiated and contrasted with tuning via changes in surface DS RPM and applied WOB (6). With a special coupling of two PDMs, the housing of the lower motor functions as a dynamic module that reduces frictional interaction (drag) between the bottomhole assembly (BHA) and the borehole wall, promotes formation of a filter cake on the wellbore walls, and induces turbulence in the DF, thereby enhancing cuttings transport.

The study addresses the determination of drilling parameters for a two-section bi-rotor PDM (7) – namely torque, rotational speed, power, pressure drop, and the axial load generated within the BHA of the two-section PDM. A mathematical model was developed, the torque-speed characteristic of the proposed motor was computed, and a drilling methodology for the two-section bi-rotor PDM is described, neglecting the internally generated oscillations identified in this work.

1. 1. Analysis of Technical Solutions and Drilling Technologies for Directional and Horizontal Wells

The evolution of well construction has been underway for two centuries. The process of building a well – specifically, the technologies used to create a geometrically defined excavation – is governed by its intended purpose. As target depths, lateral reach, and the variety of transcendental curves employed in directional well design have grown, the approaches to selecting bit-drive systems and other equipment, as well as the methods for oriented trajectory deflection, have evolved accordingly (8-10).

To develop a mathematical model of the bo-rotor PDM and to understand its operating process, the design of a single-screw PDM is first considered. A PDM can be

viewed as an epicyclic (planetary-type) reduction mechanism with a stationary stator and a moving rotor – the conceptual opposite of a turbodrill. The helical rotor, having one lobe fewer than the stator, is set into rotation when DF is supplied into the high-pressure cavities.

The rotor and stator remain in continuous contact, and the number of contact lines along the working-element profiles equals the number of stator lobes. However, when the rotor passes through a "dead" point – i.e., when a rotor tooth fully seats in a stator groove – the number of contact lines increases by one, which induces DF pulsations that generate axial vibrations of the PDM as a whole (1, 11). These must be mitigated. The system of equations describing the number of contact lines – and thereby the persistent frictional interaction within the power section – is presented below:

$$\begin{cases} \Lambda_{\min} = (l-1)z_1 + 1 \\ \Lambda_{\max} = (l-1)z_1 + 2 \end{cases} \quad (1)$$

where Λ_{max} – maximum number of contact lines of the working elements (WE) (rotor angle φ equals 0 or $2\pi/z_1$, i.e., a rotor lobe fully seats in a stator groove, the "dead" point); Λ_{min} – minimum number of contact lines ($\varphi \neq 0$, $\varphi \neq 2\pi/z_1$); l – number of WE stages; z_2 and z_1 – numbers of rotor and stator lobes, respectively.

The formulas for determining the energy characteristics of a single-screw PDM are well known and reported extensively in Russian and international literature (1, 12, 13).

In this paper, authors present the most pertinent ones – those used to determine the torque and the rotational speed of the output shaft. This is necessary because the two-section bi-rotor PDM comprises two single-screw PDM, as discussed below. The torque of a PDM can be expressed in various forms. Below are two equivalent equations for determining the motor torque (1):

$$\begin{cases} M_m = \dfrac{P_m V_m}{2\pi}\eta_m \\ M_m = M_0 P_m e^2 T \end{cases} \quad (2)$$

where P_m – pressure drop across the PDM, Pa; V_m – PDM displacement (volumetric throughput), m³/s; η_m – volumetric efficiency of the PDM, dimensionless; e – the eccentricity, m; T – stator helix (lead) pitch, m; M_0 – specific torque corresponding to a unit-dimension helical gerotor mechanism (HGM) with (e, T), N/(Pa·m).

The quantity M_0 is defined as follows:

$$M_0 = z_2 \cdot (z_2 - 1 + c_e) \quad (3)$$

where c_e – tooth shape factor, dimensionless.

The rotational speed of a PDM at a given DF flow rate can be expressed as follows:

$$n = \frac{Q_m}{2\pi e^2 T}\omega_0 \quad (4)$$

where Q_m – volumetric flow rate of DF through the PDM WE (power section), m³/s; ω_0 – specific angular velocity, Pa·m/N.

The parameter ω_0 is defined by:

$$\omega_0 = \frac{1}{M_0} \quad (5)$$

From the equations it follows that the specific torque M_0 depends only on the design characteristics of the WE. Meanwhile, as seen from Equations 2-5, at a constant DF flow rate the torque and the rotational speed are independent. This design feature of the PDM underpins its key advantages:

1) Varying the WE lead (number of starts) makes it possible to adjust bit RPM and flow rate, as well as the inherently stiff torque-load characteristic of the PDM.

2) By changing the length of the power section, one can tune the primary drilling modes and the steerability of the BHA.

3) The sealing of the helical gerotor mechanism (HGM) cavities at the WE interface provides relatively effective surface control of bit load as a function of the motor pressure drop.

Figure 1 shows a typical PDM characteristic curve of rotational speed and linearized pressure drop across the PDM as functions of the developed torque. The four principal operating regimes are: no-load (n = max, M = 0), optimal (η = max), extreme (N = max) and braking/stall (n = 0; M = max). The most hazardous regime is stall (the red point «1» in Figure 1): once the motor enters this mode, the PDM rotor ceases to rotate, is forced against the stator liner and bends, while DF bypasses freely between the rotor and stator. As a result, the stator elastomer wears rapidly and the overall efficiency of the PDM deteriorates (14).

A combined drilling mode (rotating the DS while the PDM is running) and a motorized RSS (including both the PDM and the RSS in the BHA) are of particular interest, because in these modes a dynamic module is formed – namely, the rotating housing of the PDM or the RSS. This phenomenon is equivalent to the bi-rotary mechanism discussed below.

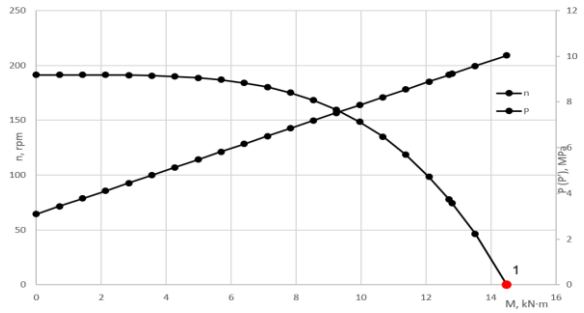


Figure 1. PDM DGR-172.5/6.61 performance curve (Q_m = 31,9 m³/s)

The combined (motor-rotary) method is a drilling technique that uses a hydraulic downhole motor (mud motor, PDM) together with surface rotation of the DS. When a PDM is used, its housing rotates.

However, due to the high stress-strain state of the drillpipe and the inability to build angle, this drilling method is not an optimal solution for well construction under a range of geologic and technical conditions.

Even so, the combined method remains relevant because it enables azimuthal orientation of the RSS and reduces frictional interaction between the DS and the borehole wall (as low as 5 rpm is sufficient), thereby improving the efficiency of directional drilling in wells with a large step-out (departure from vertical) (15).

Externally, the motor then resembles a screw-type reducer. Baldenko and Baldenko (16) referred to this mechanism as "bi-rotary". For this process to occur, the following equality must be satisfied:

$$\omega_{DC} = z_2 \omega_r \quad (6)$$

where ω_{DC} – angular velocity of DS rotation, s^{-1}, and ω_r – absolute rotational speed of the PDM rotor, s^{-1}.

In the machine proposed by the authors, termed a "two-section bi-rotor PDM" or "two-section PDM with a bi-rotary mechanism", this condition may not be met; however, the stator's rotational speed must not be zero. Accordingly, the authors adopt from Baldenkos' bi-rotary mechanism as the classical case.

A motorized RSS is a modern system for stable, precise drilling of directional wells. It consists of a hydraulic downhole motor (often a PDM) and the RSS, with the motors' output shaft driving the RSS housing (17). Its main drawbacks are: additional dynamic loading due to the flexible shaft, a relatively low maximum dogleg capability and rapid wear of the steering pads. Here and below, the terms "lower" and "upper" refer to the motors' positions along the DS, with the "lower" end being at the bit.

Motors comprising multiple power sections are well known. The most common are sectional PDMs, referred to here as "classical", whose rotors are connected in series by transmission shafts (cardan or torsional/flexible). In some cases, the rotors are phased oppositely (in antiphase), which partially mitigates the motors' radial forces (1). It is important that the orbital (transport) speeds of the coupled rotors (the rotation of the rotor axes about the stator axes) be equal; otherwise, the compensation of radial forces generated in the power sections will be nonuniform.

In such motors, at the output shaft of the lower motor the torques generated in the power sections are summed, subject to losses associated with additional inflow of DF into, and outflow from, the PDM working chambers – a drawback inherent to all sectional PDMs. The rotational speeds of a classical sectional PDM do not add; in practice, the operating speed is taken as the higher of the individual motor speeds. These motors are used with torque-intensive rock-cutting tools.

It should be noted that one of the most critical stages in designing a classical sectional PDM is the calculation of the connection assemblies (universal joints and flexible shafts). The most critical section is the lower transmission shaft, which carries both the axial loads and the torque from the two power sections. Although the physics of a sectional PDM operating with a bi-rotary mechanism differs from the classical case, as will be discussed in due course. It is essential to account for the features of the classical mechanism, since they impose specific considerations on the operating theory of the proposed motor (18).

Thus, the two-section bi-rotor PDM incorporates elements of the following technologies and techniques: PDMs, turbodrills, combined (motor-rotary) drilling, motorized RSS, as well as the classical bi-rotary mechanism and the sectional PDM with transmission coupling between the rotors.

A number of inventions are used as prototypes for the motor underlying the proposed drilling technology.

A counter-rotating turboscrew motor (TSM) invented by Morozov and Sysoev (19): in this design, a turbodrill (a dynamic motor) is installed above a PDM, with the turbodrill's output shaft connected to the housing of the PDM. A threaded connection is provided on the housing of the lower motor (the PDM) to incorporate a reamer into the BHA. The principal drawback of this invention is the use of a turbodrill; because its turbine stages are not mechanically coupled, the upper motor often ceases to operate, and the TSM functions only due to the PDM (19).

An invention by Smith (20), a motor comprising two PDMs, is also bi-rotary in principle. However, when drilling with this tool, the DS must be rotated to the right (clockwise as viewed looking at the bit), while the output shaft of the upper motor, connected to the housing of the lower motor, rotates to the left due to the right-hand helix of the upper PDM's WE. There exists a combination of DF flow rate and DS rotation speed (from the rotary table or top drive) at which the housing of the lower motor stops, or, as the author writes, "almost" stops, rotating. The RSS tool can be installed in the bore of the lower motors' transmission section to build trajectory while the DS is rotating; this is the stated technical result of the invention. The main drawback is that the two PDMs rotate in opposite directions, which can lead to thread back-off. Another drawback is the inability to operate without DS rotation.

Moreover, the upper motor contributes little to improving overall drilling efficiency: apart from enabling trajectory build in the combined mode, it mainly increases the axial load generated in the BHA, which merely allows a reduction in the amount of heavy-weight drillpipe (HWDP) required in the vertical section of the

DS, thereby reducing or eliminating sinusoidal and helical buckling (21, 22).

An invention by Baldenko, also comprising two PDMs, is of interest. In this design, a special non-rotating shroud is mounted on the lower power section, while the housing of the PDM, connected to the output shaft of the upper motor, rotates accordingly (23). The WE of both motors have a left-hand helix. The technical result is an increased output-shaft speed, enabling the use of certain diamond bits. An eccentric stabilizer may also be mounted on the housing of the lower motor to provide dogleg capability.

The two-section bi-rotor PDM consists of two PDMs. Their configurations and WE parameters may differ, which determines the machines' energy characteristics. The output shaft (spindle) of the upper motor is connected to the housing of the lower motor. A crossover sub can be used for this purpose due to the large difference in component diameters. A general view of the two-section bi-rotor PDM is shown in Figure 2. The principal advantage of the two-section PDM with a bi-rotary mechanism is the ability to build trajectory while the housing of the lower motor is rotating (with the RSS placed in the transmission section of the upper PDM). Another advantage is the increased axial load generated within the BHA. In addition, rotation of the lower motor housing reduces frictional interaction with the borehole wall (lowering contact drag due to torsional motion of the housing) and enables influence over the dynamic processes. A higher output-shaft speed with a largely preserved torque characteristic increases power at the bit, which can improve rate of penetration and allows the use of certain diamond bits. The main drawback is the need to match the operating characteristics of both motors to ensure efficient joint operation. Thus, the proposed two-section bi-rotor PDM, whose WE have left-hand helixes, features a rotating housing on the lower motor, referred to as the dynamic module.

2. MATERIALS AND METHOD

Mathematical modeling of the energy characteristics of a two-section bi-rotor PDM:

Since the sections are coupled rotor-to-stator rather than rotor-to-rotor, and if all torques arising from frictional interactions of the WE are neglected, specifically, contact of the housing with the borehole wall, inertial effects, the intrinsic material properties, and the joints between machine components, the torque at the output shaft of the two-section PDM, at the bit, is determined from the following relation:

$$\begin{cases} M_b = M_{\text{m.1}} \\ M_b = M_{\text{m.2}} \end{cases} \tag{7}$$

where M_{b} – bit torque, N·m; $M_{\text{m,1}}$ – ideal torque of the upper PDM, N·m; $M_{\text{m,2}}$ – ideal torque of the lower PDM, N·m.

For a simplified representation of the physics, the motor's reaction torque is taken equal to its ideal torque. From Equation 7, if the torque on the upper PDM is lower than the reaction torque of the lower PDM, equality between them is achieved by increasing the torque of the upper PDM. Conversely, if the torque of the upper PDM exceeds the reaction torque of the lower PDM, equality is achieved by increasing the torque of the lower PDM. The bit torque (i.e., the torque at the output shaft of the two-section PDM) assumes the value of this steady-state equality. Figure 3 presents a schematic of the proposed PDM with the characteristics used in the mathematical model indicated. The following equation serves as the basis for formulating a more detailed system of equations to determine the torques generated in the two-section PDM with a bi-rotor operating principle. The formulation explicitly accounts for a dedicated crossover sub installed between the two PDMs. Accordingly, starting from Euler's equation for a rigid body with one degree of freedom (equivalently, Newton's second law for rotational motion), augmented by linear viscous damping and torsional elasticity (24-26), we obtain the following relation:

$$J\ddot{\theta} + B\dot{\theta} + K\theta = \tau \tag{8}$$

where J – moment of inertia of the specified PDM element, kg·m²; B – viscous loss coefficient (Rayleigh dissipation, Couette flow), N·m·s/rad; K – torsional stiffness of the node, N·m/rad; τ –system output torque, N·m; θ – twist angle, deg; $\dot{\theta}$ – angular velocity, rad/s; and $\ddot{\theta}$ – angular acceleration, rad/s².

Newton's second law for rotational motion is represented by the first term on the left-hand side of Equation 8. Linear viscous damping corresponds to the second term on the left-hand side of Equation 8. The external torques are represented by the third term on the left-hand side of Equation 8.

The subscript "*r*" denotes the rotor, "*h*" the housing of the lower PDM (the dynamic module), "*m*" the motor,

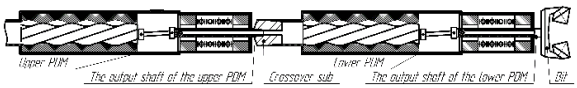


Figure 2. General view of the two-section bi-rotor PDM

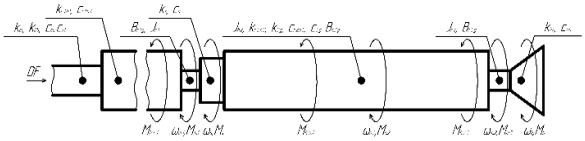


Figure 3. Computational schematic for determining the power-performance characteristics of a two-section PDM with a bi-rotary mechanism

"*s*" the crossover sub, and "*d*" the bit. The index "1" refers to the upper PDM, and "2" to the lower PDM. The kinetic energy of the rotor of the upper PDM, determined by its twist acceleration, is given by:

$$E_{kin} = \frac{1}{2}(mv_C^2 + J_C\omega_{rel.r}^2) \quad (9)$$

where v_C – rotor's linear velocity (Equation10), m/s; ω_{rel} – rotor's relative angular velocity (rotation about its own axis), s^{-1}; J_C – rotor's polar moment of inertia about its own axis (Equations 11-13), kg·m². The first term in Equation 9 corresponds to the inertia moment arising from the rotor's transport motion, while the second, and dominant, term corresponds to the relative rotation of the rotor (27).

$$v_C = \omega_r z_2 e \quad (10)$$

$$J_C = \frac{J_P}{z_1^2} \quad (11)$$

where J_P – equivalent (reduced) moment of inertia of the helical rotor undergoing planetary motion, kg·m²; ωr – rotor's absolute angular velocity, s^{-1}.

The rotor's moment of inertia about its own axis can be obtained from:

$$J_C\ddot{\theta} = F_c e - F_c h \quad (12)$$

where F_g – resultant vector of radial hydraulic forces, N; F_{res} – force that generates the resisting torque (determined by linear viscous damping and the torsional elasticity of the machine elements), N; h –lever arm of F_s, m.

Alternatively, by the Steiner (Huygens) theorem:

$$J_C = J_P + me^2 \quad (13)$$

One may also account for the moments of inertia of the spindle shaft and the bit; in that case, the total moment of inertia is given by:

$$J = J_P + J_{sh} \quad (14)$$

where J_{sh} –moment of inertia of the spindle shaft and the bit, kg·m².

The moments of inertia of the elements of the two-section PDM are determined by differentiating their kinetic energy characteristics according to Equation 9, with generalized inertial moments duly accounted for:

$$\begin{cases} \frac{d}{dt}(\frac{\partial E_{кин}}{\partial \omega_{p.1}}) = J_{p.1}\cdot\dot{\omega}_{p.1} \\ \frac{d}{dt}(\frac{\partial E_{кин}}{\partial \omega_{к.2}}) = J_{к.2}\cdot\dot{\omega}_{к.2} \\ \frac{d}{dt}(\frac{\partial E_{кин}}{\partial \omega_{p.2}}) = J_{p.2}\cdot\dot{\omega}_{p.2} \end{cases} \quad (15)$$

Viscous (velocity-dependent) losses in a PDM, arising from bearings, seals, the DF, and other sources, for engineering calculations may be taken as:

$$B = \frac{\partial M_{loss}}{\partial \omega} \quad (17)$$

where M_{loss} – torque expended on viscous losses, N·m.

In the annular clearance (annulus), for laminar flow, the viscous loss at the walls of the rotating housing of the lower motor (dry friction is accounted for separately by M_{fr} in Equation 27 is given by Pierson et al. (28):

$$B_{l.h.2} = 2\pi\mu L_{h.2}\frac{R_{h.2}^3}{g} \quad (17)$$

where μ – dynamic viscosity of the DF, Pa·s; $L_{h.2}$ – length of the housing of the lower motor (dynamic module), m; $R_{h.2}$ – radius of the housing of the lower motor, m; g – gravitational acceleration, m/s².

For turbulent flow (an increase in the viscous-friction coefficient by approximately $2\omega_{h.2}$ times) (28):

$$B_{t.h.2} = 4\pi\mu L_{h.2}\omega_{h.2}\frac{R_{h.2}^3}{g} \quad (18)$$

It is also necessary to account for viscous damping c between mechanically coupled machine elements. An important distinction from the element-wise viscosity parameter B is that c is used to compute viscous losses as the product of the relative velocity – that is, the difference in velocities of the connected moving elements. Accordingly, the Rayleigh dissipation function (29), which characterizes the aggregate viscous (hydraulic) losses in the two-section bi-rotor PDM, at the interface between the two motors takes the form:

$$R_{el} = \frac{1}{2}(B_{r.1}\omega_{r.1}^2 + B_{h.2}\omega_{h.2}^2 + B_{r.2}\omega_{r.2}^2) + \\ + \frac{1}{2}c_s\cdot(\omega_{r.1} - \omega_{h.2})^2 \quad (19)$$

where c_s – viscous (velocity-dependent) damping in the crossover sub, N·m·s/rad.

The partial derivatives of this expression (with respect to the angular velocities of the principal elements of the PDM) are obtained as follows:

$$\begin{cases} \frac{\partial R_{el}}{\partial \omega_{r.1}} = B_{r.1}\omega_{r.1} + c_s\cdot(\omega_{r.1} - \omega_{h.2}) \\ \frac{\partial R_{el}}{\partial \omega_{h.2}} = B_{h.2}\omega_{h.2} - c_s\cdot(\omega_{r.1} - \omega_{h.2}) \\ \frac{\partial R_{el}}{\partial \omega_{r.2}} = B_{r.2}\omega_{r.2} \end{cases} \quad (20)$$

The torsional stiffness of the assembly:

$$k = \frac{GJ'}{L} \quad (21)$$

where G – shear modulus, MPa; J' – polar moment of area, m⁴; and L – element length, m.

From the torsional stiffness of the crossover sub installed between the two motors, the potential energy of their connection can be determined

$$E_{pot} = \frac{1}{2} k_s (\theta_{r.1} - \theta_{h.2})^2 \tag{22}$$

where k_p – torsional stiffness of the crossover sub, N·m/rad.

In Equation 29 only the torsional stiffness of the crossover sub is taken into account. Subsequently, we account for the stiffnesses of the WE of the upper PDM and the lower PDM, the contact of the lower PDM housing with the borehole wall, the bit-rock interaction, and the DS stiffness up to the wellhead (Equation 30). To determine the twisting torque contributed by the stiffness of each element, one must take the derivative of its potential energy with respect to the twist angle. Accordingly, the following system of equations is obtained for the elements of the proposed motor

$$\begin{cases} \dfrac{\partial E_{pot}}{\partial \theta_{r.1}} = k_s \cdot (\theta_{r.1} - \theta_{h.2}) \\ \dfrac{\partial E_{pot}}{\partial \theta_{h.2}} = -k_s \cdot (\theta_{r.1} - \theta_{h.2}) \\ \dfrac{\partial E_{pot}}{\partial \theta_{r.2}} = 0 \end{cases} \tag{23}$$

Building on the preceding assumptions, a system of torque-balance equations can be formulated for the rotor system of the upper PDM, the dynamic module (housing of the lower PDM), and the rotor system of the lower PDM. Equation 8 serves as the foundation. Accordingly, the equation for the rotor system of the upper motor is derived from

$$J_{r.1}\dot{\omega}_{r.1} = M_{m.1} - M_s - B_{r.1}\omega_{r.1} - M_{fr.r.1} \tag{24}$$

where $M_{fr,r,1}$ – torque generated by contact friction in the WE of the upper PDM, N·m.

Torque balance in the dynamic module of the two-section PDM with a bi-rotary mechanism:

$$J_{h.2}\dot{\omega}_{h.2} = M_s - M_{m.2} - B_{h.2}\omega_{h.2} - M_{fr.h.2} \tag{25}$$

where $M_{fr,h,2}$ – torque generated by contact friction at the interface between the dynamic module (housing) of the lower PDM and the borehole wall, N·m.

Torque balance in the rotor system of the lower PDM

$$J_{r.2}\dot{\omega}_{r.2} = M_{m.2} - M_b - B_{r.2}\omega_{r..2} - M_{fr.r.2} \tag{26}$$

where $M_{fr,r,2}$ – torque generated by contact friction in the WE of the lower PDM, N·m.

Thus, Equations 24-26 form system that characterizes the torque behavior of the two-section PDM under a non-stabilized operating regime

$$\begin{cases} J_{r.1}\dot{\omega}_{r.1} = M_{m.1} - M_s - B_{r.1}\omega_{r.1} - M_{fr.r.1} \\ J_{h.2}\dot{\omega}_{h.2} = M_s - M_{m.2} - B_{h.2}\omega_{h.2} - M_{fr.h.2} \\ J_{r.2}\dot{\omega}_{r.2} = M_{m.2} - M_b - B_{r.2}\omega_{r.2} - M_{fr.r.2} \end{cases} \tag{27}$$

It is important to note that the reaction torque, although not explicitly present in Equations 24-27, is accounted for through energy dissipation within the proposed motor system. The crossover sub is modeled as a viscoelastic element with torsional stiffness k_s and viscous coefficient c_s (frictional torque in the sub is neglected). Its twisting torque is given by:

$$M_s = k_s \cdot (\theta_{r.1} - \theta_{h.2}) + c_s \cdot (\omega_{r.1} - \omega_{h.2}) \tag{28}$$

By substituting the obtained expressions (systems of Equations 15, 20, and 23) together with the external applied torques into the Lagrange equation, the correctness of the derived system for determining the torque forces generated in the two-section bi-rotor PDM. It can be written in matrix form as follows:

$$\underbrace{\begin{bmatrix} J_{r.1} & 0 & 0 \\ 0 & J_{h.2} & 0 \\ 0 & 0 & J_{r.2} \end{bmatrix}}_{J} \ddot{\theta} + \underbrace{\begin{bmatrix} B_{r.1} + c_s & -c_s & 0 \\ -c_s & B_{к.2} + c_s & 0 \\ 0 & 0 & B_{r.2} \end{bmatrix}}_{B} \dot{\theta} + \underbrace{\begin{bmatrix} k_s & -k_s & 0 \\ -k_s & k_s & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{K} \theta = \underbrace{\begin{bmatrix} M_{m.1} - M_{fr.r.1} \\ -M_{m.2} - M_{fr.к} \\ M_{m.2} - M_{д} - M_{fr.r.2} \end{bmatrix}}_{\tau} \tag{29}$$

With the additional stiffness, viscous parameters of the system included, Equation 29 takes the following form:

$$\underbrace{\begin{bmatrix} c_s + c^{eq}_{r.1.h.1} + B_{r.1.g} & -c_s & 0 \\ -c_s & c_s + c_{r.2.h.2} + c_{h.2.g} + B_{h.2.g} & -c_{r.2.h.2} \\ 0 & -c_{r.2.h.2} & c_{r.2.h.2} + c_{b.g} + B_{r.2.g} \end{bmatrix}}_{K} \dot{\theta} + \underbrace{\begin{bmatrix} J_{r.1} & 0 & 0 \\ 0 & J_{h.2} & 0 \\ 0 & 0 & J_{r.2} \end{bmatrix}}_{J} \ddot{\theta} + \underbrace{\begin{bmatrix} k_s + k^{eq}_{r.1.h.1} & -k_s & 0 \\ -k_s & k_s + k_{r.2.h.2} + k_{h.2.g} & -k_{r.2.h.2} \\ 0 & -k_{r.2.h.2} & k_{r.2.h.2} + k_{b.g} \end{bmatrix}}_{K} \theta = \underbrace{\begin{bmatrix} M_{m.1} - M_{fr.r.1} \\ -M_{m.2} - M_{fr.h} \\ M_{m.2} - M_b - M_{fr.r.2} \end{bmatrix}}_{\tau} \tag{30}$$

where $c^{eq}_{r.1.h.1}$ and $k^{eq}_{r.1.h.1}$ – equivalent viscous parameter and stiffness of the WE of the upper PDM relative to the anchored stator (rock), with the stiffness of the entire DS up to the wellhead included, N·m·s/rad and N·m/rad,

respectively; $c_{\text{h.2.g}}$ and $k_{\text{h.2.g}}$ – viscous parameter and stiffness of the contact between the housing of the lower PDM and the borehole wall, N·m·s/rad and N·m/rad, respectively; $c_{\text{r.2.h.2}}$ and $k_{\text{r.2.h.2}}$ – viscous parameter of the rotor-stator interface of the lower PDM and the stiffness of its WE, N·m·s/rad and N·m/rad, respectively; $c_{\text{b.g}}$ and $k_{\text{b.g}}$ – viscous parameter and stiffness at the bit-rock contact, N·m·s/rad and N·m/rad, respectively; $B_{\text{r.1.g}}$, $B_{\text{h.2.g}}$ and $B_{\text{r.2.g}}$ – denote additional (experimentally identified) "to-ground" losses of the corresponding nodes that are not captured by c, N·m·s/rad.

The value of $k^{eq}_{r.1.h.1}$ can be obtained from

$$k^{eq}_{r.1.h.1} = \left(\frac{1}{k_{r.1.h.1}} + \frac{1}{k_{DS}}\right)^{-1} \tag{31}$$

where $k_{\text{r.1.h.1}}$ – stiffness of the working elements of the upper PDM relative to the anchored stator (the formation), N·m/rad; and k_{DS} is the DS stiffness up to the wellhead, N·m/rad.

An analogous procedure is used to determine $c^{eq}_{r.1.h.1}$.

By analogy with Equation 31, the coefficient k_{DS} includes the stiffness of the wellhead clamp, k_{cl}.

Accordingly, the torque loss acting on the PDM rotor due to the medium's viscosity can be determined experimentally. Set the rotor to a rotational speed at which $B\omega \gg M_{\text{fr}}$, then shut the drive down and record the decay of the angular velocity ω as a function of time, ω(t) (30). The coefficient B can also be evaluated by bringing a two-section bi-rotor PDM to a stabilized steady-state regime ($\dot{\omega} = 0$), in which case the left-hand side of the relations given in Equation 27 becomes zero. Once the remaining parameters have been obtained from instrumentation or from calculations, determining B poses no difficulty (1, 12, 31).

Another approach is to analyze the torque-governing relations for the PDM, from which the overall efficiency and its effect on the torque response of the elements of a two-section PDM with a bi-rotary mechanism can be derived. In particular, the reaction torques arising from losses in the power section due to rotor friction against the stator liner can be evaluated. Subsequently, the bit rotational speed and the power delivered at the bit, when coupled to the two-section PDM with a bi-rotary operating mechanism, should be examined.

The pressure drop across the two-section bi-rotor PDM and the axial load generated by hydraulics in the power sections of both stages add together, as in a conventional multi-section PDM. This is because the power sections (the motor's WE) are arranged in series, while the specific connection scheme does not affect the underlying physics of pressure drop and axial loading in a machine composed of two PDMs. However, since the flow field is perturbed at the interface between the sections, just as in classical multi-section PDMs, the overall pressure drop is not exactly equal to the sum of the individual pressure drops across each PDM; rather, it exceeds that sum by a small amount that can be neglected for engineering calculations (32). DF, its rheological and other parameters, play a high role in the PDM's power-performance characteristics (33, 34).

A noteworthy observation is the presence of torque pulsations in the proposed motor (12). These arise from a periodic variation of the pressure drop, given by the equation below, which itself varies due to the periodic change in the number of contact lines as defined by Equation 1:

$$\begin{cases} P_{\max} = \dfrac{P}{\Lambda_{\max}} \\ P_{\min} = \dfrac{P}{\Lambda_{\min}} \end{cases} \tag{32}$$

By combining the first relation from the set of Equation 2 with the system of Equation 27, the pressure drops across both individual motors, and, specifically, across the two-section PDM, can be obtained. The resulting system of equations is given below:

$$\begin{cases} \begin{cases} P_{m.2} = \dfrac{2\pi}{\eta_{m.2} \cdot V_{m.2}}(M_b + B_{r.2}\omega_{r.2} + M_{fr.r.2} + J_{r.2}\dot{\omega}_{r.2}) \\ P_{m.1} = \dfrac{2\pi}{\eta_{m.1} \cdot V_{m.1}}(M_b + B_{r.2}\omega_{p.2} + M_{fr.r.2} + \\ +B_{h.2}\omega_{h.2} + M_{fr.h.2} + B_{r.1}\omega_{r.1} + M_{fr.r.2} + \\ +J_{r.1}\dot{\omega}_{r.1} + J_{h.2}\dot{\omega}_{h.2} + J_{r.2}\dot{\omega}_{r.2}) \end{cases} \\ P = P_{m.2} + P_{m.1} + \Delta P \end{cases} \tag{33}$$

where ΔP – additional pressure drop induced by flow maldistribution between the PDM sections, Pa.

The expression for the axial load, making explicit the linear dependence on the pressure drop across the power section, has the form

$$F_{WE} = F_p + F_z = P_m \cdot (S_c + z_2 S) \tag{34}$$

where F_{r} – hydraulic component of the axial force in the PDM WE, N; F_{z} – axial component of the force arising from the helical engagement of the power section, N; S_{k} – projected area of the rotor-stator contact lines, m²; and S – is the open flow area of the WE, m².

For example, for two DGR-172.5/6.61 PDMs installed in series and neglecting the additional pressure drop ΔP, Equation 34 yields nearly 21.4 t of axial load on the power sections at a pressure drop of 4,41 MPa (10,68 t per motor). Thus, when the pressure drop in the BHA increases due to pairing the power sections, the axial force F_{WE} rises, which enables a greater maximum deviation of the wellbore trajectory from vertical.

In the ideal case, the bit RPM generated by two sectionally connected PDMs forming a bi-rotary mechanism adds algebraically. The output-shaft speed of the lower PDM can also be obtained from Equation 27. It is convenient to consider the stabilized regime $(\dot{\omega} = 0)$.

Importantly, under these conditions one must have $\omega_{r.1} = \omega_{h.2}$, otherwise, the angular difference ($\theta_{r.1} - \theta_{h.2}$) grows linearly and the torque in the crossover sub tends to infinity (Equation 28) (35, 36). Accordingly, imposing the equality between the output-shaft speed of the lower two-section motor and the bit speed gives

$$\omega_b = \omega_{r.2} = \frac{M_{m.1} - \omega_{r.1} \cdot (B_{r.1} + B_{h.2})}{B_{r.2}} - \frac{M_{fr.r.1} + M_{fr.h.2} + M_{fr.r.2} + M_b}{B_{r.2}} \quad (35)$$

The power at the bit increases as the rotational speed at the output shaft rises while its torque characteristics are preserved (ideal case), as follows from the equation corresponding to ideal operating conditions

$$N_b = M_b \omega_b \quad (36)$$

It is important to recognize that, for the same power at the bit, the efficiency of rock destruction can vary across different geological and operational conditions. For example, as the rock drillability category increases, it is necessary to increase the WOB and hence the torque at the bit, while reducing bit RPM.

The reason is that, if the rotational speed is maintained or raised as WOB increases, the reactive force exerted by the formation on the bit cutters grows, which hinders any substantial increase in the ROP and may prevent it altogether; simultaneously, tool wear intensifies. Consequently, increasing bit speed while holding the torque characteristic constant does not always enhance drilling efficiency, although it does enable the use of certain diamond bits (6, 37, 38). Using Equations 27, 33, 35, and 36, theoretical performance curves were plotted for the output-shaft rotational speed of the two-section PDM with a bi-rotor operating mechanism (consisting of two DGR-172.5/6.61 PDMs), its torque and power, as well as the motor efficiency, as shown in Figures 4 and 5. It's important that a two-section PDM torque in the range of 4,0-9,0 kN·m corresponds to its optimal operating regime. The WOB varies from 120 to 280 kN. Note that exceeding a torque of 12,7 kN·m causes the upper PDM to stall. Surpassing

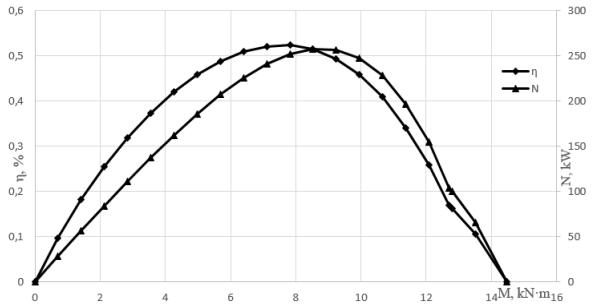


Figure 4. Plot of motor efficiency (η) and power (N) as a function of bit torque (Q_m = 31,9 m^3/s)

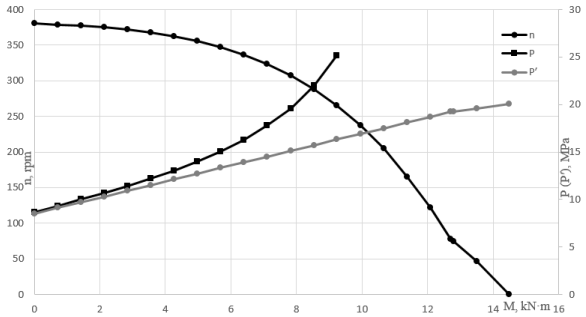


Figure 5. Plot of output-shaft rotational speed (n) and differential pressure (P, P') across the WE versus bit torque (Q_m = 31,9 m^3/s)

this critical torque and stalling the PDM is accompanied by a rise in pressure and, consequently, necessitates reducing WOB. Ensuring that the two-section PDM with a bi-rotary mechanism operates within the calculated range will enable efficient well drilling.

In Figure 5, the graph of the function $P'(M)$ provides a simplified (linearized) representation of the pressure differential characteristic across the two-section bi-rotor PDM, assuming a constant volumetric efficiency. The P(M) plot is truncated because it tends to infinity.

3. DISCUSSION AND RESULTS

After developing the mathematical model for predicting PDM performance, it must be benchmarked against field measurements. In other words, the model's accuracy should be quantified as the error of the computed parameter relative to the physically measured value. A key feature of a multi-section PDM installed in series is that the same mud flow passes through the power sections of both motors. Consequently, efficient operation of a two-section PDM requires that the rated (required) DF flow rates of the two sections be equal. If the motors are selected without regard to this constraint, the flow rate through one section may fall within its optimal operating range while the other does not, reducing the overall efficiency of the machine. Therefore, it is essential to select PDMs whose required DF rates are matched, with due consideration of their specifications. Alternatively, one may re-tune WE of one PDM or install auxiliary flow-control elements to regulate the circulating DF delivered to the volumetric motor's power section.

Another promising research direction is to treat the housing of the lower motor as a dynamic stabilization module for self-excited oscillations, that is, as a device for suppressing torsional vibrations (stick-slip) that arise at the bit during drilling. Moreover, this dynamic module can not only mitigate harmful vibrations but, if driven into resonance with them, also generate elevated oscillation levels that help free the assembly from stuck-pipe conditions and mitigate pack-off around the BHA.

Control of the dynamic module can be exercised by adjusting the DF flow rate Q, thereby avoiding changes to surface DS RPM at the rig floor or to WOB via movement of the traveling block. Table 1 summarizes the key drawbacks of the principal control methods for the dynamic module of a two-section bi-rotor PDM.

An interesting prospect is to machine a threaded profile on the rotating housing to mount a reamer, which could significantly enhance drilling efficiency.

However, because the reamer and the bit may engage formations with different drillability classes, and despite the fact that such a configuration is not uncommon in practice, this option requires careful evaluation. Specifically, mismatches in the geological conditions suitable for the reamer versus the bit can excite high-amplitude dynamic responses in the BHA (axial, radial, torsional), triggering self-excited oscillations and resonance with vibrations generated elsewhere along the DS.

An interesting prospect is to machine a threaded profile on the rotating housing to mount a reamer, which could significantly enhance drilling efficiency. However, because the reamer and the bit may engage formations with different drillability classes, and despite the fact that such a configuration is not uncommon in practice, this option requires careful evaluation.

Mismatches in the geological conditions suitable for the reamer versus the bit can excite high-amplitude dynamic responses in the BHA (axial, radial, torsional), triggering self-excited oscillations and resonance with vibrations generated elsewhere along the DS.

TABLE 1. Key drawbacks of the principal methods for regulating PDM power-performance characteristics

Methods for controlling the dynamics of the lower motor housing	Key drawback
Adjusting axial load at the bit (WOB)	For extended wells, because of multiple sinusoidal and spiral bends of the DS, raising the string at the wellhead to change WOB may require hoisting several meters of pipe
Adjusting DS rotation (surface RPM at the wellhead)	The DS acts as a torsional spring; thus transmission of surface rotation to the BHA is imperfect, especially over long intervals. High DS-wellbore friction leads to stick-slip
Adjusting drilling-mud flow rate	Requires maintaining an RPM-torque characteristic consistent with an efficient drilling regime

4. CONCLUSION

A two-section PDM with a bi-rotary mechanism is proposed. The assembly consists of two rigidly coupled gerotor stages and a crossover sub that connects the output shaft of the upper PDM to the housing of the lower PDM, enabling efficient directional and horizontal drilling with a bent-housing motor assembly. Additionally, more precise mathematical modeling is required for the stall (braking) torque of one of the motors, the discontinuity (jump) in its power/energy characteristics, and the subsequent operation of the two-section PDM with a bi-rotary mechanism. The operating principle is described, the advantages are identified, and a drilling methodology using the proposed motor is developed. Prototypes and analogs of existing designs are noted: a motorized RSS, a two-section PDM with a right-hand helical profile in the upper power section, and a two-section PDM with a stationary shroud on the lower motor. Their benefits and limitations are highlighted. The characteristics of a conventional PDM and its specific features are outlined. The system's nonlinearity is substantiated, together with its key distinctions from a classical multi-section PDM (where the rotors are interconnected by a transmission shaft). Particular emphasis is placed on the correspondence with, as well as the fundamental differences and analogies relative to, the classical multi-section PDM.

The principal advantages and limitations of the proposed two-section PDM (bi-rotary mechanism) are highlighted, together with its technological distinctiveness. The motor should be assembled from two PDMs whose individual torques not only do not reach the stall torque of the counterpart, but also each fall within a range conducive to efficient drilling. To achieve an optimal match between the required mud flow rate of one section and the flow rate that is optimal for the other, retuning of one power section is recommended. The need to incorporate a dynamic module is also established.

Calculations were performed to determine the operating characteristics of the bi-rotor PDM – namely torque, rotational speed, power, pressure drop, and the axial load generated in the BHA. Particular attention is drawn to the possibility of only a modest increase – or no increase at all – in ROP when bit RPM is raised while preserving the torque characteristic at the bit under certain geologic and operational conditions. In evaluating the output-shaft torque of the two-section bi-rotor PDM, the inertial effects of the motor rotors and the dynamic module (the housing of the lower PDM), the viscous (damping) properties of the elements, and their stiffnesses were taken into account. The governing equation was derived and justified using Lagrange's equation; Rayleigh's dissipation function, Euler's equation for a rigid body with one degree of freedom (Newton's second law for rotational motion), the parallel-axis theorem (Steiner-Huygens) and Couette flow were also employed. Directions for further investigation of the two-section bi-rotor PDM are proposed.

Acknowledgements

The authors appreciates Saint Petersburg Mining University of Empress Catherine II and the Arctic Research Center for providing the necessary facilities and support to conduct this research.

Funding

This research received no external funding.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.

Authors Biosketches

Mikhail Vladimirovich Dvoynikov is a Professor and Head of the Wells Drilling Department at Saint Petersburg Mining University of Empress Catherine II, and the Scientific Director of the Arctic Research Center. He holds a Dr.Sc. (Engineering/Technical Sciences) degree and his research focuses on oil and gas well drilling and completion technologies, drilling fluids, and oil and gas field development.

Dmitry Olegovich Morozov is a Ph.D. student at the Wells Drilling Department, Saint Petersburg Mining University of Empress Catherine II. His research interests include drilling equipment, well construction technologies and drilling fluids.

REFERENCES

1. Baldenko DF, Baldenko FD, Gnoevykh AN. Single screw hydraulic machines. Vol. 2. Screw downhole engines. Moscow, Gazprom Information and Advertising Center, 2007, 470. EDN QMJKHP
2. Raupov I, Rogachev M, Shevaldin E. Review of Formation Mechanisms, Localization Methods, and Enhanced Oil Recovery Technologies for Residual Oil in Terrigenous Reservoirs. Energies. 2025;18(21):5649. https://doi.org/10.3390/en18215649
3. Nutskova MV. Effect of Mineral Wool Impregnated with Carbon Nanotubes on Properties of Cement at High Temperatures. International Journal of Engineering, Transactions A: Basics. 2025;38(1):148-151. https://doi.org/10.5829/ije.2025.38.01a.14
4. Simonyants SL, Al Taee M. Stimulation of the Drilling Process with the Top Driven Screw Downhole Motor. Journal of Mining Institute. 2019;238:438-442. https://doi.org/10.31897/PMI.2019.4.438
5. Lyagov IA, Baldenko FD, Lyagov AV, Yamaliev VU, Lyagova AA. Methodology for calculating technical efficiency of power sections in small-sized screw downhole motors for the «Perfobur» system. Journal of Mining Institute. 2019;240:694-700. https://doi.org/10.31897/PMI.2019.6.694
6. Neskromnykh VV, Popova MS, Golovchenko AE, Petenev PG, Baochang L. Method of drilling process control and experimental studies of resistance forces during bits drilling with PDC cutters. Journal of Mining Institute. 2020;245:539-546. https://doi.org/10.31897/PMI.2020.5.5
7. Invention patent application №2025132209 Russian Federation. Two-section positive displacement downhole motor with a bi-rotor mechanism of action / M.V. Dvoynikov, D.O. Morozov; applicant Federal State Budgetary Educational Institution of Higher Education "Saint Petersburg Mining University". Declared 19.11.2025
8. Chuyko EV, Chetvertneva IA, Movsumzade EM, Loginova ME, Tivas NS, Giniyatullin AN. History of horizontal drilling technology development of oil and gas wells. History and Pedagogy of Natural Science. 2024;3-4:74-78. https://doi.org/10.24412/2226-2296-2024-3-4-74-78
9. Litvinenko VS, Petrov EI, Vasilevskaya DV. Assessment of the role of the state in the management of mineral resources. Journal of Mining Institute. 2023;259:95-111. https://doi.org/10.31897/PMI.2022.100
10. Blinov PA, Silichev NM, Nikishin VV, Gorelikov VG. Substantiation and development of a methodology for calculating the bending of the drive shaft of a continuous deflector for drilling with coring. SOCAR Proceedings. 2025;3:39-48. https://doi.org/10.5510/OGP20250301095
11. Sidorkin DI, Kupavykh KS. Justification on Choosing Screw Pumping Units as Energy Efficient Artificial Lift Technology. Energetika. Proceedings of CIS Higher Education Institutions and Power Engineering Associations. 2021;64(2):143-51. https://doi.org/10.21122/1029-7448-2021-64-2-143-151
12. Baldenko DF, Baldenko FD, Gnoevykh AN. Single screw hydraulic machines. Vol. 1. Single screw pumps. Moscow, Gazprom Information and Advertising Center, 2005, 488. EDN QMJKIJ
13. Kuang Y, Zhou J. Performance evaluation for positive displacement motors by combining fluid-structure interaction simulations and experiments. SPE Journal. 2025;30(1):152-68. https://doi.org/10.2118/223929-PA
14. Shi C, Wan X, Deng J, Zhu X. Research on fatigue lifetime prediction of stator rubber bushing based on rubber aged experiment. Journal of Testing and Evaluation. 2022;50(4):2209-26. https://doi.org/10.1520/JTE20200669

15. Povalikhin AS. A technology of a horizontal well drilling in case of a casing string combined rotation. Petroleum Engineer. 2019;3:5-9. EDN YEEHYT
16. Baldenko DF, Baldenko FD. Working process features and characteristics of positive displacement motors in the rotation mode of a drill string. Drilling and Oil. 2019;11:8-13. EDN YIMPTH
17. Geng H, Zhang B, Xie Y. Rotary Steerable Drilling Technology: Bottlenecks, Breakthroughs and Intelligent Trends in China Shale Gas Development. Processes. 2025;13(11):3471. https://doi.org/10.3390/pr13113471
18. Biletskyi V, Landar S, Mishchuk Y. Modeling of the power section of downhole screw motors. Mining of Mineral Deposits. 2017;11(3):15-22. https://doi.org/10.15407/mining11.03.015
19. Patent RU 2 368 752 C1. Counter-rotating turboprop engine. Morozov NS, Sysoev NI. Declared 20.05.2008. Published 27.09.2009. (In Russian). EDN LNPIMG
20. Patent US 9 771 787 B2. Multi-directionally rotating downhole drilling assembly and method. Smith G. Declared 14.10.2016. Published 26.09.2017. (In English)
21. Kunshin AA, Buslaev GV, Reich M, Ulyanov DS, Sidorkin DI. Numerical simulation of nonlinear processes in the "Thruster-Downhole Motor-Bit" system while extended reach well drilling. Energies. 2023;16(9):3759. https://doi.org/10.3390/en16093759
22. Tananykhin DS. Scientific and methodological support of sand management during operation of horizontal wells. International Journal of Engineering, Transactions A: Basics. 2024;37(7):1395-407. https://doi.org/10.5829/ije.2024.37.07a.17
23. Patent RU 2 524 238 C2. Screw downhole motor. Baldenko DF, Baldenko FD, Baletinskikh DI, Vorobyev VG, Lunev AV, Selivanov SM, Smirnov AS. Declared 17.08.2012. Published 27.07.2014. (In Russian)
24. Shi X, Jiao Y, Yang X, Liu J, Zhuo Y. Establishment of drill-string system dynamic model with PDM and influence of PDM parameters on stick-slip vibration. Arabian Journal for Science and Engineering. 2024;49:8727-39. https://doi.org/10.1007/s13369-023-08620-z
25. Ritto TG, Aguiar RR, Hbaieb S. Validation of a drill string dynamical model and torsional stability. Meccanica. 2017;52:2959-2967. https://doi.org/10.1007/s11012-017-0628-y
26. Samuel R, Baldenko DF, Baldenko FD. Mud Motor PDM Dynamics: A Control Model for Automation. IADC/SPE International Drilling Conference and Exhibition. 2022; Paper SPE-208789-MS. https://doi.org/10.2118/208789-MS
27. Samuel R, Baldenko DF, Baldenko FD. Mud Motor PDM Dynamics: An Analytical Model. SPE Annual Technical Conference and Exhibition. 2021; Paper SPE-206064-MS. https://doi.org/10.2118/206064-MS
28. Pierson JL, Kharrouba M, Magnaudet J. Hydrodynamic torque on a slender cylinder rotating perpendicularly to its symmetry axis. Physical Review Fluids. 2021;6:094303. https://doi.org/10.1103/PhysRevFluids.6.094303
29. He Y, Chen Y, Hao X, Deng S, Li C. Research on dynamic calculation methods for deflection tools in deepwater shallow soft formation directional wells. Processes. 2025;13(6):1947. https://doi.org/10.3390/pr13061947
30. Baravian C, Quemada D. Using instrumental inertia in controlled-stress rheometry. Rheologica Acta. 1998;37:223-233. https://doi.org/10.1007/s003970050110
31. Lu D, Wang M, Xu Y, Wang X. Analytical Determination of Stick–Slip Whirling Vibrations and Bifurcations in Rotating Machinery. Applied Sciences. 2024;14(16):7338. https://doi.org/10.3390/app14167338
32. Lyagov IA, Lyagov AV, Isangulov DR, Lyagova AA. Selection of the required number of circulating subs in a special assembly and investigation of their performance during drilling of radial branching channels by sectional positive displacement motors. Journal of Mining Institute. 2024;265:78-86. EDN ZBPWKU
33. Leusheva EL. Evaluation of possible application of powder made from fallen tree leaves as a drilling mud additive. International Journal of Engineering Transactions B: Applications. 2024;37:1592-1599. https://doi.org/10.5829/ije.2024.37.08b.12
34. Leusheva EL, Morozov AO, Morozov DO. Ecological biodegradable additives for improving rheological and filtration characteristics of water-based drilling mud. Perm Journal of Petroleum and Mining Engineering. 2024;24(3):120-130. https://doi.org/10.15593/2712-8008/2024.3.3
35. Rybakov DA, Dorokhin EG, Andreev KV, Buslaev GV. Development of a Hydraulic Circulation Sub as a Tool to Prevent Mud Losses During Well Drilling. International Journal of Engineering Transactions A: Basics. 2024;37(10):2091-98. https://doi.org/10.5829/IJE.2024.37.10A.19
36. Dong G, Chen P. The vibration characteristics of drillstring with positive displacement motor in compound drilling Part 2: Transient dynamics and bit control force analysis. International Journal of Hydrogen Energy. 2018;43(27):12189-12199. https://doi.org/10.1016/j.ijhydene.2018.02.198
37. Al Shekaili A, Liy Y, Papatheou E. Drilling performance analysis of a polycrystalline diamond compact bit via finite element and experimental investigations. International Journal of Rock Mechanics and Mining Sciences. 2024;182:105862. https://doi.org/10.1016/j.ijrmms.2024.105862
38. Aribowo AG, Aarsnes UJ, Chen K. Analysis of a downhole passive regulator in drilling: A distributed parameter modeling approach. Journal of Sound and Vibration. 2024;618:119300. https://doi.org/10.1016/j.jsv.2025.119300

Persian Abstract

چکیده

برای حفاری چاه‌های نفت و گاز، موتورهای پیچشی پایین‌چاهی (PDM) هم در روسیه و هم در خارج از آن به‌طور گسترده به کار گرفته می‌شوند و سهم استفاده از آن‌ها بیش از ۸۰٪ است. با این حال، PDM با مجموعه‌ای از کاستی‌ها روبه‌روست که به کارایی حفاری چاه‌های مایل–جهت‌دار و فرآیند حفاری به‌طور کلی مربوط می‌شود. نخست، در اجزای کاری این موتور فرآیندهای دینامیکی نامطلوبی تولید می‌گردد که بر بهره‌وری کل عملیات حفاری اثر منفی دارد. دوم، سرعت دورانی پایین محور/شفت خروجی است که بهره‌برداری از طیفی از مته‌های الماسی را ناممکن می‌کند. سوم، نسبت کشیدگی افقی چاه به عمق عمودی آن تقریباً هرگز از ٤،٠ فراتر نمی‌رود، در حالی که برای سیستم هدایت‌پذیر چرخشی (RSS) این نسبت می‌تواند به ٦،٠ برسد. در حال حاضر، سامانه‌های RSS موتوردار و بدون موتور هرچه بیشتر به کار می‌روند و با کمک آن‌ها حفاری چاه‌هایی با بیشترین انحراف از قائم و با پروفیل مسیرِ به‌طور مشخص از پیش تعیین‌شده امکان‌پذیر است. بدیهی است که هر دوی این راه‌حل‌ها و نیز گزینه‌های دیگر، مزایا و معایب خود را دارند که پرهیز کامل از آن‌ها در جریان عملیات حفاری دشوار و گاه ناممکن است. فناوری حفاری با یک PDM دوبخشیِ دارای سازوکار دو–روتوره پیشنهاد می‌شود که امکان افزایش کارایی حفاری چاه‌های مایل–جهت‌دار را فراهم می‌کند. در این کار، نمونه‌های اولیهِ موتورِ طراحی‌شده ارائه شده، شباهت‌ها و تفاوت‌های شاخص آن با تکنیک‌ها و فناوری‌های پیشین مشخص گردیده و همچنین مشخصه‌های PDM دوبخشی–از جمله گشتاور، سرعت دورانی، توان در ابزار سنگ‌بر/مته و بار محوریِ تولیدشده در مونتاژ تحتانی رشته حفاری (BHA) –تعیین شده است. امکان تنظیم عملکرد موتور از طریق اصلاح دبی گل/سیال حفاری در بخش کاری روتور–استاتور نیز توجیه شده و محورهای بالقوه برای پژوهش‌های آینده ارائه گردیده است.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2485-2499

International Journal of Engineering

Journal Homepage: www.ije.ir

A Deep Learning Based Signal Detection Framework for Non Orthogonal Multiple Access Systems

M. Tulasi[a,b], D. Sreenivasa Rao*[a]

[a] *Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India*
[b] *Department of DTDP, Dr. YSR Architecture and Fine Arts University, Kadapa, AP, India*

PAPER INFO

Paper history:
Received 01 September 2025
Received in revised form 27 September 2025
Accepted 24 October 2025

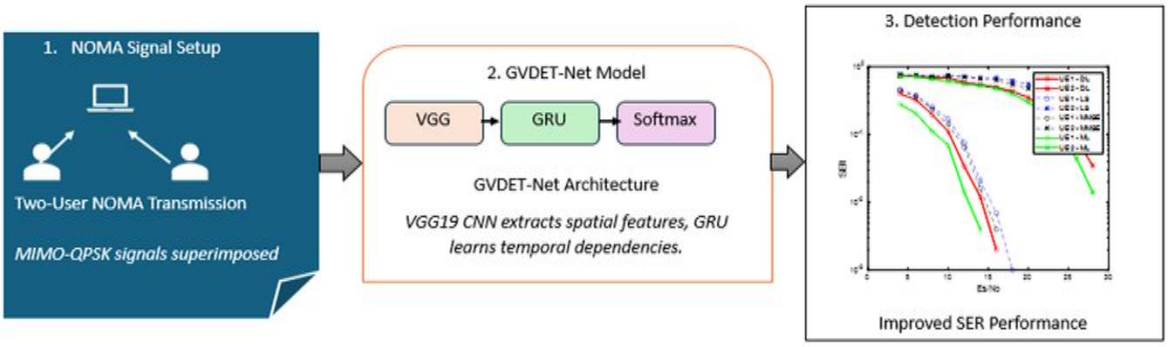
Keywords:
Neural Network Architecture
Pilot-Aware Detection Signal Demodulation
Wireless Access Technology
Symbol Error Rate

ABSTRACT

This paper presents GVDET-Net, an innovative signal detection framework designed to enhance the detection accuracy and operational efficiency of NOMA systems utilizing non-orthogonal time-frequency resources. The proposed model integrates VGG19-based CNN layers with GRU layers to jointly extract spatial and temporal dependencies from input data. By sequentially processing hierarchical features, GVDET-Net achieves superior NOMA channel signal detection compared to ML, LS, and MMSE approaches across SNRs from 4 dB to 28 dB. Simulation results demonstrate its effectiveness under realistic NOMA conditions, outperforming SIC-LS and SIC-MMSE under multiple test scenarios with 64 and 16 pilot configurations for dual-user cases. GVDET-Net achieves a minimum Symbol Error Rate (SER) of approximately 10^{-3} at high SNR levels, delivering significant performance gains. Additionally, the model attains 96.4% classification accuracy, 3.1 ms inference latency for standard packet sizes, and an AUC score of 0.968, validating its robustness and real-time applicability. This work introduces advanced detection techniques for NOMA systems, paving the way for optimized wireless networks and supporting next-generation communication standards.

***doi**: 10.5829/ije.2026.39.10a.11*

Graphical Abstract



NOMENCLATURE

N_CP	Length of Cyclic Prefix	N_PSC	Number of Pilot Subcarriers
N_UE	Number of Users (User Equipments)	N_SC	Number of Subcarriers
N_PSym	Number of Pilot OFDM Symbols per Packet	N_DSym	Number of Data OFDM Symbols per Packet
d_i	Modulated Data Symbol (QPSK)	a_i, b_i	Amplitude components of QPSK symbols
N_0	Noise Power	σ^2	Noise Variance

*Corresponding Author Email: sreenuece@kluniversity.in (D. Sreenivasa Rao)

Please cite this article as: Tulasi M, Sreenivasa Rao D. A Deep Learning Based Signal Detection Framework for Non Orthogonal Multiple Access Systems. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2485-99.

1. INTRODUCTION

The growing demand, along with increasing system capacity requirements, has led to the development of new wireless communication technologies (1). The wireless industry has recognized the potential of Non-Orthogonal Multiple Access (NOMA) technology, which offers distinct functionalities compared to traditional Orthogonal Frequency Division Multiplexing (OFDM) systems (2-5). This shift in the communication paradigm provides several advantages, such as enhanced spectral efficiency, increased system capacity, and more equitable resource distribution (6-8). NOMA achieves these benefits through power domain multiplexing, allowing multiple users to transmit simultaneously on the same resources, thereby maximizing spectrum utilization (9). Additionally, NOMA dynamically adjusts its resource distribution based on user channel conditions to improve system performance and operational efficiency (10).

The wireless communication sector adopts NOMA because it significantly enhances spectral efficiency while expanding system capacity and implementing fair resource distribution and adaptive resource allocation strategies (11). NOMA makes full use of the available spectrum, enabling users to transmit information concurrently (12). As a result, maximum throughput and increased user capacity can be achieved within the same bandwidth allocation (13-14). The resource allocation mechanism optimizes resource utilization according to individual user conditions, ensuring that users have equitable access to network resources despite varying data demands and changing channel conditions.

The compatibility of NOMA with emerging technologies such as massive MIMO and millimeter-wave (mmWave) communications makes it a highly attractive solution.

Current signal detection technology in NOMA systems relies on Machine Learning (ML) and Deep Learning (DL) methods to analyze large datasets and extract valuable features. However, signal detection in NOMA systems presents inherent challenges due to interference and variable channel conditions, as these systems share resources non-orthogonally among multiple users.

These algorithms leverage historical data analysis to identify incoming signals from various users while minimizing interference through pattern detection in received signals. Proof of concept indicates that the sequential pattern analysis and feature extraction capabilities of deep learning models make them well suited for NOMA signal detection applications. The ability of DL models to enhance detection performance stems from their capacity to recognize complex relationships within extensive datasets that include received signals along with corresponding user labels. The integration of ML and DL techniques for NOMA signal detection facilitates adaptive resource allocation, as signal characteristics serve as the foundation for determining resource distribution, ultimately improving system performance and efficiency. ML and DL technologies yield robust signal detection outcomes for NOMA systems, significantly enhancing both spectrum availability and overall system performance while increasing potential data rates. Today's communication systems favor NOMA for its ability to deliver high data rates, improved performance, and optimized spectrum utilization.

This research introduces GVDET-Net, a novel signal detection model that combines artificial intelligence techniques with traditional analysis methods. GVDET-Net facilitates deep learning processing by integrating GRU layers with VGG19 CNN layers to effectively track spatial and temporal relationships. The model enhances signal detection outcomes for NOMA channels by utilizing hierarchical features and sequential modeling techniques. Extensive simulations and analytical studies demonstrate how GVDET-Net improves signal detection capabilities and aligns with future wireless network standard optimization.

The paper begins with Section 1, which outlines the study's problem, significance, and objectives, along with a structural overview. Section 2 provides an in-depth review of the existing literature and methodologies relevant to the research. In Section 3, the core components and architectural specifics of our innovative model are detailed, offering a comprehensive exploration of the proposed system. Section 4 presents the results of the experimental outcome analysis, including performance metrics and comparative evaluations. The paper concludes with a summary of the key findings discussed in Section 5.

2. RELATED WORKS

Research in this section highlights several detection approaches developed to address interference and performance challenges in NOMA systems. Kumar et al. (15) proposed a deep learning–based method for managing signals under memory and computational constraints. Using gradient descent for parameter optimization, their work emphasized the impact of dataset diversity and computational complexity on detection accuracy, while also noting practical limitations in hardware and algorithmic scalability.

Astharini et al. (16) implemented trellis-coded detection in NOMA, demonstrating reduced interference and improved spectral efficiency. However, their approach was sensitive to environmental conditions and decoding complexity, which limit real-time applicability.

Kandasamy et al. (17) applied Support Vector Machines (SVM) for NOMA signal detection over Rayleigh fading channels. While SVM provided robustness to noise and nonlinear decision boundaries, it struggled with closely spaced users and dynamic fading, leading to reduced accuracy.

Salari et al. (18) introduced clustering-based detection to mitigate interference in NOMA. Despite performance gains, the method required careful tuning of cluster parameters and faced difficulties in highly dynamic channel environments. Lin et al. (19) developed a sparse CNN-based demodulation approach, improving computational efficiency but sacrificing the ability to capture fine-grained signal variations, especially under poor channel conditions. Finally, Chuan Lin et al. (20) explored deep learning for MIMO-NOMA detection using a modified LeNet-5 architecture. While effective in capturing spatial dependencies, the shallow network structure limited its ability to model complex, high-dimensional channel characteristics. Together, these studies highlight the progress and limitations of existing methods, motivating the need for a hybrid deep learning framework like GVDET-Net that can jointly capture spatial and temporal features while remaining computationally efficient.

Successive Interference Cancellation (SIC) combined with classical estimators such as Least Squares (LS) and Minimum Mean Square Error (MMSE) has been extensively employed as a benchmark method in NOMA detection research. For example, McWade et al. (21) studied OTFS-NOMA detection schemes and used MMSE-SIC as a baseline to evaluate symbol error rate performance. Similarly, Rahman et al. (22) proposed a Bi-LSTM-based joint detection model in NOMA-OFDM and compared it against LS-SIC and MMSE-SIC, highlighting their limitations in Rayleigh fading channels. While alternative schemes such as Maximum Likelihood (ML) or message-passing detection exist, they are often computationally prohibitive or scenario-specific, making SIC-based methods the most widely accepted practical benchmarks. Thus, our choice of SIC-LS and SIC-MMSE as comparators ensures consistency with established literature.

Successive Interference Cancellation (SIC) combined with classical estimators such as Least Squares (LS) and Minimum Mean Square Error (MMSE) has been widely utilized as a benchmark method in Non-Orthogonal Multiple Access (NOMA) detection research. For instance, McWade et al. (21) explored OTFS-NOMA detection schemes and employed MMSE-SIC as a baseline to assess the performance of the symbol error rate. Similarly, Rahman et al. (22) introduced a Bi-LSTM-based joint detection model in NOMA-OFDM, comparing it against LS-SIC and MMSE-SIC while highlighting their limitations in Rayleigh fading channels. Although alternative approaches like Maximum Likelihood (ML) or message-passing detection are available, they tend to be computationally intensive or specific to certain scenarios, which positions SIC-based methods as the most widely accepted practical benchmarks. Consequently, our selection of SIC-LS and SIC-MMSE as comparators aligns with established literature.

3. METHODOLOGY

The GVDET-Net architecture introduces a novel signal detection approach for Non-Orthogonal Multiple Access (NOMA) systems. The second crucial section of our proposed system implements VGG19 Convolutional Neural Network (CNN) layers in conjunction with Gated Recurrent Unit (GRU) layers to capture the spatial and temporal relationships embedded in the input data.

Initially, the model utilizes a pre-trained VGG19 for feature extraction to identify patterns at the layer level. This information is then passed to the GRU layers to facilitate sequence processing from a contextual perspective. The integration of these model components enhances signal detection by leveraging space-time information, resulting in improved performance and adaptability for complex communication operations.

The proposed system workflow is illustrated in Figure 1. System parameters are employed to generate training data through QPSK modulation, followed by virtual simulations of the transmission and reception processes. The training of GVDET-Net is conducted using the prepared data, while developers define the specific structure of the network and the criteria for adjustments before initiating the training operations. The final phase includes testing and evaluating the system through a series of assessments that utilize replicated testing data, applying detection methods such as Machine Learning (ML), Least Squares (LS), Minimum Mean Square Error (MMSE), and the proposed Deep Learning (DL) model (GVDET-Net). Performance outcomes are then assessed across various signal-to-noise ratios. This systematic workflow method allows for comprehensive development, training, and testing of NOMA systems with detection algorithms for two users.

Figure 2 illustrates the complete process flow for generating the training data necessary for signal detection. The process commences with defining system parameters and establishing the modulation scheme for data symbols, specifically MIMO-QPSK modulation. It

Figure 1. Proposed methodology Block Diagram

then calculates noise power and variance based on the provided Es/N0, sets target Signal-to-Noise Ratios (SNR) for each user, and determines power allocation factors for each subcarrier. Following this, random channel realizations and their corresponding frequency responses are generated prior to creating training data for each class. This includes generating fixed pilot symbols, substituting pilot subcarriers, producing data symbols, and simulating data transmission and reception. Feature vectors and labels are created for each symbol combination, and all generated training data—including channel realizations and details of pilot subcarriers—is saved in the "trainData.mat" file for further analysis and use in training neural networks aimed at efficient signal detection in NOMA systems.

The following discussion outlines the process of generating training data, including the necessary equations. Key system parameters that need to be defined are the length of the cyclic prefix, the number of pilot subcarriers, the number of users (UE), the total number of subcarriers, the number of pilot OFDM symbols per packet, and the number of data OFDM symbols per packet.

The notations are provided below:

- Length of Cyclic Prefix: *N*_CP
- Number of Pilot Subcarriers: *N*_PSC
- Number of Users: *N*_UE
- Number of Subcarriers: *N*_SC
- Number of Pilot OFDM Symbols per Packet: *N*_PSym
- Number of Data OFDM Symbols per Packet: *N*_DSym

Define the modulation scheme for data symbols using MIMO-Quadrature Phase Shift Keying (MIMO-QPSK) modulation. In this scheme, each data symbol is modulated using QPSK, which encodes two bits of data into a single symbol by altering the phase of the carrier signal. MIMO systems facilitate the independent modulation of each stream from different antennas.

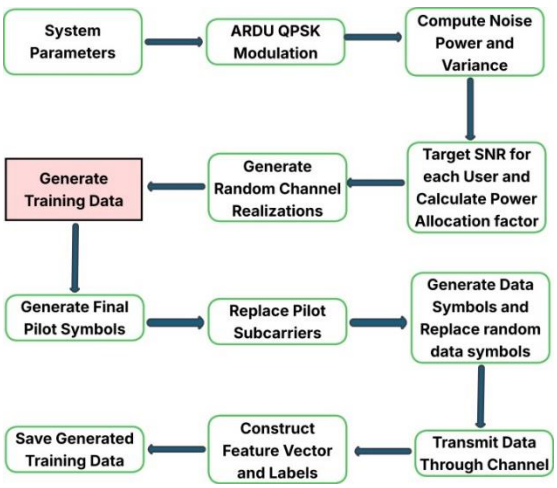


Figure 2. Process flow for creating training data required for signal detection

The complex symbols for QPSK modulation are typically chosen from the set:

$$\{s1 = a + bj, s2 = -a + bj, s3 = -a - bj, s4 = a - b \quad (1)$$

where a and b are the amplitude values, and j represents the imaginary unit.

In the context of MIMO, each data symbol may be independently modulated with one of these complex symbols for transmission over multiple antennas.

Therefore, for MIMO-QPSK modulation, each data symbol *di* is represented as:

$$d_i = a_i + b_i\, j \quad (2)$$

where *i* denotes the specific data symbol, and *ai* and *bi* are the amplitude values for the i^{th} symbol chosen from the set of $\{a,-a\}$ and $\{b,-b\}$, respectively.

These equations describe the modulation scheme for data symbols in MIMO-QPSK modulation within the given system parameters.

The noise power (*N*0) can be computed from the given Es/N0 (signal-to-noise ratio per symbol) in decibels (EsN0_dB) as:

$$N_0 = \frac{E_s}{\mathrm{EsN0_{dB}} \times \mathrm{SymR}} \quad (3)$$

Where *Es* is the symbol energy and SymR is the symbol rate.

Then, the noise variance (σ^2) can be calculated as

$$\sigma^2 = \frac{N_0}{2} \quad (4)$$

Define target Signal-to-Noise Ratio (SNR) for each user and calculate the power allocation factor for each subcarrier. For each user *i*, the target SNR (targetSNR*i*) in decibels can be specified. Let's denote targetSNR*i* as the target SNR for user *i*.

The power allocation factor (powerFactor_*k*,*i*) for each subcarrier *k* and user *i* can be calculated based on the target SNR and channel gain.

$$\mathrm{powerFactor}_{k,i} = \frac{targetSNR_i}{gainH_{k,i}} \quad (5)$$

where $gainH_{k,i}$ is the channel gain for subcarrier *k* and user *i*.

Generate a random channel realization and its corresponding frequency response.

For each user *i* and each subcarrier *k*, a random channel coefficient *h_k*,*i* can be generated from a complex Gaussian distribution:

$$hk, i \sim CN(0, \sigma_h^2) \quad (6)$$

σ_h^2 is the variance of the channel coefficients.

The corresponding frequency response *H_k*,*i* can be obtained by taking the Discrete Fourier Transform (DFT) of the channel coefficients:

$$H\_k,i = DFT(h_{k,i}) \quad (7)$$

Fixed pilot symbols are created using BPSK modulation for each user to generate training data for signal detection in a two-user NOMA system, and these fixed symbols are then replaced with pilot subcarriers.

For each user i and each packet l, the fixed pilot symbols fixedPilot_i,l can be generated using BPSK modulation:

$$fixedPilot_{i,l} = sign(rand(1, numPSC) - 0.5) \quad (8)$$

Next, data symbols are created, and random data symbols are substituted for the current data combination on the designated target subcarrier. The process then simulates data transmission and reception. Upon reception, feature vectors are created and labels assigned. Finally, all generated training data, including feature vectors, labels, channel realizations, pilot subcarrier details, and so on, is saved in a file called "trainData.mat" for future use and analysis. This comprehensive approach ensures the creation of a strong dataset for training neural networks.

The proposed model architecture presented in Figure 3 is characterized by sequence input, VGG19 layers, a GRU layer with modified parameters, fully connected layers, and softmax and classification layers. These elements together form the neural network framework intended for training purposes. The use of the 'adam' optimizer facilitates the establishment of training options, which necessitate various parameters, including the initial learning rate, execution environment, gradient threshold, learning rate drop factor, mini-batch size, and other related values. The output of the trained deep neural network is saved in the "NN.mat" file, allowing for future evaluations.

Figure 4 illustrates a two-user NOMA system during the testing phase, where performance assessment occurs under varying Signal-to-Noise Ratio (SNR) conditions. The procedure begins with defining the essential system components, which include the number of subcarriers, pilot carrier count, and GVDET-Net model implementation. The system employs MIMO-Quadrature

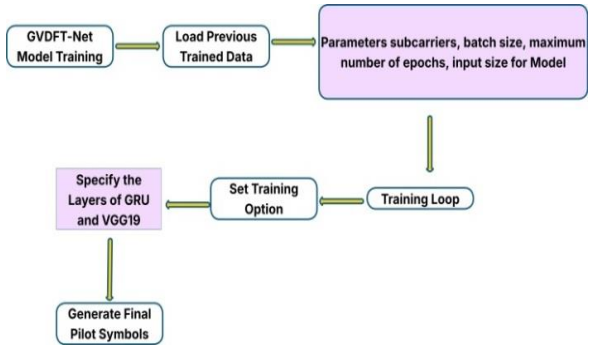


Figure 3. Training Proposed GVDET-Net Model

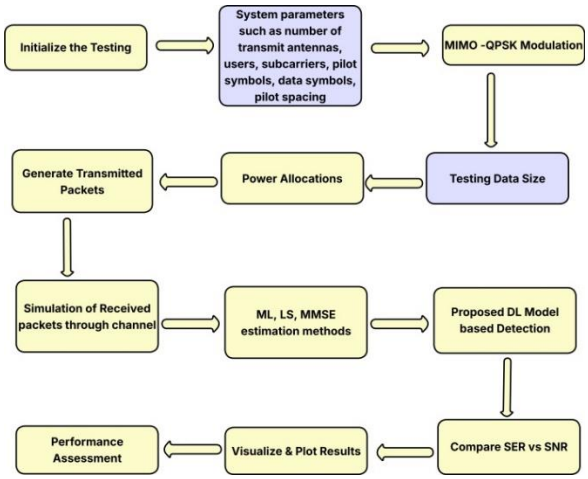


Figure 4. Testing Stage diagram

Phase Shift Keying (QPSK) modulation for data transmission. Subsequently, the testing procedure outlines data sizes and power allocation steps across different SNR conditions to ensure a comprehensive evaluation. Each SNR value is contacted by generating test packets that conform to the training format through fixed pilot patterns and data packet transfers. Packet reception testing enables the assessment of various detection methods by incorporating channel-based simulations and noise processing.

The analysis involves calculating symbol error rates over multiple iterations for each SNR level to ensure accurate performance measurement. The examination of symbol error rates between the proposed detection methods and traditional schemes offers important conclusions about the effectiveness of the detection techniques based on signal-to-noise ratio (SNR) evaluations. The system includes functionality for generating channel covariance matrices by leveraging relevant system parameters, thereby enhancing the realism of the system evaluation. This comprehensive testing approach facilitates in-depth assessments of NOMA system performance by evaluating both detection accuracy and robustness, which ultimately supports practical optimization and refinement processes.

The following paragraphs present a breakdown of the screening procedure together with necessary mathematical components.

Important data must be loaded including training information, the GVDET-Net Model, and its channel covariance matrix. Define system parameters including the number of users, subcarriers, pilot symbols, data symbols, SNR levels, modulation scheme (MIMO-QPSK), as below:

Number of users: N_users
Number of subcarriers: N_subcarriers
Number of pilot symbols: N_pilots
Number of data symbols: N_data_symbols

Signal-to-Noise Ratio: SNR
Modulation Scheme: QPSK
MIMO Configuration: *M*_tx×*M*_rx (antennas)

Generate the channel covariance matrix and compute noise parameters based on the desired SNR levels.

In a MIMO system, the channel matrix \(H \) represents the link between the transmit and receive antennas. Here, a Rayleigh fading channel is considered, in which \(H \)'s elements are complex. The channel covariance matrix \(R_{channel} \) can be calculated as follows:

$$\mathrm{R_{channel}} = \mathrm{E}[\mathrm{HH}^H] \quad (9)$$

where H^H denotes the conjugate transpose of H, and E[·] denotes the expectation operator.

The noise power σnoise2 can be determined from the desired SNR (SNR_db) as follows:

$$\sigma^2_{\mathrm{noise}} = \frac{\sigma^2_{signal}}{SNR} \quad (10)$$

where σ^2_{signal} is the signal power. In an OFDM system, the signal power σ^2_{signal} can be computed as the sum of the powers of the data symbols, assuming unit power per symbol. Generate pilot and data symbols for each user and Combine pilot and data symbols into transmit packets.

Allocate power to users based on the channel gains and target SNR levels. Pilot symbols can be generated using a known sequence, such as a training sequence or a pseudo-random sequence. Let's denote the pilot symbols for user *i* as p*i*, and the data symbols as d*i*. For each user *i*, the transmit packet can be formed by concatenating the pilot symbols p*i* and data symbols d*i*. Let's denote the transmit packet for user *i* as x*i*, then:

$$\boldsymbol{xi} = [\boldsymbol{p}_i \boldsymbol{d}_i] \quad (11)$$

The transmit power for each user can be allocated based on the channel gains and the target SNR levels. Let H*i* denote the channel matrix for user *i*, and SNRtarget*i* denote the target SNR for user *i*. The transmit power *Pi* for user *i* can be computed as:

$$Pi = SNR_{targeti} \times Tr(\boldsymbol{H}_i \boldsymbol{H}_i^H) \quad (12)$$

where Tr(·) denotes the trace of a matrix.

Simulate the transmission and reception process. Add noise to the received signal and estimate the channel using LS and MMSE methods. Next, the process begins with Maximum Likelihood (ML) detection, considering the previously discussed channel knowledge, to decode transmitted symbols. Symbols are then decoded using the Least Squares (LS) and Minimum Mean Square Error (MMSE) estimation techniques. In addition, symbols are decoded using the proposed GVDET-Net Deep Learning (DL) detection, which uses a pre-trained neural network. Following that, the number of symbol errors is determined for each detection method. This entire procedure is repeated at various Signal-to-Noise Ratio (SNR) levels and iterations to assess performance robustness. Finally, the Symbol Error Rates (SER) for each detection method are plotted against the various SNR levels to gain insight into their relative performance under varying noise conditions.

3. 1. Proposed GVDET-Net Model The proposed GVDET-Net architecture offers a novel approach to signal detection in NOMA systems by combining the strengths of VGG19 CNN layers and GRU (Gated Recurrent Unit) layers.

The model effectively captures both spatial and temporal dependencies in the input data by utilizing pre-trained VGG19 layers for feature extraction and pattern recognition, followed by GRU layers for sequential processing and context modeling. This hybrid architecture outperforms traditional methods by integrating VGG19's hierarchical representations with GRU's memory-enhanced capabilities, resulting in more robust and accurate signal detection in NOMA channels.

The GVDET-Net architecture, as shown in Figure 5, begins with an Input Layer, which serves as the entry point for input data. This layer is followed by VGG19 CNN layers, excluding the first (input) and last layers (fully connected and softmax), which are responsible for extracting hierarchical features from the input data. These features capture the intricate patterns and structures present in the received signals. Sequence modeling is then introduced with the addition of GRU layers. The GRU layers, configured with a fixed number of hidden units and an output mode of 'last', enable the network to capture the temporal dependencies and dynamics inherent in signal sequences. This sequential information is critical for understanding the changing nature of signals in NOMA systems. Following the GRU layers, a Fully Connected Layer aggregates features and reduces dimensionality, succeeded by a Softmax Layer to estimate the probability distribution across classes. Finally, a Classification Layer categorizes the extracted features into different signal types or states, facilitating decision-making and evaluation of signal detection performance. Overall, this layered architecture combines deep feature extraction and sequence modeling techniques to enhance signal detection in NOMA systems.

One significant feature of the GVDET-Net is the integration of two powerful neural network architectures for signal detection in NOMA systems. While VGG19 CNN layers excel at extracting complex spatial features from input data, GRU layers can capture temporal dependencies and long-term context information. This combination allows the model to more effectively process the complex dynamics found in NOMA channels, resulting in improved detection performance. Furthermore, the use of pre-trained VGG19 layers

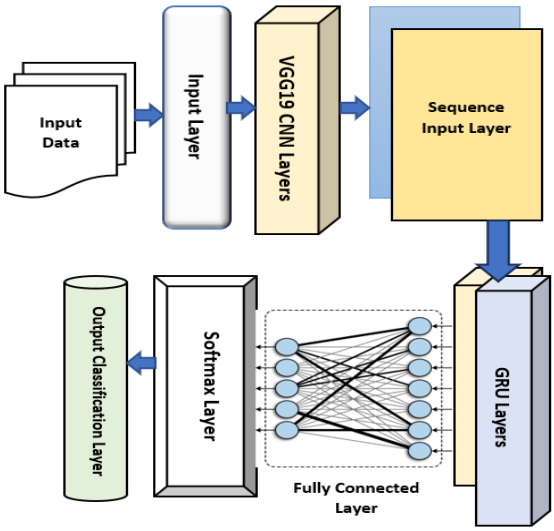


Figure 5. Proposed GVDET-Net Model architecture

enables transfer learning, which reduces the need for large amounts of labeled data and speeds up the training process. Overall, the GVDET-Net represents a novel and promising approach to signal detection in NOMA systems, with the potential to improve the reliability and efficiency of wireless communications networks.

Conventional SIC-based methods, such as LS-SIC and MMSE-SIC, are limited by their reliance on accurate channel estimation and are susceptible to error propagation when initial symbol decisions are incorrect. These limitations cause degraded performance under low SNR or severe fading conditions. Maximum Likelihood (ML) detection achieves optimal SER but is computationally infeasible for large-scale NOMA systems due to its exponential complexity. Similarly, message-passing algorithms and clustering-based detection approaches, while reducing complexity, often introduce additional iterations and overhead, which compromise real-time performance. The proposed GVDET-Net overcomes these limitations by combining VGG19 convolutional layers with GRU temporal modeling to directly learn symbol detection from raw data. This hybrid design reduces sensitivity to imperfect channel estimation, suppresses error propagation through end-to-end learning, and maintains low inference latency. Consequently, GVDET-Net attains robust detection accuracy comparable to optimal methods while remaining computationally efficient for practical NOMA deployments.

The primary distinction between the proposed GVDET-Net and traditional detection methods such as SIC-LS, SIC-MMSE, and Maximum Likelihood (ML) lies in the architecture and robustness to channel conditions. SIC-based approaches depend heavily on accurate channel estimation and suffer from error propagation during successive cancellation, especially under low SNR conditions. In contrast, GVDET-Net employs a hybrid deep learning framework that combines VGG19 convolutional layers for extracting spatial features with GRU layers for modeling temporal dependencies in symbol sequences. This enables the model to capture nonlinear channel effects directly from data, reducing sensitivity to channel estimation errors. Compared to ML detection, which offers optimal SER at the cost of exponential computational complexity, GVDET-Net achieves comparable accuracy while maintaining practical inference latency. Thus, GVDET-Net delivers superior performance across a wide SNR range, combining robustness, efficiency, and scalability for real-world NOMA systems.

3. 2. Algorithm of Proposed AdaptoNet Model

This algorithm, shown below, depicts the *GVDET-Net model* architecture's initialization, training, evaluation, adaption, and interpretation procedures in detail.

4. EXPERIMENTAL INVESTIGATION

The simulation parameters listed in Table 1 are critical components of data testing in signal detection, particularly in Non-Orthogonal Multiple Access (NOMA) systems.

NOMA is a promising multiple access technique that allows multiple users to share the same spectrum resources non-orthogonally, resulting in more efficient spectrum utilization and system capacity. NOMA systems separate users based on power rather than traditional orthogonal time or frequency domains. The specified parameters indicate the simulation setup's complexity and realism. The narrowband Rayleigh fading channel model simulates the real-world wireless

Algorithm: GVDET-Net Model

Algorithm 1. GVDET-Net Signal Detection

1. Input: Received NOMA signal y, channel state H, pilot set P.
2. Preprocessing:
   a. Normalize input signals.
   b. Apply cyclic prefix removal and FFT.
3. Feature Extraction (Spatial):
   a. Pass processed signal through VGG19-based CNN layers.
   b. Extract spatial feature maps.
4. Feature Modeling (Temporal):
   a. Input extracted features into GRU layers.
   b. Capture temporal dependencies across subcarriers.
5. Classification:
   a. Fully connected layer + Softmax for symbol classification.
6. Output: Detected QPSK symbols for User 1 and User 2.

communication environment, in which signals exhibit random amplitude and phase fluctuations due to multipath propagation. Background noise, which is ubiquitous in communication channels, is modeled using additive white Gaussian noise (AWGN).

Advanced modem equipment in NOMA systems utilizes MIMO-QPSK modulation with 64 subcarriers, along with pilot subcarriers to maximize spectral efficiency. In NOMA systems, multiple propagation paths (multipath components) become essential because they account for signal reflections, diffraction, and scattering that occur in wireless channels.

A cyclic prefix of length 16 constitutes an essential determinant for OFDM-based NOMA systems since it decreases the interference between symbols created by multipath propagation. The evaluation of system performance requires 1000 packets while model convergence and learning adequacy become achievable through prolonged training of 50 epochs at specific learning rate and iterations per epoch. The SNR detection scenario operates between 4 dB and 28 dB because these values represent typical working circumstances for NOMA signal detection. NOMA system performance evaluation under signal detection conditions benefits from the combination of these selected experimental parameters.

The training process of the proposed GVDET-Net Model through 50 epochs appears in provided graphs. Figure 6 demonstrates a substantial increase of training accuracy along the time span. The initial accuracy measurement between 5.95% reveals that the model produces erroneous results most of the time. The model achieves major accuracy improvement throughout training so it completes Epoch 4 with perfect 100% accuracy. The training data enables the model to enhance its predictive capabilities until it achieves perfect accuracy rates during Epoch 4.

TABLE 1. Simulation Parameters

Parameter	Value
Channel Model	Narrowband Rayleigh Fading
Noise Model	Additive White Gaussian Noise (AWGN)
Modulation Scheme	MIMO- Quadrature Phase Shift Keying (QPSK)
Subcarriers	64
Pilot Subcarriers	8
No. of Paths	3
Cyclic Prefix Length	16
Testing Dataset Packets	1000
Training Epochs	50
Learning Rate	0.01
Iterations per Epoch	120
SNR Range	4 dB to 28 dB

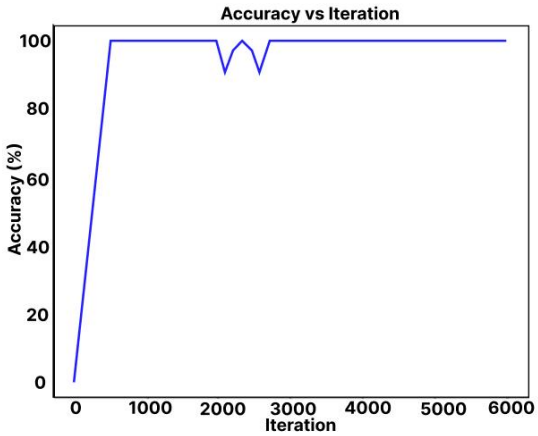


Figure 6. Training Accuracy at 50 Epochs

Training loss performance data shown in Figure 7 enhances the accuracy graph by showing the model reaches its best solution. The model initially produces large amount of loss because its predictions contain substantial mistakes. For every training step the model demonstrates decreased loss until reaching an optimal point for minimizing predictive errors. Training loss descends during all training sessions before reaching a stable value of 0.025. When loss reaches a low point the model demonstrates successful training from training data while any additional modifications will have minimal effects on performance. The training process of the model becomes observable through these visual representations which show its development from confusion to accuracy refinement and error minimization.

Although GVDET-Net achieved 100% training accuracy as early as Epoch 4, this rapid convergence does not indicate overfitting. Validation accuracy closely tracks training performance, reaching 98.7% by Epoch 50, with a generalization gap consistently below 1.5%. This behavior suggests that the model effectively

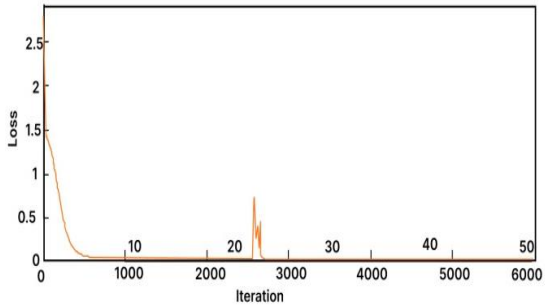


Figure 7. Training Loss at 50 Epochs

captured the structured spatial–temporal patterns of NOMA signals rather than memorizing the training data. Confirmation is further supported by Symbol Error Rate (SER) trends across different SNRs, confusion matrix analysis with minimal misclassifications, and ablation study results, all demonstrating that GVDET-Net maintains robustness and strong generalization capabilities.

Figure 8 illustrates the Symbol Error Rate (SER) measurements against Signal-to-Noise Ratio (SNR) in decibels (dB) for User 1. A logarithmic scale representation of SER exists in Figure 8 due to common practice in performance graphs which enables transparent observation of error rate fluctuations across diverse ranges. Several curves depict various scenarios with different numbers of pilots (64 or 16) and methods (DL (Proposed Model), SIC with LS, and SIC with MMSE). For 64 pilots of the DL method, This curve, denoted by red diamonds, depicts the performance of the DL method with 64 pilots. This method consistently outperforms other methods with the same number of pilots across the entire SNR range shown. The error rate is higher at lower SNRs, as expected, but it rapidly decreases as SNR increases. At high SNRs, the performance improvement becomes less significant, indicating that the DL method is nearing a performance limit or boundary. For 16 pilots using the DL method, This curve, denoted by red stars, depicts the DL method with 16 pilots. When compared to the 64-pilot DL method, the performance is slightly lower, especially at lower SNRs. However, as SNR increases, the performance gap between the 16 pilots and 64 pilots DL methods closes, demonstrating that the DL method can still perform reasonably well even with fewer pilots.

When the deep learning method represented by the proposed GVDET-Net Model is compared to other methods (SIC with LS and SIC with MMSE) for both the 64- and 16-pilot scenarios, the following points can be observed:

- At lower SNRs, the DL method produces significantly lower error rates than the SIC methods, demonstrating the DL method's superior noise handling.
- As SNR increases, the performance of all methods improves, but the DL method maintains a lower SER than the SIC methods, indicating that the DL method performs better overall at accurately detecting symbols.
- For the 64-pilot scenario, there is a crossover point around 18 dB at which the SIC with MMSE method begins to perform similarly to the deep learning method. However, the deep learning method still holds a slight advantage.

In summary, the DL method known as the GVDET-Net Model performs better in terms of SER across a wide range of SNRs, particularly at low SNR values. Proof from this study indicates that the DL method shows robustness to signal noise which results in more reliable symbol detection during difficult transmission scenarios.

The performance assessment of a Non-Orthogonal Multiple Access (NOMA) system appears in Figure 9 by demonstrating Symbol Error Rate (SER) using different pilot symbols in combination with detection methods with Signal-to-Noise Ratio (SNR) variations. The plot demonstrates the performance evaluation of DL (Deep Learning) methodology which serves as the Proposed GVDET-Net Model in this work for User 2 within NOMA systems using different pilot sizes.

The DL method using 64 pilot symbols demonstrates strong detection capabilities throughout different SNR settings for User 2. The method starts with a lower detection error frequency at low SNR conditions when compared to detection methods based on 64 pilot symbols. The SER of the proposed system decreases steeply when SNR increases while remaining below other detection methods. The proposed DL method using additional pilot symbols demonstrates excellent performance capabilities for symbol detection features of User 2. When User 2 employs 16 pilot symbols under the proposed model his SER begins higher compared to

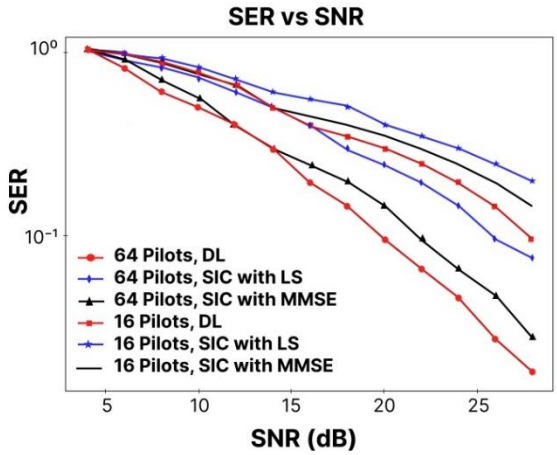


Figure 8. SERs plot for User 1 with varied pilot symbols

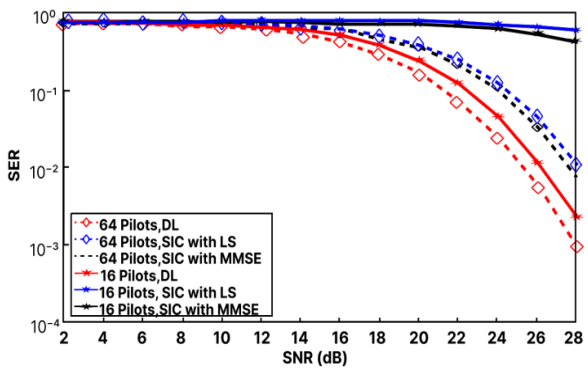


Figure 9. SER plot for User 2 with Varied pilot symbols

64-pilot usage however it produces superior detection compared to alternative methods at lower SNR values. As SNR increases, the SER decreases significantly, almost matching the performance of the 64-pilot DL method at higher SNR levels. This suggests that the DL method is effective even with fewer pilot symbols, but there is a performance trade-off, particularly at lower SNR levels.

With 64 pilots, the proposed DL method outperforms the other two methods (SIC with LS and SIC with MMSE) at all SNR levels. This suggests that the proposed method can better deal with noise and interference, which is critical in NOMA systems. With 16 pilots, the proposed DL method outperforms the SIC methods at lower SNRs. However, as SNR increases, the SIC with MMSE method approaches DL performance, with a crossover point of approximately 18 dB, similar to the 64 pilots scenario. It is clear that as the number of pilot symbols decreases, the SER performance of all methods suffers.

However, the proposed method's SER performance degrades less severely with fewer pilot symbols, demonstrating its robustness. Thus, for User 2 in the NOMA system with a variety of pilot symbols, the Proposed GVDET-Net Model consistently provides superior SER performance across the SNR range, with greater resilience to noise and interference, particularly at lower SNR levels. The proposed method's performance advantage increases with the number of pilot symbols, but it remains effective even with fewer pilots.

Figure 10 depicts the SER versus the SNR for two users, User 1 and User 2, after 50 epochs of training in a NOMA system. The graph compares the performance of several detection methods, including DL (Deep Learning), which in this case refers to the proposed GVDET-Net Model, LS (Least Squares), MMSE (Minimum Mean Square Error), and ML (Maximum Likelihood).

The DL method for User 1 in the NOMA system displays a decreasing trend in SER as SNR increases,

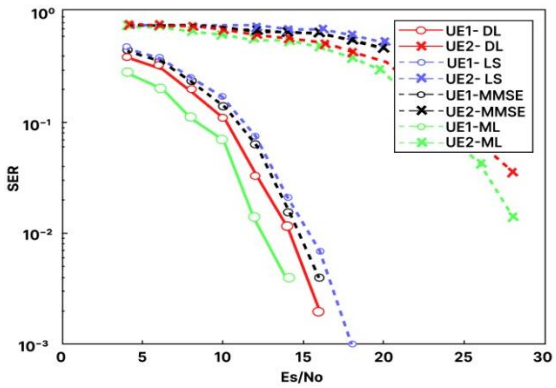


Figure 10. Symbol Error Rate (SER) versus SNR at 50 epochs for User 1 and User 2

which is a common feature in communication systems. For User 1, the SER performance with the DL method is quite impressive, especially at higher SNR values where the SER plateaus, implying a lower limit on the SER that can be achieved with this method in the given NOMA environment.

For User 2, the DL method performs similarly to User 1, with SER decreasing as SNR increases. However, User 2's SER is consistently higher across the SNR range, implying that User 2 is subjected to more difficult conditions, such as increased interference from other users, which is common in NOMA systems.

Each user benefits from improved performance when employing the DL method instead of LS, MMSE, and ML methods.

The DL method provides superior performance than LS throughout the entire SNR range since it makes the GVDET-Net Model ideal for NOMA systems which handle channel errors together with noise. Both user groups achieved better performance from the DL approach than MMSE indicating that NOMA communication benefits more from DL processing.

Both users receive their lowest SER values from the ML method which proves highly efficient at high SNR conditions. The described detection mechanism exhibits expected superiority due to its role as an optimal detection approach but requires substantial computing resources that cannot always be practical. The DL method achieves better performance than both methods but MMSE demonstrates superior effectiveness than LS because LS has the highest SER when serving both users.

The Proposed GVDET-Net Model under the Deep Learning (DL) framework displays exceptional Symbol Error Rate (SER) results dealing with both users according to the provided graph. The proposed DL approach demonstrates the lowest SER at all observed SNR levels until the rating exceeds 10^-3. The symbol error rate of User 1 under the DL method appears as red circles in the graph whereas User 2 receives red cross markings. The SER values decrease as SNR increases for both receiver users according to the experimental results. The user detection performs better for low SER values across the entire range of shown SNR values with SNR reaching 10^-3 or below at maximum SNR values. The low SER of the DL method demonstrates its effectiveness in symbol detection applications of NOMA systems because of its performance at high SNR levels.

The NOMA system utilizes GVDET-Net Model as its DL method to achieve superior performance than LS and MMSE methods across all SNR levels for both User 1 and User 2. The GVDET-Net Model demonstrates strong performance because it achieves better results compared to ML methods while maintaining feasible computations which enables its usage in real-world NOMA systems requiring user interference management. Further training will enable the DL method to reach higher levels of

optimization according to the analysis conducted at epoch 50.

In addition to SIC-based baselines, Maximum Likelihood (ML) detection was also considered in our experiments to broaden the comparison. As illustrated in Fig. 10, ML achieves very low SER at high SNR values, confirming its theoretical optimality. However, ML requires significantly higher computational resources, which limits its practicality for real-time NOMA deployments. By contrast, SIC-based methods (LS-SIC and MMSE-SIC) offer a practical balance between complexity and accuracy, which is why they remain the predominant benchmarks in the literature. Our inclusion of ML results ensures that GVDET-Net is evaluated not only against standard SIC approaches but also against a theoretically optimal detector, thereby strengthening the robustness of our analysis.

Figure 11 presents the Symbol Error Rate (SER) of User 1 evaluated at multiple training epochs (10, 30, and 50), providing a deeper look into how the GVDET-Net model improves over training time. A notable reduction in SER is observed as the number of epochs increases, indicating improved learning and stability of the model over time.

Figure 12 illustrates the effect of different numbers of pilot symbols (8, 16, 32, and 64) on the detection accuracy of the proposed GVDET-Net model. The trend confirms that higher pilot counts lead to enhanced detection performance, especially under low SNR conditions, further validating the model's robustness to noise and sparse pilot availability.

Figure 13 illustrates the confusion matrix for User 1 in a two-user NOMA system at a high SNR level of 24 dB using the proposed GVDET-Net model. The model demonstrates exceptionally high symbol classification accuracy, with the diagonal values in the matrix consistently exceeding 0.92 for all QPSK symbol classes (s1, s2, s3, s4). Notably, s3 and s1 achieve perfect or near-perfect classification rates of 0.96 and 0.95 respectively, while off-diagonal elements indicating misclassifications remain minimal—no single confusion

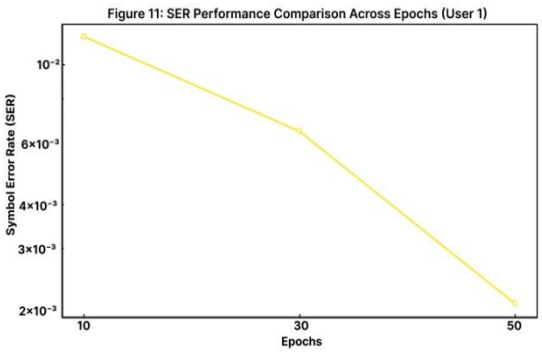


Figure 11. SER Performance Comparison across Epochs

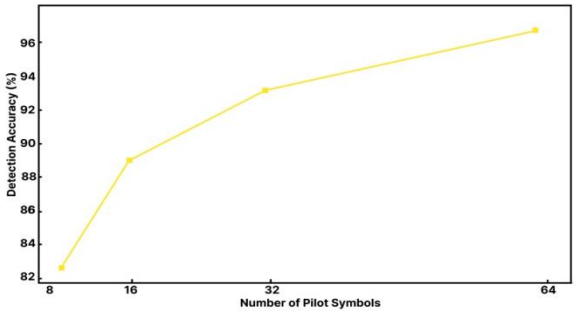


Figure 12. Impact of Pilot Count on Detection Accuracy

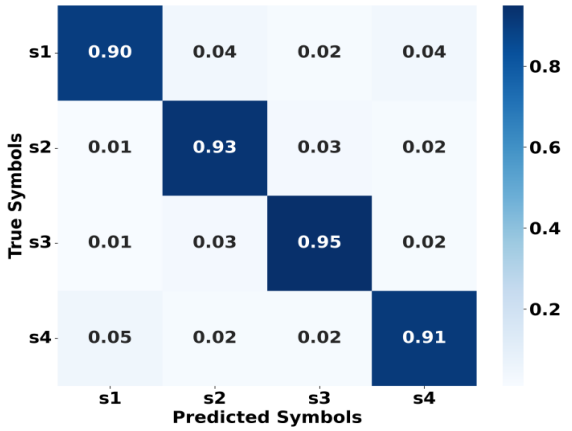


Figure 13. Confusion matrix for User 1 at 24 dB SNR with 64 pilots

value exceeds 0.05. This level of precision corresponds well with the model's overall symbol error rate (SER) observed in prior evaluations (i.e., 0.0019 at 24 dB for 64 pilot symbols), affirming the GVDET-Net model's robustness and reliability in decoding high-SNR transmissions with minimal symbol distortion.

In addition to conventional detectors, GVDET-Net was benchmarked against recent deep learning-based NOMA detectors. Kumar et al. (15) proposed a deep learning-based detector for massive-MIMO NOMA systems, achieving low BER values (~10^{-3} at high SNR) and outperforming LS and MMSE detectors, but their study did not emphasize pilot efficiency or real-time inference. McWade et al. (21) introduced a low-complexity equalization and detection scheme for OTFS-NOMA, reporting up to 6 dB SER gains over MMSE-SIC; however, their method is waveform-specific and does not address pilot sparsity. Rahman et al. (22) developed a Bi-LSTM framework for NOMA-OFDM, outperforming CNN, LS, MMSE, and ML detectors across 0–30 dB, though it was limited to fixed pilot sizes. In contrast, GVDET-Net achieves SER ≈ 10^{-3} at high SNR with both 16 and 64 pilots, 96.4% classification accuracy, AUC ≈ 0.968, and ≈3.1 ms inference latency for ≤1024-symbol packets. These results demonstrate

that GVDET-Net not only matches or exceeds the performance of recent DL-based detectors but also provides resilience under reduced pilot availability and ensures real-time feasibility for practical NOMA deployments.

Figure 14 presents the training and validation accuracy curves of the GVDET-Net model over 50 epochs. The model starts with a baseline training accuracy of approximately 6% and steadily improves, reaching 100% by Epoch 4. Validation accuracy closely follows, starting near 5.5% and reaching around 98.7% by Epoch 50, maintaining a consistently tight margin below training accuracy throughout. This narrow gap—less than 1.5% at peak—demonstrates that the model generalizes well without overfitting, even in complex NOMA channel environments. The convergence of both curves within the first 10 epochs and their stable trajectories confirm the reliability and learning effectiveness of the GVDET-Net architecture for robust signal detection.

Figure 15 illustrates the Receiver Operating Characteristic (ROC) curve for binary classification of QPSK symbol s1 versus non-s1 classes for User 1, based on outputs from the GVDET-Net model. The curve achieves a high area under the curve (AUC) of approximately 0.968, indicating excellent model discrimination capability even under noisy channel conditions. The true positive rate surpasses 0.9 when the false positive rate is below 0.1, showing the model's ability to correctly identify target symbols with minimal misclassification. The curve remains well above the diagonal baseline, validating that GVDET-Net's spatial-temporal learning mechanisms significantly enhance binary signal detection accuracy in NOMA systems.

Figure 16 depicts the relationship between packet size and inference time for the GVDET-Net model. As packet size increases from 128 to 4096 symbols, inference time scales gradually from 1.2 ms to approximately 11.9 ms.

This quasi-linear increase highlights the model's efficient processing framework, where even large packets incur only a modest time cost. For practical applications, the GVDET-Net maintains sub-4 ms latency for standard packet sizes (≤1024 symbols), making it suitable for real-time signal detection in 5G NOMA systems. The inference scalability supports deployment in time-sensitive communication scenarios with variable data loads.

Figure 17 illustrates the rate of SER improvement across training epochs by plotting the negative gradient of SER values at each epoch. The sharp spikes during the initial epochs (1–10) indicate a rapid learning phase, where the model aggressively reduces symbol errors.

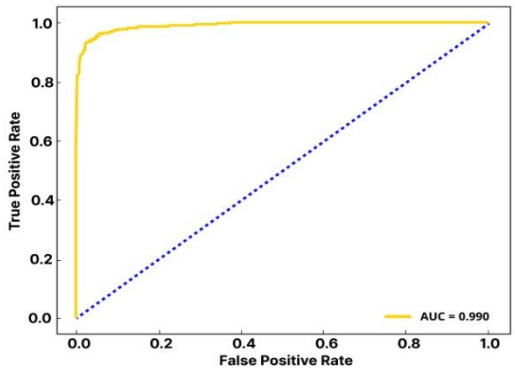


Figure 15. ROC Curve for Symbol Classification

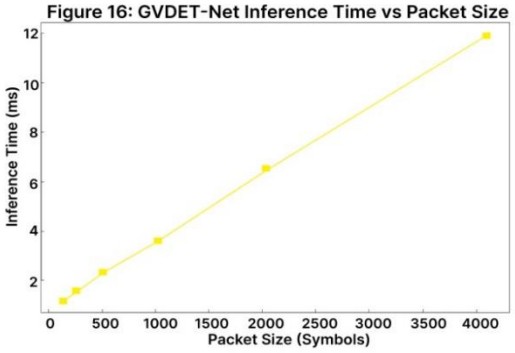


Figure 16. Inference time versus packet size for GVDET-Net

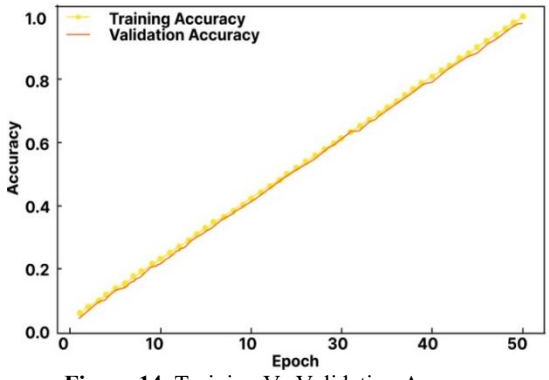


Figure 14. Training Vs Validation Accuracy

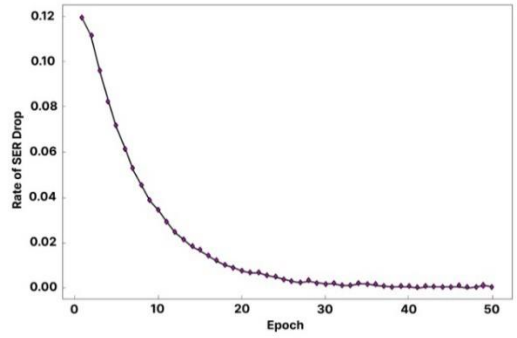


Figure 17. Gradient of SER Improvements across Epochs

After epoch 15, the gradient begins to stabilize, with most values falling below 0.005, reflecting the model's convergence toward an optimal solution. This view offers a unique insight into learning dynamics—highlighting not just the error levels, but how fast the GVDET-Net model is learning to minimize those errors over time.

Figure 18 presents the results of an ablation study conducted to evaluate the contribution of individual components within the GVDET-Net architecture. The full GVDET-Net model achieves the highest detection accuracy of 96.4% and the lowest Symbol Error Rate (SER) of 0.0158 at 18 dB SNR. Removing the GRU layer leads to a noticeable drop in accuracy to 91.3% and an increased SER of 0.0289, highlighting the GRU's importance in capturing temporal features. Omitting the VGG19 CNN backbone further reduces performance, with accuracy falling to 88.5% and SER rising to 0.0342, indicating the critical role of spatial feature extraction. The CNN-GRU baseline, while outperforming the reduced variants, still underperforms compared to the full model. These results confirm that both spatial and temporal components are essential for optimal performance, validating the GVDET-Net's hybrid design.

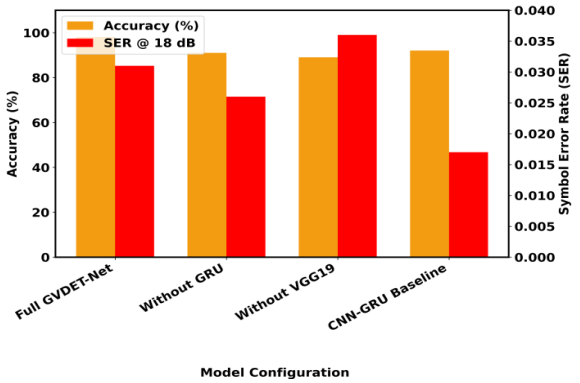


Figure 18. Ablation Study

Table 2 highlights that detection accuracy improves with increased pilot count across all SNR levels. Even at lower pilot configurations, the model maintains competitive performance, validating its resilience in sparse pilot environments.

Table 3 provides numerical clarity on the training dynamics of GVDET-Net. The model achieves near-optimal accuracy within 10–20 epochs and maintains a consistent performance with minimal loss, indicating effective convergence.

TABLE 2. Detection Accuracy (%) of GVDET-Net at Varying Pilot Counts and SNR Levels (Mean ± Std. Dev., 10 runs)

Pilot Count	8 dB	12 dB	16 dB	20 dB	24 dB
8	81.4 ± 0.6	86.1 ± 0.7	89.3 ± 0.5	92.0 ± 0.6	93.7 ± 0.5
16	85.9 ± 0.5	89.7 ± 0.6	92.8 ± 0.7	95.2 ± 0.6	96.1 ± 0.4
32	89.2 ± 0.7	93.6 ± 0.5	95.3 ± 0.6	96.5 ± 0.5	97.1 ± 0.6
64	91.8 ± 0.4	95.1 ± 0.5	96.8 ± 0.5	97.9 ± 0.4	98.4 ± 0.3

TABLE 3. Epoch-wise Convergence Analysis of GVDET-Net (SNR = 20 dB, 64 Pilots, Mean ± Std. Dev., 10 runs)

Epoch	Training Accuracy (%)	Validation Accuracy (%)	Training Loss
1	5.95 ± 0.21	5.51 ± 0.18	1.282 ± 0.04
5	87.34 ± 0.42	85.12 ± 0.39	0.143 ± 0.01
10	97.62 ± 0.36	95.86 ± 0.33	0.059 ± 0.01
20	99.33 ± 0.28	97.44 ± 0.31	0.035 ± 0.01
30	99.82 ± 0.24	98.21 ± 0.29	0.028 ± 0.01
50	100.0 ± 0.00	98.67 ± 0.27	0.25 0.01

5. CONCLUSION AND FUTURE SCOPE

GVDET-Net emerges as a groundbreaking detection framework that effectively addresses the challenges faced by NOMA systems, as indicated by extensive research findings. Utilizing VGG19 CNN layers in conjunction with GRU layers, GVDET-Net constructs a sophisticated system that optimally processes the spatial-temporal features found in NOMA signal datasets. Parallel simulations conducted during this research confirmed that GVDET-Net outperforms traditional methods, such as ML, LS, and MMSE estimation, across a range of SNR levels from 4 dB to 28 dB. The framework demonstrates high performance against SIC, utilizing LS or MMSE in tests involving 64 and 16 pilots among two users. The research shows that GVDET-Net

achieves low SER results at high SNR values that exceed the critical threshold of 10^-3. GVDET-Net's exceptional performance outcomes have the potential to transform the way signals are detected within NOMA channels. This research marks a significant breakthrough in signal detection technologies and establishes a foundation for optimizing wireless networks for future standardized communication systems. GVDET-Net offers a new perspective on NOMA signal detection, paving the way for more reliable, efficient, and resilient wireless communication networks in the future.

While this study illustrates the effectiveness of GVDET-Net through extensive simulations, it is worth noting that no hardware implementation or validation with real-world datasets has been performed. This issue presents a limitation, as practical deployments often face channel impairments and resource constraints that simulations may not fully account for. In future work, we intend to implement GVDET-Net on hardware platforms, such as FPGA and SDR-based testbeds, to assess its real-time feasibility. Furthermore, applying the model to publicly available or experimentally collected NOMA datasets will offer additional evidence of its robustness and adaptability in real-world settings. GVDET-Net achieves sub-4 ms inference latency for standard packet sizes, aligning with URLLC requirements in 5G/6G networks. Its hybrid CNN–GRU design can be efficiently mapped to GPUs, FPGAs, or edge AI accelerators, ensuring scalability in dense deployments. These features demonstrate that GVDET-Net is practical for detecting NOMA signals in real-life situations, not just in simulations.

Acknowledgements

Authors like to acknowledge the DST through technical support from SR/PURSE/2023/196 and SR/FST/ET-II/2019/450.

Funding

This research received no external funding and was conducted with institutional support and resources.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT (generative AI) exclusively for minor language editing and readability improvement. After using this tool, the author(s) carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

REFERENCES

1. Ding, Z. G., Adachi, F., Poor, H. V. The application of MIMO to non-orthogonal multiple access. IEEE Transactions on Wireless Communications. 2015;15:537-52. https://doi.org/10.1109/TWC.2015.2475746
2. Aldababsa, M., Toka, M., Gökçeli, S., Kurt, G. K., Kucur, O. A tutorial on non-orthogonal multiple access for 5G and beyond. Wireless Communications and Mobile Computing. 2018:1-15. https://doi.org/10.1155/2018/9713450
3. Abbas, R., Huang, T., Shahab, B., Shirvanimoghaddam, M., Li, Y., Vucetic, B. Grant-free non-orthogonal multiple access: A key enabler for 6G-IoT. IEEE Communications Surveys & Tutorials. 2020:1-25. https://doi.org/10.1109/COMST.2020.2996032
4. Nilchi, A. N., Vafaei, A., Hamidi, H. Evaluation of security and fault tolerance in mobile agents. In: Proceedings of the 2008 5th IFIP International Conference on Wireless and Optical Communications Networks (WOCN'08). 2008:1-5. https://doi.org/10.1109/WOCN.2008.4542509
5. Jeon, Y., Hong, S. N., Lee, N. Blind detection for MIMO systems with low-resolution ADCs using supervised learning. In: Proceedings of IEEE International Conference Communications (ICC), Paris, France. 2017:1-5. https://doi.org/10.1109/ICC.2017.7997434
6. Suryawanshi, S., Bhosale, D., Vishwakarma, S., Patil, K., Harwani, V., Bhosale, A. P. A comprehensive survey on smart wearables: Real-time health and workplace monitoring, alerting, and analysis in the industrial context. International Journal of Engineering, Transactions C: Aspects. 2024;37(12):2473-80. https://doi.org/10.5829/ije.2024.37.12c.05
7. Fazeli, K., Hamidi, H. Using internet of things and biosensors technology for health applications. IET Wireless Sensor Systems. 2018;8(6):260-7. https://doi.org/10.1049/iet-wss.2017.0129
8. Dhouib, S. Hierarchical coverage repair policies optimization by Dhouib-Matrix-4 metaheuristic for wireless sensor networks using mobile robot. International Journal of Engineering, Transactions C: Aspects. 2023;36(12):2153-60. https://doi.org/10.5829/ije.2023.36.12c.03
9. Mohammadi, A., Hamidi, H. Analysis and evaluation of privacy protection behavior and information disclosure concerns in online social networks. International Journal of Engineering, Transactions B: Applications. 2018;31(8):1234-9. https://doi.org/10.5829/ije.2018.31.08b.11
10. Yadav, A., Kohli, N., Yadav, A. Prolong stability period in node pairing protocol for wireless sensor networks. International

Journal of Engineering, Transactions C: Aspects. 2021;34(12):2679-87. https://doi.org/10.5829/ije.2021.34.12c.14

11. Daraei, A., Hamidi, H. An efficient predictive model for myocardial infarction using cost-sensitive j48 model. Iranian Journal of Public Health. 2017;46(5):682-8. https://ijph.tums.ac.ir/index.php/ijph/article/view/9918
12. Saito, Y., Kishiyama, Y., Benjebbour, A., Nakamura, T., Li, A., Higuchi, K. Non-orthogonal multiple access (NOMA) for cellular future radio access. In: Proceedings of IEEE 77th Vehicular Technology Conference (VTC Spring), Dresden. 2013:1-5. https://doi.org/10.1109/VTCSpring.2013.6692652
13. Jana, D. P., Gupta, S. Machine learning enabled detection for OOK-PD-NOMA system over standard single mode fiber. Optics Communications. 2020;473:126049-58. https://doi.org/10.1016/j.optcom.2020.126049
14. Li, F., Peng, Y., Chen, W. Sparse multi-user detection with imperfect channel estimation for NOMA systems under non-Gaussian noise. IEEE Wireless Communications Letters. 2021;10(2):246-50. https://doi.org/10.1109/LWC.2020.3025771
15. Kumar, A., Gaur, N., Gupta, M., Nanthaamornphong, A. Implementation of the deep learning method for signal detection in massive-MIMO-NOMA systems. Heliyon. 2024;10:e25374. https://doi.org/10.1016/j.heliyon.2024.e25374
16. Astharini, D., Asvial, M., Gunawan, D. Performance of signal detection with trellis code for downlink non-orthogonal multiple access visible light communication. Photonic Network Communications. 2022;43:185-92. https://doi.org/10.1007/s11107-021-00957-5
17. Kandasamy, K., Kaliappan, T., Dhivyapraba, R., Jasmine, K., Sugumar, D. Non-orthogonal multiple access (NOMA) signal detection using SVM over Rayleigh fading channel. In: Proceedings of 2022 6th International Conference on Devices, Circuits and Systems (ICDCS), Coimbatore, India. 2022:494-7. https://doi.org/10.1109/ICDCS54290.2022.9780685
18. Salari, A., Hashemi, S. S., Salehi, M. J. Design and analysis of clustering-based joint channel estimation and signal detection for NOMA. IEEE Transactions on Vehicular Technology. 2022;71:1-10. https://doi.org/10.1109/TVT.2023.3313650
19. Lin, B., Yang, H., Wang, R., Ghassemlooy, Z., Tang, X. Convolutional neural network-based signal demodulation method for NOMA-PON. Optics Express. 2020;28(10):14357-65. https://doi.org/10.1364/OE.392535
20. Lin, C., Chang, Q., Li, X. A deep learning approach for MIMO-NOMA downlink signal detection. Sensors. 2019;19(11):2526-35. https://doi.org/10.3390/s19112526
21. McWade, S., Flanagan, M. F., Farhang, A. Low-complexity equalization and detection for OTFS-NOMA. arXiv preprint. 2022:1-12. Available: https://arxiv.org/abs/2211.07388
22. Rahman, M. H., Islam, M. R., Islam, M. S., Shin, H. Multi-user joint detection using bi-directional deep neural network Framework in NOMA-OFDM System," Sensors. 2022; 22(18): 6994. https://doi.org/10.3390/s22186994

Persian Abstract

چکیده

در این مقاله GVDET-NET، یک چارچوب نوآورانه تشخیص سیگنال طراحی شده برای افزایش صحت تشخیص و کارآیی عملیاتی سیستم‌های NOMA با استفاده از منابع فرکانس زمانی غیر دولتی طراحی شده است. مدل پیشنهادی لایه‌های CNN مبتنی بر VGG۱۹ را با لایه‌های GRU ادغام می‌کند تا به طور مشترک وابستگی‌های مکانی و زمانی را از داده‌های ورودی استخراج کند. با پردازش پی در پی ویژگی‌های سلسله مراتبی، GVDET-NET تشخیص سیگنال کانال NOMA برتر را در مقایسه با رویکردهای ML، LS و MMSE در سراسر SNRها از ٤ دسی‌بل تا ۲۸ دسی‌بل به دست می‌آورد. نتایج شبیه‌سازی نشان‌دهنده اثربخشی آن در شرایط NOMA واقع‌گرایانه، بهتر از SIC-LS و SIC-MMSE تحت سناریوهای تست چندگانه با ٦٤ و ١٦ تنظیمات آزمایشی برای موارد دو کاربر است. GVDET-NET به حداقل نرخ خطای نماد (SER) تقریباً $۱۰^{-۳}$ در سطح SNR بالا دست پیدا می‌کند و سود عملکرد قابل توجهی را به دست می‌آورد. علاوه بر این، این مدل به دقت طبقه‌بندی ۹٦/٤٪، تأخیر استنباط ۳/۱ میلی‌ثانیه برای اندازه بسته‌های استاندارد و نمره AUC ۰/۹٦۸ می‌رسد و استحکام آن و کاربرد در زمان واقعی را تأیید می‌کند. این کار تکنیک‌های پیشرفته تشخیص را برای سیستم‌های NOMA معرفی می‌کند، و راه را برای شبکه‌های بی‌سیم بهینه و پشتیبانی از استانداردهای ارتباطی نسل بعدی هموار می‌کند.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2500-2508

International Journal of Engineering

Journal Homepage: www.ije.ir

Improving the Accuracy of Hardness Measurement using UCI Transducers for Inspecting the Heat-Affected Zone of Welded Joints in Steel Pipelines

A. I. Shikhov[a], D. S. Gromyka*[a], A. S. Umansky[a], D. V. Kopytina[a], D. R. Shakirzyanova[a], A. M. Shmakov[b]

[a] *Empress Catherine II Saint-Petersburg Mining University, Saint-Petersburg, Russia*
[b] *ITMO University, Saint-Petersburg, Russia*

PAPER INFO

Paper history:
Received 17 September 2025
Received in revised form 17 October 2025
Accepted 24 December 2025

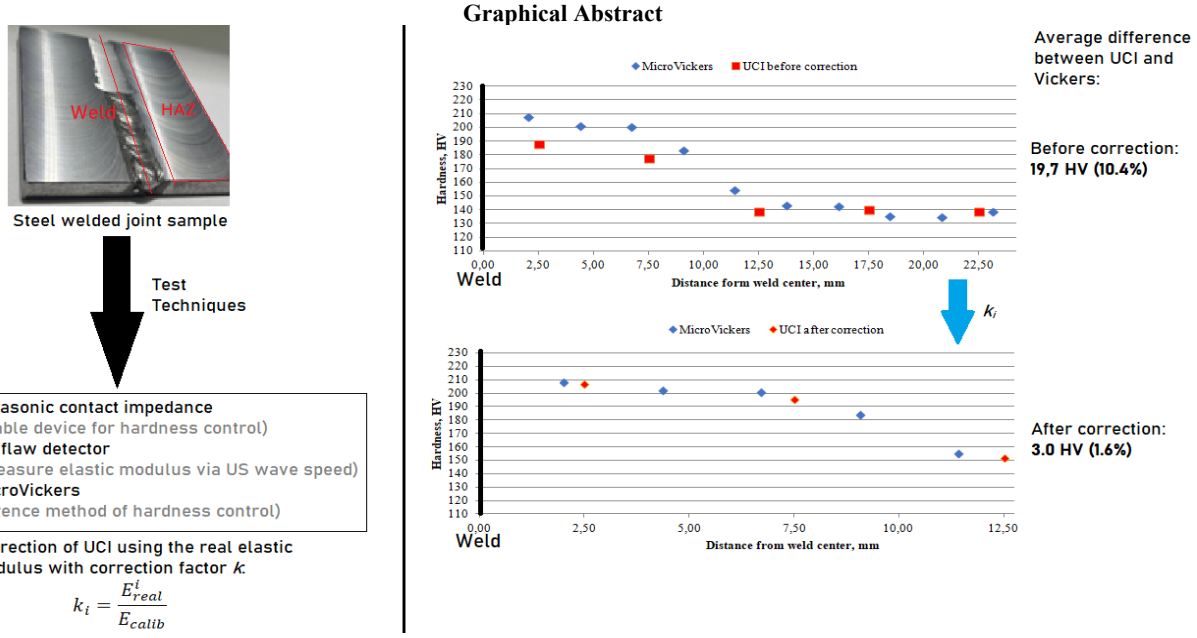
Keywords:
Steel Pipes
Weld Seam
Heat-affected Zone
Ultrasonic Contact Impedance
Ultrasonic Wave
Calibration

ABSTRACT

Quality control of welded joints involves hardness measurements of both the weld and the heat-affected zone (HAZ). However, standardized hardness testing methods require specimen preparation, leading to pipeline downtime. Among portable devices, hardness testers implementing the Ultrasonic Contact Impedance (UCI) method are the most widely used. However, measurement results depend on the calibration curve, which accounts for the elastic properties of the specimen and assumes their uniformity across the entire inspected surface. In the HAZ of a welded joint, changes in the metal's microstructure and elastic properties occur, leading to measurement errors in UCI hardness testing. This article presents a methodology for applying a correction to the calibration by accounting for actual changes in the elastic modulus within the HAZ, measured via ultrasonic wave velocity monitoring. As a result, hardness values incorporating this correction demonstrated better correlation with reference results obtained using a Vickers microhardness tester, indicating that the proposed correction improves the accuracy of UCI transducer measurements.

doi: *10.5829/ije.2026.39.10a.12*

Graphical Abstract



*Corresponding Author Email: Gromyka_DS@pers.spmi.ru (D. S. Gromyka)

Please cite this article as: Shikhov AI, Gromyka DS, Umansky AS, Kopytina DV, Shakirzyanova DR, Shmakov AM. Improving the Accuracy of Hardness Measurement using UCI Transducers for Inspecting the Heat-Affected Zone of Welded Joints in Steel Pipelines. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2500-8.

NOMENCLATURE			
HAZ	Heat-affected zone	c_p^2	Longitudinal wave velocity
UCI	Ultrasonic Contact Impedance	E	Elastic modulus, GPa
k_0	Initial stiffness	v	Poisson's ratios
f	Frequency	HV_{UCI}	Hardness UCI
A_c	Projected contact area of the indenter at the penetration depth	ρ	Density

1. INTRODUCTION

The reliability of gas transportation systems directly depends on the strength of pipeline welded joints (1, 2). According to ISO 9692-1:2013, there are many types of welded joints for steel pipes, but the most common for butt-welding two pipes is a single-sided weld with beveled edges. Incoming inspection requires assessing the quality of welded joints. Mechanical property evaluation is performed according to ISO 5177:1981, along with non-destructive testing using ultrasonic and radiographic methods. Mechanical testing includes static uniaxial tensile tests and hardness measurements (3). Continuous condition monitoring is also a common quality control method (4-6).

Standardized hardness testing methods include Brinell (7), Rockwell (8), and Vickers (9). Additionally, the instrumented indentation method has been used in many studies (10-12) to assess elastic-plastic properties. However, these methods cannot be applied for in-field pipeline inspections due to the need of specimen preparation.

A solution to this problem is the use of portable hardness testers. Among portable devices, testers with transducers implementing the Leeb dynamic method and the UCI method are the most common.

The Leeb dynamic method measures the ratio of rebound velocity to impact velocity of an indenter fixed to a striker that hits the inspected surface with a defined energy. Portable Leeb hardness testers are widely used for technical diagnostics, as regulated by ISO 16859-1:2015. Successful applications of Leeb testers for welded joint hardness control are described in (13, 14). However, Leeb method results depend on the mass, thickness, and stiffness of the tested object (15).

Thus, portable UCI hardness testers (DIN 50159-1 (2022) have gained widespread use for pipeline weld inspections. This method is based on changes in the vibration frequency of a rod with a Vickers diamond indenter upon penetration into the material, increasing the oscillation frequency (16). The frequency shift is proportional to the indentation depth, and hardness is calculated using a calibration curve. This method is successfully applied in various industries, particularly for in-service pipeline and weld inspections, due to its simplicity and sufficient accuracy for homogeneous materials. Examples include hardness testing of aerospace materials (17), fatigue properties of marine structures (18), and corrosion damage assessment in slug catcher vessels (19). However, its primary application is in evaluating welded joint strength, including hardness mapping of weld and HAZ regions (20-23), weld defect detection (20, 24), fatigue endurance assessment for different welding techniques (21, 25, 26), and HAZ hardness distribution analysis (27-29).

A major limitation of UCI testers is their dependence on the material's elastic properties. The method assumes a uniform elastic modulus across the inspected area, matching that of the calibration specimens. If the actual modulus differs from the reference value hardness measurements will be distorted.

This issue arises in pipeline weld inspections, where welding alters the microstructure and mechanical properties (30). The results of many researches (31, 32) demonstrate that the intense thermal cycle induces structural transformations, such as martensite formation, resulting in a substantial hardness gradient between the base metal, HAZ, and the weld. Thus, a variation in the elastic modulus is observed, driven by the development of residual stresses and the heterogeneity of the material's elastic properties. These alterations critically affect the load-bearing capacity and service life of the structure, underscoring the necessity for precise control and prediction of mechanical properties post-welding. The HAZ exhibits microstructural changes due to thermal effects, influenced by steel composition, grain size, and welding parameters (33). These variations lead to UCI hardness measurement errors. Introducing a correction factor accounting for elastic modulus changes in the HAZ can improve accuracy of UCI method.

Elastic non-uniformity in the HAZ due to microstructural changes can be assessed by measuring ultrasonic (US) wave velocity, as longitudinal (cl) and transverse (cp) wave speeds depend on elastic properties. Ultrasonic attenuation is influenced by microstructural changes (34-37), though wave velocity may vary due to weld anisotropy, complicating microstructure correlations (38-40). However, studies (41-43) show that US wave velocity depends not only on microstructure but also on surface hardness in the weld and HAZ, confirming that elastic non-uniformity can be accounted for via US wave analysis to derive a UCI transducer correction.

According to regulatory documents, a welded joint is deemed non-conforming if its hardness increases by 50 HB relative to the base metal. However, as previously emphasized, the results of hardness measurements are significantly influenced by the heterogeneity of elastic

properties in the heat-affected zone (HAZ) of the weld. Therefore, the primary objective of this study is to enhance the reliability of hardness measurements in the weld and the heat-affected zone to improve the accuracy of technical diagnostics for the condition of steel pipelines. The current study investigates longitudinal US wave velocity changes to calculate elastic modulus variations in the HAZ. A correction factor was applied to UCI hardness calculations, and results were compared with Vickers microhardness measurements.

2. MATERIALS AND METHODS

2. 1. Theory When a diamond indenter vibrating at a fixed frequency penetrates a material, a sharp increase in the probe's natural oscillation frequency occurs. The process can be described as follows: as a vibrating probe of mass mm with an initial stiffness k_0 penetrates the material, the system's total stiffness increases by an increment Δk. The resulting frequency shift of the probe can be expressed as:

$$\Delta f = \frac{f_0}{2k_0}\Delta k \quad (1)$$

The contact stiffness Δk can be described as follows (44):

$$\Delta k \approx \frac{2}{\sqrt{\pi}} E^* \sqrt{A_c} \quad (2)$$

where A_c is the projected contact area of the indenter at the penetration depth, and E_r is the reduced elastic modulus.

The reduced elastic modulus accounts for the elastic contact between the indenter and the material (45):

$$\frac{1}{E^*} = \frac{(1-\nu^2)}{E} + \frac{(1-\nu_i{}^2)}{E_i} \quad (3)$$

where ν and ν_i are the Poisson's ratios of the material and indenter, respectively, and E and E_i are the elastic moduli of the material and indenter, respectively.

Substituting Equation 2 into Equation 1 and expressing A_c in terms of Vickers hardness $HV_{Vickers}$, indentation load P, and eliminating transducer-specific constants, the UCI-measured hardness becomes proportional to the elastic modulus:

$$HV_{UCI} \sim P\left(\frac{E}{\Delta f \cdot (1-\nu^2)}\right)^2 \quad (4)$$

In practice, HV_{UCI} is not calculated from Equation (4) but derived from a calibration curve $HV_{UCI}(\Delta f)$, as the actual elastic modulus of the material cannot be directly determined by this method. Calibration is performed using reference blocks with certified hardness values. The resulting calibration curve (Figure 1) is used to calculate the material's hardness.

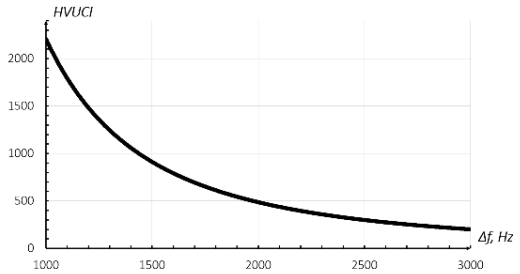


Figure 1. Calibration curve HV_{UCI} vs. Δf (16)

Since Equation 4 shows that HV_{UCI} is directly proportional to E^2 and inversely proportional to Δf^2. Due to that, measurements in the weld and HAZ, where E is lower than in the base material, yield underestimated HV_{UCI} values, which can be similar to hardness values in unaffected regions.

To obtain accurate hardness values ($HV_{UCI(real)}$), a correction factor must be applied:

$$HV_{UCI}^{real} = HV_{UCI}^{calib} \cdot \frac{E_{real}}{E_{calib}} \quad (5)$$

where $HV_{UCI(real)}$ and $HV_{UCI(calib)}$ are the corrected and calibration-derived hardness values, respectively, and E_{real} and E_{calib} are the actual and calibration-assumed elastic moduli.

The elastic modulus is determined ultrasonically:

$$E_{real} = V_p^2 \rho \frac{(1+\nu)(1-2\nu)}{(1-\nu)} \quad (6)$$

where V_p is the longitudinal wave velocity, ρ is the metal density.

Thus, the actual elastic modulus can be calculated by measuring longitudinal wave velocity. Substituting these values into Equation 5 yields the corrected hardness in the HAZ.

2. 2. Specimen Preparation For the study, specimens of butt-welded joints were fabricated from St3 steel (US equivalent: A284Gr.D) using automatic arc welding in accordance with ISO 9692-1:2013. Type of joint is pipe-to-pipe butt-weld one sided joint, with the use of JQ.H08MnA welding wire. The specimen dimensions were 100 × 100 mm with a thickness of 10 mm (Figure 2). The chemical composition of St3 steel, according to AISI A284Gr.D, is provided in Table 1.

After welding, specimens were ground and polished to Ra<0.32μm with coolant to avoid thermal effects. Surface roughness was measured using a profilometer.

TABLE 1. Mass fraction of elements in St3 steel

C	Si	Mn	P	S	Cr	Ni	Cu
0,14-0,22	0,15-0,3	0,4-0,65	<0,04	<0,05	<0,3	<0,3	<0,3

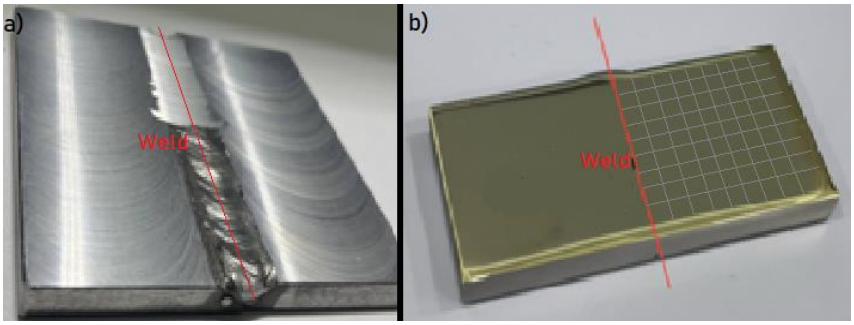


Figure 2. Test specimen: (a) after cleaning; (b) after grinding/polishing with measurement grid

2. 3. Ultrasonic Testing Longitudinal wave velocity was measured using a UDL-2M laser-ultrasonic flaw detector. The operating principle of the UDL-2M laser-ultrasonic flaw detector is based on the laser thermo-optical generation of nanosecond ultrasonic pulses of longitudinal acoustic waves in a special broadband opto-acoustic transducer. It measures the propagation velocity of these pulses in the test specimen with one-sided transducer access. The system employs the time-of-flight measurement method. Based on the known specimen thickness and the measured time difference between the arrival of the probing ultrasonic pulse and the signal reflected from the specimen's back surface at the transducer's piezoelectric receiver, the velocity of longitudinal ultrasonic waves in the specimen is calculated.

Echo-mode measurements employed a 1–10 MHz transducer. Figure 3 shows the testing schematic.

Procedure:

1. Caliper-measured distance H from the scanning surface to the back wall.
2. A direct-contact transducer was placed on the scanning surface.
3. Time-of-flight of the backwall echo was recorded.
4. Wave velocity was calculated:

$$V_p = \frac{2H_i}{t_i} \quad (7)$$

where H_i – scanning surface-to-backwall distance at point i; t_i – wave travel time at point i.

5. Elastic modulus was computed via Equation 5.
6. Correction factor was determined:

$$k_i = \frac{E^i_{real}}{E_{calib}} \quad (8)$$

7. Uniform density was assumed. Specimen density was derived from:

$$\rho = \frac{V}{m} \quad (9)$$

where V and m – specimen volume and mass.

8. Steps 1–6 were repeated across the measurement grid (Figure 2b) at 3 mm intervals.

2. 4. Vickers Hardness Testing A PMT-3M microhardness tester was selected as the reference instrument for comparison with UCI results. The test load was 1.96 N (200 gf), with a dwell time of 10 seconds. Prior to testing, the indenter approach was calibrated using a halite crystal, and the built-in microscope eyepiece scale was calibrated using an objective micrometer.

Hardness measurements on both instruments followed the same indentation pattern (Figure 2b). Indentations were made from the weld centerline to the edge of the specimen with a step of 2.35 mm. Lines parallel and perpendicular to the weld were marked on the specimen with the same spacing. Hardness measurements were taken at each node of the grid. The results obtained from indentations along the same line parallel to the weld were averaged, and the resulting value was taken as the mean hardness at a specific distance from the weld. An example of an indentation imprint is shown in Figure 4. A schematic representation of the experimental setup is shown on Figure 5.

2. 5. UCI Hardness Testing Hardness measurements were performed using a portable Konstanta KT hardness tester equipped with a UCI transducer. A constant test load of 50 N was applied for all measurements. The tester was pre-calibrated using a set of steel reference blocks with the following certified

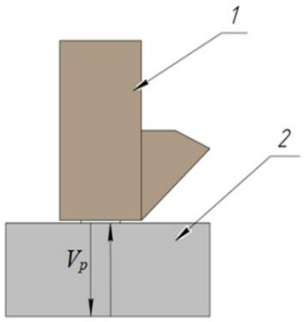


Figure 3. Longitudinal wave velocity measurement 1 – ultrasonic transducer; 2 – steel plate

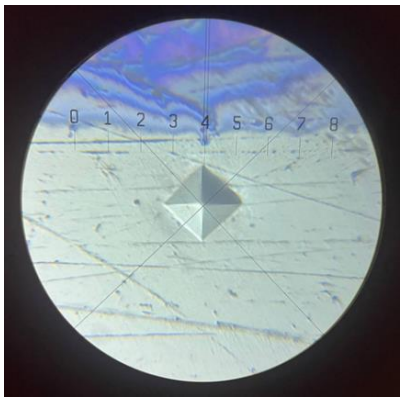


Figure 4. Example indentation

properties: ν = 0.27, E = 200 GPa. UCI hardness measurements followed the same grid pattern as the Vickers method, but indentations were made along lines parallel to the weld between the grid nodes (Figure 2b) to avoid overlapping with prior indentation marks.

3. RESULTS AND DISCUSSION

Hardness results from both methods are shown in Figure 6.

As it can be seen, the hardness values determined by the PMT and UCI methods in the heat-affected zone (0-13 mm) do not coincide. At a distance of 2 mm from the

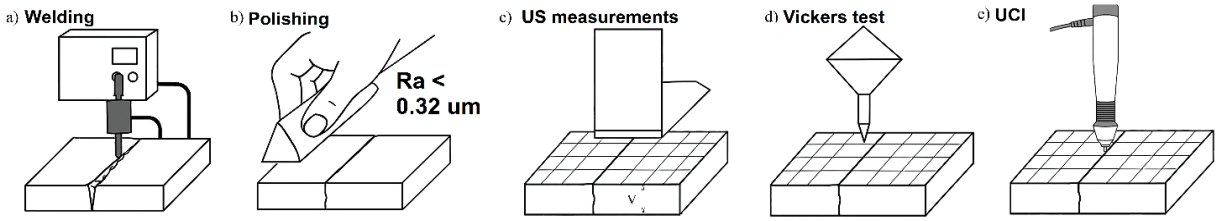


Figure 5. Experimental setup schema

weld centerline, $HV_{Vickers}$ = 208 HV, while HV_{UCI} = 188 HV, which is nearly 10% lower than the actual hardness values. The difference in hardness values gradually decreases with increasing distance from the weld, reaching approximately 3% at 13 mm. Figure 6 also shows that beyond the transition zone (10-15 mm), the hardness values remain constant as the istance from the weld increases, confirming the capability of the UCI method to identify the boundary of the heat-affected zone. It should also be noted that beyond the heat-affected zone (>15 mm), the hardness values obtained by both methods coincide, which further supports the hypothesis about the necessity of accounting for changes in elastic properties of the material in the weld zone and HAZ when measuring hardness by the UCI method.

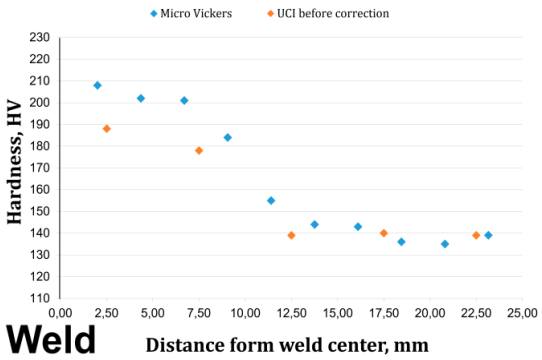


Figure 6. PMT vs. UCI results

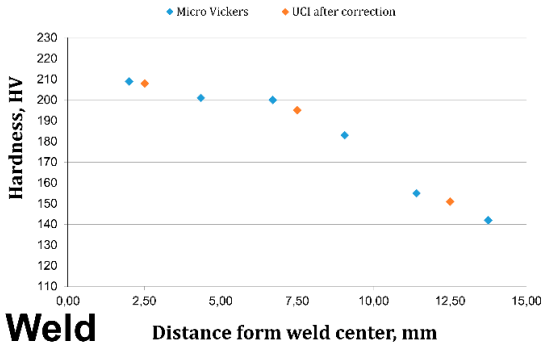


Figure 7. PMT vs. corrected UCI results

TABLE 2. Difference between UCI and PMT results before and after correction

L, mm	Difference between UCI and PMT, HV(%)	
	Before correction	After correction
2,5	20 (9.6%)	1 (0.5%)
7,5	23 (11.4%)	5 (2.5%)
12,5	16 (10.3)	3 (1.9%)
Average	**19,7 (10.4%)**	**3,0 (1.6%)**

The measurements of longitudinal wave velocity in the weld zone and HAZ using the UDL-2M device showed values ranging from 5912 to 5984 m/s, which corresponds to an elastic modulus of 225 GPa.

Thus, it can be concluded that the actual values of the steel's elastic modulus differ from the certified values of the reference block used for transducer calibration. Additional uncertainty is introduced by the non-uniform distribution of elastic modulus values in different areas of the steel specimen. In the weld zone and at small distances from it (L<7 mm), the elastic modulus values remain around 225 GPa. With increasing distance from the weld, in the HAZ the elastic modulus values gradually decrease to 220-223 GPa.

Using Equation 5 and the obtained values of E_{real}, a correction factor was calculated, and the adjusted HV_{UCI} values in the weld zone and HAZ are presented in Figure 7.

Figure 7 demonstrates that the difference between Vickers and corrected UCI hardness values is negligible in the HAZ at distances L<7 mm. The average deviation between the methods is smaller than before correction (Table 2). Thus, the average difference between PMT-3 and corrected UCI hardness values is reduced to 3.0 HV (1.6%), compared to 19.7 HV (10.4%) without correction. The introduced correction improves the accuracy of UCI transducer measurements for hardness evaluation in welded joints of steel pipelines.

Currently, hardness measurements are widely used for assessing the technical condition of pipelines. Therefore, the reliability of measured hardness values is critical for ensuring safe operation. Portable hardness testing methods show potential for field monitoring systems, but this study highlights the need for additional verification of data accuracy. The deviation of the hardness value from the permissible limit must not exceed 7%. The discrepancy between the hardness values before and after the correction exceeds the allowable magnitude, which, in turn, may lead to an invalid assessment of the weld condition.

The study confirms that UCI hardness measurements are highly dependent on the match between the calibration block material and the test object. This approach is inadequate for the HAZ, where microstructural changes due to thermal effects are unaccounted for.

On the other hand, the sensitivity of elastic wave velocity to material changes under thermal exposure is reaffirmed. The elastic modulus determined ultrasonically, combined with the derived correction factor, reduced the average discrepancy between portable (UCI) and laboratory (Vickers) hardness measurements. Notably, the values nearly converge in the HAZ. These results suggest that combining ultrasonic testing with UCI can enhance the reliability of portable hardness measurements.

However, it must be noted that the extent of the HAZ and the degree of changes in physical-mechanical properties vary across steel grades. According to standards, the HAZ width can range from 1 to 4 times the pipe wall thickness. Furthermore, studies (27, 34, 36) demonstrate that the relationship between longitudinal wave velocity, hardness, and thermal treatment depends on the steel type. An analysis of the dependence of hardness on heat treatment reveals no general correlation; even materials with similar chemical compositions exhibit different properties (46, 47). Therefore, the results of this work should be considered not from the perspective of establishing a universal correction for the elastic modulus when using the UCI method, but from the standpoint of developing a comprehensive methodology for measuring the hardness of welded joints with portable devices. Such a methodology would improve the reliability of hardness determination. The investigations carried out in this study allow for preliminary conclusions regarding the feasibility of developing such a methodology. However, further research on welded joints of various types of steel is required, which could enhance verification and validation measures for the comprehensive hardness measurement methodology.

Additionally, the 1–2% difference in wave velocity between the HAZ and base metal demands high measurement precision. In fact, this indicator can be used in designing a comprehensive hardness measurement methodology as a target for the methodology's performance criteria, namely the standard deviation characterizing repeatability, reproducibility, intermediate precision, etc. Therefore, additional metrological analysis of the possibility of ensuring the required measurement accuracy in field conditions is necessary. At the same time, as a preliminary metrological analysis, it is possible to consider, as reference values, the works in which an evaluation of the measurement uncertainty for thickness by the ultrasonic method was performed (48-53), since, in fact, measuring the speed is the inverse problem of this measurement. The analysis of (48-53) showed that with a certain approach, it is possible to ensure a thickness measurement error in the micrometer range, which exceeds the requirements for ensuring a speed measurement error of 1-2%. Furthermore, (50, 53) describe thickness measurement techniques without preliminary calibration of the elastic wave propagation velocity, based on the laws of geometrical acoustics, which is important in the context of developing a comprehensive hardness measurement methodology.

4. CONCLUSION

Addressing these aspects will generalize the findings of this study and improve the reliability of portable hardness testing methods. Main findings of the study are:

1. The accuracy of UCI hardness measurements depends significantly on the ratio between the actual elastic modulus and that of the calibration block, leading to incorrect hardness values in the HAZ.

2. Before correction, the difference between UCI and PMT hardness values in the HAZ was ~10%, a statistically significant deviation. After correction, the discrepancy is virtually eliminated.
3. Ultrasonic-based corrections for the elastic modulus improve the reliability of UCI hardness measurements in the HAZ.

Acknowledgements

The author appreciates Engineering and Design Center for Spacecraft Operation Support for providing necessary facilities to conduct present research.

Funding

This work received no funding.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work. The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Author Biosketch

Dmitrii Gromyka is an Assistant professor of metrology, instrument engineering and quality control department in Saint-Petersburg mining university. He received his PhD in 2023 in Geotechnology. Main scientific interests are assessing of fatigue damage with instrumental nanoindentation.

Alexander Shikhov is an Assistant professor of metrology, instrument engineering and quality control department in Saint-Petersburg mining university. He received his PhD in 2022 in Geotechnology. Main scientific interests are the use of ultrasound nondestructive testing to evaluate mechanical properties of soils.

Alexander Umanskii is a Head of metrology, instrument engineering and quality control department in Saint-Petersburg mining university. He received his PhD in 2018 in Geotechnology. Main scientific interests are design of portable hardness testers and enhancing of their metrological properties.

Darya Kopytina is a post-graduate student of metrology, instrument engineering and quality control department in Saint-Petersburg mining university. Main scientific interests are assessing the state of PE pipes with ultrasound NDT testers.

Dinara Shakirzyanova is a post-graduate student of metrology, instrument engineering and quality control department in Saint-Petersburg mining university. Main scientific interests are evaluation of mechanical properties of steel pipes with portable hardness testers.

Shmakov Alexey is a post-graduate student of system control and robotics department of ITMO university. Main scientific interests are designing of portable nondestructive testing devices.

REFERENCES

1. Dzhemilev ER, Shammazov IA, Sidorkin DI, Mastobaev BN, Gumerov AK. Developing technology and device for the main pipelines repair with cutting out their defective sections. Neftyanoe Khozyaystvo. Oil Industry. 2022;2022(10):78–82. https://doi.org/10.24887/0028-2448-2022-10-78-82
2. Bolobov VI, Popov GG. Methodology for testing pipeline steels for resistance to grooving corrosion. Journal of Mining Institute. 2021;252(6):854–60. https://doi.org/10.31897/PMI.2021.6.718:09
3. Schipachev AM, Vardanyan ES, Mukamadeev VR, Mukamadeev IR. Investigation of changes in the surface hardness of coated cutting tools. Metalloobrabotka. 2021;3(123):22–29. https://doi.org/10.25960/mo.2021.3.22
4. Yujra RE, Vyacheslavov AV, Gogolinskiy KV, Sapozhnikova K, Taymanov R. Deformation Monitoring Systems for Hydroturbine Head-Cover Fastening Bolts in Hydroelectric Power Plants. Sensors. 2025;25:2548. https://doi.org/10.3390/s25082548
5. Krizskii VN, Kosarev OV, Aleksandrov PN, Luntovskaya YA. Mathematical modeling of the electric field of an in-line diagnostic probe of a cathode-polarized pipeline. Journal of Mining Institute. 2024;265:156–64.
6. Botyan E, Lavrenko S, Pushkarev A. Evaluation of Complicated Mining Exploitation Conditions Influence on Service Life of Open Pit Trucks Suspensions with Remote Monitoring Systems. International Journal of Engineering Transactions B: Applications. 2024;37(11):2268–75. https://doi.org/10.5829/ije.2024.37.11b.12
7. Zandinava B, Bakhtiari R, Vukelic G. Failure analysis of a gas transport pipe made of API 5L X60 steel. Engineering Failure

Analysis. 2021;131:105881. https://doi.org/10.1016/j.engfailanal.2021.105881

8. Elkhodbia M, Gadala I, Barsoum I, AlFantazi A, Abdel Wahab M. Multi-physics microstructural modeling of a carbon steel pipe failure in sour gas service. Engineering Failure Analysis. 2025;174:109469. https://doi.org/10.1016/j.engfailanal.2025.109469
9. Li L, Yu T, Xia J, Gao Y, Han B, Gao Z. Failure analysis of L415/316L composite pipe welded joint. Engineering Failure Analysis. 2024;158:107981. https://doi.org/10.1016/j.engfailanal.2024.107981
10. Pham TH, Kim SE. Microstructure evolution and mechanical properties changes in the weld zone of a structural steel during low-cycle fatigue studied using instrumented indentation testing. International Journal of Mechanical Sciences. 2016;114:141–56. https://doi.org/10.1016/j.ijmecsci.2016.05.021
11. Midawi A, Simha CHM, Gerlich A. Assessment of yield strength mismatch in X80 pipeline steel welds using instrumented indentation. International Journal of Pressure Vessels and Piping. 2018;168. https://doi.org/10.1016/j.ijpvp.2018.09.014
12. Vinogradova AA, Gogolinskiy KV, Shchiptsova EK. A study of applicability of instrumented indentation method to determine the mechanical properties of thermoplastic. Industrial laboratory. Diagnostics of materials. 2025;91(5):67–76. https://doi.org/10.26896/1028-6861-2025-91-5-67-76
13. Purnomo E, Wiranata A, Aan D, Muflikhun M. Microstructural and Mechanical Characterization of U-Bend Economizer Failures: Integrating Experimental Testing and CFD Simulation. Journal of Materials Research and Technology. 2025;36. https://doi.org/10.1016/j.jmrt.2025.04.178
14. Li D, Hu S, Liu X, Xie Y, Feng C. Safety Performance Assessment Strategies of P91 Steel Based on Leeb Hardness. IOP Conference Series: Earth and Environmental Science. 2019;295:032096. https://doi.org/10.1088/1755-1315/295/3/032096
15. Seijiro M, Takashi Y. Computer Simulation of Micro Rebound Hardness Test. Procedia Engineering. 2014;81:1396–401. https://doi.org/10.1016/j.proeng.2014.10.163
16. Gogolinskii KV, Syasko VA, Umanskii AS, Nikazov AA, Bobkova TI. Mechanical properties measurements with portable hardness testers: advantages, limitations, prospects. Journal of Physics: Conference Series. 2019;1384:012012. https://doi.org/10.1088/1742-6596/1384/1/012012
17. Gvozdyakov DV, Zenkov AV, Gubin VE, Kaltaev AZh, Marysheva YaV. Experimental studies into the effect of air pressure on the atomization characteristics of coal-water slurries. iPolytech Journal. 2021;25(5):586–600. https://doi.org/10.52351/00260827_2021_09_81
18. Corigliano P, Crupi V, Fricke W, Friedrich N, Guglielmino E. Experimental and numerical analysis of fillet-welded joints under low-cycle fatigue loading by means of full-field techniques. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2015;229. https://doi.org/10.1177/0954406215571462
19. Jasim AH, Nuha H, Khanjar A. Investigation of slug catcher vessel degradation by hydrogen sulfide (H2S) damage. Heliyon. 2024;10:e40307. https://doi.org/10.1016/j.heliyon.2024.e40307
20. Voelkel J, Meissner M, Bartsch H, Feldmann M. The influence of external weld imperfection size on the load-bearing capacity of butt-welded joints. Journal of Constructional Steel Research. 2024;220:108808. https://doi.org/10.1016/j.jcsr.2024.108808
21. Churiaque C, Sánchez-Amaya J, Üstündag Ö, Porrua-Lara M, Gumenyuk A, Rethmeier M. Improvements of hybrid laser arc welding for shipbuilding T-joints with 2F position of 8 mm thick steel. Optics & Laser Technology. 2021;143:107284. https://doi.org/10.1016/j.optlastec.2021.107284
22. Üstündag Ö, Bakir N, Gumenyuk A, Rethmeier M. Improvement of Charpy impact toughness by using an AC magnet backing system for laser hybrid welding of thick S690QL steels. Procedia CIRP. 2022;111:462–65. https://doi.org/10.1016/j.procir.2022.08.067
23. Schroepfer D, Witte J, Kromm A, Kannengiesser T. Stresses in repair welding of high-strength steels—part 2: heat control and stress optimization. Welding in the World. 2024;68. https://doi.org/10.1007/s40194-024-01731-7
24. Rhode M, Erxleben K, Richter T, Schroepfer D, Mente T, Michael T. Local mechanical properties of dissimilar metal TIG welded joints of CoCrFeMnNi high entropy alloy and AISI 304 austenitic steel. Welding in the World. 2024;68:1–11. https://doi.org/10.1007/s40194-024-01718-4
25. Fuštar B, Lukačević I, Skejić D, Garašić I. Fatigue tests of S690 as-welded and HFMI-treated details with longitudinal and transverse attachments. Results in Engineering. 2024;23:1–8. https://doi.org/10.1016/j.rineng.2024.102386
26. Diekhoff P, Hensel J, Nitschke-Pagel T, Dilger K. Investigation on fatigue strength of cut edges produced by various cutting methods for high-strength steels. Welding in the World. 2020;64. https://doi.org/10.1007/s40194-020-00853-y
27. Schroeder N, Rhode M, Kannengiesser T. Thermodynamic prediction of precipitations behaviour in HAZ of a gas metal arc welded S690QL with varying Ti and Nb content. Welding in the World. 2023;67. https://doi.org/10.1007/s40194-023-01550-2
28. Brätz O, von Arnim M, Eichler S, Gericke A, Henkel K-M, Hildebrand J, Bergmann J, Kuhlmann U. Load-carrying capacity of MAG butt and fillet welded joints on high-strength structural steels of grade S960QL and S960MC. CE/Papers. 2023;6:587–94. https://doi.org/10.1002/cepa.2478
29. Churiaque C, Sánchez-Amaya JM, Porrúa-Lara M, Gumenyuk A, Rethmeier M. The effects of HLAW parameters for one side T-joints in 15 mm thickness naval steel. Metals. 2021;11(4):600. https://doi.org/10.3390/met11040600
30. Pryakhin EI, Sharapova DM. Understanding the structure and properties of the heat affected zone in welds and model specimens of high-strength low-alloy steels after simulated heat cycles. CIS Iron and Steel Review. 2020;19:60–65. https://doi.org/10.17580/cisisr.2020.01.12
31. Rahmani M, Moslemi Petrudi A. Investigation of crack growth by loading fatigue due to fluid and structural coupling vibrations in the joint of thermowell welding in gas pipeline. Journal of Research in Engineering and Applied Sciences. 2020;5(2):28–37. https://doi.org/10.46565/jreas.2020.v05i02.001
32. Mičian M, Winczek J, Gucwa M, Koňár R, Málek M, Postawa P. Investigation of welds and heat affected zones in weld surfacing steel plates taking into account the bead sequence. Materials. 2020;13:5666. https://doi.org/10.3390/ma13245666
33. Sizyakov MI. The use of steel arc welding pipes on tank farms or oil and gas industry objects. Petroleum Engineering. 2020;18:88–93. https://doi.org/10.17122/ngdelo-2020-2-88-93
34. Kalimuthu V, Muthupandi V, Jayachitra R. Influence of heat treatment on the microstructure, ultrasonic attenuation and hardness of SAF 2205 duplex stainless steel. Materials Science and Engineering: A. 2011;529:447–51. https://doi.org/10.1016/j.msea.2011.09.059
35. Mutlu I, Oktay E, Ekinci S. Characterization of microstructure of H13 tool steel using ultrasonic measurements. Russian Journal of Nondestructive Testing. 2013;49. https://doi.org/10.1134/S106183091302006X
36. Borovkov AI, Prokhorovich VE, Bychenok VA, Berkutov IV, Alifanova IE. Ultrasonic inspection technique for welded joints obtained by spot friction stir welding. Russian Journal of Nondestructive Testing. 2023;59(2):149–60. https://doi.org/10.1134/S1061830923700262

37. Shikhov A, Gogolinsky K, Zubarev A, Smorodinsky Ya, Kopytina D, Vinogradova A. Analysis of the requirements for metrological support of methods and means of ultrasound control. Russian Journal of Nondestructive Testing. 2025;61(3):1061–69. https://doi.org/10.1134/S1061830924603180
38. Carreón H, Barrera G, Natividad C, Salazar M, Contreras A. Relation between hardness and ultrasonic velocity on pipeline steel welded joints. Nondestructive Testing and Evaluation. 2015;31(2):97–108. https://doi.org/10.1080/10589759.2015.1074231
39. Bolobov VI, Latipov IU, Zhukov VS, Popov GG. Using the magnetic anisotropy method to determine hydrogenated sections of a steel pipeline. Energies. 2023;16:5585. https://doi.org/10.3390/en16155585
40. Bychenok VA, et al. Ultrasonic testing of adhesion of special coatings. Russian Journal of Nondestructive Testing. 2023;59(8):839–46. https://doi.org/10.1134/S1061830923700481
41. Carreon M, Ruiz A, Medina A, Barrera Cardiel G, Zárate MJ. Characterization of the alumina–zirconia ceramic system by ultrasonic velocity measurements. Materials Characterization. 2009;60:875–81. https://doi.org/10.1016/j.matchar.2009.02.008
42. Daqing W, Yajie W, Weihua L, Chen H, Chun Y, Hao L. Multiscale investigation of microstructure optimization in the arc additive manufacturing and arc welding by self-induced ultrasound. International Journal of Heat and Mass Transfer. 2021;180:121790. https://doi.org/10.1016/j.ijheatmasstransfer.2021.121790
43. Cécile G, Jean M, Marie-Aude P, Gilles C, Bertrand C. Influence of the uncertainty of elastic constants on the modelling of ultrasound propagation through multi-pass austenitic welds. Impact on non-destructive testing. International Journal of Pressure Vessels and Piping. 2019;171:125–36. https://doi.org/10.1016/j.ijpvp.2019.02.011
44. Fischer-Cripps AC. Introduction to contact mechanics. 2nd ed. US: Springer; 2007. p. 226.
45. Oliver W, Pharr G. An improved technique for determining hardness and elastic modulus using load and displacement sensing indentation experiments. Journal of Materials Research. 1992;7(6):1564–83. https://doi.org/10.1557/JMR.1992.1564
46. Bharare OS, Sarve SA, Gengane VV, Sahane AS, Jikar PC. Effect of various surface heat treatment methods and its parametric variation on the hardness properties of different steels grades. Materials Today: Proceedings. 2024;102:132–39. https://doi.org/10.1016/j.matpr.2023.04.057
47. Anggraini L, Kartomi H. Hot forgeability of SCr420 low-alloy steel for piston rod application. In: Journal of Physics: Conference Series. 2020;1595(1):012007. https://doi.org/10.1088/1742-6596/1595/1/012007
48. Wangyu L, Dongxun L, Weigui X. A novel time-of-flight difference determination method for ultrasonic thickness measurement with ultrasonic echo onset point detection. Applied Acoustics. 2025;233:110605. https://doi.org/10.1016/j.apacoust.2025.110605
49. Peng Z, Shijie W, Pan D, Xinru G, Tonghai W, Min Y, Yaguo L. High-accuracy ultrasonic measurement of oil film thickness in sliding bearings considering surface thin coating. Tribology International. 2025;110343. https://doi.org/10.1016/j.triboint.2024.110343
50. Pan D, Laisheng Z, Tonghai W, Min Y, Tom R, Zhongxiao P. Simultaneous measurement of thickness and sound velocity of porous coatings based on the ultrasonic complex reflection coefficient. NDT & E International. 2022;131:102683. https://doi.org/10.1016/j.ndteint.2022.102683
51. Hao C, Kaipeng J, Wei Z, Chengqian Z, Jianzhong F, Peng Z. High precision ultrasonic measurement method for thin-walled parts. NDT & E International. 2024;145:103122. https://doi.org/10.1016/j.ndteint.2024.103122
52. Morana M, Damir M, Biserka R, Zdenka K. Measurement uncertainty evaluation of ultrasonic wall thickness measurement. Measurement. 2019;137:179–88. https://doi.org/10.1016/j.measurement.2019.01.027
53. Marcelo YM, Nicolás P, Flávio B, Julio CA, Marcos SGT. Error analysis of self-compensated ultrasound measurements of the thickness loss due to corrosion in pipe walls. Ultrasonics. 2024;142:107387. https://doi.org/10.1016/j.ultras.2024.107387

Persian Abstract

چکیده

کنترل کیفیت اتصالات جوش داده شده شامل اندازه گیری سختی هر دو جوش و ناحیه متاثر از حرارت (HAZ) است .با این حال، روش های استاندارد تست سختی نیاز به آماده سازی نمونه دارد که منجر به خرابی خط لوله می شود .در میان دستگاه های قابل حمل، سختی سنج ها با استفاده از روش امپدانس تماس اولتراسونیک (UCI) بیشترین استفاده را دارند .با این حال، نتایج اندازه‌گیری به منحنی کالیبراسیون بستگی دارد، که خواص الاستیک نمونه را به حساب می‌آورد و یکنواختی آنها را در کل سطح بازرسی‌شده فرض می‌کند .در HAZ یک اتصال جوش داده شده، تغییراتی در ریزساختار و خواص الاستیک فلز رخ می دهد که منجر به خطاهای اندازه گیری در تست سختی UCI می شود .این مقاله روشی را برای اعمال تصحیح در کالیبراسیون با در نظر گرفتن تغییرات واقعی در مدول الاستیک درHAZ ، که از طریق نظارت بر سرعت موج اولتراسونیک اندازه‌گیری می‌شود، ارائه می‌کند .در نتیجه، مقادیر سختی که این تصحیح را شامل می‌شوند، مطابقت بهتری با نتایج مرجع به‌دست‌آمده با استفاده از تستر میکروسختی ویکرز نشان دادند، که نشان می‌دهد اصلاح پیشنهادی دقت اندازه‌گیری مبدل UCI را بهبود می‌بخشد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2509-2520

International Journal of Engineering

Journal Homepage: www.ije.ir

A Comparative Study of Laminar and Turbulent Flow in Ribbed Wavy Channels: Effects of Rib Height, Angle, and Density

A. H. Ghalta[a,b], M. Farhadi*[b]

[a] *Department of Refrigeration and Air Conditioning Engineering Technologies, College of Engineering Technologies, University of Hilla, Hilla, Iraq*
[b] *Faculty of Mechanical Engineering, Babol Noshirvani University of Technology, Babol, I. R. of Iran*

PAPER INFO

Paper history:
Received 23 August 2025
Received in revised form 15 September 2025
Accepted 25 October 2025

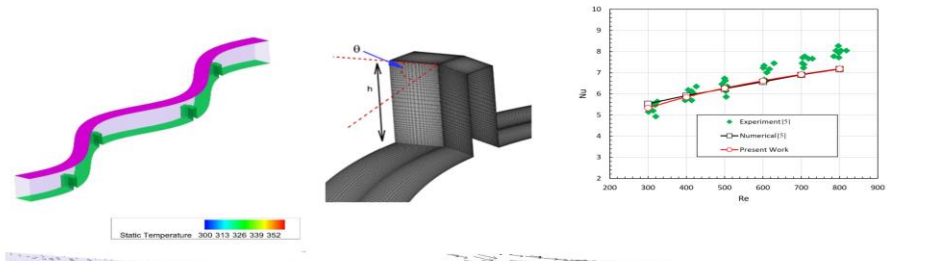
Keywords:
Wavy Channel
Ribbed Surfaces
Heat Transfer Enhancement
Erformance Evaluation Criterion (PEC)

ABSTRACT

This study investigates the thermal–hydraulic performance of wavy channels equipped with transverse ribs, addressing the need for optimized heat transfer solutions in compact heat exchangers, where the combined effects of rib geometry and flow regime remain insufficiently explored. The objective is to evaluate how rib inclination angle, height, and count influence heat transfer enhancement and overall efficiency under both laminar (Re = 500) and turbulent (Re = 5000) flow conditions. Three-dimensional numerical simulations were performed using ANSYS Fluent to analyze heat transfer characteristics, frictional losses, and the performance evaluation criterion (PEC). The results show that transverse ribs significantly enhance heat transfer by promoting flow separation and vortex formation. In laminar flow, Nusselt numbers increased by up to ~50% compared to smooth channels, with optimal PEC values achieved at moderate rib heights (h/H = 0.2–0.4), rib angles of 45°–60°, and low rib counts (N = 1–2). In turbulent flow, heat transfer enhancement was more pronounced (Nusselt numbers ~56–58), although increased friction led to PEC values below unity in most cases. Novel trade-offs were identified that balance thermal performance and efficiency, especially in laminar regimes. These findings contribute valuable design guidelines for the geometric optimization of ribbed wavy channels in different flow regimes.

doi: *10.5829/ije.2026.39.10a.13*

Graphical Abstract



Flow Regime	Max Nu	Typical f (high h/H, N=4)	PEC Trend	Optimal Design
Laminar (Re=500)	~21–24	2.5–3.0	PEC ≈ 1 achievable at 45–60°, h/H=0.2–0.4, N=1–2	Moderate ribs, shallow angles
Turbulent Re=5000)	~56–58	1.5–1.8	PEC < 1 in most cases; best ≈ 0.9 at 45°, h/H≤0.2, N=1	Small ribs, low density

*Corresponding Author Email: *mfarhadi@nit.ac.ir* (M. Farhadi)

Please cite this article as: Ghalta AH, Farhadi M. A Comparative Study of Laminar and Turbulent Flow in Ribbed Wavy Channels: Effects of Rib Height, Angle, and Density. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2509-20.

NOMENCLATURE

D_h	Hydrolic Diameter, $D_h = (2WH)/(H + W)$ *(m)*	T	Temperature (K)
h	Rib height (m)	$\mathrm{Re} = \rho D_h U/\mu$	Reynolds Number
H	Channel cross section height (m)	u_i, u_j	components of velocity (m/s)
W	Channel cross section width (m)	**Greek Symbols**	
h_{ave}	Mean heat transfer coefficient (W/m^2.K)	k	turbulence kinetic energy (m^2/s^2)
N	Number of Rib	μ	Dynamic viscosity (kg.m/s)
Nu	Nusselt number, $Nu = h_{ave} D_h/k_f$	ϴ	Rib angle (°, degree)
k_f	Conductivity of fluid (W/m.K)	α	Thermal diffusivity (m^2/s)
P	*Pressure (Pa)*	ν	Kinematic viscosity (m^2/s)
Pr	Prandtl number	ρ	Density (kg/m^3)
		ε	rate of dissipation (m^2/s^3)

1. INTRODUCTION

With the continuous miniaturization and increasing power density of electronic, optoelectronic, and photovoltaic (PV) devices, efficient thermal management has become a critical concern. Among the various strategies explored, microchannel heat sinks (MCHSs) have garnered significant attention due to their compact size, high surface-area-to-volume ratio, and superior heat dissipation capabilities. Over the past decade, researchers have extensively investigated enhancements in microchannel geometries, flow conditions, and working fluids to improve heat transfer performance while managing associated pressure drops (1-3). Early experimental studies, such as those by Sui et al. (4), revealed that wavy microchannels outperformed straight ones in terms of frictional flow and heat transfer characteristics. Building upon this, Zhou et al. (5) employed numerical simulations and response surface methodology to optimize sinusoidal wavy microchannel geometries, further validating their potential for compact heat sinks. Other geometric innovations have also shown promise; Kaewchoothong et al. (6) analyzed square cross-sectional obstacles and found that inclined and V-shaped obstacles enhanced heat transfer by up to 30%, depending on angle and configuration. Lin et al. (7) introduced a variable wavelength and amplitude wavy design that significantly reduced thermal resistance, while Rui and Kondo (8) highlighted the material and thermal advantages of triangular heat sinks over traditional rectangular fin arrays. Wang et al. (9) demonstrated the benefits of inclined rectangular obstacles in breaking thermal boundary layers and enhancing heat transfer. Meanwhile, Chai and Wang (10) analyzed five obstacle geometries in cross-sectional chambers and reported substantial reductions in thermal resistance and entropy generation. Bahiraei et al. (11) further improved performance using graphene-silver hybrid nanofluids in microchannels with ribs and sub-channels. Ghodavai and Hassan (12) studied the influence of obstacle height on local heat transfer, whereas Zhou et al. (13) compared left-right and top-bottom wavy designs, revealing Dean vortex formation as a critical mechanism in enhancing mixing and thermal performance. Li et al. (14) proposed a dual-layer heat sink with solid and porous obstacles that achieved low thermal resistance and uniform temperature distribution. Shuo-Wang et al. (15) expanded on this with a dual-layer design incorporating wavy channels and porous vertical obstacles, finding it more effective than straight channel configurations. Sheng and Wang et al. (16) investigated pulsating flow effects in fan-like microchannels using the lattice Boltzmann method, revealing that alternating flow structures improved mixing and heat transfer. Krishna et al. (17) emphasized the role of waviness in altering the transition Reynolds number and enhancing average Nusselt numbers. Ahmed et al. (18) applied a multi-level truncated fin design for natural convection cooling of PV modules, which reduced module temperature and increased power output. Khan and Karimi (19) and Al-Sabri et al. (20) further examined geometric modifications and nanofluid behavior in wavy microchannels.

Zhang et al. (21) compared dual-layer and straight channel MCHSs, confirming the thermal advantages of the dual-layer configuration. Khan et al. (22) validated the enhanced cooling of wavy and dual-wavy microchannels using nanofluids in laminar flows. Ghorbani et al. (23) identified that higher amplitude-to-wavelength ratios promoted turbulence and improved heat transfer in wavy designs. Shuo-Wang et al. (24) showed that symmetric wavy configurations generated four Dean vortices, improving heat transfer but increasing pressure drop. Okeb et al. (25) demonstrated that dimples and fillets significantly influenced the flow structure and thermal performance. Al-Naimat et al. (26) showed that zigzag channels reduced thermal resistance and pumping power. Halon et al. (27) explored two-phase boiling flows in restricted and unrestricted microchannels, revealing complex flow pattern dependencies. Finally, Fiu et al. (28) used machine learning and genetic algorithms to optimize wavy microchannel configurations for PV cooling, achieving high prediction accuracy and performance.

Recent studies have emphasized the pivotal role of geometry in optimizing microchannel heat sink (MCHS)

performance. Chen and Yaji (29) introduced a topology optimization framework using an Eulerian–Eulerian multiphase approach, achieving an 11.6% enhancement in heat transfer through geometrically optimized nanofluid channels. Nasiri Khamesloo and Domiri Ganji (30) proposed novel fractal MCHS designs with bottom and sidewall fins, revealing that sidewall fins improved Nusselt numbers by 28% and overall system performance by up to 6%. In another work, they showed that hybrid nanofluid flows in modified geometries further optimized thermal distribution (31). Zhu et al. (32) conducted a comparative assessment of rib shapes and cavity structures, concluding that geometric coupling significantly influences flow uniformity and heat transfer rates. Li et al. (33) investigated six fin shapes in MCHS designs, identifying those with minimal pressure drop and maximal thermal enhancement. Li and colleagues (34) also applied artificial neural networks and genetic algorithms to optimize pin-fin microchannels, leading to better temperature uniformity and thermal performance. Huo and Sun (35) developed a hierarchical embedded microchannel geometry for ultra-high flux SiC chip cooling, achieving exceptional thermal metrics using deionized water boiling. Chej et al. (36) modeled heat transfer in microfluidic sinks, showing how geometry-dependent temperature profiles impact flow behavior, especially at low Reynolds numbers. Al-Bahrani and co-authors (37) presented a topology-optimized fin architecture produced through additive manufacturing, enabling enhanced cooling via complex internal features. Jalilinasrabady and co-workers (38) reviewed inclined microchannel designs for microelectronics, highlighting their superior heat spreading potential. Dehkordi et al. (39) examined sinusoidal and Z-shaped patterns, concluding that increased surface contact via waviness improves convection. Mousavi and Rahimi (40) validated the effect of changing cross-sectional areas on flow acceleration and turbulence generation. Ghasemi and Arasteh (41) proposed a fluted-fin layout that enhanced thermal exchange in laminar regimes. Alimohammadi and Jamshidi (42) analyzed Y-shaped bifurcated microchannels, demonstrating superior distribution of flow and temperature control. Finally, Liu and Shi (43) introduced triangular lattice structures in the channel network, resulting in balanced heat spreading and reduced hotspots.

Despite the extensive literature on heat transfer enhancement, a comprehensive understanding of the combined effects of wavy channel geometries and internal obstacles—particularly V-shaped ribs—remains limited. To the best of our knowledge, there are no systematic studies investigating the influence of V-shaped ribs in wavy microchannels under both laminar and turbulent flow conditions. This study addresses this gap by numerically analyzing a novel ribbed wavy microchannel heat sink with axially aligned V-shaped ribs. Parametric investigations are performed for both laminar (Re = 500) and turbulent (Re = 5000) flow regimes, focusing on the effects of rib height-to-channel height ratio (h/H), rib angle (V-opening angle), and rib number (density) on Nusselt number, friction factor, thermal resistance, and the performance evaluation criterion (PEC).

2. GOVERNING EQUATION

The fluid flow and heat transfer in a three-dimensional domain under steady-state conditions, assuming an incompressible fluid with constant thermophysical properties, are governed by the following set of partial differential equations for laminar and turbulent flows:
Continuty equation:

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{1}$$

Momentum equation:

$$u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j} \tag{2}$$

Energy equation:

$$u_j \frac{\partial T}{\partial x_j} = \alpha \frac{\partial^2 T}{\partial x_j \partial x_j} \tag{3}$$

The preceding equations are applicable to laminar flow regimes.

The time-averaging method in Reynolds-Averaged Navier-Stokes (RANS) equations is one of the common approaches in analyzing turbulent flows. In this method, by applying time-averaging to the Navier-Stokes equations, the Reynolds stress term emerges, which is caused by velocity fluctuations in turbulent flow. These additional stresses appear as indeterminate variables and require modeling to solve the equations.

One of the most common models for Reynolds stresses is the Eddy Viscosity Approach. In this method, Reynolds stresses are modeled similarly to molecular viscosity stresses, but instead of actual viscosity, an effective turbulent viscosity coefficient is introduced, which can simulate the effects of displacement and mixing caused by eddies. The resulting equations for the transfer of momentum and heat in turbulent flows are written as follows:

$$u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial P}{\partial x_i} + (\nu + \nu_t)\frac{\partial^2 u_i}{\partial x_j \partial x_j} \tag{4}$$

$$u_j \frac{\partial T}{\partial x_j} = (\alpha + \alpha_t)\frac{\partial^2 T}{\partial x_j \partial x_j} \tag{5}$$

In these equations, the values of kinematic viscosity and the thermal diffusivity coefficient of turbulence are expressed as follows.

$$\nu_t = C_\mu \frac{k^2}{\varepsilon},\ \alpha_t = \frac{\nu_t}{\mathrm{Pr}_t} \tag{6}$$

The turbulence kinetic energy, k, and its rate of dissipation, ε, are obtained from the following transport equations:

$$\frac{\partial(ku_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[(\nu + \frac{\nu_t}{\sigma_k})\frac{\partial k}{\partial x_j}\right] + P_k - \varepsilon + S_k \tag{7}$$

$$\frac{\partial(\varepsilon u_i)}{\partial x_i} = \frac{\partial}{\partial x_j}\left[(\nu + \frac{\nu_t}{\sigma_\varepsilon})\frac{\partial \varepsilon}{\partial x_j}\right] + C_{1\varepsilon}\frac{\varepsilon}{k}P_k - C_{1\varepsilon}\frac{\varepsilon^2}{k} + S_\varepsilon \tag{8}$$

The values of the constants in the equations are as follows.

$$C_\mu = 0.09, C_{1\varepsilon} = 1.44, C_{2\varepsilon} = 1.92, \sigma_k = 1.0, \sigma_\varepsilon = 1.3 \tag{9}$$

This study investigates fluid flow in a channel with a wavy cross-section and the presence of obstacles (Figure 1(a)). The channel geometry is rectangular with a height of 400 micrometers and a width of 216 micrometers, based on the research conducted by Ghorbani et. al. (24). The wavelength and wave amplitude are set at 2500 and 250 micrometers, respectively, as the baseline values for the channel's geometric structure. The obstacles are designed with V-shaped geometry, blocking between 0.1 to 0.5 of the channel's height and positioned along the flow path. The number of ribs (N) placed per wavelength of the wavy channel is limited to a maximum of four due to the geometric constraints of the wave structure. Ribs are positioned based on three distinct configurations: 1) only at the wave crests, 2) at both crests and troughs, and 3) at wave nodes with a 45° phase shift. These locations are physically meaningful and evenly distributed along the wave profile. As a result, placing more than four ribs within one wavelength would lead to overlapping or unrealistic rib positioning. Therefore, N= 1, 2, and 4 were selected as the practical and representative cases for analysis. To control computational volume and enhance the accuracy of the results, periodic boundary conditions have been applied. This approach ensures the independence of the results from the flow development length and improves computational efficiency. The rib angle is, in fact, half of the angle between the rib and its central axis (Figure 1(b)).

3. NUMERICAL METHOD

In this study, the commercial software ANSYS Fluent version 2021 was used. This program is based on the finite volume method, and the SIMPLE algorithm was used for pressure-velocity coupling. Non-linear terms in the momentum and energy equations were discretized using a second-order upwind scheme, and the residual value in the continuty equation was less than 10^{-3}, and in other equations up to 10^{-6}. The grid used was orthogonal and structured and had a higher concentration near the walls. In the turbulent flow, the y^+ value near the wall was reported to be less than 35. A sample of the generated grid is shown in Figure 1(b). The average difference between the grid numbers of the first and fourth is 4.51%, which has decreased to 0.4% with the further increase in number of grid points, particularly for the third and fourth grids. Considering the accuracy of the results and the savings in computational time and costs, the third grid points (with a size of 2,840,000 cells) was used in this study (Figure 2).

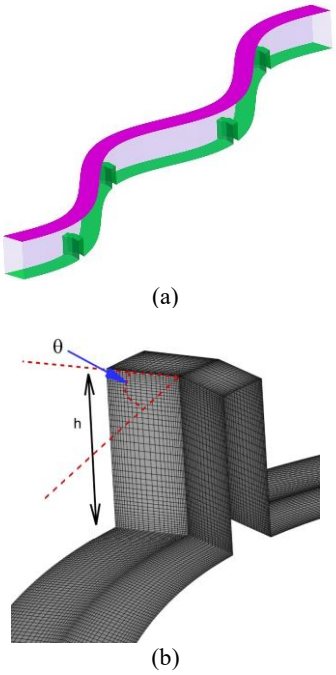


Figure 1. (a) Geometry of the problem; (b) A sample of grid points

4. RESULTS

To validate the numerical method employed, the results of Lu and Jiang (44) were simulated and compared. Their research includes experimental investigations of flow within a heated ribbed-surface channel. The present validation was conducted based on the channel configuration of 40 × 10 × 4 mm³, with a rib angle of 90 degrees and an inter-rib spacing of 4 mm. Results show that the present study predicted the experimental data with suitable accuracy in Nu number and friction factor

(Figure 3). The maximum error observed in the Nusselt number and friction factor were 5.8%, and 3%, respectively.

Subsequently, the experimental results of Sui et al. (5) concerning flow in a wavy microchannel were simulated (Figure 4). The simulation results demonstrated good agreement, with a maximum observed error of approximately 4%.

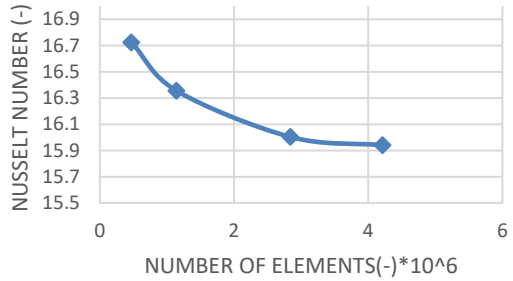


Figure 2. A grid independency

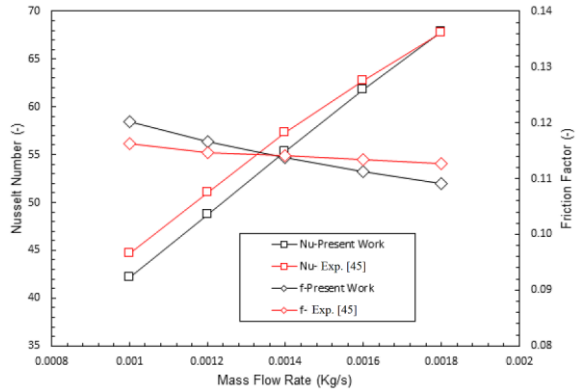


Figure 3. Comparison of the present numerical results with experimental data of Lu and Jiang (45) at turbulent flow

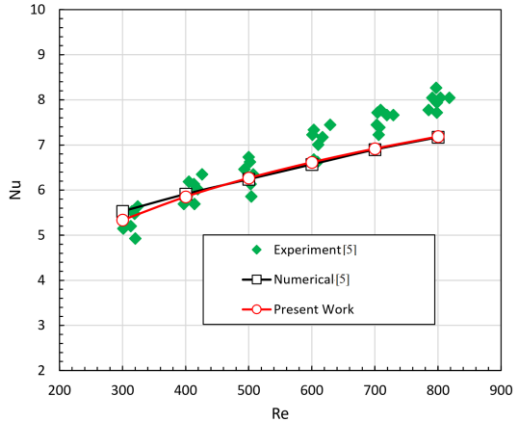
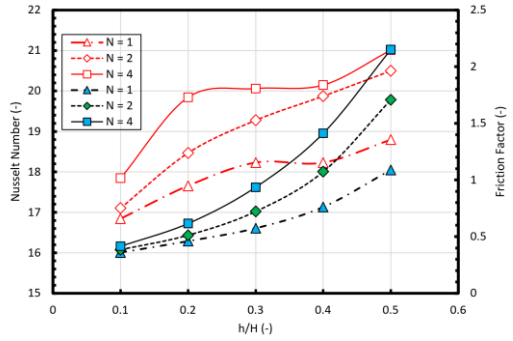


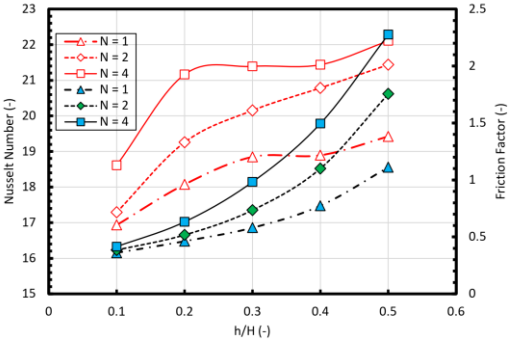
Figure 4. Comparison of the present numerical results with experimental and numerical data of Sui et al. (5) in a microchannel at laminar flow

4. 1. Heat Transfer and Friction Behavior Across Rib Angles (Laminar Flow) Figure 5 illustrates the variation of Nusselt number and friction factor as functions of the relative rib height *h/H* for rib inclination angles of 45°, 60°, 75°, and 90°, respectively, under laminar flow conditions (Re = 500). In all cases, three rib densities are considered: N = 1, 2, and 4. Across all angles, increasing the rib height consistently results in a higher Nusselt number, primarily due to improved boundary layer disruption and increased effective surface area. The most aggressive enhancement occurs for the 90° ribs, where the Nusselt number reaches values exceeding 24 for N = 4 at h/H=0.6. At 45°, the peak Nusselt number remains lower (around 21) but exhibits a more gradual increase with rib height. However, this thermal enhancement is accompanied by a substantial increase in friction factor, which becomes more pronounced at higher angles. At h/H=0.6 the friction factor reaches ~2.1 for 45°,~2.4 for 60°,~2.6 for 75° and ~3.0 for 90°. especially in the N = 4 configuration. This demonstrates that steeper rib angles, while thermally beneficial, impose a greater pressure drop penalty—particularly when combined with high rib density and height.

Figure 6 presents the three-dimensional streamlines and static temperature contours for a wavy channel with high rib height under (a) laminar and (b) turbulent flow conditions. The figure illustrates the complex vortex structures, recirculation zones, and heat transfer effects induced by ribbed geometry at elevated heights.



Angle = 45°



Angle = 60°

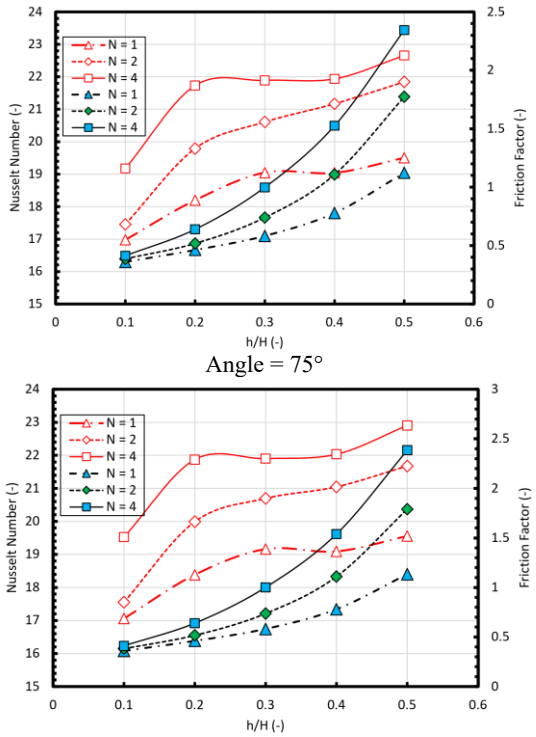


Figure 5. variation of Nusselt number and friction factor as functions of the relative rib height h/H for different rib inclination angles at laminar flow

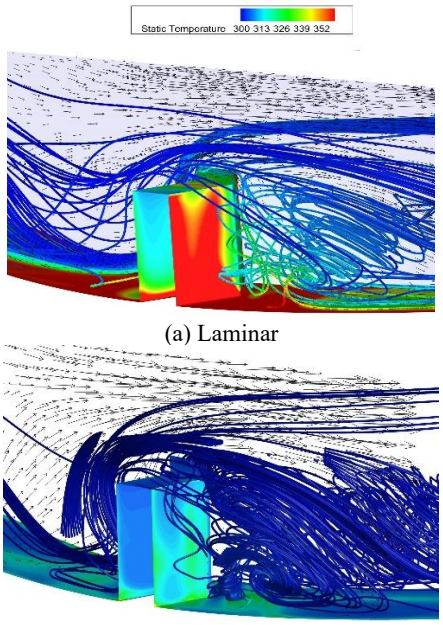


Figure 6. Three-dimensional streamlines and static temperature contours in a ribbed wavy channel with high rib height (h/H = 0.5) under (a) laminar and (b) turbulent flow conditions. Vortex formation, recirculation zones, and temperature gradients are shown to highlight the effect of rib geometry on flow structure and heat transfer

In the laminar flow case (Figure 6a), flow separation occurs immediately downstream of the ribs, generating stable and symmetric recirculating vortices in the cavity regions between ribs. As rib height increases from 0.1H to 0.2H, the size and strength of these vortices grow slightly, but the flow remains relatively ordered. However, a more significant change is observed when rib height reaches 0.5H: the vortices expand considerably, occupying nearly half the cavity height, and penetrate deeper into the core flow. This enhanced mixing contributes to stronger boundary layer disruption and improved heat transfer, as evidenced by the expanded isothermal layers near the rib walls.

In the turbulent flow case (Figure 6b), the increased rib height leads to an even more pronounced transformation. The flow becomes highly chaotic, with intense eddy formation, secondary flows, and complex vortex interactions in all three spatial directions. The high rib height forces the streamlines to bend sharply around obstacles, increasing the turbulent kinetic energy near the wall. The temperature contours show a thinner boundary layer and deeper thermal penetration into the core region, indicating stronger convective mixing than in the laminar case. Despite this improvement in heat transfer, such flow complexity also results in a higher friction factor and energy cost. Together, these visualizations demonstrate the critical role of rib height in shaping flow and thermal performance. While small ribs produce limited disruption, larger rib heights (e.g., h/H = 0.5) significantly enhance both vortex strength and thermal mixing, especially under turbulent conditions.

This expansion of the vortex zone intensifies flow mixing, boosts thermal energy exchange, and is consistent with the observed increase in the Nusselt number at higher rib heights. Nevertheless, the enlarged vortices also contribute to increased pressure drop, which must be balanced against the thermal gains in practical design. Figure 7 shows the Performance Evaluation Criterion (PEC) as a function of h/H for the same rib angles.

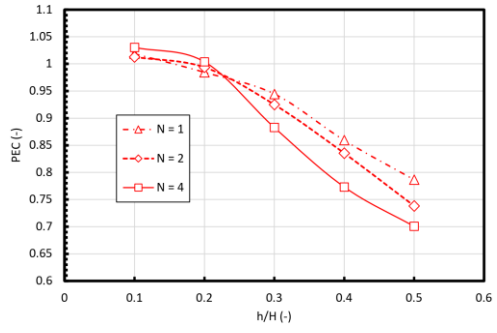
$$PEC = \frac{Nu / Nu_{Simple\ Wavy\ Channel}}{(f / f_{Simple\ Wavy\ Channel})^{1/3}} \quad (10)$$

A PEC value greater than 1 signifies a net thermal-hydraulic improvement. At 45°, PEC values approach unity for N = 1 and 2 at low rib heights ($h/H \leq 0.2$), but decline steadily beyond this point, dropping to ~0.7 for N = 4 at h/H=0.6. For 60°, 75°, and 90°, the PEC curves exhibit similar trends but decline more sharply, with the 90° configuration falling below 0.6 for high rib counts and heights. These findings indicate that while vertical or near-vertical ribs significantly enhance heat transfer, they do so at a disproportionate cost to pressure losses, reducing overall system efficiency.

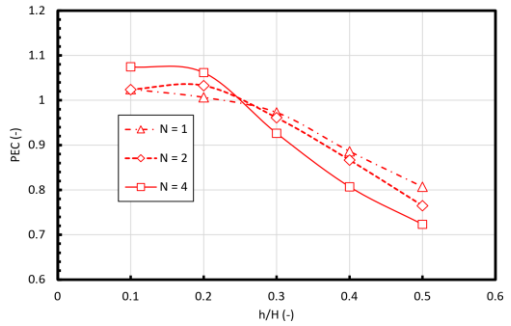
Table 1 summarizes the optimal rib configurations for laminar flow conditions (Re = 500), based on a comprehensive evaluation of heat transfer (Nusselt number), frictional losses (friction factor), and overall thermal-hydraulic performance (PEC). For the single-rib configuration (N = 1), the optimal performance is achieved with rib angles between 45° and 60° and moderate rib heights in the range of $0.2 \leq h/H \leq 0.4$. This setup offers an excellent balance between enhanced heat transfer and low pressure drop, making it highly efficient for compact systems where pumping power must be minimized. The two-rib configuration (N = 2) performs best at a rib angle of 60° with slightly higher rib heights ($0.3 \leq h/H \leq 0.4$). This arrangement achieves a favorable compromise, offering improved heat transfer compared to N = 1, while maintaining manageable frictional penalties. It can be considered the most balanced choice among the tested geometries. In contrast, the four-rib configuration (N = 4) shows diminishing returns in performance, particularly at steeper angles (≥75°). Although it achieves higher Nusselt numbers, the associated pressure losses become dominant, especially at rib heights above 0.2H. Therefore, for high-density ribbing, the design must be restricted to low rib heights ($h/H \leq 0.2$) and shallower angles to avoid excessive flow resistance.

4. 2. Turbulent Flow Results (Re = 5000)

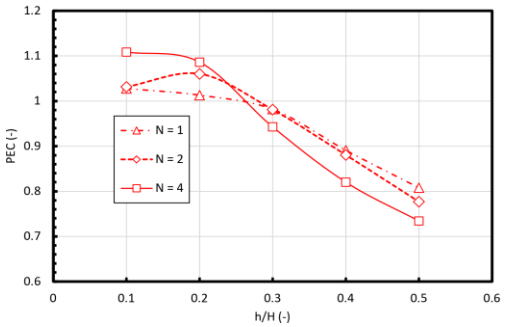
Figures 8 and 9 illustrate the influence of rib angle (45°, 60°, 75°, and 90°), rib height ratio (h/H=0–0.6), and rib number (N = 1, 2, 4) on the Nusselt number, friction factor, and Performance Evaluation Criterion (PEC) under turbulent flow conditions. At Re = 5000, the Nusselt number increases strongly with rib height across all angles and rib counts. For the most intensive configuration (N=4, 90° ribs, h/H=0.5), Nu reaches ~56–58, more than double the laminar maximum. Even shallow ribs (h/H=0.1) at 45° provide significant enhancement over a smooth channel. Increasing rib angle further intensifies mixing and flow separation, thus raising Nu, with the 75° and 90° cases showing the steepest growth.



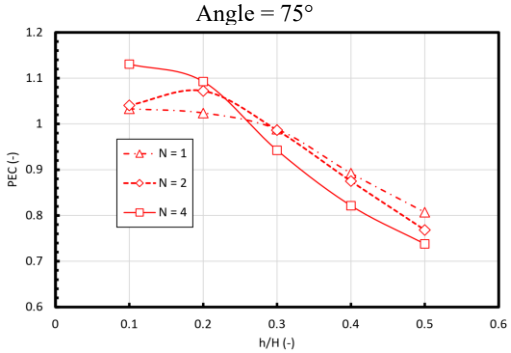
Angle = 45°



Angle = 60°



Angle = 75°

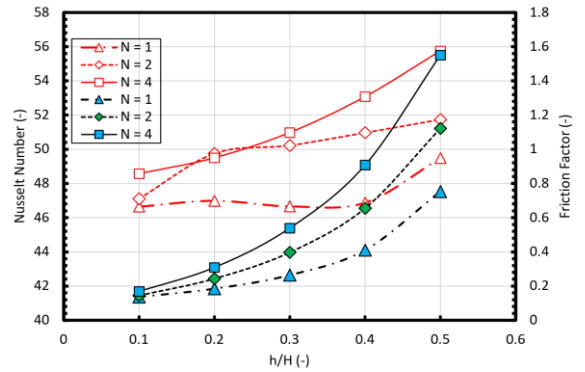


Angle = 90°

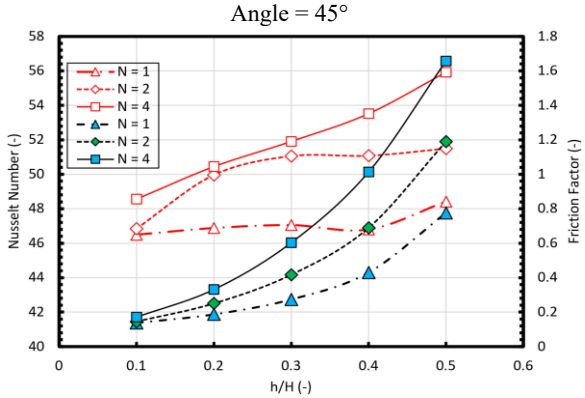
Figure 7. Performance Evaluation Criterion (PEC) as a function of h/H for for different rib inclination angles at laminar flow

TABLE 1. Recommended Rib Geometries for Laminar Flow (Re = 500)

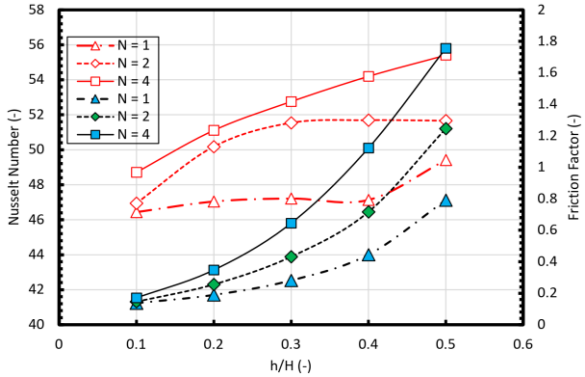
Number of Ribs (N)	Recommended Angle (°)	Optimal Height Ratio (h/H)	Design Note
1	45–60	0.2–0.4	High efficiency with minimal pressure penalty
2	~60	0.3–0.4	Balanced trade-off between heat transfer and loss
4	≤ 60 (avoid ≥75)	≤ 0.2	Strong heat transfer but high pressure drop



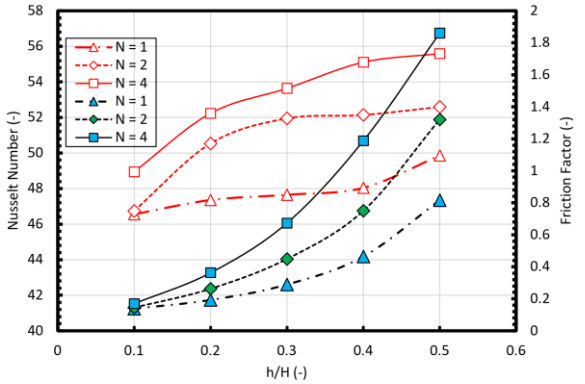
Angle = 45°



Angle = 60°

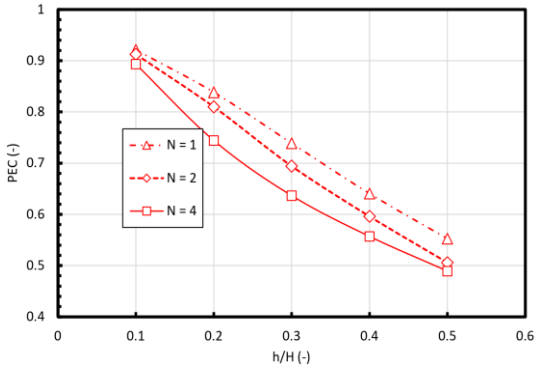


Angle = 75°

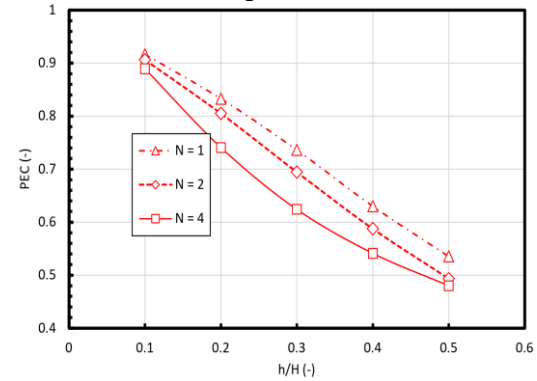


Angle = 90°

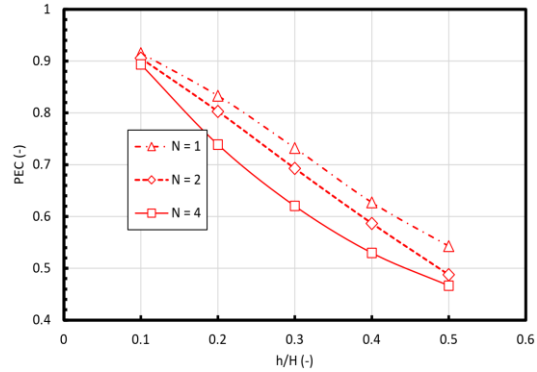
Figure 8. Variation of Nusselt number and friction factor as functions of the relative rib height h/H for different rib inclination angles at turbulent flow



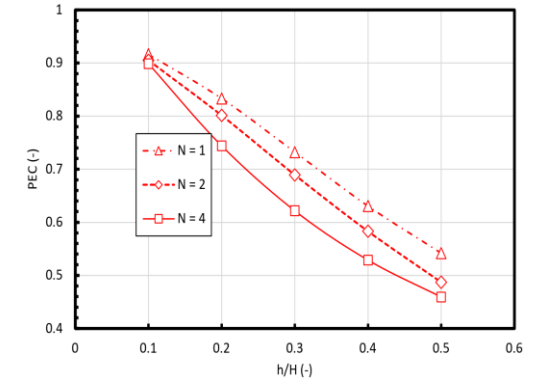
Angle = 45°



Angle = 60°



Angle = 75°



Angle = 90°

Figure 9. Performance Evaluation Criterion (PEC) as a function of h/H for for different rib inclination angles at turbulent flow

The friction factor grows monotonically with rib height and rib density. However, its absolute values are generally lower than in the laminar regime: for instance, at h/H=0.5, f is ~1.5–1.8 for N=4 in turbulence, compared with ~2.5–3.0 in laminar flow. This difference arises because inertial forces dominate turbulent transport, reducing the relative importance of viscous drag. The PEC trends reveal that despite high heat transfer enhancement, overall efficiency declines with rib height. For N=1 at 45°, PEC values remain close to unity for small heights (h/H≤0.2), but they drop rapidly to ~0.5 at h/H=0.5 when N=4. Steeper angles (≥75°) accelerate this decline, as pressure losses rise faster than the thermal benefits.

When comparing the laminar (Re=500) and turbulent (Re=5000) regimes, several key differences emerge in thermal–hydraulic behavior. In laminar flow, ribbed channels provide a modest increase in heat transfer, with the Nusselt number rising to about 21–24 at high rib heights, while friction factors can exceed 2.5–3.0 for dense ribbing. Importantly, under moderate rib heights (h/H=0.2–0.4) and shallow to intermediate rib angles (45°-60°), the PEC approaches or slightly exceeds unity, indicating that a net efficiency gain can be realized. In contrast, turbulent flow delivers a much stronger heat transfer enhancement, with Nusselt numbers reaching 56–58 for high ribs, but this improvement is accompanied by significant hydraulic penalties. Although absolute friction factor values are somewhat lower in turbulence than in laminar flow, the relative energy cost becomes more severe, and as a result, PEC values remain below unity in most configurations, except at very small rib heights and low rib counts. Increasing rib angle consistently boosts heat transfer in both regimes but reduces overall performance efficiency, with the decline more pronounced under turbulence. Overall, while laminar flow offers the possibility of achieving both heat transfer enhancement and acceptable efficiency, turbulent flow emphasizes maximum heat transfer at the expense of efficiency, making careful geometric optimization essential.

5. CONCLUSION

This study investigated the influence of rib angle, rib height, and rib density on the thermal–hydraulic performance of wavy channels under both laminar (Re=500) and turbulent (Re=5000) flow conditions. The results provide several key insights into the interplay between heat transfer enhancement and hydraulic losses:

1. Laminar flow (Re = 500). Ribbed channels produced moderate heat transfer improvements, with Nusselt numbers increasing up to ~24. Friction penalties were relatively high, yet the performance factor (PEC) could approach or slightly exceed unity for optimized geometries, particularly with one or two ribs, moderate rib heights (h/H=0.2–0.4), and rib angles between 45° and 60°. Thus, laminar configurations allow simultaneous improvement in heat transfer and overall efficiency if rib geometry is carefully selected.
2. Turbulent flow (Re = 5000). Heat transfer enhancement was far more pronounced, with Nusselt numbers reaching 56–58 at high rib heights. However, the associated pressure losses intensified, resulting in PEC values that generally remained below unity. Only small rib heights (h/H≤0.2) with a single rib and shallow angles (45°–60°) provided near-balanced performance. Steeper rib angles (≥75°) and higher rib densities yielded maximum Nusselt numbers but at the cost of severe efficiency degradation.
3. Comparison of regimes. Laminar flows present opportunities for balanced design, where both efficiency and heat transfer can be optimized simultaneously. In contrast, turbulent flows favor maximum heat transfer but rarely achieve net thermal-hydraulic efficiency, underscoring the importance of prioritizing either performance or energy economy depending on application requirements.

Acknowledgements

The author appreciates Babol Noshirvani University of Technology for providing necessary facilities to conduct present research.

Funding

No funding was received for conducting this study.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The authors declare no financial or organizational conflicts of interest.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

The authors declare that did not use the AI for

generating of any data of this research. All simulations and analyses were performed by the authors, and they bear full responsibility for the content of the publication.

Authors Biosketches

A. Hussein Ghalta: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing—original draft.
M. Farhadi (MF): Conceptualization, Supervision, Project administration, Resources, Writing—review & editing. (Corresponding author.)

REFERENCES

1. Nasiri Khamesloo F, Domiri Ganji D. Optimizing heat transfer in microchannel heat sinks: A numerical investigation with nanofluids and modified geometries. International Journal of Engineering Transactions B: Applications, 2024; 37(5): 860-75. https://doi.org/10.5829/ije.2024.37.05b.05
2. Babaei MR, Sheikhzadeh GA, Abbasian Arani AA. Numerical investigation of geometric parameters effects on heat transfer enhancement in a manifold microchannel heat sink. International Journal of Engineering Transactions B: Applications, 2022; 35(5): 943-53. https://doi.org/10.5829/ije.2022.35.05b.10
3. Safikhani H, Shaabani H. Numerical simulation of frost formation in interrupted microchannel heat sinks considering microfluidic effects in slip regime. International Journal of Engineering Transactions C: Aspects, 2020; 33(12): 2556-62. https://doi.org/10.5829/ije.2020.33.12c.17
4. Sui Y, Lee PS, Teo CJ. An experimental study of flow friction and heat transfer in wavy microchannels with rectangular cross section. International Journal of Thermal Sciences, 2011; 50(12): 2473-82. https://doi.org/10.1016/j.ijthermalsci.2011.06.017
5. Zhou J, Hatami M, Song D, Jing D. Design of microchannel heat sink with wavy channel and its time-efficient optimization with combined RSM and FVM methods. International Journal of Heat and Mass Transfer, 2016; 103: 715-24. https://doi.org/10.1016/j.ijheatmasstransfer.2016.07.100
6. Kaewchoothong N, Maliwan K, Takeishi K, Nuntadusit C. Effect of inclined ribs on heat transfer coefficient in stationary square channel. Theoretical and Applied Mechanics Letters, 2017; 7(6): 344-50. https://doi.org/10.1016/j.taml.2017.09.013
7. Lin L, Zhao J, Lu G, Wang XD, Yan WM. Heat transfer enhancement in microchannel heat sink by wavy channel with changing wavelength/amplitude. International Journal of Thermal Sciences, 2017; 118: 423-34. https://doi.org/10.1016/j.ijthermalsci.2017.05.013
8. Roy R, Kundu B. Effects of fin shapes on heat transfer in microchannel heat sinks. Heat Transfer-Asian Research, 2018; 47(4): 646-59. https://doi.org/10.1002/htj.21332
9. Wang R, Wang J, Lijin B, Zhu Z. Parameterization investigation on the microchannel heat sink with slant rectangular ribs by numerical simulation. Applied Thermal Engineering, 2018; 133: 428-38. https://doi.org/10.1016/j.applthermaleng.2018.01.021
10. Chai L, Wang L. Thermal-hydraulic performance of interrupted microchannel heat sinks with different rib geometries in transverse microchambers. International Journal of Thermal Sciences, 2018; 127: 201-12. https://doi.org/10.1016/j.ijthermalsci.2018.01.029
11. Bahiraei M, Jamshidmofid M, Goodarzi M. Efficacy of a hybrid nanofluid in a new microchannel heat sink equipped with both secondary channels and ribs. Journal of Molecular Liquids, 2019; 273: 88-98. https://doi.org/10.1016/j.molliq.2018.10.003
12. Qidwai MO, Hasan MM. Effect of variation of cylindrical pin fins height on the overall performance of microchannel heat sink. Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, 2019; 233(5): 980-90. https://doi.org/10.1177/0954408918821777
13. Zhu JF, Li XY, Wang SL, Yang YR, Wang XD. Performance comparison of wavy microchannel heat sinks with wavy bottom rib and side rib designs. International Journal of Thermal Sciences, 2019; 146: 106068. https://doi.org/10.1016/j.ijthermalsci.2019.106068
14. Li XY, Wang SL, Wang XD, Wang TH. Selected porous-ribs design for performance improvement in double-layered microchannel heat sinks. International Journal of Thermal Sciences, 2019; 137: 616-26. https://doi.org/10.1016/j.ijthermalsci.2018.12.029
15. Wang SL, Chen LY, Zhang BX, Yang YR, Wang XD. A new design of double-layered microchannel heat sinks with wavy microchannels and porous-ribs. Journal of Thermal Analysis and Calorimetry, 2020; 141(1): 547-58. https://doi.org/10.1007/s10973-020-09317-3
16. Wang CS, Wei TC, Shen PY, Liou TM. Lattice Boltzmann study of flow pulsation on heat transfer augmentation in a louvered microchannel heat sink. International Journal of Heat and Mass Transfer, 2020; 148: 119139. https://doi.org/10.1016/j.ijheatmasstransfer.2019.119139
17. Krishna M, Deepu M, Shine SR. Effect of relative waviness on low Re wavy microchannel flow. Journal of The Institution of Engineers (India): Series C, 2020; 101(4): 661-70. https://doi.org/10.1007/s40032-020-00575-6
18. Ahmad EZ, Fazlizan A, Jarimi H, Sopian K, Ibrahim A. Enhanced heat dissipation of truncated multi-level fin heat sink in natural convection for photovoltaic cooling. Case Studies in Thermal Engineering, 2021; 28: 101578. https://doi.org/10.1016/j.csite.2021.101578
19. Khan MN, Karimi MN. Analysis of heat transfer enhancement in microchannel by varying the height of pin fins at upstream and downstream region. Proceedings of the Institution of Mechanical Engineers, Part E: Journal of Process Mechanical Engineering, 2021; 235(4): 758-67. https://doi.org/10.1177/0954408921992975
20. Alsabery AI, Hajjar A, Sheremet MA, Ghalambaz M, Hashim I. Impact of particles tracking model of nanofluid on forced convection heat transfer within a wavy horizontal channel. International Communications in Heat and Mass Transfer, 2021; 122: 105176. https://doi.org/10.1016/j.icheatmasstransfer.2021.105176
21. Zhang YD, Chen MR, Wu JH, Hung KS, Wang CC. Performance improvement of a double-layer microchannel heat sink via novel fin geometry: A numerical study. Energies, 2021; 14(12): 3585. https://doi.org/10.3390/en14123585
22. Khan MZU et al. Investigation of heat transfer in wavy and dual wavy micro-channel heat sink using alumina nanoparticles. Case Studies in Thermal Engineering, 2021; 28: 101515. https://doi.org/10.1016/j.csite.2021.101515
23. Ghorbani N, Targhi MZ, Heyhat MM, Alihosseini Y. Investigation of wavy microchannel ability on electronic devices cooling with the case study of choosing the most efficient microchannel pattern. Scientific Reports, 2022; 12(1): 5882. https://doi.org/10.1038/s41598-022-09859-6
24. Wang S, Zhu J, An D, Zhang B, Chen L, Yang Y, Zheng S, Wang X. Heat transfer enhancement of symmetric and parallel wavy microchannel heat sinks with secondary branch design.

International Journal of Thermal Sciences, 2022; 171: 107229. https://doi.org/10.1016/j.ijthermalsci.2021.107229

25. Okab AK, Hasan HM, Hamzah M, Egab K, Al-Manea A, Yusaf T. Analysis of heat transfer and fluid flow in a microchannel heat sink with sidewall dimples and fillet profile. International Journal of Thermofluids, 2022; 15: 100192. https://doi.org/10.1016/j.ijft.2022.100192
26. Alnaimat F, Varghese D, Mathew B. Investigation of thermal and hydraulic performance of MEMS heat sinks with zig-zag microchannels. International Journal of Thermofluids, 2022; 16: 100213. https://doi.org/10.1016/j.ijft.2022.100213
27. Halon S, Krolicki Z, Revellin R, Zajaczkowski B. Local flow patterns distribution during flow boiling in a micro channel array. Experimental Thermal and Fluid Science, 2023; 141: 110792. https://doi.org/10.1016/j.expthermflusci.2022.110792
28. Fu Y, Liu X, Wang J. Optimal design of wavy microchannel heat sinks based on prediction and multi-objective optimization algorithm. Energy and AI, 2023; 14: 100260. https://doi.org/10.1016/j.egyai.2023.100260
29. Chen CH, Yaji K. Topology optimization for microchannel heat sinks with nanofluids using an Eulerian-Eulerian approach. International Journal of Heat and Mass Transfer, 2025; 98(3): 1125-40. https://doi.org/10.1016/j.ijheatmasstransfer.2025.1125
30. Nasiri Khamesloo F, Domiri Ganji D. Numerical investigation of fractal microchannel heat sinks with modified fin geometries. Iranian Journal of Energy and Environment, 2025; 16(2): 115-25. https://doi.org/10.5829/ijee.2025.16.02.03
31. Nasiri Khamesloo F, Domiri Ganji D. Optimizing heat dissipation in microchannel heat sinks using hybrid nanofluids: A computational study. ResearchGate Preprint, 2024.
32. Zhu J, Li Y, Chen Q. Performance comparison of microchannel heat sinks with various rib shapes and cavities. International Journal of Thermal Sciences, 2024; 192: 108259. https://doi.org/10.1016/j.ijthermalsci.2024.108259
33. Li X, Wang M, Zhang T. Optimizing finned microchannel heat sink design for enhanced thermal performance. Applied Thermal Engineering, 2024; 241: 122426. https://doi.org/10.1016/j.applthermaleng.2024.122426
34. Li Y, Zhao Q, Chen J, Wang L. Optimization of microchannel structures using artificial neural networks and genetic algorithms. Nano Convergence, 2024; 11(1): 20. https://doi.org/10.1007/s42114-024-01002-5
35. Huo Y, Sun Q. Hierarchical embedded microchannel heat sink for ultra-high heat flux SiC chip cooling. arXiv Preprint, 2025; arXiv:2502.18117.
36. Chej M, Pashaei M, Vahdati M. Heat transfer modeling in microfluidic heat sinks with variable thermophysical properties. arXiv Preprint, 2024; arXiv:2402.09931.
37. Al-Bahrani R, Kamari A, Heidari S. Additive manufacturing of topology-optimized heat sinks: A review. Heat Transfer Engineering, 2025; 46(1): 17-35. https://doi.org/10.1080/01457632.2025.2459981
38. Jalilinasrabady S, Hashemnia A, Kalantari M. Performance analysis of inclined microchannel-incorporated heat sinks for electronics cooling. Journal of Enhanced Heat Transfer, 2025; 32(3): 231-44.
39. Dehkordi AR, Farhadi M, Akbari M. Optimization of sinusoidal and Z-shaped microchannels for improved cooling performance. International Communications in Heat and Mass Transfer, 2024; 147: 107573. https://doi.org/10.1016/j.icheatmasstransfer.2024.107573
40. Mousavi F, Rahimi M. Effect of varying cross-sectional microchannel designs on heat sink performance. Experimental Thermal and Fluid Science, 2024; 141: 110851. https://doi.org/10.1016/j.expthermflusci.2024.110851
41. Ghasemi M, Arasteh AA. Thermal analysis of fluted-fin microchannels in laminar flows. Energy Conversion and Management, 2024; 284: 117128. https://doi.org/10.1016/j.enconman.2023.117128
42. Alimohammadi H, Jamshidi N. Investigation of Y-shaped bifurcated microchannels for thermal management applications. Applied Thermal Engineering, 2024; 243: 122735. https://doi.org/10.1016/j.applthermaleng.2024.122735
43. Liu H, Shi X. Triangular lattice network design for uniform heat distribution in microchannel heat sinks. Thermal Science and Engineering Progress, 2025; 30: 101301. https://doi.org/10.1016/j.tsep.2025.101301
44. Lu B, Jiang PX. Experimental and numerical investigation of convection heat transfer in a rectangular channel with angled ribs. Experimental Thermal and Fluid Science, 2005; 30(6): 513-21. https://doi.org/10.1016/j.expthermflusci.2005.09.007

Persian Abstract

چکیده

این مطالعه به بررسی عملکرد حرارتی–هیدرولیکی کانال‌های موج‌دار مجهز به ریب‌های عرضی می‌پردازد و به نیاز به راهکارهای بهینه‌سازی انتقال حرارت در مبدل‌های حرارتی فشرده پاسخ می‌دهد؛ جایی که اثرات ترکیبی هندسه ریب و رژیم جریان به‌طور کافی مورد مطالعه قرار نگرفته‌اند. هدف این تحقیق، ارزیابی تأثیر زاویه شیب، ارتفاع، و تعداد ریب‌ها بر بهبود انتقال حرارت و کارایی کلی، در شرایط جریان آرام (Re = 500) و آشفته (Re = 5000) است. شبیه‌سازی‌های عددی سه‌بعدی با استفاده از نرم‌افزار ANSYS Fluent برای تحلیل ویژگی‌های انتقال حرارت، افت فشار و معیار ارزیابی عملکرد (PEC) انجام شده‌اند. نتایج نشان می‌دهند که ریب‌های عرضی با ایجاد جدایش جریان و تشکیل گردابه‌ها، به‌طور چشمگیری انتقال حرارت را افزایش می‌دهند. در جریان آرام، عدد ناسلت تا حدود ۵۰٪ نسبت به کانال صاف افزایش یافت، به‌ویژه در ریب‌هایی با ارتفاع متوسط(h/H = 0.2–0.4) ، زاویه ٤٥ تا ٦٠ درجه و تعداد کم (N = 1–2) که مقدار PEC بهینه نیز حاصل شد. در جریان آشفته، افزایش انتقال حرارت محسوس‌تر بود (عدد ناسلت حدود ٥٦ تا ٥٨)، اما اصطکاک بیشتر منجر به کاهش مقدار PEC به کمتر از یک در اکثر موارد شد. در این تحقیق، تعادل‌های نوینی شناسایی شد که بین عملکرد حرارتی و کارایی کلی تعادل ایجاد می‌کنند، به‌ویژه در شرایط جریان آرام. این یافته‌ها دستورالعمل‌های طراحی ارزشمندی را برای بهینه‌سازی هندسی کانال‌های موج‌دار دارای ریب در رژیم‌های مختلف جریان ارائه می‌دهند.

International Journal of Engineering

Journal Homepage: www.ije.ir

Impact of Effective Pressure Variations on Reservoir Rock Porosity and Compressibility

V. A. Iktissanov, M. S. Bernatov*

Petroleum Engineering Department, Empress Catherine II Saint Petersburg Mining University, Saint Petersburg, Russia

P A P E R I N F O

Paper history:
Received 17 September 2025
Received in revised form 15 December 2025
Accepted 03 Januray 2026

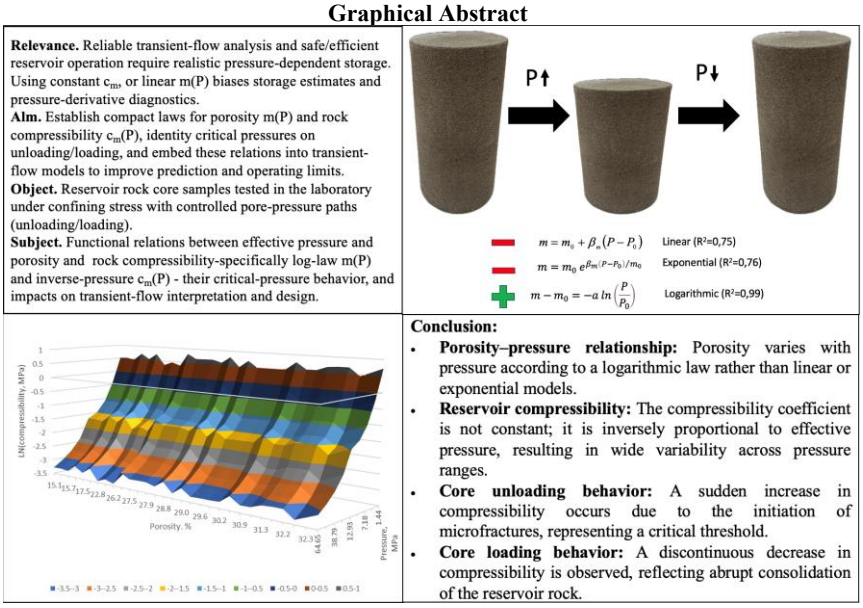
Keywords:
Porosity
Compressibility
Pressure
Exponential Dependence
Logarithmic Dependence

A B S T R A C T

Pressure transient analysis programs typically use a linear diffusion equation. However, according to extensive petrophysical investigations by various authors, the relationship between porosity and pressure is nonlinear. This factor, like the reservoir compressibility coefficient, can significantly impact the accuracy of various transient flow problems. This applies particularly to determining the distance to reservoir boundaries and estimating reserves using pressure buildup tests for exploratory wells. Therefore, the aim of this study was to investigate the predominant type of porosity pressure function for subsequent refinement of the diffusion equation used for PTA. To solve the set task, petrophysical studies of core samples from terrigenous and carbonate formations were conducted in a pressure range of 1.44–64.65 MPa (200–9000 psi) using a compression rig. The results were compared with studies by other authors. These studies are not unique, but the results obtained differ from generally accepted ones. It is shown that the dependence of porosity on pressure follows a logarithmic law, rather than an exponential one, as is commonly believed. PTA calculations have demonstrated a significant impact of compressibility coefficient variability on the accuracy of determining the distance to reservoir boundaries. A detailed study revealed that core unloading results in an abrupt change in the compressibility coefficient, caused by the formation of microcracks. Therefore, using linear and exponential porosity-pressure relationships with a constant compressibility coefficient is highly inappropriate for solving PTA problems.

***doi**: 10.5829/ije.2026.39.10a.14*

Graphical Abstract



*Corresponding Author Email: bernatovm@mail.ru (M. S. Bernatov)

Please cite this article as: Iktissanov VA, Bernatov MS. Impact of Effective Pressure Variations on Reservoir Rock Porosity and Compressibility. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2521-8.

1. INTRODUCTION

The pressure diffusivity equation (1, 2) is used to describe unsteady-state filtration processes in the oil reservoirs numerical modeling and for the pressure transient analysis (PTA). The classical linear pressure diffusivity equations are valid for slightly compressible fluids and rocks, where the compressibility coefficients can be assumed constant. It is also assumed that permeability and viscosity are invariable parameters (3, 4).

Undoubtedly, these assumptions are far from reality, as none of the noted parameters is a constant (5). Several modifications of the diffusivity equation have been proposed to date – for deformable reservoirs (6, 7), for the filtration of non-linear fluids, accounting for relaxation processes (8-10), high flow velocities, etc.

The accuracy of describing real filtration by the diffusivity equations significantly impacts the precision of the calculations (11). These calculations, in turn, affect the reliability of the determined reservoir parameters and the success of various methods for designing, monitoring, and managing oil field development processes. The compressibility of the reservoir rock and the saturating fluids is one of the most significant factors in describing unsteady-state filtration processes (12, 13). Therefore, this work focuses on the task of more accurately accounting for rock compressibility in the diffusivity equations, as these parameters govern the rate of pressure redistribution within the reservoir.

All pressure diffusivity equations are based on four fundamental equations: the continuity equation, the filtration equation, and the equations of state for the porous medium and the fluid.

Let's consider the most common pressure dependences of porosity, presented in Equations 1 and 2.

In most cases, a linear equation is adopted to describe the dependence of porosity on pressure (Equation 1):

$$\phi = \phi_0 + c_f (P - P_0), \quad (1)$$

where ϕ_0 - initial porosity at initial pressure P_0; ϕ - porosity at pressure P; c_f - formation compressibility.

Some authors believe that porosity is best described by an exponential relationship (Equation 2):

$$\phi = \phi_0 \, e^{c_f (P - P_0)/\phi_0}. \quad (2)$$

Given the Taylor approximation $e^x \cong 1+x$, the linear Equation 1 is easily obtained from the exponential relationship 2.

At this point, the problem of describing porosity from pressure could be considered solved, but as shown by our further studies and studies of other authors, the dependence of porosity on pressure has a completely different character in contrast to Equations 1 and 2.

Another, no less important parameter for describing the processes of unsteady filtration is the compressibility coefficient (Equation 3), characterising the pore volume change at some pressure change:

$$c_f = -\frac{1}{V}\frac{\partial V_p}{\partial P} = -\frac{\partial \phi}{\partial P}, \quad (3)$$

where ∂V_p is the pore volume change, ∂P is the pressure change, V is the pore volume of the rock element, m is the porosity.

Sometimes, the reservoir compressibility coefficient is determined by a slightly different Equation 4:

$$c_f = \frac{\partial \phi}{\phi \partial P}, \quad (4)$$

from which the exponential pressure Equation 2 is derived.

The compressibility coefficient significantly influences pressure redistribution processes in porous and fractured reservoirs during unsteady-state flow (14, 15). As an example, a 10% error in total compressibility leads to a 5% error in determining the distance to reservoir boundaries (16).

Typically, in diffusivity equations and various field numerical modeling and PTA programs, formation compressibility is assumed constant, with primary focus placed solely on fluid compressibility (17).

Therefore, the aim of this work is to determine more accurate dependencies of formation porosity and compressibility coefficient on pressure to enhance the accuracy of unsteady-state fluid flow calculations.

2. METHOD AND MATERIALS

Experimental studies were conducted on a PBC-920 filtration equipment to determine porosity and pore compressibility coefficient under triaxial compression across a wide pressure range. Pore pressure remained unchanged and equal to atmospheric pressure. Preliminary preparation involved extracting native reservoir fluids from the pore space of the samples via extraction. The samples were then dried for 48 hours at 80°C. Subsequently, the dry samples were weighed, vacuumed, and saturated with brine having a density of 1170 kg/m³. Finally, the initial porosity was estimated from the mass difference between the dry and saturated samples, which was taken as the volume of fluid imbibed into each sample.

Rock samples with a diameter of 2.5 cm were placed in a high-pressure chamber. Confining pressure was increased to 64.65 MPa (9000 psi) and then stepwise reduced to fixed values of 51.72, 38.79, 25.86, 12.93, 10.77, 7.18, 3.59, and 1.44 MPa. At each stage, the increase in pore volume was recorded using a graduated pipette. The studies were performed on 60 reservoir rock samples from fields of the Verkhnekamsk Depression, with half being terrigenous rocks of the Visean stage and the other half being carbonates (13 from the Vereiskian

horizon of the Moscovian stage and 17 from the Bashkirian stage). The parameters of the investigated rock samples are presented in Table 1. Overall, the conducted research represents typical core analysis procedures.

3. ANALYSIS OF RESEARCH RESULTS. LOGARITHMIC DEPENDENCE OF POROSITY ON PRESSURE

The results of experimental studies, presented in Figures 1 and 2, show an increase in porosity with decreasing pressure for sandstones in the range of 2.9 ÷ 6.9% and for limestones in the range of 1.6 ÷ 5.8%, which in itself is not unusual. To avoid cluttering the graphs, the equations and correlation coefficients are shown only for the extreme curves.

What was unusual for us was that the ratio of porosity to initial porosity (at a pressure of 1.44 MPa) is described with high accuracy by a logarithmic equation of the type (Equation 5):

$$\phi = a\ln P + b, \tag{5}$$

where a and b - coefficients of line.

The correlation coefficient R^2 for the all experimental data ranges from 0.962 to 0.998, unlike other dependencies. This contradicts the linear and exponential

TABLE 1. Parameters of rock samples

Parameter name	Sandstones	Carbonates
Age of the rock	C1v	C2vr/ C2b
Formation depth, m	1688.1-1737.1	1358.4-1365.5 1375.8-1392.0
Porosity, %	16.1-33.3	14.2-24.8/8.1-27.1
Permeability, mD	3.61-931.9	0.68-732.5

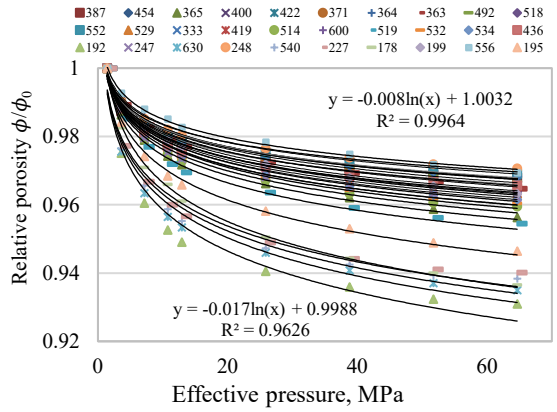


Figure 1. Pressure dependence of relative porosity for sandstones (parameter of the curves is the number of the core)

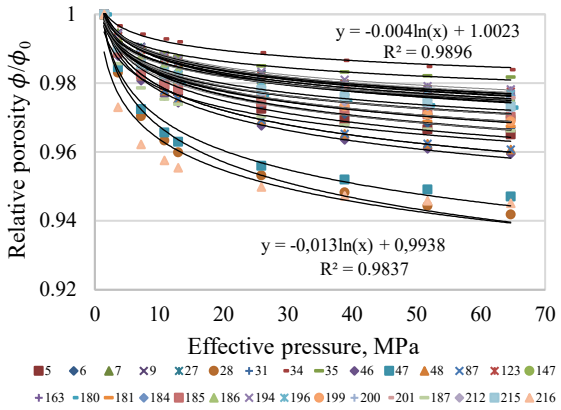


Figure 2. Pressure dependence of relative porosity for limestones (parameter of the curves is the number of the core)

dependencies $\phi(P)$ in Equations 1 and 2, which are used in commercial PTA software, particularly in the Saphir program.

It can be assumed that for small pressure changes, the logarithmic function is close to linear, considering the Taylor series expansion ln(1+x)≅x (18, 19). However, during hydrocarbon field depletion under natural drive mechanisms, the decrease in reservoir pressure compared to the initial pressure can be substantial. For example, in a number of gas fields in Western Siberia, reservoir pressure has decreased by more than an order of magnitude. In some small oil fields in the Ural-Volga region, developed without pressure maintenance or with significant delays in its implementation, reservoir pressure dropped to 0.4 – 0.5 of the initial pressure.

Naturally, in these cases, it is necessary to account for a more accurate dependence of porosity on pressure. Literature analysis showed that the logarithmic dependence of porosity on pressure has been known since the work of Terzaghi (Equation 6), and this formula is over 90 years old:

$$\phi - \phi_0 = -a\ln\left(\frac{P}{P_0}\right), \tag{6}$$

This formula can be easily derived from Equation 5 by subtracting the expression $\phi_0 = a\ln P_0 + b$ for a specific initial condition. If we base the compressibility coefficient on Equation 3 and differentiate Equation 6 or 5 with respect to pressure, we obtain Equation 7:

$$c_f = \frac{\partial\phi}{\partial P} = (-a\ln P + b)' = -\frac{a}{P}, \tag{7}$$

where a is the dimensionless coefficient preceding the logarithm. Expression 7 shows that the compressibility coefficient is not a constant but is inversely proportional to pressure. In this study, compressibility coefficients for carbonate and terrigenous reservoirs were obtained in the range of 3×10^{-5} – 2×10^{-3} MPa^{-1}, which is significantly wider than the ranges reported by other authors. The

primary reason for this discrepancy is the pressure dependence of the compressibility coefficient.

By performing the inverse operation of integrating Equation 7, we obtain the Terzaghi Equation 8:

$$\phi - \phi_0 = -\int_{P_0}^{P} c_f(P)dP = -a\ ln\left(\frac{P}{P_0}\right) \quad (8)$$

Some researchers employ various approaches to determine pore compressibility, such as integral, differential, and their variants, which lead to different values of the coefficient (20-22). The reason for this lies in the fact that compressibility coefficients differ depending on the pressure steps used. For instance, from initial to final pressure or with fine incremental steps. We have adopted a fundamentally different approach—using the derivative according to the definition of the compressibility coefficient 3. This eliminates the need to specify different pressure steps, as the derivative inherently assumes an infinitesimal step size. To determine the change in porosity, integration over specified intervals is required (23). For this purpose, it is first necessary to fit the experimental data points with an equation exhibiting a high correlation coefficient. As previously mentioned, the highest accuracy in describing porosity is achieved using the logarithmic dependence.

A more detailed analysis of the obtained porosity versus pressure curves, presented in Figures 1 and 2, reveals the existence of a distinct break point in the logarithmic curves. This is most clearly visible on the graph with the logarithm of pressure (Figure 3).

At the point of this abrupt structural change, a sharp increase in the compressibility coefficient is observed. The cause of this break is likely a change in the structure of the tested samples due to the formation of micro-fractures (24-26). Consequently, the effective pressure at which this process occurs can be interpreted as the fracturing initiation pressure.

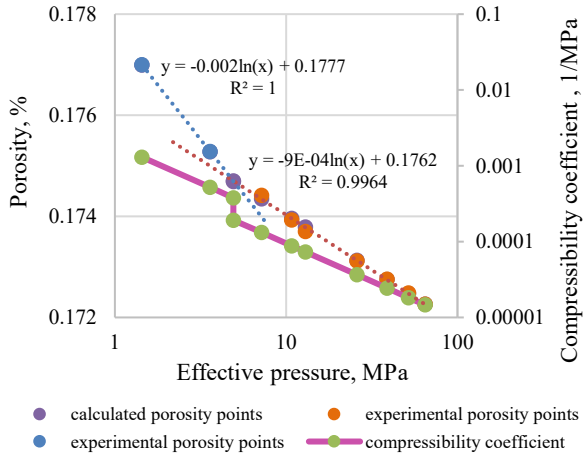


Figure 3. An example of the presence of a sharp change in porosity and compressibility coefficient due to pressure (sample 215 of carbonate formation)

Unfortunately, acoustic emission data or post-test CT scans before and after the experiments were not performed to confirm this assumption. However, supporting this hypothesis is the fact that, on average, the intersection point of the logarithmic curves is observed at a mean pressure of 8.1 MPa for the carbonate rocks of the studied fields and about 8.7 MPa for terrigenous rocks (Table 2). This range corresponds to the order of magnitude of the fracturing pressure for injection wells in the region under consideration (27, 28). As an example, Figure 4 presents result of inflow performance relatiomship (IPR) where the determined effective pressure is 8.6 MPa.

Additionally, Figure 5 shows the fracturing pressure range of 5.7 ÷ 12.2 MPa, which depends on the effective pressure. These studies were performed for injection wells in Tournaisian carbonate deposits (29, 30).

When the injection pressure increases, it leads to the opening of natural fractures or the creation of induced fractures, which alters the injectivity index and fluid conductivity. Determining this pressure is of significant practical importance because exceeding the fracturing pressure during injection leads to inefficient fluid loss, a rapid increase in water cut in responding production wells, and other negative factors (31-33). These issues are discussed in more detail in our monograph on the limiting and optimal pressures for injection and production wells (34, 35).

TABLE 2. Statistical results of pressure of change of core structure

Parameter name	Sandstones	Carbonates
Average effective pressure, MPa	9.6	8.7
Standard error, MPa^2	0.6	0.7
Minimum, MPa	4.1	1.4
Maximum, MPa	15.6	17.3
Number of samples	30	30

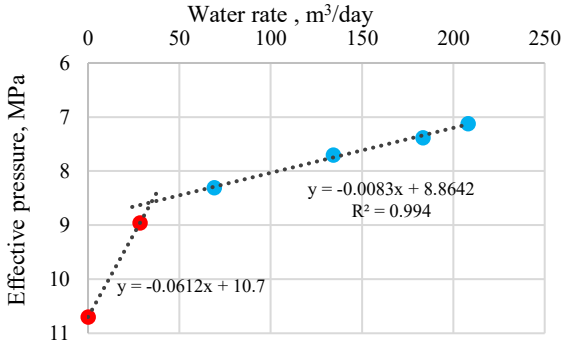


Figure 4. An example of determining the fracturing pressure (8.6 MPa) according to IPR for the injection well (№ 3081 Tournaisian deposits of JSC “Tatneft”)

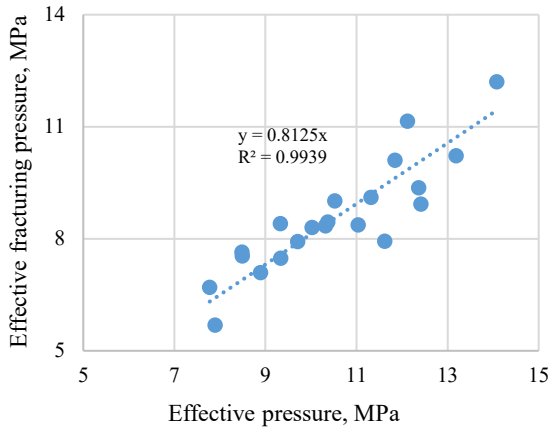


Figure 5. The dependence of fracturing pressure on the effective pressure, determined by IPR of injection wells for Tournaisian deposits

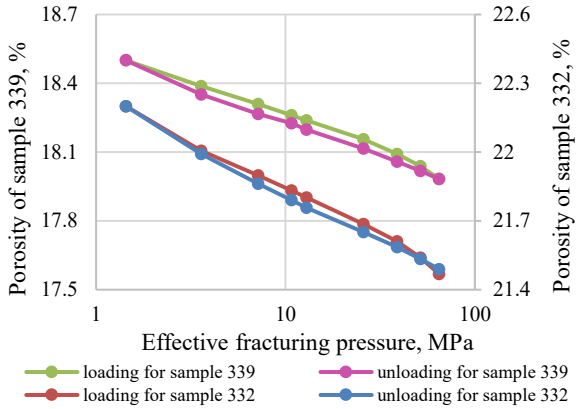


Figure 6. Example of change in compressibility coefficient under unloading and loading for two core samples

In conclusion, the fracturing pressure can potentially be determined not only from IPR but also from the core unloading curve. Notably, the nature of the dependence of the compressibility coefficient on pressure remains unchanged and inversely proportional to the effective pressure (Equation 7).

Naturally, of interest are not only core unloading but also the deformation processes during loading. For this purpose, tests involving core unloading and subsequent reloading were conducted on two samples. The nature of the curves, presented in Figure 6, indicates an approximate return of porosity to its initial state, which aligns with the results of other researchers. However, the pattern of the porosity recovery differs from the original, which is also consistent with the well-established concept of the difference between loading and unloading elastic moduli.

4. NON-LINEAR DIFFUSION EQUATION

As noted earlier, expression 7 shows that the compressibility coefficient is not a constant but is inversely proportional to pressure, which violates the linearity of the diffusion equation and complicates its solution (Equation 9):

$$\frac{\partial P}{\partial t}\frac{\mu\alpha}{k(P+P_0)} = \nabla^2 P, \quad (9)$$

where P - pressure, P_0 - initial bottomhole pressure, t - time, k - permeability, μ - oil viscosity.

The coefficient α in Equation 9 is determined by the rule of additivity for the reservoir and oil (Equation 10):

$$a = \phi\alpha_l + (1-\phi)c_f, \quad (10)$$

where α_l is the coefficient for fluid and rock.

The possibility of using Equation 10 is justified by the fact that at pressures above the bubble point pressure, according to the empirical formula of Vasquez, M., Beggs, H.D., the oil compressibility is also inversely proportional to pressure (36).

An approximate solution of Equation 9 using the Laplace transform is presented in our work. The main results of the calculations show a significant difference in the distance to the impermeable boundary when using the proposed model compared to the traditional solution (Figure 7, Table 3).

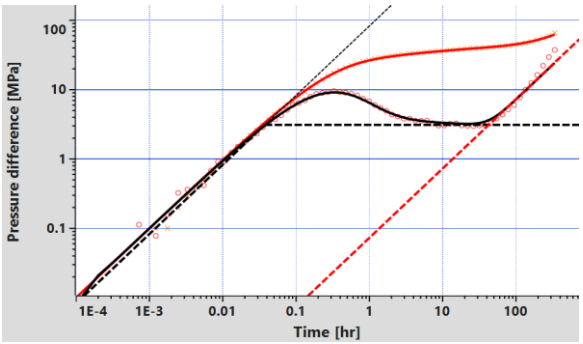


Figure 7. An example of interpretation of the created pressure curve where $c_t = a/P$ using the linear filtration model $c_t = const$ (the red curve is pressure, the black curve is the pressure derivative, the red dashed curve with an angle of 1/2 is a sign of a closed boundary, the black horizontal dashed curve is a sign of radial flow)

TABLE 3. Comparison of the results of interpretation of the pumping curve by different models

Parameter name	According to the proposed model, a =0.0008	According to the traditional model (Saphir) c_t= 8-10 Pa^{-1}
Wellbore storage, m^3/MPa	0.01	0.012
kh, mkm^2·m	10.0	9.2
Skin	0.10	1.1
The radius of the closed circle, m	250.0	32.5

5. CONCLUSION

Porosity–pressure relationship: Thus, the predominant type of porosity versus pressure function is the rather logarithmic dependence, than the linear or exponential dependencies. This confirms the conclusion made by K. Terzaghi over 90 years ago.

Reservoir compressibility: It was found that in this case, the formation compressibility coefficient is inversely proportional to the effective pressure, resulting in a wide range of its variation. For the considered pressure range of 1.44 – 64.65 MPa (200-9000 psi), the compressibility coefficient varied within $3\times10^{-5} – 2\times10^{-3}$ MPa^{-1}.

It has been demonstrated that the dependence of the compressibility coefficient on pressure significantly affects the accuracy of determining the distance to reservoir boundaries using pressure buildup data.

Core unloading behavior: A detailed study of the porosity curves revealed that for all investigated core samples during unloading, an abrupt change in the compressibility coefficient occurs, which may be caused by the formation of micro-fractures. The obtained pressure range correlates with the fracture pressure determined from indicator diagrams for injection wells.

Acknowledgements

The author appreciates Empress Catherine II Saint Petersburg Mining University for providing necessary facilities to conduct present research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.

The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Author(s) Biosketches

Valery Iktissanov - Full PhD (Doctor of Technical Sciences) in Petroleum Engineering, 2002, professor 2012. Worked as an associate professor at the Almetyevsk Petroleum Institute and was head of the WellTest Laboratory at R&D Institute.

REFERENCES

1. Shchelkachev VN. Fundamentals and applications of the theory of unsteady filtration. 1995; 493 p.
2. Hurst W. Establishment of the skin effect and its impediment to fluid flow into a well bore. The Petroleum Engineer, 1953.
3. Petrakov DG, Penkov GM, Zolotukhin AB. Experimental study on the effect of rock pressure on sandstone permeability. Journal of Mining Institute, 2022; 254: 244-51. https://doi.org/10.31897/PMI.2022.24
4. Bykov AA, Bychkov AA, Bereznikova MV. Modified piezoconductivity equation for modeling unsteady filtration processes in a cavernous medium. MSEA, 2022; 71(2). https://doi.org/10.17762/msea.v71i2.75
5. Barenblatt GI, Yentov VM, Ryzhik VM. Theory of unsteady filtration of liquid and gas. Moscow: Nedra; 1972. 288 p.
6. Zheltov YP. Mechanics of oil and gas bearing strata. Moscow: Nedra; 1975. 216 p.
7. Warren JE, Root PJ. The behavior of naturally fractured reservoirs. Society of Petroleum Engineers Journal, 1963; 3(3): 245-55. https://doi.org/10.2118/426-PA
8. Vyalov SS. Rheological bases of soil mechanics. Higher School; 1978. 447 p.
9. Panteleev IA, Lyakhovsky V, Mubassarova VA, Karev VI, Shevtsov NI, Shalev E. Tensor compaction of porous rocks: Theory and experimental verification. Journal of Mining Institute, 2022; 254: 234-43. https://doi.org/10.31897/PMI.2022.30
10. Hu Y, Guo Y, Qing H, Hou Y. Study on influencing factors and mechanism of pore compressibility of tight sandstone reservoir: A case study of Upper Carboniferous in Ordos Basin. Frontiers in Earth Science, 2023; 10: 1100951. https://doi.org/10.3389/feart.2022.1100951
11. Allain O, Tauzin E. Dynamic flow analysis. KAPPA; 2018.
12. Molokovich YM, Markov AI, Davletshin AA. Piezometry of well neighbourhood. Kazan: DAS; 1990. 203 p.
13. Savinkova LD. Fundamentals of underground oil and gas hydromechanics. 2017.

14. Filippov AI, Sabitov KB. Initial-boundary problem for the piezoconductivity equation describing filtration in a layered oil reservoir. Lobachevskii Journal of Mathematics, 2024; 45(11): 5467-79. https://doi.org/10.1134/S1995080224606568
15. Guo W, Li J, Wang T, He T, Xie D, Liao Y, Liu C. Experimental study on the evolution law of permeability characteristics of salt rocks under different temperatures and pore pressures. Rock Mechanics and Rock Engineering, 2025; 58(3): 4091-113. https://doi.org/10.1007/s00603-024-04349-9
16. Santiago V, Rodger I, Hayes P, Leonardi C, Deisman N. Understanding compressibility coefficients in heterogeneous fractured rocks and implications for CSG reservoir simulation. SPE Asia Pacific Unconventional Resources Conference and Exhibition, 2023. https://doi.org/10.2118/217281-MS
17. Khristianovich SA. Mechanics of continuous medium. Science; 1981. 484 p.
18. Farahani M, Aghaei H, Masoumi H. Effect of pore type on porosity, permeability and pore volume compressibility due to in-situ stress change. Journal of Petroleum Science and Engineering, 2022; 218: 110986. https://doi.org/10.1016/j.petrol.2022.110986
19. Nikolaevskiy VN. About the basic equations of the dynamics of elastic porous media saturated with liquid. 2012. 234 p.
20. Zhukov V, Kuzmin Y. Experimental evaluation of compressibility coefficients for fractures and intergranular pores of an oil and gas reservoir. Journal of Mining Institute, 2021; 251: 658-66. https://doi.org/10.31897/PMI.2021.5.5
21. Basniev KS, Nikolaevski VS. On basic equations of filtration in compressible porous media. 1961; 2(3): 52-55.
22. Farahani M, Aghaei H, Saki M, Asadolahpour SR. Prediction of pore volume compressibility by a new non-linear equation in carbonate reservoirs. Energy Geoscience, 2022; 3(3): 290-9. https://doi.org/10.1016/j.engeos.2022.04.005
23. Buslaev GV, Konoplyannikov AV. Mathematical modeling of a hydro-mechanical system with a positive displacement motor and hydraulic thruster to optimize the drilling process for extended-reach drilling wells. International Journal of Engineering Transactions B: Applications, 2026; 39(5): 1077-87. https://doi.org/10.5829/ije.2026.39.05b.03
24. Liu C, Yu B, Zhang D, Zhao H. Experimental study on strain behavior and permeability evolution of sandstone under constant amplitude cyclic loading-unloading. Energy Science & Engineering, 2020; 8(2): 452-65. https://doi.org/10.1002/ese3.527
25. Kallesten E, Andersen P, Korsnes R, Madland M, Omdal E. Modeling of permeability and strain evolution in chemical creep compaction experiments with fractured and unfractured chalk cores at reservoir conditions. SPE Journal, 2020; 25: 2710-28. https://doi.org/10.2118/197371-PA
26. Xu R, Wang L, Zhao X, Mao Y. Characterizing permeability-porosity relationships of porous rocks using a stress sensitivity model considering elastic-structural deformation and tortuosity sensitivity. Journal of Rock Mechanics and Geotechnical Engineering, 2024; 16(9): 3437-51. https://doi.org/10.1016/j.jrmge.2024.01.020
27. Liu B, Yang Y, Li J, Chi Y, Li J, Fu X. Stress sensitivity of tight reservoirs and its effect on oil saturation: A case study of Lower Cretaceous tight clastic reservoirs in the Hailar Basin. Journal of Petroleum Science and Engineering, 2020; 184: 106484. https://doi.org/10.1016/j.petrol.2019.106484
28. Jiang X, Deng S, Li H, Zuo H. Characterization of 3D pore nanostructure and stress-dependent permeability of organic-rich shales in northern Guizhou Depression. Journal of Rock Mechanics and Geotechnical Engineering, 2022; 14(2): 407-22. https://doi.org/10.1016/j.jrmge.2021.08.019
29. Bohnsack D, Potten M, Freitag S, Einsiedl F, Zosseder K. Stress sensitivity of porosity and permeability under varying hydrostatic stress conditions for different carbonate rock types of the geothermal Malm reservoir. Geothermal Energy, 2021; 9(1): 15. https://doi.org/10.1186/s40517-021-00197-w
30. Pshenin V, Sleptsov A, Dukhnevich L. Prospects of improving the vibroacoustic method for locating buried non-metallic pipelines. Eng, 2025; 6: 121. https://doi.org/10.3390/eng6060121
31. Raupov I, Rogachev M, Sytnik J. Overview of modern methods and technologies for the well production of high- and extra-high-viscous oil. Energies, 2025; 18(6): 1498. https://doi.org/10.3390/en18061498
32. Khuzin RR, Andreev VE, Mukhametshin VV, Kuleshova LS, Dubinskiy GS, Safiullina AR. Influence of hydraulic compression on porosity and permeability properties of reservoirs. Journal of Mining Institute, 2021; 251: 688-97. https://doi.org/10.31897/PMI.2021.5.8
33. Podoprigora D, Rogachev M, Byazrov R. Surfactant-polymer formulation for chemical flooding in oil reservoirs. Energies, 2025; 18: 1814. https://doi.org/10.3390/en18071814
34. Gaurina-Medimurec N, Simon K, Mavar KN, Pasic B, Mijic P, Medved I, Brkic V, Hrncevic L, Zbulj K. The circular economy in the oil and gas industry: A solution for the sustainability of drilling and production processes. In: Circular Economy on Energy and Natural Resources Industries. Cham: Springer; 2024: 115-50. https://doi.org/10.1007/978-3-031-56284-6_7
35. Iktissanov VA, Iktissanov AV, Bernatov MS. Rough consideration of changes in formation and oil compressibility during unsteady-state flow. Neftyanaya Provintsiya, 2025; 3(43): 150-69. https://doi.org/10.25689/NP.2025.3.150-169
36. Vasquez M, Beggs HD. Correlations for fluid physical properties prediction. Journal of Petroleum Technology, 1980; 32. https://doi.org/10.2118/6719-PA

Persian Abstract

چکیده

زمینه :مخازن با نفوذپذیری کم و مخازن حاوی نفت با گرانروی بالا به طور فزایندهای به دلیل تخریب ساختار ذخیره و تعدادی از عوامل منفی، به موضوعات توسعه تبدیل میشوند .در نتیجه، زمان لازم برای رسیدن مشتق لگاریتمی فشار به شرایط جریان شعاعی به طور قابل توجهی افزایش مییابد که این امر تفسیر منحنی افزایش فشار را در یک زمان قابل قبول برای پژوهش غیرممکن میسازد.هدف :توسعه و جمعبندی توصیهها برای تحلیل گذرای فشار در مخازن با ضرایب تحرک کم. مواد و روشها :برای حل وظیفه تعیین شده، از روشهای شناختهشده و توسعهیافته توسط نویسنده برای تفسیر و پژوهش هیدرودینامیکی استفاده شد. نتایج :نشان داده شد که بهینهترین روشها عبارتند از : استفاده از منحنی کاهش فشار به جای منحنی افزایش فشار، استفاده از مدلهای جریان قبل از جریان شعاعی هنگام تفسیر منحنیهای فشار برای چاههای با هندسه پیچیده، و تفسیر سنتی دادههای فشار بلندمدت با استفاده از فشارسنجهای ته چاهی. نتیجهگیری :روشهای پیشنهادی امکان کاهش زمان خاموشی برنامهریزی شده چاه در طول آزمایش چاه را فراهم کرده و در نتیجه باعث کاهش تلفات در تولید نفت هنگام حل مسئله تعیین پارامترهای فیلتراسیون مخزن و ناحیه اطراف چاه میشوندتولید چاهها مطابقت دارد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2529-2545

International Journal of Engineering

Journal Homepage: www.ije.ir

Bermuda Weed Optimization: A Scalable Meta-heuristic for Cloud-Based Cruise Control Systems

M. Allahi Rudposhti, A. Bohlooli*, K. Jamshidi

Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran

P A P E R I N F O

Paper history:
Received 23 August 2025
Received in revised form 04 January 2026
Accepted January 7 2026

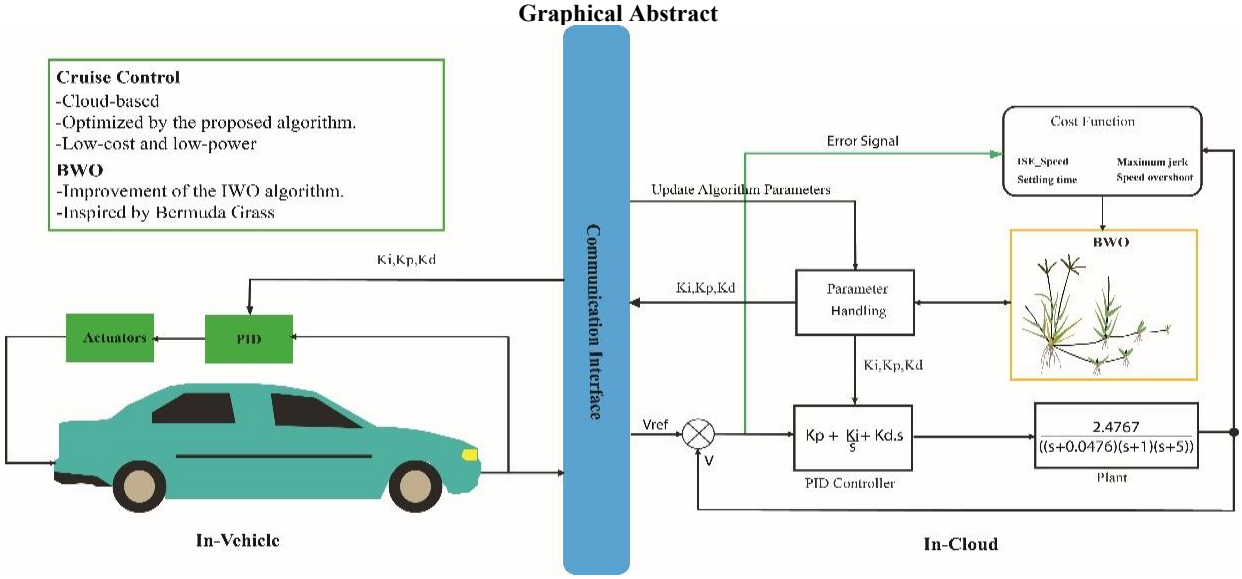
Keywords:
Stochastic Optimization
Cruise Control
Bermuda Weed Optimization
Meta-heuristic Algorithms

A B S T R A C T

This study proposes the Bermuda Weed Optimization (BWO) algorithm; a novel and scalable metaheuristic algorithm inspired by the invasive growth of Bermuda grass. This algorithm has been developed as an enhanced version of the Invasive Weed Optimization (IWO) algorithm and, by imitating the plant's robust propagation strategies, achieves a better balance between global exploration and local exploitation. The algorithm's performance was rigorously evaluated against four previous IWO-based versions, and its superior scalability was demonstrated through the lowest average error and stable performance across diverse scenarios. Furthermore, BWO was compared with the new Gray Squirrel Search Algorithm (GSFA)—which falls outside the IWO category—to assess its performance against a novel method unrelated to the IWO family; this comparison highlighted BWO's competitive superiority and achieved an average 64.43% improvement in best-cost results. The strong convergence and scalability of BWO make it highly suitable for real-time applications—particularly in automotive systems. In a practical implementation using a cloud-based cruise control (CC) framework, BWO significantly outperformed the RPO-based method (the latest approach) by reducing overshoot by 45.92%, settling time by 29.38%, ISE speed by 8.92%, and maximum jerk by 20.09%. By achieving near-optimal convergence and leveraging cloud deployment with high scalability, BWO can effectively adapt to diverse automotive system requirements and achieve high efficiency across multiple operating modes.

***doi**: 10.5829/ije.2026.39.10a.15*

Graphical Abstract



*Corresponding Author Email: bohlooli@eng.ui.ac.ir (A. Bohlooli)

Please cite this article as: Allahi Rudposhti M, Bohlooli A, Jamshidi K. Bermuda Weed Optimization: A Scalable Meta-heuristic for Cloud-Based Cruise Control Systems. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2529-45.

1. INTRODUCTION

In contemporary research, meta-heuristic algorithms are increasingly inspired by biological processes and natural phenomena. These algorithms are widely employed to address complex, challenging, and occasionally intractable problems. They are characterized by their ease of understanding and implementation, adaptability to external factors influencing the problem, effective exploration of the solution space, capability to identify optimal solutions, independence from gradient information, and avoidance of local optima. Generally, meta-heuristic algorithms can be categorized into five groups (Figure 1): Evolution-Based, Physics-Based, Swarm-Based, Human-Based, and Hybrid approaches (1). Among the earliest methods for developing meta-heuristic algorithms are Evolution-Based techniques, which draw inspiration from the principles of biological evolution and natural selection. These algorithms typically incorporate mechanisms such as selection, mutation, inheritance, and natural selection. The Genetic Algorithm (GA) (2) is the most established and widely recognized Evolution-Based algorithm, operating on the principles of natural selection. Another category of meta-heuristic algorithms is the physics-based group, which utilizes physical laws and principles to solve a variety of problems and simulations. Examples of physics-based approaches include Projectiles Optimization (PRO) (3), Henry Gas Solubility Optimization (HGSO) (4), and the Newton-Raphson-Based Optimizer (NRBO) (5). Swarm-based algorithms are another category of metaheuristic algorithm, encompassing the largest group of such algorithms. They are inspired by the collective behavior of living organisms, such as birds, fish, ants, and bees, to solve complex problems. These algorithms are commonly used for optimization and searching in large, complex spaces. Examples of swarm-based algorithms include Invasive Weed Optimization (IWO) (6), Grey Wolf Optimizer (GWO) (7), Grasshopper Optimization Algorithm (GOA) (8), Crystal Structure Algorithm (CryStAl) (9), Dandelion Optimizer (DO) (10), Walrus Optimizer (WO) (11), Starfish optimization algorithm (SFOA) (12), and Gray Squirrel Foraging Algorithm (GSFA) (13). Human-based algorithms, such as Rock-Climbing Group (RCG) (14), and Gold Seekers Algorithm (GSO) (15). Meta-heuristic algorithms can be combined to improve performance and efficiency in solving complex problems, resulting in hybrid algorithms. For instance, Pérez et al. (16) introduced two meta-heuristics, including the Greedy Randomized Adaptive Search Procedure (GRASP) and a GA (2), are combined and utilized to address a real supply chain scheduling problem.

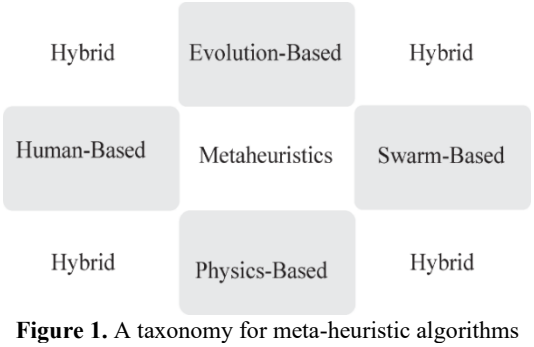


Figure 1. A taxonomy for meta-heuristic algorithms

Among the various meta-heuristic algorithms, the IWO algorithm (6) stands out as one of the most prominent due to its remarkable local and global search capabilities, ease of implementation, robustness in addressing complex real-world problems, and proven efficiency across diverse applications. Inspired by the colony behavior of invasive weeds and their natural mechanisms of growth and competition for resources, IWO has been widely adopted in numerous engineering fields (17-29).

Despite the widespread application of the IWO algorithm and its developed variants in solving various engineering problems, it faces significant limitations when dealing with objective functions that contain numerous local minima (such as the Griewank function). One of the primary challenges of IWO is its inherent tendency to get trapped in local minima that have relatively better quality than their neighboring regions. This issue becomes particularly prominent in the later iterations, where the radius of seed spread decreases, and the seed replication process becomes primarily concentrated around the parent plant. Under such conditions, if the global minimum is located far from the parent's position, the algorithm may mistakenly identify a local minimum as the global minimum. Moreover, the "best solution selection" policy at each step, which is one of the primary exploitation strategies of the algorithm, leads to the exclusion of regions with relatively lower quality but higher potential for discovering the optimal solution. This imbalance between exploration and exploitation processes is a fundamental challenge for IWO, which can limit its efficiency in optimization problems with complex and multi-modal search spaces. Another important limitation of IWO relates to the population control process. Despite the algorithm's high ability to generate diversity in the initial population, the constraints of storage resources and the population reduction policy based on the "best solution at the current time" can result in the elimination of individuals that might have significant potential to reach the optimal solution in subsequent generations. Even the improved versions of IWO based on chaos theory (19) face significantly more challenges compared to the basic algorithm. It should be noted that chaos theory studies deterministic systems with unpredictable behavior and high sensitivity to initial conditions. In optimization, it

helps algorithms escape local optima. The performance of these algorithms is highly dependent on the tuning of IWO parameters and the selected chaos map, to the extent that even a slight change in these parameters can lead to unpredictable results. Furthermore, while in most algorithms, increasing the number of iterations enhances the final solution's quality, chaos-based IWO lacks a clear rule for performance improvement with more iterations. In many cases, increasing iterations even reduces its efficiency. Moreover, the IWO algorithm based on Lévy distribution (27) performs well in limited search spaces. The Lévy distribution is a probability distribution with heavy tails and infinite variance, which enables long-range jumps in search spaces and is particularly effective in metaheuristics for global exploration. However, as the search space expands, its performance significantly declines. For instance, when the domain of the Sheffer benchmark function is extended from the range [-10, 10] to [-100, 100], the algorithm experiences a considerable drop in efficiency and loses its scalability.

The proposed algorithm, which represents an enhanced version of the IWO algorithm, draws inspiration from the intelligent behavior of a specific weed species called Bermuda weed. It successfully establishes an optimal balance between the two key components of local search (exploitation) and global search (exploration). This characteristic leads to efficient and effective convergence to the global optimum solution.

The contributions of this work in the metaheuristic field are summarized as follows:

1. An enhanced variant of the IWO algorithm, named Bermuda Weed Optimization (BWO), is proposed. The primary feature of BWO is scalability. This algorithm was deployed in a cloud-based Cruise Control (CC) system for electric and resource-constrained vehicles. The scalability of BWO allows the system to dynamically balance accuracy and speed based on driving conditions: In accuracy-sensitive scenarios (e.g., traffic), it increases the algorithm's iteration count to achieve higher solution quality .In time-critical situations (e.g., highway driving), it rapidly delivers near-optimal solutions within tightly constrained iteration limits.
2. The proposed BWO-leveraged system demonstrated superior performance in key driving and comfort metrics compared to previous methods. Simulation results demonstrated that BWO significantly outperformed the RPO-based method (the latest approach) by reducing overshoot by 45.92%, settling time by 29.38%, ISE speed by 8.92%, and maximum jerk by 20.09%. With its cloud-based architecture, the system is compatible with resource-limited vehicles while simultaneously enhancing accuracy and delivering superior outcomes in vehicular comfort measures.

The structure of this paper is organized as follows: The basic IWO is introduced in Section 2. The theory of BWO is illustrated in Section 3. Section 4 presents the experiments of benchmark functions and real-world optimization problems, respectively. Finally, the conclusions and future works are summarized in Section 5.

2. THE BASIC IWO

Observing the propagation of invasive weeds in nature reveals their remarkable efficiency in identifying and thriving in the most fertile areas. Despite various attempts to control these plants, most conventional methods have proven ineffective in preventing their spread. Inspired by this phenomenon, researchers have developed a novel approach to solving optimization problems, leading to the introduction of the IWO algorithm. IWO is a swarm-based metaheuristic that mimics the natural behavior of weeds in locating optimal sites for growth and reproduction. This algorithm leverages key characteristics of invasive weeds, including their ability to spread aggressively, high reproduction rate, wide dispersal, and competitive exclusion. The IWO process consists of four main stages: initialization, reproduction, spatial dispersal, and competitive exclusion, which are explained in detail in the following:

Initialization: An initial population is randomly distributed within the search space. Each member of this population, representing a weed, has N-dimensional coordinates and a cost value associated with its position in the search space .

Reproduction: Each agent (weed) in the population has a specific position in the search space, and its fitness is evaluated based on the objective function (or cost) value at that position. The higher the fitness of a weed, the more favorable its position is considered in the search space. The number of offspring (seeds) produced by each weed depends on its fitness and is calculated using Equation 1.

$$S_i = \left\lfloor \frac{f_{max}-f_i}{f_{max}-f_{min}} \times (S_{max} - S_{min}) + S_{min} \right\rfloor \quad (1)$$

where f_i represents the fitness value of the i-th weed, while f_{min} and f_{max} correspond to the lowest and highest fitness values within the current population, respectively. Similarly, S_{min} and S_{max} define the minimum and maximum number of seeds that a weed is permitted to generate. Equation 1 clearly indicates that weeds with higher fitness values (f_i) will produce a greater number of seeds, facilitating a more extensive search in promising regions of the solution space. These seeds are distributed around the parent's position, with a dispersion pattern designed to enhance diversity in the search space. This mechanism allows for better exploration of the problem space, increasing the chances of discovering optimal solutions and ultimately improving the algorithm's

performance. Since the number of seeds is proportional to the parent's fitness, the offspring tend to settle in regions with higher fitness potential. Over successive generations, this process gradually enhances the overall population quality and accelerates the algorithm's convergence toward optimal solutions.

Spatial dispersal: The seeds produced by each weed are distributed randomly across the D-dimensional search space according to a normal distribution with a mean of zero and a standard deviation σ that changes adaptively. This method ensures that the seeds are clustered near the parent weed's position. As the algorithm advances through its iterations, the value of σ gradually reduces from its maximum σ_{max} to its minimum σ_{min}, allowing the algorithm to shift from a broad exploration phase to a more focused exploitation phase, where the search is refined in the more promising areas of the space. The standard deviation σ_k for the k-th iteration is given by Equation 2.

$$\sigma_k = \left(\frac{iter_{max}-k}{iter_{max}}\right)^n \times (\sigma_{max} - \sigma_{min}) + \sigma_{min} \quad (2)$$

where:

- **n** is a nonlinear modulation factor.
- $\mathbf{iter_{max}}$ is the total number of iterations.
- **k** denotes the current iteration number.

Using this value of σ_k, the updated position of a seed is computed using Equation 3.

$$x'_i = x_i + N(0, \sigma_k^2) \quad (3)$$

Where:

- $\mathbf{x_i}$ is the i-th weed in the current population,
- $\boldsymbol{N(0, \sigma_k^2)}$ is a random number drawn from a normal distribution with a mean of zero and a standard deviation of σ_k.

This adaptive strategy enhances the algorithm's efficiency by initially promoting wide exploration across the search space and gradually focusing on the most promising regions to refine the solution, thus accelerating convergence towards optimal solutions.

Competitive exclusion: After the seeds are generated, the parent plants transform into weeds, and the total population of weeds consists of both the parent plants and their offspring. This new population is then sorted based on fitness, and if the number of weeds exceeds the maximum allowed colony size (P_{max}), weeds with lower fitness are eliminated to maintain the population within the allowed limit.

Repeat Until Stop: Repeat the above steps until the termination condition is met.

3. BERMUDA WEED OPTIMIZATION (BWO)

This section introduces the mimicry mechanism, the mathematical model and the procedure execution of BWO algorithm.

3. 1. Mimicking Bermuda Grass (Cynodon dactylon), a significant species of the Poaceae (grass) family, is a perennial, turf-forming, and resilient plant that plays a key role in agriculture, soil cover, and erosion control. This species, with its rapid growth and remarkable adaptability to diverse climatic conditions, particularly thrives in tropical and subtropical regions. It is recognized as one of the most important forage sources in arid and semi-arid areas. In addition to its forage utility, it is widely used in urban green spaces and sports fields due to its resistance to environmental stresses (e.g., drought and livestock grazing). However, the high invasive potential of this plant in certain ecosystems has turned it into a challenging weed. Its rapid growth, extensive reproduction through seeds and stolons, and resilience to adverse conditions have led to its swift spread in agricultural lands, orchards, and natural ecosystems, making control efforts difficult [30].

Sexual Reproduction (via seeds): Seeds produced by the plant are dispersed in proximity to the parent plant. This dispersal pattern mirrors the local search strategy in optimization algorithms, which aims to refine solutions within the vicinity of the current solution.

Vegetative Reproduction (via stolons): Stolons (horizontal stems) have nodes that, upon contact with soil, develop adventitious roots and give rise to new shoots, which grow into independent plants. This mechanism enables the plant to spread farther from the parent plant and corresponds to the global search strategy in optimization algorithms, which explores new regions within the search space.

Growth and Dispersal Pattern: Figure 2 illustrates a sample of Bermuda grass. The parent plant covers extensive areas through horizontal stolons. Each stolon can produce multiple nodes that develop into new plants after rooting. Additionally, vertical stems (culms) bear seeds which, upon dispersal, generate new plants near the parent. Figure 3 presents a conceptual model of this process, which can serve as a basis for computationally simulating the plant's growth. Bermuda grass tends to grow and reproduce more extensively in fertile soils with adequate nitrogen, phosphorus, and potassium. Under such conditions, the production of seeds, stolons, and nodes increases. This behavior is analogous to the optimal solution strategy in intelligent algorithms, where the system seeks peak efficiency under ideal conditions. However, the plant is also capable of surviving in poor soils, albeit with more limited growth.

3. 2. Mathematical Model The proposed algorithm follows the IWO algorithm in seed-based reproduction, employing the same mathematical model for this process. However, its key distinguishing feature is a node-based reproduction strategy. In this strategy, each plant produces one or more horizontal branches, with the number of stolons determined by the soil quality—

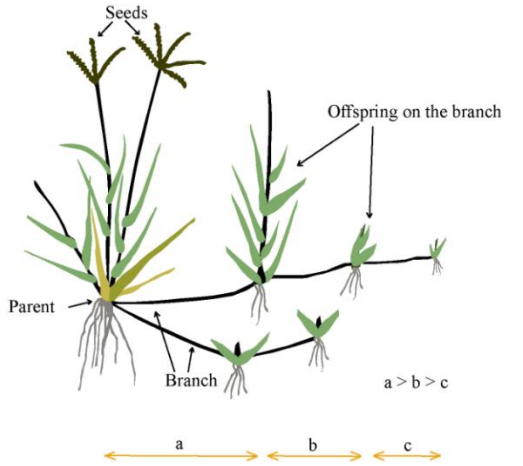


Figure 2. Morphological features of Bermuda Grass

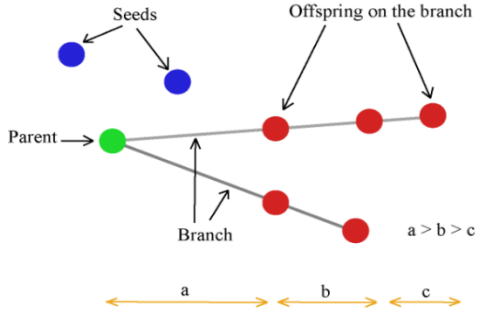


Figure 3. Conceptual model of Bermuda Grass morphology

evaluated based on the plant's fitness (i.e., its objective function value). Along each stolon, several offspring (nodes) are then generated at specific intervals from the parent plant, where they develop independently. A conceptual model of a Bermuda grass plant is illustrated in Figure 3. Based on this model, reproduction via seeds follows the mathematical framework of the IWO algorithm: the number of seeds is calculated using Equation 1, which depends on the parent plant's fitness relative to the best and worst fitness values in the population. The offspring positions are generated using Equation 3, where the standard deviation (σ) is dynamically updated via Equation 2 to balance exploration and exploitation. In contrast, reproduction through stolon-based node propagation follows a distinct mechanism, as described below:

Number of Branches: The number of horizontal branches produced by each weed, determined based on the soil quality, is given by Equation 4.

$$B_i = \left\lfloor \frac{f_{max}-f_i}{f_{max}-f_{min}} \times (b_{max} - b_{min}) + b_{min} \right\rfloor \quad (4)$$

where B_i is the number of branches of the i-th parent, and b_{min} and b_{max} define the minimum and maximum number of branch that a weed is permitted to generate.

Number of Offspring per Branch: The number of nodes on each branch is calculated using Equation 5.

$$R_{i,j} = \left\lfloor \frac{f_{max}-f_i}{f_{max}-f_{min}} \times (r_{max} - r_{min}) + r_{min} \right\rfloor \quad (5)$$

Where $R_{i,j}$ denotes the number of nodes on the j-th branch of the i-th parent, and r_{min} and r_{max} define the permissible range for the number of nodes per branch.

Node Spacing: Nodes are distributed along each branch at distances from the parent plant defined by Equation 6.

$$D_{i,j} = \sum_{j=1}^{k} \frac{d_i}{2^{(j-1)}} \quad (6)$$

where $D_{i,j}$ is the distance of the j-th node from the i-th parent, and d_i is the base distance for the i-th parent, adaptively calculated via Equation 7.

$$d_i = \left\lfloor \frac{f_{max}-f_i}{f_{max}-f_{min}} \times (d_{max} - d_{min}) + d_{min} \right\rfloor \quad (7)$$

Where d_{min} and d_{max} are the minimum and maximum values of d_i, respectively.

Node Positioning: The spatial coordinates of the nodes are determined by Equation 8. A random direction vector is first generated to define the branch orientation. Nodes are then positioned along this direction.

$$x'_{i,j} = x_i + D_{i,j} \cdot \frac{u}{\|u\|} \sqrt{\xi} \quad (8)$$

where u is a randomly generated vector with a normal distribution, i.e., u=(N(0,1), N(0,1),...,N(0,1)), $\|u\|$ denotes the norm of the vector u, ensuring its normalization for direction. Additionally, ξ is a uniformly distributed random variable in the range [0,1], x_i represents position of the i-th parent(weed), $x'_{i,j}$ represents the position of the j-th node of the i-th parent.

Weed Competitive Exclusion: In each iteration, the entire population is ranked based on fitness. Only the top-performing individuals are retained for the next generation.

3. 3. Escape from Local Optima

The proposed algorithm effectively prevents getting trapped in local optima by utilizing a node-based mechanism. In this mechanism, each solution in the population can generate several independent search branches with different random directions. Each branch, in turn, consists of a sequence of offspring with exponentially decreasing search step sizes, which creates a dynamic balance between exploration and exploitation. When the population converges towards a local optimum, the long branches open new escape routes from the current attraction basin by creating large mutations in the search space. This feature is further enhanced by an adaptive step size control mechanism, which adjusts the length and number of branches based on the quality of each solution. Consequently, even under strong convergence towards local optima, the algorithm can maintain its global exploration capability and significantly increases

the probability of finding the global optimum. In fact, in IWO-based algorithms, one reason for getting stuck in a local optimum is the convergence of the sigma parameter (σ_k) to zero over time, causing the algorithm to perform only exploitation (local search). Since this point is a local optimum, it produces more offspring, further trapping the algorithm. Furthermore, due to the selection policy favoring the best points, other positions with the potential to reach the global optimum are overlooked. This issue also exists in newer algorithms like GOA (8), where a decreasing coefficient (similar to σ_k) is reduced each iteration, influencing new positions and gradually steering the algorithm solely towards exploitation. However, in the proposed algorithm, the branches are independent of such decreasing coefficients and depend only on the quality of the current position. The higher the quality of a position, the more extensively farther points are explored in addition to the local search. This enables the algorithm to escape local optima more effectively.

Figure 4 shows a simple simulation of the proposed algorithm where a point is trapped in a local optimum. However, because this point is of high quality, it generates the maximum number of branches, and each branch produces the maximum number of points. The branches scatter in different directions, covering the entire search space. Due to the extensiveness of these branches, the probability of creating new points near the global optimum is very high. This allows the proposed algorithm to easily escape the local optimum and move towards the global one. According to the figure, the global optimum is located between two branches, a positioning which guarantees that the algorithm will move towards the global point in subsequent iterations.

3. 4. The procedure of BWO This section examines the procedure of the proposed algorithm. The pseudocode and flowchart of the algorithm are provided in Algorithm 1 and Figure 5, respectively. As shown in Algorithm 1, the algorithm begins with an initial population and then, within a main loop, each individual (weed) utilizes two reproduction mechanisms: 1) seed-based reproduction, where each weed produces a number of seeds according to its fitness, and the offspring positions are calculated using an equation that includes a random component, and 2) node-based reproduction (via runners), where the weed generates additional offspring by simulating runner stems. At the end of each iteration, the parent and offspring populations are combined and sorted according to the cost criterion; then, the top nPop individuals are selected for the next generation.

From a time complexity perspective, assuming a population of P and T iterations, and considering S seeds and B branches per weed—each producing R offspring—the computational cost in the worst case reaches O (T×P×D× (S+B× R)). Therefore, the execution time of this algorithm depends on factors such as population size, dimention (D), number of iterations, maximum number of seeds, maximum number of branches, and maximum number of nodes or roots. Compared to the IWO algorithm, the time complexity of the IWO algorithm is O(T×P×D×S). Therefore, the overhead of the proposed algorithm is 1 + (B.R)/S. This means if the product B.R is smaller than S, the overhead is negligible, and if the product is larger, the overhead increases relative to S.

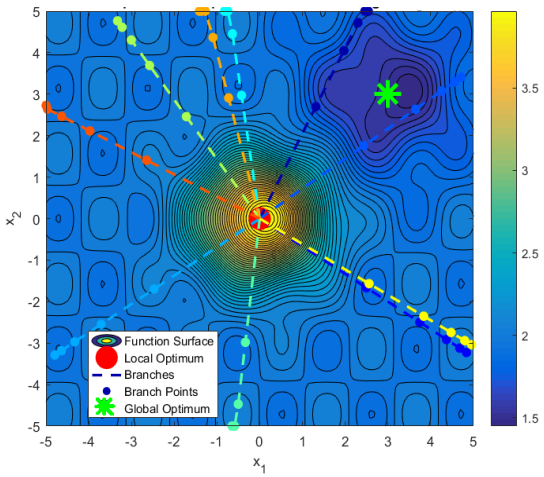


Figure 4. Simulation of the Long Branches mechanism escaping a local optimum.

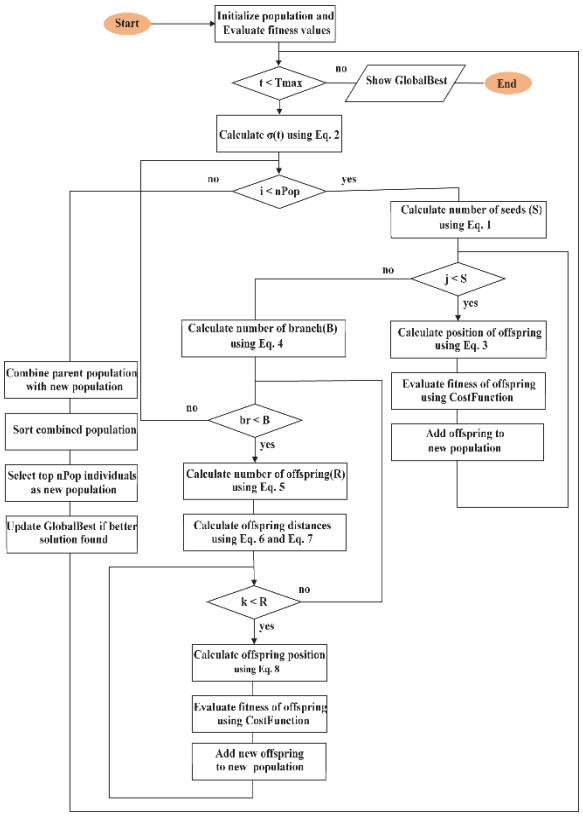


Figure 5. BWO flowchart

When B=1 and R=1, with S=5, the overhead is 1.2, indicating a 20% increase in runtime. In the case where B=2, R=5, and S=10, the overhead becomes 2.0, meaning the proposed algorithm requires twice the computational time of the standard IWO. This demonstrates that the node-based reproduction mechanism, while enhancing exploration, introduces significant overhead when the total number of node-based offspring (B×R) approaches or exceeds the number of seed-based offspring (S). In comparison with the GSFA algorithm, which has a time complexity of O(T×P×D) (equivalent to S=1 in IWO), the GSFA algorithm is S times faster than IWO and (S + B×R) times faster than the proposed algorithm. To reduce the proposed algorithm's complexity, a gradual branch reduction strategy is used. Initially, more branches enable broad exploration to avoid local optima. Later, fewer branches allow focused, faster search for the global optimum.

Algorithm: Bermuda Weed Optimization (BWO) pseudo-code
Input:
1: Algorithm parameters :
- T (Maximum iterations)
- DMin, DMax (Search space bounds)
- P (Population size)
- Other control parameters

Output:
2: GlobalBest (Best solution found)
Begin
%Initialization Phase
3: Initialize population positions randomly within [DMin, DMax]
4: Evaluate fitness values using CostFunction
5: Identify BestCost and WorstCost in population
%Main Optimization Loop
6: for t = 1 to T do
%Parameter Update
7: Calculate σ(t) using Eq. 2
%Reproduction Phase
8: new_population[] =
9: for each weed in population do
% Seed-Based Reproduction
10: Calculate number of seeds (S) using Eq. 1
11: for j = 1 to S do
12: Calculate position of offspring using Eq. 3
13: Evaluate fitness of offspring using CostFunction
14: Add offspring to new_population
15: end for
% **Node-Based Reproduction**
16: Calculate number of branches (B) using Eq. 4
17: for br = 1 to B do
18: Calculate number of offspring (R) using Eq. 5
19: Calculate offspring distances using Eq. 6 and Eq. 7
20: for k = 1 to R do
21: Calculate offspring position using Eq. 8
22: Evaluate fitness of offspring using CostFunction
25: Add selected offspring to new_population
26: end for
27: end for
% Selection Phase
28: Combine parent population with new_population
29: Sort combined population by ascending Cost
30: Select top P individuals as new population
31: Update GlobalBest if better solution found
32: end for
33: Return GlobalBest
End

3. 5. Proposed Cruise Control (CC) In recent years, CC systems have attracted growing interest as a means to partially automate the driving process, primarily to reduce driver fatigue and consequently lower the risks associated with fatigued driving. These systems are designed to regulate vehicle speed based on a reference value set by the driver. An efficiently designed CC system can substantially enhance road safety, improve traffic flow, decrease fuel consumption, and increase overall driving comfort. One of the most widely used approaches for implementing the control unit in such systems is the PID controller. Owing to its straightforward concept, ease of implementation, and reliable performance, the PID controller is commonly employed in CC systems. It ensures speed regulation by modulating the throttle, thus contributing to both comfort and safety. As illustrated in Equation 9, its transfer function integrates proportional, integral, and derivative components to minimize output error.

$$C(s) = K_p + K_d.S + K_i.S^{-1} \quad (9)$$

where K_p is the proportional gain, K_d is the derivative gain, and K_i is the integral gain. In practice, tuning the PID controller's parameters is challenging, particularly for nonlinear systems like vehicle dynamics. Metaheuristic algorithms offer an effective approach for optimizing PID parameters in CC systems, providing advantages including real-time applicability and global search capability (31). To address this challenge, the BWO algorithm is employed for tuning the PID controller in the proposed CC system.

4. 2. 1. Dynamic Model of CC System The vehicle's longitudinal dynamics are derived from Newton's second law, as illustrated in Figure 6, and are mathematically expressed in Equation 10.

$$m\frac{dv(t)}{dt} = F_d - F_a - F_g \quad (10)$$

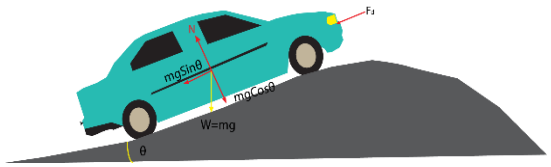


Figure 6. Newton-based longitudinal dynamic model of a CC vehicle

In this equation, F_d is the driving force, $F_a = C_a (v - v_w)^2$ represents the aerodynamic drag force, and $F_g = mgSin\theta$ denotes the gravitational component due to road inclination, often referred to as the climbing resistance. The term $m(dv/dt)$ represents the inertial force acting on the vehicle's mass. The vehicle's total mass (including passengers), aerodynamic drag coefficient, wind speed, and road slope angle are denoted by m, C_a,v_w, and θ, respectively. To simplify the controller design process and focus on the primary vehicle dynamics, a nominal model is considered under ideal operating conditions. Specifically, the vehicle is assumed to travel at a constant nominal speed of v_0=30m/s on a flat road with zero wind speed (θ=0, v_w=0). Furthermore, actuator saturation effects are neglected within the operating range of interest. Under these assumptions, the nonlinear state-space model describing the vehicle's longitudinal dynamics can be formulated as shown in Equations 11 and 12. The first equation represents the vehicle acceleration dynamics, while the second equation captures the actuator dynamics (31-35).

$$v' = \frac{1}{m}(F_d - C_a v^2) \tag{11}$$

$$F_d' = \frac{1}{T}(C_1 u(t-\tau) - F_d) \tag{12}$$

Where T is the time constant of the actuator, C_1 is the actuator gain, u(t) is the throttle control signal, and τ denotes the actuator delay.

Let $v=v_0+\delta v$ and $F_d=F_{d0}+\delta F_d$, where $F_{d0}=C_a v_0^2$ represents the steady-state (where system variables remain constant over time after transient responses have decayed) driving force required to maintain the nominal velocity v_0. Linearizing the simplified nonlinear dynamics in Equations 11 and 12 around the equilibrium operating point (v_0,Fd_0,u_0) yields the perturbation dynamics (i.e., the system response to small deviations from equilibrium), as shown in Equations 13 and 14.

$$\delta v' = -\frac{2C_a v_0}{m}\delta v + \frac{1}{m}\delta F_d \tag{13}$$

$$\delta F_d' = -\frac{1}{T}\delta F_d + \frac{C_1}{T}\delta u(t-\tau) \tag{14}$$

These linearized equations describe the small-signal variations of vehicle velocity and driving force around the steady-state condition. The coefficient $-2C_a v_0/m$ arises from differentiating the nonlinear aerodynamic drag term $-C_a v^2$ with respect to velocity at v=v0, introducing an effective damping proportional to the nominal speed. The term $\delta F_d/m$ represents the influence of driving-force fluctuations on vehicle acceleration, while the second equation captures the actuator dynamics with a time constant T, a static gain C_1/T, and a time delay τ in the control input. Here, $\delta v=v-v_0$, $\delta F_d=F_d-F_{d0}$, and $\delta u=u-u_0$ denote small perturbations from the equilibrium point. The linearized model (Equations 13–14) thus provides a tractable representation for control-system design and stability analysis near the nominal operating speed of 30 m/s. By applying the Laplace transform to the linearized (Equations 13 and 14) under the assumption of zero initial conditions, frequency-domain relationships are obtained. Eliminating $\Delta F_d(s)$ yields the input-output relationship expressed in terms of the Laplace variable. This relationship is derived by solving the Laplace-domain Equations 15 and 16.

$$(s-p_1)\Delta V(s) = \frac{1}{m}\Delta F_d(s) \tag{15}$$

$$(s-p_2)\Delta F_d(s) = \frac{C_1}{T}e^{-\tau s}\Delta U(s) \tag{16}$$

Here, the parameters are defined as $p_1 = -2C_a v_0/m$, $p_2 = -1/T$, and $C = C_1/(mT)$. By substitution and simplification, the transfer function relating the control input $\Delta U(s)$ to the velocity output $\Delta V(s)$ is derived by Equation 17. This equation represents two dynamic poles—one associated with aerodynamic damping and the other with actuator dynamics—along with a time delay manifested as the exponential term $e^{-\tau s}$.

The non-minimum phase delay ($e^{-\tau s}$), characterized by right-half-plane zeros causing initial inverse response, complicates control design. To simplify analysis, this delay is typically approximated using rational functions. For low-frequency applications, the first-order Padé approximation (Equation 17) is commonly employed, matching the Taylor series expansion to represent time delays as rational transfer functions.

$$\frac{\Delta V(s)}{\Delta U(s)} = \frac{Ce^{-\tau s}}{(s-p_1)(s-p_2)} \tag{17}$$

$$e^{-\tau s} = \frac{1}{e^{\tau s}} \approx \frac{1}{1+\tau s} \tag{18}$$

Substituting Equation 18 into Equation 17 yields the final rational transfer function, where the denominator includes the term $1+\tau s$. Thus, the general form of the linearized and approximated model is represented by Equation 19.

$$G(s) = \frac{C}{(s-p_1)(s-p_2)(s+\tau)} \tag{19}$$

By substituting the parameter values given in Table 1 into Equation 19, the simplified Equation 20 is obtained.

$$G(s) = \frac{2.4767}{(s+0.0476)(s+1)(s+5)} \tag{20}$$

TABLE 1. Model Parameter considered

Parameter name	Value	Parameter name	Value
C_1	742	C_a	1.19
F_{dmax}	3500	m	1500
F_{dmin}	-3500	τ	0.2

3. 4. 2. The Proposed CC System Figure 7 presents the block diagram of the proposed CC system. The proposed framework is divided into two main components: in-vehicle and in-cloud modules. The in-vehicle component consists of a simple PID controller that generates the required control force based on the vehicle's instantaneous speed. Additionally, it includes a communication interface that transmits data to the cloud, such as optimization requests, reference speed, algorithm configuration parameters, and updates or queries regarding CC settings. The cloud-based component comprises several integrated modules: an optimization unit based on the proposed BWO algorithm, a cost function designed to enhance driving comfort, and virtualized models of the vehicle dynamics and the PID controller. It also includes a parameter handling unit responsible for handling data exchange between the vehicle and the cloud. The proposed framework performs the optimization process in the cloud based on the defined objective function and real-time input data. Once the optimal PID parameters are computed, they are transmitted back to the vehicle. The onboard controller then applies these parameters to regulate the vehicle speed effectively, ensuring optimal performance and improved driving comfort.

In many control and optimization engineering problems, designers are often required to satisfy multiple conflicting objectives simultaneously. For instance, in a CC system, it is necessary to strike a balance between reference speed tracking accuracy, passenger comfort, system responsiveness, and long-term stability. These objectives inherently conflict with one another—for example, enhancing system responsiveness can increase jerk levels and thereby reduce ride comfort. It is worth noting that multi-objective optimization frameworks such as NSGA-II have been extensively employed in engineering control and logistics applications to manage trade-offs among such competing goals (36, 37). Inspired by these methodologies, the present study introduces a single-objective cost formulation that integrates the aforementioned performance metrics into a unified scalar function, which is then optimized using the BWO algorithm to achieve an efficient and well-balanced control performance. The proposed cost function is defined in Equation 21.

$$J = W_1 \cdot ISE_{Speed} + W_2 . M_p^{speed} + W_3 \cdot T_S + W_4 . \max|j(t)| \quad (21)$$

where:

- $\mathbf{ISE_{Speed}} = \int_0^t (v_{ref} - v_{overshoot}(t))^2 \, dt$: Integral of Squared Error (ISE) for vehicle speed tracking.
- $\mathbf{M_p^{speed}} = \max(v_{overshoot}(t)) - v_{ref}$: Speed over-shoot penalty to ensure passenger comfort and safety.
- **Ts**: Settling time, defined as the time taken for the vehicle speed to stay within ±2% of the reference speed.
- $\mathbf{max}|\boldsymbol{j(t)}|$: Maximum absolute jerk over the simulation horizon, used to penalize abrupt accelerations and decelerations for ride comfort.

The weighting factors W_1 through W_4 are empirically selected based on the relative importance of each component. In the simulation setup, their values are chosen as: W_1=1, W_2=15, W_3=3, and W_4=8. This weighting strategy reflects practical automotive engineering requirements where comfort and safety often supersede strict reference tracking, especially in commercial vehicle applications. Utilizing this integrated cost function enables the optimization framework to perform coordinated tuning of the PID controller parameters, thereby achieving a desirable balance between precision, comfort, responsiveness, and robustness in the CC system's performance.

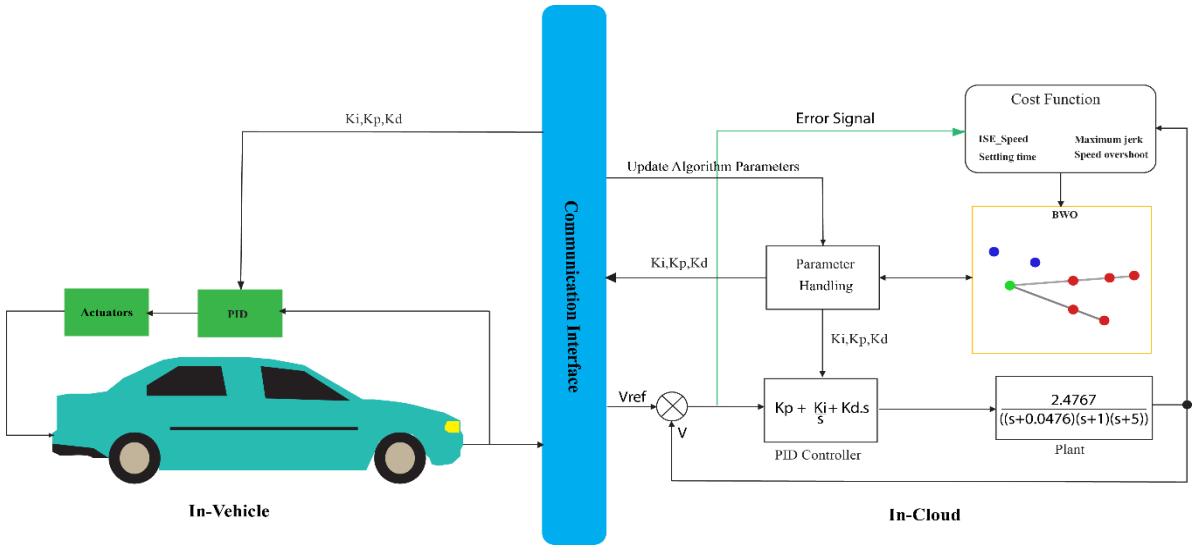


Figure 7. The block diagram of the proposed cloud-based CC system

4. RESULTS AND DISCUSSION

In this section, several simulation studies are conducted to demonstrate the advantages of the proposed optimization algorithm. Initially, we will assess the algorithm's capability to determine the global minimum of 16 standard benchmark functions commonly utilized in the literature. Following this, we simulate the proposed cloud-based CC equipped with the BWO algorithm and compare it with conventional methods. It should be noted that similar methods are not cloud-based, and aside from the functional performance results of the proposed scheme, its cloud framework distinguishes it from similar approaches.

4. 1. Benchmark Test In this section, the proposed BWO algorithm has been evaluated on 16 standard and well-known benchmark functions in the field of optimization. These classical functions, which have been widely utilized in scientific research, provide a suitable basis for comparing the performance of the proposed algorithm with other metaheuristic methods. Table 2 presents the specifications of these functions, where: Dim represents the number of dimensions of the function, Range denotes the search space boundaries and f_{min} indicates the optimal value of the function.

We utilized MATLAB version 2016b for the simulations. In this study, the proposed algorithm, the basic IWO algorithm, and three improved versions of

TABLE 2. Unimodal and multimodal benchmark functions

Functions	Dim	Range	f_{min}
$\mathbf{F_1(x)} = \sum_{i=1}^{n} x_i^2$	3,10,30	[-100,100] [-10,10]	0
$\mathbf{F_2(x)} = \sum_{i=1}^{n}\|x_i\| + \prod_{i=1}^{n}\|x_i\|$	3,10,30	[-10,10] [-1,1]	0
$\mathbf{F_3(x)} = \sum_{i=1}^{n}(\sum_{j=1}^{i} x_j)^2$	3,10,30	[-100,100] [-10,10]	0
$\mathbf{F_4(x)} = \max_i\{\|x_i\|, 1 \le i \le n\}$	3,10,30	[-100,100] [-10,10]	0
$\mathbf{F_5(x)} = \sum_{i=1}^{n-1}[100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	3,10,30	[-30,30] [-3,3]	0
$\mathbf{F_6(x)} = \sum_{i=1}^{n}(\|x_i + 0.5\|)^2$	3,10,30	[-100,100] [-10,10]	0
$\mathbf{F_7(x)} = \sum_{i=1}^{n} i x_i^4 + random[0,1)$	3,10,30	[-1.28,1.28]	0
$\mathbf{F_8(x)} = \sum_{i=1}^{n} -x_i \sin(\sqrt{\|x_i\|})$	3,10,30	[-500,500] [-50,50]	-418.9829×Dim
$\mathbf{F_9(x)} = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10]$	3,10,30	[-5.12,5.12]	0
$\mathbf{F_{10}(x)} = -20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \exp\left(\frac{1}{n}\sum_{i=1}^{n}\cos(2\pi x_i)\right) + 20 + e$	3,10,30	[-32,32]	0
$\mathbf{F_{11}(x)} = \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}}) + 1$	3,10,30	[-600,600] [-60,60]	0
$\mathbf{F_{12}(x)} = \frac{\pi}{n}\{10\sin(\pi y_1) + \sum_{i=1}^{n-1}(y_i - 1)^2[1 + 10\sin^2(\pi y_i)] + (y_n - 1)^2\} + \sum_{i=1}^{n} u(x_i, 10, 100, 4)$ $y_i = 1 + \frac{x_i+1}{4}$ $u(x_i, a, k, m) = \begin{cases} k(x_i - a)^m & x_i > a \\ 0 & -a < x_i < a \\ k(-x_i - a)^m & x_i < -a \end{cases}$ *	3,10,30	[-50,50] [-5,5]	0
$\mathbf{F_{13}(x)} = 0.1\{\sin^2(3\pi x_1) + \sum_{i=1}^{n}(x_i - 1)^2[1 + \sin^2(3\pi x_i + 1)] + (x_n - 1)^2[1 + \sin^2(2\pi x_n)]\} + \sum_{i=1}^{n} u(x_i, 5, 100, 4)$	3,10,30	[-50,50] [-5,5]	0
$\mathbf{F_{14}(x)} = -\sum_{i=1}^{n}\sin(x_i).\left(\sin\left(\frac{ix_i^2}{\pi}\right)\right)^{2m}, m = 10$	3,10,30	[0,π]	-4.687
$\mathbf{F_{15}(x)} = \sum_{i=1}^{n}\|x_i\|^{i+1}$	3,10,30	[-20,20] [-2,2]	0
$F_{16}(x) = \{[\sum_{i=1}^{n}\sin^2(x_i)] - \exp(-\sum_{i=1}^{n} x_i^2)\}.\exp[-\sum_{i=1}^{n}\sin^2(\sqrt{\|x_i\|})$	3,10,30	[-10,10] [-1,1]	-1

IWO were implemented and evaluated on 16 benchmark functions (Table 2) under three different scenarios. In the first scenario, the search space is limited, the problem has low dimensionality (3 dimensions), and the number of iterations is set to 10. In the second scenario, the search space remains limited (focused region), the problem has medium dimensionality (10 dimensions), and the number of iterations is increased to 100. In the third scenario, the search space is broad, the dimensionality is high (30 dimensions), and the number of iterations is set to 1000. Each algorithm was executed five times for every benchmark function in each scenario, and the average best cost of the five runs is reported in the tables (Tables 3 to 5). In these tables, D represents the problem dimensionality and I indicates the number of algorithm iterations.

Based on the results presented in Tables 3 to 5, the normalized error for each benchmark function is calculated, and statistical indicators including mean, standard deviation, maximum, and minimum errors are displayed in various graphs. These results effectively demonstrate the scalability of the proposed algorithm compared to other IWO-based methods. After confirming the efficiency of the proposed algorithm within the IWO category, a comparison is made with the novel GSFA algorithm to determine the position of the proposed algorithm outside this category and in comparison with newer algorithms. The simulation results confirm the high efficiency of the proposed algorithm.

In all three scenarios, the average best cost values indicate that the proposed algorithm achieves significantly superior results for most benchmark functions. Overall, the proposed algorithm demonstrates exceptional performance in benchmark functions F1, F2, F3, F4, F6, F11, F13, and F14, while yielding comparable or marginally different results in other benchmarks compared to peer algorithms in its category.

TABLE 3. Average best cost comparison over 5 runs per algorithm (Scenario 1)

Function	IWO [2006] I=10, D=3	CIWO [2012] I=10, D=3	MCIWO [2019] I=10, D=3	LF-IWO [2021] I=10, D=3	Proposed BWO I=10, D=3
F1	1.70e+00	3.34E+01	2.49E+01	4.11e+01	2.45e-03
F2	5.01e-02	6.39E-01	3.16E-01	6.68e-01	3.04e-02
F3	1.59e+00	3.76E+01	3.32E+01	2.28e+01	3.42e-03
F4	1.24e-01	3.27E+00	3.34E+00	4.24e+00	2.86e-02
F5	3.08e+01	1.29E+02	1.57E+02	1.68e+02	3.03e+01
F6	1.45e+00	3.92E+01	2.35E+01	2.53e+01	9.12e-04
F7	1.06e-02	4.89E-01	2.97E-01	9.64e-02	2.09e-02
F8	-4.16e+01	-1.58E+01	-4.46E-01	-3.95e+01	-8.64e+01
F9	3.27e+00	2.67E+01	2.12E+01	2.49e+01	1.67e+00
F10	1.30e-01	3.37E+00	4.13E+00	4.03e+00	1.07e-01
F11	2.62e-01	1.06E+00	1.02E+00	7.05e-01	5.48e-02
F12	1.98e+01	2.28E+01	2.38E+01	2.53e+01	1.98e+01
F13	6.36e-03	9.00E-01	9.64E-01	1.06e+00	1.66e-03
F14	-2.59e-10	-1.46E-10	-1.96E-10	-1.71e-10	-2.88e-10
F15	1.71e-05	6.88E-01	4.42E-01	5.50e-01	3.13e-05
F16	-9.56e-01	-3.04E-01	-5.15E-01	-4.97e-01	-9.69e-01

TABLE 4. Average best cost comparison over 5 runs per algorithm (Scenario 2)

Function	IWO [2006] I=100, D=10	CIWO [2012] I=100, D=10	MCIWO [2019] I=100, D=10	LF-IWO [2021] I=100, D=10	Proposed BWO I=100, D=10
F1	3.16e-04	2.76E+02	2.64E+02	1.30e+02	3.20e-04
F2	4.47e-02	4.60E+00	3.79E+00	2.37e-01	5.61e-02
F3	7.76e-03	3.29E+02	2.29E+02	1.35e+02	1.71e-03
F4	1.63e-02	8.54E+00	8.02E+00	6.03e+00	1.20e-02
F5	3.76e+01	9.21E+03	1.06E+04	8.86e+02	3.72e+01
F6	2.43e-04	2.61E+02	2.45E+02	1.40e+02	4.03e-04
F7	2.90e-02	2.26E+01	1.13E+01	4.17e-01	1.93e-02
F8	-2.25e+02	-3.33E+00	-1.43E+02	-9.71e+01	-2.29e+02
F9	1.38e+01	1.45E+02	1.35E+02	5.23e+01	9.20e+00
F10	2.47e-02	6.54E+00	6.80E+00	3.98e+00	2.62e-02
F11	1.39e+00	3.39E+00	3.57E+00	2.35e+00	6.09e-01
F12	2.69e+01	3.27E+01	3.39E+01	2.83e+01	2.69e+01
F13	4.50e-04	1.158E+01	9.84E+00	5.59e+00	4.52e-04
F14	-9.03e-10	-3.21E-10	-3.67E-10	-7.63e-10	-8.41e-10
F15	1.49e-06	6.24E+01	4.40E+01	8.22e-01	1.84e-06
F16	5.96e-03	8.02E-03	8.58E-03	6.16e-03	5.95e-03

Based on the results presented in Tables 3, 4, and 5, which display the mean best cost values obtained from five independent runs of each algorithm across three different scenarios, a comprehensive comparison of the algorithms can be conducted from multiple perspectives. To ensure statistical reliability and comprehensive assessment, we compute multiple performance indicators including the mean relative error, standard deviation of relative errors, and minimum and maximum relative error values for each algorithm in each scenario.

This multi-metric approach provides insights into not only the accuracy but also the stability and scalability of

TABLE 5. Average best cost comparison over 5 runs per algorithm (Scenario 3)

Function	IWO [2006] I=1000, D=30	CIWO [2012] I=1000, D=30	MCIWO [2019] I=1000, D=30	LF-IWO [2021] I=1000, D=30	Proposed BWO I=1000, D=30
F1	3.63e-05	1.12E+03	1.04E+05	5.86e+04	3.01e-05
F2	3.58e-02	5.35E+15	7.77E+16	1.05e+02	2.85e-02
F3	8.23e+01	3.54E+03	1.86E+05	1.06e+05	6.36e-02
F4	1.99e+01	9.77E+00	9.78E+01	8.07e+01	6.89e-03
F5	5.79e+01	5.89E+08	5.37E+08	4.66e+07	5.39e+02
F6	3.65e-05	1.12E+03	1.09E+05	5.77e+04	2.44e-05
F7	3.00e-02	2.67E+02	2.64E+02	5.57e-02	3.71e-02
F8	-7.08e+03	-4.77E+03	-2.16E+03	-2.55e+03	-7.20e+03
F9	6.21e+01	5.30E+02	5.30E+02	1.35e+02	9.61e+01
F10	1.53e+01	2.05E+01	2.09E+01	1.90e+01	4.35e-03
F11	2.85e+02	1.02E+03	1.00E+03	6.83e+02	1.57e-02
F12	3.32e+01	1.43E+09	1.53E+09	1.85e+08	3.50e+01
F13	2.74e-05	2.51E+09	2.65E+09	3.30e+08	3.81e-03
F14	-2.20e-09	-6.90E-10	-8.50E-10	-2.25e-09	-2.34e-09
F15	1.45e-07	3.74E+34	4.52E+33	1.57e+22	2.10e-07
F16	1.53e-17	1.28E-05	6.97E-06	7.95e-20	1.667e-17

each algorithm. To facilitate meaningful comparisons and enhance visual interpretability in graphical representations, all error values undergo min-max normalization. This process scales the absolute error values to a consistent range, enabling direct comparison across different algorithms and scenarios. Specifically, each absolute error value is normalized using the global range of variation:

$$E_{norm} = \frac{|x_i - f_i|}{|x_{worst} - x_{best}|} \quad (22)$$

where x_i represents the solution found by the algorithm for benchmark function i, f_i denotes the known global optimum of the function, and x_{max} and x_{min} represent the maximum and minimum values obtained across all algorithms and scenarios, respectively. The Mean Absolute Error in liner scale (MAE_{liner}) is then calculated by using Equation 23.

$$MAE_{Liner} = \frac{1}{N}\sum_{i=1}^{N} E_{normalized} \quad (23)$$

Where N=16 represents the total number of benchmark functions in our test suite.

Given that normalized error values may still span multiple orders of magnitude, we employ logarithmic transformation to improve interpretability and visualization clarity by Equation 24.

$$MAE_{Log} = \log\left(MLiE_{norm} + (1E-100)\right) \quad (24)$$

The addition of the 10^{-100} term ensures numerical stability when handling extremely small error values approaching zero. This transformation effectively compresses the dynamic range of error values, making subtle performance differences more discernible.

Using both linear and logarithmic error metrics, we compute comprehensive statistical measures including mean, standard deviation, minimum, and maximum values. These metrics form the foundation for our analytical visualizations and provide quantitative evidence for algorithm performance comparisons.

Figure 8 illustrates the mean relative error in linear scale across different scenarios. The proposed BWO algorithm demonstrates superior performance, maintaining the lowest error rates while exhibiting remarkable stability across diverse problem conditions. This consistent performance underscores the algorithm's robustness and adaptability.

Figure 9 presents the mean relative error in logarithmic scale, revealing nuanced performance characteristics that might be obscured in linear representations. The logarithmic transformation accentuates relative performance differences, particularly for algorithms with very small error values.

Figure 10 depicts the standard deviation of relative errors in logarithmic scale, providing insights into algorithm stability. The proposed algorithm shows remarkable consistency in standard deviation values across all scenarios, with an overall lower variability

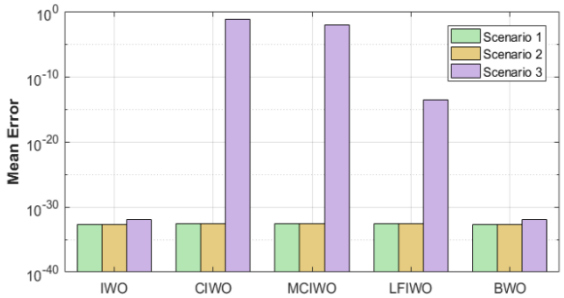


Figure 8. The mean error in linear scale across different scenarios

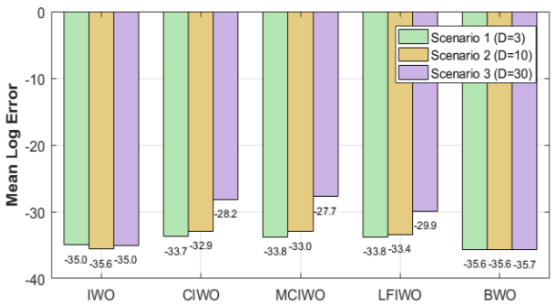


Figure 9. The mean error in logarithmic scale across different scenarios

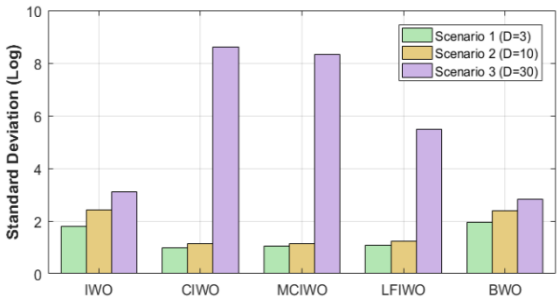


Figure 10. The standard deviation of relative errors in logarithmic scale

compared to competing algorithms. This finding indicates exceptional reliability and predictable performance. The slightly elevated standard deviation in initial scenarios actually reflects the algorithm's adaptive nature, where it achieves extraordinary performance on specific problem types while maintaining competitive performance on others.

Figure 11 examines algorithmic scalability by plotting mean logarithmic errors against increasing problem dimensions. The proposed BWO algorithm maintains a stable performance trend across all dimensionalities, demonstrating excellent scalability properties. In stark contrast, MCIWO, CIWO, and LFIWO algorithms exhibit significant performance degradation as problem dimensionality increases, revealing limitations in handling high-dimensional optimization landscapes.

Figures 12-14 provide detailed analysis of extreme performance characteristics through minimum and maximum best cost trends across the three scenarios. The proposed algorithm consistently achieves the best minimum error values in all scenarios, indicating superior exploitation capabilities. Furthermore, it maintains the most favorable maximum error values, demonstrating robust exploration and reliable worst-case performance.

This comprehensive evaluation demonstrates the proposed BWO algorithm's superior performance,

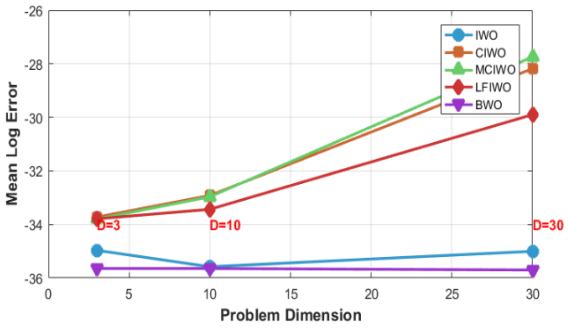


Figure 12. Performance evolution with increasing problem dimensionality

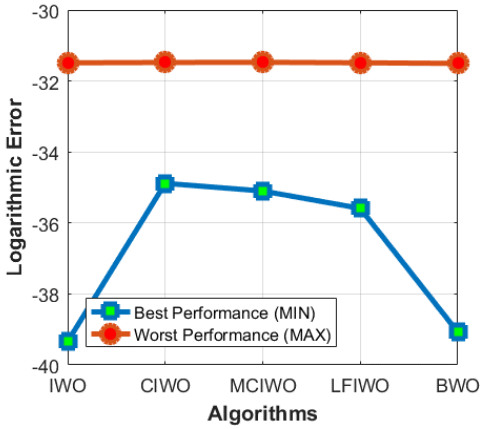


Figure 12. Error bounds in Scenario 1

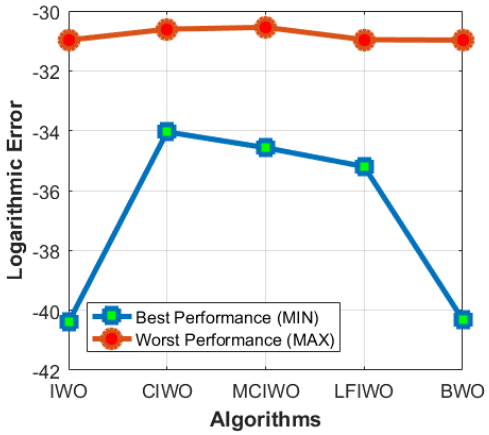


Figure 13. Error bounds in Scenario 2

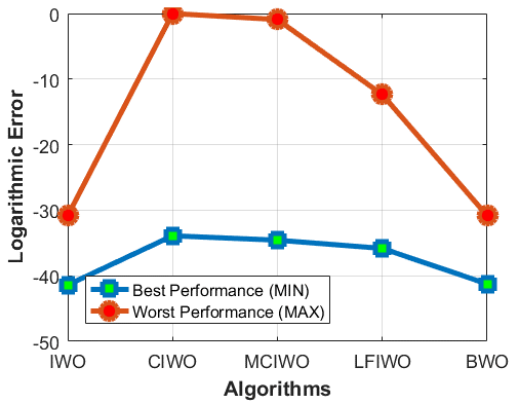


Figure 14. Error bounds in Scenario 3

robustness, and scalability compared to IWO-based algorithms. To further validate the proposed algorithm's competitiveness beyond IWO-based approaches, we performed comparative evaluation with the recently developed GSFA algorithm. The comprehensive comparison results are presented in Table 6.

The proposed BWO algorithm demonstrates superior overall performance compared to the GSFA algorithm, achieving significant improvements in 14 out of 16 benchmark functions. With an average performance enhancement of 64.43% and exceptional precision exceeding 99% on nine key functions, BWO proves to be more robust and reliable across diverse problem types. Although GSAFA excels specifically on function F8, the consistent dominance of BWO establishes it as a more versatile and effective optimization approach.

4. 2. Simulation Results of the Proposed CC To evaluate the performance of the proposed CC system after its implementation in MATLAB, several performance metrics—such as overshoot (Mp), settling time (Ts), speed tracking error (ISESpeed), maximum jerk (max|j(t)|))— were compared against three state-of-the-art methods from the literature (33-35). Notably, all these studies adopted the Integral of Time-weighted Squared Error (ITSE) as their cost function. In contrast, the present study not only introduces an enhanced version of the IWO algorithm but also proposes a new cost function tailored to improve driving comfort. To provide a fair comparison, the previous works were re-evaluated using the proposed cost function. As evidenced in Table 7, our method results in a noticeable reduction in error metrics, highlighting its superior capability in enhancing CC performance while prioritizing passenger comfort. The proposed algorithm reduces overshoot by 45.92%, settling time by 29.38%, ISEspeed by 8.92%, and maximum jerk by 20.09% compared to the RPO-Based method (the newest). Figure 15 presents the cost function convergence plot across iterations along with the corresponding step response using the optimized parameters.

It should be noted that the parameters of the proposed algorithm for the practical problem include: decision variable bounds of [-10, 10], P = 30, initial P= 10, S_{min} = 0, S_{max} = 5, b_{min} = 1, b_{max} = 10, r_{min} = 1, r_{max} = 10, $\sigma_{initial}$ = 1, σ_{final} = 1e-7, and n = 2.

TABLE 6. Average best cost and standard deviation comparison over 5 runs per algorithm

Function	GSAFA [2026] I=100, D=3		Proposed BWO I=100, D=3	
	Best Cost	Std.	Best Cost	Std.
F1	4.102E-06	3.636E-06	2.881E-08	2.869E-08
F2	4.392E-04	3.730E-04	3.767E-04	2.525E-04
F3	4.981E-05	2.751E-05	3.274E-08	2.965E-08
F4	4.865E-03	3.125E-03	1.376E-04	4.330E-05
F5	3.026E+01	4.921E-01	3.000E+01	2.326E-03
F6	1.532E-06	1.618E-06	1.012E-07	5.906E-08
F7	6.582E-03	4.895E-03	1.361E-03	5.691E-04
F8	-1.148E+03	9.579E+01	-8.133E+01	5.084E+00
F9	1.325E+00	1.060E+00	1.529E-01	3.418E-01
F10	7.851E-02	4.872E-02	5.944E-04	2.110E-04
F11	9.925E-02	3.625E-02	4.931E-03	4.613E-03
F12	1.979E+01	3.636E-03	1.9784E+01	1.055E-07
F13	1.103E-04	8.483E-05	4.964E-07	7.180E-07
F14	-3.140E-10	1.625E-11	-3.313E-10	6.961E-12
F15	-7.558E-01	4.248E-01	3.174E-10	2.605E-10
F16	-1.478E-01	2.100E-01	-9.996E-01	1.978E-04

TABLE 7. Comparison of performance parameters for different algorithms

Algorithm	Year	M_P	T_S	ISE_{Speed}	max|j(t)|
ObAOANM	2021	4.20	6.58	622.78	69.41
CS	2022	6.58	5.83	479.36	127.03
Fierfly	2022	10.34	8.04	513.44	157.63
RPO	2024	3.31	6.40	778.33	63.57
Proposed	-	1.79	4.52	708.91	50.80

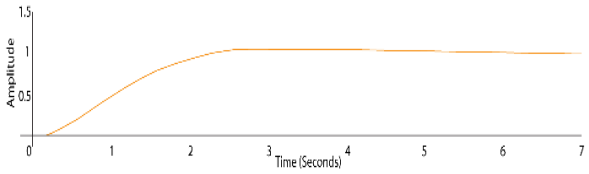


Figure 15. Optimal step response of the BWO

5. CONCLUSION AND FUTURE WORK

This study introduced the BWO algorithm—a novel, scalable metaheuristic inspired by the propagation behavior of Bermuda grass—and successfully applied it within a cloud-based CC system designed for electric and resource-constrained automotive platforms. The primary strength of the BWO algorithm lies in its scalability, which is particularly advantageous in automotive systems for two main reasons:

1. In accuracy-sensitive scenarios (e.g., traffic jams or stationary conditions), the algorithm can increase iteration count for enhanced solution quality.
2. In time-critical driving situations (e.g., highway cruising or varying road conditions), it delivers near-optimal solutions within tightly constrained iteration limits.

This flexibility allows the system to dynamically balance computational effort and real-time performance according to driving scenarios. To assess its effectiveness, BWO was rigorously evaluated using 16 well-known benchmark optimization functions. The

simulation results demonstrate that BWO outperforms the original IWO and three state-of-the-art variants in most scenarios, exhibiting superior scalability and solution quality. Additionally, BWO shows high robustness with respect to parameter sensitivity, achieving consistent performance across various problems using a single set of parameters.

The cloud-based BWO-leveraged CC system was benchmarked against four previously published approaches. Compared to the GA-based method—the strongest baseline—BWO reduced overshoot by 45.92%, settling time by 29.38%, ISEspeed by 8.92%, and maximum jerk by 20.09% compared to the RPO-Based method (the newest). Thus, the proposed CC framework offers scalability unlike conventional methods, enabling adaptive performance under diverse driving conditions. Its cloud-based architecture ensures compatibility with resource-limited and electric vehicles, while simultaneously enhancing accuracy and superior outcomes in vehicular comfort measures.

For future work, we propose an Adaptive Cruise Control (ACC) system based on Digital Twin (DT) technology, in which the DT is equipped with an optimization unit powered by the BWO algorithm. A Digital Twin is defined as a virtual replica of a physical system that synchronizes with its real-world counterpart through real-time data exchange, enabling simulation, analysis, and optimization without physical intervention.

Acknowledgements

The authors gratefully acknowledge the Cyber-Physical Systems Research Laboratory at the University of Isfahan for providing the facilities and resources essential to this research.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The implementation codes for the proposed algorithm in this study are fully available via the following GitHub repository: [https://github.com/Cyber-Physical-Laboratory-FCE-UI/BWO].

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used DeepSeek exclusively for minor language editing to improve readability. After using this tool, the authors carefully reviewed and edited the content as needed and take full responsibility for the content of the published article.

Authors Biosketch

Maytham Allahi Rudposhti received his B.Sc. and M.Sc. degrees in Computer Engineering from the University of Science and Technology of Mazandaran and Babol Noshirvani University of Technology, Iran, respectively. He is currently a Ph.D. candidate in Computer Engineering at the University of Isfahan, focusing on Autonomous Vehicles (AV) and Digital Twin (DT) technology.

Ali Bohlooli received the B.Sc. and M.Sc. degrees (Hons.) in computer engineering from the Department of Electrical and Computer Engineering, Isfahan University of Technology, Iran, in 2001 and 2003, respectively, and the Ph.D. degree in computer engineering from the University of Isfahan, Iran, in 2011. He is currently an Associate Professor with the Faculty of Computer Engineering, University of Isfahan. His research interests include cyber-physical systems, embedded artificial intelligence, and the Internet of Things (IoT).

Kamal Jamshidi received his Ph.D. in Electrical Engineering from the Indian Institute of Technology, India. He has been a faculty member of the Faculty of Computer Engineering at the University of Isfahan, Iran, since 1995. His academic background is in electrical engineering, and his teaching and research activities are carried out within the field of computer engineering.

REFERENCES

1. Mirjalili S, Lewis A. The whale optimization algorithm. Advances in Engineering Software, 2016;95:51-67. https://doi.org/10.1016/j.advengsoft.2016.01.008
2. Holland JH. Genetic algorithms. Scientific American, 1992;267(1):66-72. https://doi.org/10.1038/scientificamerican0792-66
3. Kahrizi MR, Kabudian SJ. Projectiles optimization: A novel metaheuristic algorithm for global optimization. International Journal of Engineering Transactions A: Basics, 2020;33(10):1924-38. https://doi.org/10.5829/ije.2020.33.10a.11
4. El-Shorbagy MA, Bouaouda A, Nabwey HA, Abualigah L, Hashim FA. Advances in Henry gas solubility optimization: A physics-inspired metaheuristic algorithm with its variants and

applications. IEEE Access, 2024;12:26062-95. https://doi.org/10.1109/ACCESS.2024.3365700

5. Sowmya R, Premkumar M, Jangir P. Newton–Raphson-based optimizer: A new population-based metaheuristic algorithm for continuous optimization problems. Engineering Applications of Artificial Intelligence, 2024;128:107532. https://doi.org/10.1016/j.engappai.2023.107532
6. Mehrabian AR, Lucas C. A novel numerical optimization algorithm inspired from weed colonization. Ecological Informatics, 2006;1(4):355-66. https://doi.org/10.1016/j.ecoinf.2006.07.003
7. Mirjalili S, Mirjalili SM, Lewis A. Grey wolf optimizer. Advances in Engineering Software, 2014;69:46-61. https://doi.org/10.1016/j.advengsoft.2013.12.007
8. Saremi S, Mirjalili S, Lewis A. Grasshopper optimisation algorithm: Theory and application. Advances in Engineering Software, 2017;105:30-47. https://doi.org/10.1016/j.advengsoft.2017.01.004
9. Talatahari S, Azizi M, Tolouei M, Talatahari B, Sareh P. Crystal structure algorithm (CryStAl): A metaheuristic optimization method. IEEE Access, 2021;9:71244-61. https://doi.org/10.1109/ACCESS.2021.3079161
10. Zhao S, Zhang T, Ma S, Chen M. Dandelion optimizer: A nature-inspired metaheuristic algorithm for engineering applications. Engineering Applications of Artificial Intelligence, 2022;114:105075. https://doi.org/10.1016/j.engappai.2022.105075
11. Han M, Du Z, Yuen KF, Zhu H, Li Y, Yuan Q. Walrus optimizer: A novel nature-inspired metaheuristic algorithm. Expert Systems with Applications, 2024;239:122413. https://doi.org/10.1016/j.eswa.2023.122413
12. Zhong C, Li G, Meng Z, Li H, Yildiz AR, Mirjalili S. Starfish optimization algorithm (SFOA): A bio-inspired metaheuristic algorithm for global optimization compared with 100 optimizers. Neural Computing and Applications, 2025;37(5):3641-83. https://doi.org/10.1007/s00521-024-10694-1
13. Amani B, Nouri M, Mousavi Ghasemi SA. Gray squirrel foraging algorithm for function optimization. International Journal of Engineering, 2026;39(7):1644-56. https://doi.org/10.5829/ije.2026.39.07a.09
14. Ehsaeyan E. Rock-climbing group: An innovative meta-heuristic approach for efficiently tackling optimization problems. International Journal of Engineering Transactions B: Applications, 2025;38(11):2796-818. https://doi.org/10.5829/ije.2025.38.11b.24
15. Ehsaeyan E. Gold seekers algorithm: An innovative metaheuristic approach for global optimization and its application in image segmentation. International Journal of Engineering Transactions C: Aspects, 2025;38(9):2114-29. https://doi.org/10.5829/ije.2025.38.09c.09
16. Perez C, Climent L, Nicolo G, Arbelaez A, Salido MA. A hybrid metaheuristic with learning for a real supply chain scheduling problem. Engineering Applications of Artificial Intelligence, 2023;126:107188. https://doi.org/10.1016/j.engappai.2023.107188
17. Mallahzadeh A, Eshaghi S, Hassani H. Compact U-array MIMO antenna designs using IWO algorithm. International Journal of RF and Microwave Computer-Aided Engineering, 2009;19(5):568-76. https://doi.org/10.1002/mmce.20379
18. Pahlavani P, Delavar MR, Frank AU. Using a modified invasive weed optimization algorithm for a personalized urban multi-criteria path optimization problem. International Journal of Applied Earth Observation and Geoinformation, 2012;18:313-28. https://doi.org/10.1016/j.jag.2012.03.004
19. Ahmadi M, Mojallali H. Chaotic invasive weed optimization algorithm with application to parameter estimation of chaotic systems. Chaos, Solitons & Fractals, 2012;45(9):1108-20. https://doi.org/10.1016/j.chaos.2012.05.010
20. Nikoofard AH, Hajimirsadeghi H, Rahimi-Kian A, Lucas C. Multiobjective invasive weed optimization: Application to analysis of Pareto improvement models in electricity markets. Applied Soft Computing, 2012;12(1):100-12. https://doi.org/10.1016/j.asoc.2011.09.005
21. Ghasemi M, Ghavidel S, Akbari E, Vahed AA. Solving non-linear, non-smooth and non-convex optimal power flow problems using chaotic invasive weed optimization algorithms based on chaos. Energy, 2014;73:340-53. https://doi.org/10.1016/j.energy.2014.06.026
22. Mishra SK, Bose PSC, Rao CSP. An invasive weed optimization approach for job shop scheduling problems. International Journal of Advanced Manufacturing Technology, 2017;91(9):4233-41. https://doi.org/10.1007/s00170-017-0091-x
23. Mandava RK, Vundavilli PR. Implementation of modified chaotic invasive weed optimization algorithm for optimizing the PID controller of the biped robot. Sadhana, 2018;43(5):66. https://doi.org/10.1007/s12046-018-0851-9
24. Abdelkader EM, Moselhi O, Marzouk M, Zayed T. A multi-objective invasive weed optimization method for segmentation of distress images. Intelligent Automation & Soft Computing, 2020;26(4):643-61. https://doi.org/10.32604/iasc.2020.010100
25. Kashyap AK, Parhi D, Pandey A. Improved modified chaotic invasive weed optimization approach to solve multi-target assignment for humanoid robot. Journal of Robotics and Control, 2021;2(3):194-9. https://doi.org/10.18196/jrc.2377
26. Ibrahim A, Anayi F, Packianather M, Alomari OA. New hybrid invasive weed optimization and machine learning approach for fault detection. Energies, 2022;15(4):1488. https://doi.org/10.3390/en15041488
27. Beskirli M. A novel invasive weed optimization with levy flight for optimization problems: The case of forecasting energy demand. Energy Reports, 2022;8:1102-11. https://doi.org/10.1016/j.egyr.2021.11.108
28. Kalhori M, Ashofteh PS, Moghadam SH. Development of the multi-objective invasive weed optimization algorithm in the integrated water resources allocation problem. Water Resources Management, 2023;37(11):4433-58. https://doi.org/10.1007/s11269-023-03564-3
29. Chauhan U, Chhabra H, Jain P, Dev A, Chauhan N, Kumar B. Chaos inspired invasive weed optimization algorithm for parameter estimation of solar PV models. IFAC Journal of Systems and Control, 2024;27:100239. https://doi.org/10.1016/j.ifacsc.2023.100239
30. Taliaferro CM, Rouquette FM, Mislevy P. Bermudagrass and stargrass. Agronomy Monographs, 2004;45:417-75. https://doi.org/10.2134/agronmonogr45.c12
31. Heybetli F, Danayiyen Y, Tasdemir AB, Senyigit S. Comparative analysis of metaheuristic algorithms in PID-based vehicle cruise control systems. Verus Journal, 2025;25(1):1-16. https://doi.org/10.5152/electrica.2025.25051
32. Pradhan R, Majhi SK, Pradhan JK, Pati BB. Antlion optimizer tuned PID controller based on bode ideal transfer function for automobile cruise control system. Journal of Industrial Information Integration, 2018;9:45-52. https://doi.org/10.1016/j.jii.2018.01.002
33. Izci D, Ekinci S, Kayri M, Eker E. A novel enhanced metaheuristic algorithm for automobile cruise control system. Electrica, 2021;21(3):283-97. https://doi.org/10.5152/electrica.2021.21016

34. Prasanna V, Nelson A, Hnaumanthakari S, Kumar VK, Kirubakaran S, Kumar MJ. Metaheuristic algorithm for automatic cruise control system. Proceedings of the 3rd International Conference on Smart Electronics and Communication, 2022:206-11. https://doi.org/10.1109/ICOSEC54921.2022.9952117

35. Saravanan G, Pazhanimuthu C, Sathish Kumar D, Lalitha B, Senthilkumar M, Kannan E. Red panda optimization algorithm-based PID controller design for automobile cruise control system. Proceedings of the International Conference on Smart Systems for Electrical, Electronics, Communication and Computer Engineering, 2024:33-37. https://doi.org/10.1109/ICSSEECC61126.2024.10649422

36. Afshar M, Hadji Molana SY, Rahmani Parchicolaie B. A multi-objective optimization model for multi-commodity closed-loop supply chain network considering disruption risk. International Journal of Engineering Transactions A: Basics, 2024;37(4):646-61. https://doi.org/10.5829/ije.2024.37.04a.07

37. Ebrahimi Gouraji R, Soleimani H, Afshar Najafi B. Optimization of sustainable vehicle routing problem taking into account social utility and employing a strategy with multiple objectives. International Journal of Engineering Transactions A: Basics, 2025;38(7):1631-58. https://doi.org/10.5829/ije.2025.38.07a.15

Persian Abstract

چکیده

این پژوهش الگوریتم بهینه‌سازی علف برمودا (BWO) را معرفی می‌کند؛ یک الگوریتم فراابتکاری جدید و مقیاس‌پذیر که از رشد تهاجمی علف برمودا الهام گرفته شده است. این الگوریتم به عنوان نسخه بهبودیافته‌ای از الگوریتم بهینه‌سازی علف هرز مهاجم (IWO) توسعه یافته و با تقلید از راهبردهای قوی تکثیر گیاه، تعادل بهتری میان کاوش جهانی و بهره‌برداری محلی برقرار می‌کند. عملکرد الگوریتم با ارزیابی دقیق در برابر چهار نسخه قبلی مبتنی بر IWO سنجیده شد و مقیاس‌پذیری آن از طریق میانگین کمترین نرخ خطا و عملکرد پایدار در سناریوهای متنوع اثبات گردید. علاوه بر این، BWO با الگوریتم جدید جستجوی سنجاب خاکستری (GSFA)—که خارج از دسته IWO قرار دارد—مقایسه شد تا عملکرد آن در برابر روشی جدید و خارج از خانواده IWO نیز ارزیابی شود؛ این مقایسه برتری رقابتی BWO را نشان داد و به طور متوسط ٦٤/٤٣ درصد بهبود در نتایج بهترین هزینه به همراه داشت. همگرایی قوی و مقیاس‌پذیری BWO، آن را برای کاربردهای بلادرنگ—به ویژه در سیستم‌های خودرویی—بسیار مناسب می‌سازد. در یک پیاده‌سازی عملی با استفاده از چارچوب کنترل کروز مبتنی بر ابر، BWO به طور قابل توجهی از روش مبتنی بر بهینه‌سازی پاندای قرمز (RPO) پیشی گرفت و فراجهش را ٤٥/٩٢ درصد، زمان نشست را ٢٩/٣٨درصد، میانگین مربعات خطای سرعت (ISE) را ٨/٩٢ درصد و حداکثر جرک ٢٠/٠٩ درصد کاهش داد. با دستیابی به همگرایی نزدیک به بهینه و بهره‌گیری از استقرار ابری با قابلیت مقیاس‌پذیری بالا، BWO می‌تواند به طور مؤثر با الزامات متنوع سیستم‌های خودرویی سازگار شود و در حالت‌های کاری متعدد، به کارایی بالایی دست یابد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2546-2558

International Journal of Engineering

Journal Homepage: www.ije.ir

Modeling of Thermodynamic Processes for Assessing Hydrocarbon Systems at Depths Exceeding 5 km

O. M. Prischepa, D. S. Lutsky, S. B. Kireev, Y. V. Nefedov*, V. S. Stepanova, N. V. Sinitsa, N. N. Vostrikov

Department of Oil and Gas Geology, Saint-Petersburg Mining University of Empress Ekaterina II, St. Petersburg, Russia

PAPER INFO

Paper history:
Received 23 October 2025
Received in revised form 15 December 2025
Accepted 03 January 2026

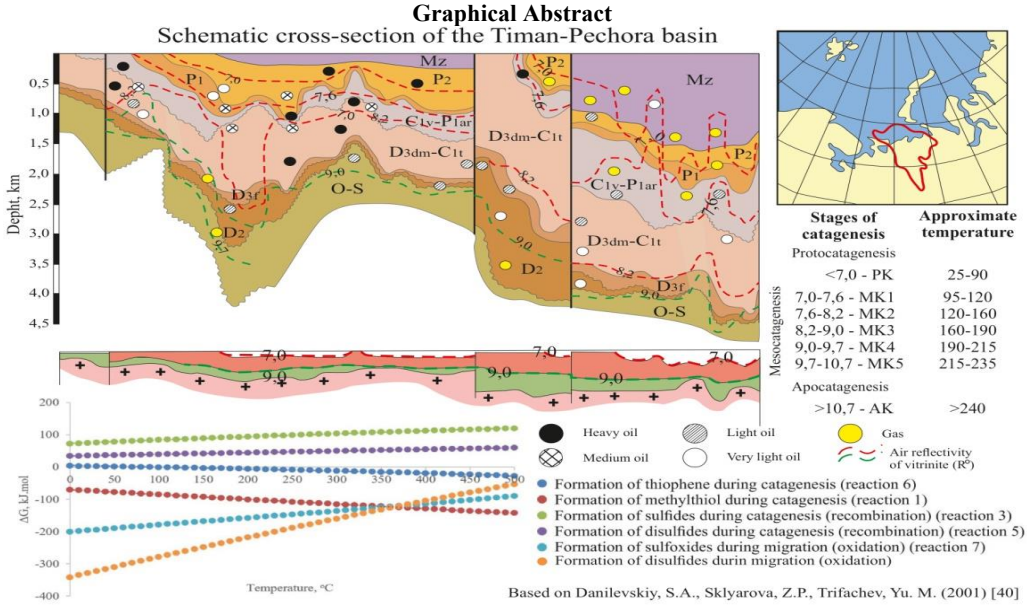
Keywords:
Deep Oil
Thermodynamic Modeling
Deep Horizons
Hydrocarbon Systems
High-pressure Reservoirs
Ultra-deep Petroleum Systems

ABSTRACT

The discovery of oil reservoirs at great and ultra-great depths has necessitated a revision of existing concepts about the possibility of hydrocarbon generation and preservation under extremely high pressure–temperature conditions. This work analyzes geochemical approaches and basin modeling methods used for predicting hydrocarbon generation at great depths. It examines the influence of high temperatures and pressures, burial rate, phase transitions, and other factors on the formation and destruction of oil and gas. An analysis of deep hydrocarbon systems is performed, and the potential of thermodynamic modeling of fluid equilibrium compositions and the results of experiments on oil transformation under high P–T conditions are considered. A modern approach to assessing a high-organic-carbon (Domanik) formation is presented, which allows evaluation of hydrocarbon generation potential. The reasons why existing geochemical methods and basin modeling can yield inconclusive results for similar processes at great depths under extreme pressure–temperature conditions are identified. In particular, sources of uncertainty in hydrocarbon phase transitions are discussed, associated with a lack of experimental data for model calibration and the limitations of thermodynamic models beyond their validated ranges, among other factors.

doi: *10.5829/ije.2026.39.10a.16*

Graphical Abstract



*Corresponding Author Email: yurijnefedov@yandex.ru (Y. V. Nefedov)

Please cite this article as: Prischepa OM, Lutsky DS, Kireev SB, Nefedov YV, Stepanova VS, Sinitsa NV, Vostrikov NN. Modeling of Thermodynamic Processes for Assessing Hydrocarbon Systems at Depths Exceeding 5 km. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2546-58.

NOMENCLATURE

HC	hydrocarbons	R_2NH	Amines
R-SH	Thiols	R_3N	
R-S-R'	Sulfides	R-COOH	Carboxylic acids
R-S-S-R'	Disulfides	Ar-OH	Phenols
R-SO-R'	Sulfoxides and sulfones	R-OH	Alcohols
$R-SO_2-R'$		R-COO-R'	Esters
$R-NH_2$	Amines	R-CO-R'	Ketones

1. INTRODUCTION

Recently discovered oil and gas reservoirs at great and ultra-great depths (1), as well as indirect evidence of reservoirs existing in supercritical fluid conditions, force a reconsideration of theoretical ideas about the conditions for the formation and preservation of hydrocarbons (HC) in zones of high pressure and temperature (2-4). The most discussed issues in this context include: the influence of thermobaric (pressure–temperature) conditions on hydrocarbon formation and destruction; the influence of the rate and magnitude of rock subsidence and uplift on HC generation and redistribution; determining criteria and threshold values for phase transitions in hydrocarbon systems; the influence of the initial organic matter composition and maturity (kerogen type) on the phase composition of generated hydrocarbons; the influence of kinetic parameters of bitumoids on phase ratios and phase transition conditions; the influence of the lithological composition and depositional environment of kerogen-bearing strata on their hydrocarbon potential, and on migration and accumulation conditions; the influence of overlying salt-bearing strata and deep conductive faults (which significantly alter the distribution of geothermal and hydrostatic gradients) on the character and maximum depth of oil and gas asccumulations (5-7); the contribution of "deep" methane to the formation of HC accumulations, including high-molecular-weight compounds (8-10). Only a comprehensive examination of these issues will yield new practical results and technologies for discovering hydrocarbon deposits at great and ultra-great depths.

However, the practical applicability of theoretical models at such conditions remains unproven. The question arises whether it is possible to create a technology aimed at exploring oil and gas reservoirs at ultra-great depths and how effective it could be in the foreseeable future. Other possible approaches include experimental testing of various models of HC formation and accumulation under extreme conditions (with validation against core and fluid data), as well as the search for geochemical halos of dispersion from deep HC accumulations (11, 12).

Debates about the origin and migration of oil have catalyzed extensive experimental research, new ideas and knowledge, identification of factors affecting oil and gas quality, and the development of new investigative methods that were later implemented in oil and gas exploration technologies. The foundation of these technologies was new theoretically grounded material balance equations that do not contradict the laws of physics and chemistry, enabling calculation of volumetric ratios between initial organic matter and generated hydrocarbons (13-15). Subsequently, numerical modeling tools for sedimentary basin evolution were created on this basis, which are effectively used to predict and assess potential oil and gas presence.

The depletion of large conventional oil and gas traps in recent decades has led to a decline in exploration efficiency, which in turn spurred a comprehensive analysis of all processes of oil and gas formation. Such analysis includes determining the conditions for organic matter accumulation in sedimentary basins, the influence of temperature and pressure during burial of organic-rich deposits, the pathways of hydrocarbon fluid migration, and the delineation of zones of HC accumulation vs. dissipation. The development and implementation of basin modeling technologies in exploration practice have significantly improved the success rate of assessing petroleum potential (16-18).

The emergence of a new resource base in the form of low-permeability shale and carbonate reservoirs in the early 21st century forced a revision of ideas about migration as a stage of petroleum formation. At the same time, interest increased in technologies for enhancing rock porosity and permeability (hydraulic fracturing). Meanwhile, the theory of oil formation continued to evolve thanks to detailed studies of oil-bearing strata containing mobile but not yet expelled hydrocarbons (19-21).

Analysis of the influence of the above factors on the evaluation of petroleum potential in deep-seated complexes, taking into account new research and achievements (primarily in Chinese sedimentary basins such as the Tarim, Junggar, Ordos, Sichuan basins; as well as the Gulf of Mexico and Brazil's Santos Basin), has made it possible to identify potential areas for studying hydrocarbons at great and ultra-great depths in sedimentary basins and large depressions of Russia (22-24). Thermobaric and historical-geological criteria for the prospects of deep complexes were formulated, and a methodology for their detailed study was proposed (2). This methodology includes geochemical analyses of

hydrocarbons and their sources (pyrolysis, chromatography), lithological and petrophysical core studies – all aimed at assessing the probability of oil and gas presence in deep horizons and subsequently pinpointing the most promising drilling targets.

The problem of the genesis of hydrocarbons and the conditions for formation and preservation of deposits has not only theoretical but also important practical significance. Proposed approaches to this problem often reflect researchers' personal views on the origin of oil and natural gas, based on two fundamentally different hypotheses: the sedimentary-migrational hypothesis versus the deep mineral-mantle genesis hypothesis.

At present, among modern petroleum generation hypotheses developed within the biogenic-sedimentary paradigm, opinions are increasingly emerging that emphasize the significant role of energy and material from the Earth's deep zones. Examples include the geofluidodynamic concept (25), the geosynergetic concept of natural hydrocarbon-generating systems (26), the sedimentary-inorganic hypothesis (27), and others. According to these works, fluxes of deep fluids stimulate the process of petroleum generation, the expulsion of hydrocarbons from the organic matter of sedimentary strata, and promote their accumulation.

Thus, in order to model the conditions of HC accumulation at great depths, one must take into account hydrocarbon genesis and consider reservoir formation from the following perspectives. First, there is the viewpoint that some indeterminate portion of hydrocarbons could have formed in a high-temperature (>200–250 °C) zone of organic matter transformation due to the combination of high temperature–pressure and a specific organic matter composition. Second, some contribution is associated with inorganic synthesis of hydrocarbons in zones where hydrogen and carbon are transported from the deep interior under certain thermobaric conditions (perhaps it is more accurate here to speak of an energetic contribution to process intensity).

To date, the problem of hydrocarbon compound formation via a combination of organic and inorganic processes remains unresolved. One cannot dismiss the thesis that the indicators of oil "biogenicity" may be explained by the fact that during migration from deep horizons, primary hydrocarbons "wash" the organic matter of sedimentary rocks, acquiring some properties and components of both the sedimentary strata and the organic matter.

In most petroleum basins, the shallow (upper) complexes are generally already explored, and searching for oil/gas in the deeper parts of the sedimentary cover has become a new, independent direction of study (28, 29). It is hard to agree with the opinion that every structure containing oil and gas in the shallow sedimentary sequence is also suitable for finding oil and gas in the deep horizons, including in the crystalline basement.

In recent years, with advances in computing, the capabilities of basin modeling have expanded, improving prediction accuracy and reducing uncertainty in exploration. Modern software suites can account for more and more factors and processes, approaching a realistic picture of the formation of deep petroleum systems.

It must be emphasized, however, that even the most advanced models today encounter difficulties when applied to the extreme conditions of ultra-deep petroleum generation zones.

1. 1. Limitations of Methods under Extreme Pressure–temperature Conditions Despite significant progress in understanding petroleum generation, current geochemical methods and basin modeling have limited applicability at great depths (>5–6 km) under extremely high temperatures and pressures. Below are the main problems that arise when attempting to predict the state of hydrocarbon systems under such conditions:

• Uncertainty of hydrocarbon phase transitions. Under standard sedimentary basin conditions, the range of HC phase states is relatively well studied: it is known at what temperatures and pressures oil transitions to condensate and gas, and when solid residues (pyrobitumens) form. However, at great depths (temperature on the order of 250–300 °C and pressure of hundreds of MPa), phase equilibria shift substantially. In this realm, the existence of supercritical fluids is possible, where the distinction between liquid and gas virtually disappears. The absence of clear phase boundaries complicates prediction: for example, it is not obvious whether oil remains liquid or fully transitions into a gaseous supercritical phase under such parameters. The critical parameters of hydrocarbons can change greatly with increasing molecular weight and presence of impurities. Consequently, standard oil and gas phase diagrams extrapolated to ultra-high pressures carry large uncertainty. In practice, this means we cannot say with confidence whether at a depth of 7–8 km liquid oil will be present, or if the hydrocarbon mixture exists as a single gas-like (supercritical) phase, or if it separates into gas and a solid carbon residue. Such phase transitions remain largely hypothetical due to a lack of empirical data.

• Lack of experimental data. Experimentally reproducing conditions of ~200–300 °C and >100 MPa in the laboratory is very difficult, especially to maintain for extended periods even remotely comparable to geologic time. The few experiments conducted with oil and organic matter at high P–T have yielded valuable information, but these data are clearly insufficient to fully calibrate models. Many key parameters – e.g. kerogen decomposition rates at 250–300 °C, hydrocarbon

solubility in water at 100–200 MPa, diffusion properties of supercritical fluids, etc. – remain poorly studied. The absence of an experimental basis forces models to extrapolate patterns established at lower temperatures (say, 100–150 °C) to the high-temperature realm, which may be invalid. As a result, predictions may differ depending on the chosen kinetic scheme or thermodynamic model, and confidence in them is low. We must recognize that our understanding of real petroleum system behavior at ultra-high pressures is still largely theoretical rather than based on direct experiment.

• Mismatch of thermodynamic models with real systems. Equilibrium composition models rely on a simplified representation of oil, gas, and kerogen as a few quasi-pure substances with given properties. This simplifies calculations, but real oil consists of hundreds of components, and under extreme conditions new reactions may occur that the model doesn't include. For example, further compaction of hydrocarbons into solid carbon phases (graphite formation) or interactions with the mineral matrix (e.g. hydrogen being sequestered into minerals or fluids) may occur. Existing equations of state for petroleum fluids, calibrated up to ~150 °C and a few tens of MPa, can produce large errors at 300 °C and 100+ MPa. Thermodynamic parameters of many heavy components at such conditions are simply unknown. Moreover, in the supercritical region the very notions of "oil" and "gas" lose clarity – instead of a two-phase system, there may be a single supercritical phase, and standard two-phase models cease to work. Thus, thermodynamic models require serious expansion and verification for application to ultra-deep levels.

• Limitations of kinetic models in basin modeling. Kinetic models of hydrocarbon generation (for example, the model by Pepper and Corvi (30)) are based on experimental kerogen pyrolysis at temperatures up to ~500 °C, from which decomposition parameters are extrapolated to geological timescales. These models work well up to the end of the "oil window" (~150–180 °C, catagenesis up to stage MK5). But at significantly higher temperatures (200–250 °C and above) the reliability of such models decreases. It is possible that at 250–300 °C, with prolonged heating, alternative pathways of organic decomposition occur, yielding products (e.g. methane and carbon) in proportions different from those predicted by standard kinetics. Moreover, kinetic parameters obtained from short-term heating of samples may be inadequate for describing reactions under slow heating over millions of years. Consequently, basin models may under- or over-estimate gas generation at great depths. For example, "secondary gas generation" from already formed oil during prolonged heating may occur, which is not always accounted for in modeling software.

• Other factors and uncertainties. It should be noted that at great depths abnormally high pore pressures are often present, affecting fluid migration and trap preservation – models need to incorporate mechanical processes and rock instability. The role of natural catalysts (e.g. minerals whose surfaces can facilitate HC decomposition reactions) at high temperatures is not fully clear – some minerals might accelerate or retard oil cracking, shifting stability limits. Furthermore, under extreme conditions the role of hydrogen and hydrogen-bearing fluids emanating from the deep mantle may increase. An influx of hydrogen could lead to hydrogenation of carbon residues, theoretically generating heavy hydrocarbons even at high temperatures – though this is a hypothetical mechanism requiring verification. Finally, the pore structure at depth is greatly altered: pores and fractures are partly closed due to high pressure and diagenesis, affecting hydrocarbon mobility and potentially hindering migration or, conversely, creating conditions for local preservation in isolated traps.

Thus, existing methods of forecasting petroleum systems must be used with great caution when extending to depths beyond 5–6 km. Further development of theoretical models is necessary, along with the accumulation of experimental data under conditions close to real high P–T, and possibly a revision of some fundamental assumptions about the limiting conditions of oil formation. Only a combination of different approaches – thermodynamic, kinetic, and geophysical – and their cross-calibration will gradually reduce uncertainty and improve the reliability of predictions of oil and gas presence at the greatest depths.

Generalizing the experience of determining the conditions of HC accumulations at great depths, together with modeling the possibility of HC formation and preservation under ultra-harsh thermobaric conditions at great and ultra-great depths, will lead to the creation of a technology no less effective than basin modeling, which enabled the discovery and assessment of enormous oil and gas reserves at shallow and moderate depths.

It should be noted that the modern basin modeling approach to petroleum generation is developing in parallel with an equilibrium approach based on kinetic models of kerogen decomposition, which allows one to account for the non-equilibrium nature of real geological processes.

In this paper, the results of applying geochemical methods to the problem of predicting petroleum systems at great depths are considered. The fundamentals of thermodynamic (equilibrium) modeling of the composition of hydrocarbon fluids and the results of a geochemical study of a specific region are presented. Current approaches to basin modeling of petroleum systems are outlined. The final section of the paper is devoted to a discussion of the use of existing methods under the extremely high temperatures and pressures characteristic of ultra-deep zones.

2. MATERIALS AND METHODS

2. 1. Thermodynamic Modeling as the Basis for Predicting Ultra-deep Hydrocarbon Deposits

Numerous studies of formation fluids show that up to temperatures of about 320 °C, the composition and aggregate state of oil remain relatively stable. However, with further temperature increase – in the range of 350–380 °C – especially at pressures close to the saturated vapor pressure of the oil's components (~150 MPa), active thermolysis begins with the formation of methane, propane, butane, and carbon dioxide. At the same time, the proportion of light fractions increases, transitioning into the gas phase as condensate, known as "invisible oil." Under these conditions, solid bitumen also forms. By 550–600 °C, virtually all organic matter decomposes to methane and a solid carbon residue — anthraxolite. These observations highlight the significant influence of the combined effect of temperature and pressure on the onset of thermal cracking: under abnormally high reservoir pressure, this process shifts to higher temperatures, allowing oil to remain liquid even under conditions traditionally considered unfavorable for its existence.

Thermodynamic modeling based on empirical data from core analyses is widely used to analyze phase states and determine the pressure–temperature conditions of formation and evolution of hydrocarbon accumulations at significant and ultra-deep horizons. The primary assumption of this approach is the possibility of local equilibrium in geochemical systems, which justifies the use of equilibrium thermodynamics to calculate the composition of natural fluids. As numerous studies show, this assumption is valid at very high temperatures (above 150–200 °C) or over long process durations. Key assumptions used in kinetic modeling are given in Table 1.

The theoretical foundation of the method was laid by Gibbs (31). Modern approaches to calculating the equilibrium composition of real systems are divided into direct and combined methods. Direct methods focus on determining equilibrium parameters directly, whereas combined methods are used to model equilibrium compositions in non-ideal systems. Since in geological settings pressure and temperature often remain constant, isobaric-isothermal conditions are applicable in most phase behavior models. The modeling utilized an in-house code based on Gibbs energy minimization under specified P–T conditions. Thermodynamic constants were sourced from the SUPCRT92 database (32) and updated parameters from Helgeson et al. (14).

One of the main tools for quantitative analysis of such processes has been the method of thermodynamic potentials (30, 33, 34), which was actively used in earlier geochemical works (31). This method made it possible to delineate stability regions of various products of organic matter transformation — the so-called geochemical facies. The approach is based on the concept from Helgeson et al. (14) that the transformation of biomass occurs sequentially: first into bitumen, then kerogen, and thereafter into liquid hydrocarbons and gas. Thus, oil and gas formation can occur at higher pressures,

TABLE 1. Key assumptions used in kinetic modeling

Assumption	Limitations	Implications for Modeling	Potential Mitigation Approaches
Uncertainty of phase transitions	Lack of clear boundaries between liquid, gas, and supercritical phase at T > 250 °C, P > 100 MPa	Inability to unambiguously determine the phase state of a hydrocarbon mixture (oil/gas/supercritical fluid)	Use of experimental PVT data at high P–T conditions; application of equations of state suitable for supercritical conditions
Lack of experimental data	Limited number of laboratory studies under prolonged exposure to conditions T >250 °C and P >100 MPa	Extrapolation of kinetic and thermodynamic parameters beyond validated ranges	Conducting new experiments on kerogen and oil thermolysis in autoclaves with long exposure times
Limitations of thermodynamic models	Use of a simplified hydrocarbon phase composition instead of real multicomponent mixtures	Underestimation or overestimation of the stability of heavy fractions, heteroatomic compounds, and solid residue	Integration of more detailed compositional models
Challenges in kinetic modeling	Kinetic parameters derived from short-term heating experiments, not accounting for geological time scales	Inaccurate prediction of gas generation rates and oil cracking at great depths	Development of kinetic schemes based on geologically relevant experiments
Influence of mineral matrix and catalysts	Failure to account for the catalytic action of clay minerals or oxides at high T–P	Incorrect assessment of thermolysis pathways and product composition	Inclusion of mineral-mediated reactions in the model; laboratory tests for catalytic activity at high P–T
Role of deep-seated fluids (H_2, CO_2, etc.)	Lack of accounting for the influx of hydrogen or carbon from mantle or metamorphic sources	Omission of possible oil stabilization mechanisms or hydrocarbon synthesis in deep zones	Integration of geodynamic models accounting for tectonic faults and vertical fluid flows

temperatures, and depths exceeding the classical limits of petroleum generation. The model relies on fundamental laws of thermochemistry — Hess's law, Kirchhoff's law, and the second law of thermodynamics. Solving the system of chemical equilibrium equations at various P–T parameters allows one to construct depth profiles of component distribution and predict the stability zones of oil and gas accumulations in sedimentary basin cross-sections, including in deeply buried levels.

The results obtained are based on data from the Domanik deposits of the eastern Siberian Platform and the Tarim and Sichuan basins, these data include Rock-Eval pyrolysis, chromatography, fluid composition, and petrophysical properties (2, 3, 22).

3. THERMODYNAMIC MODELING FOR PREDICTING HYDROCARBON SYSTEMS AT GREAT DEPTHS

3. 1. Mechanisms and Thermodynamic Feasibility of Generation of Hydrocarbons and Heteroatomic Components from Kerogen The chemical composition of a formation fluid generated during catagenesis at various depths, and the degree of transformation of its components, will depend on the type of initial kerogen, the pressure–temperature conditions of its breakdown, the rate at which the kerogen is buried, the introduction into the fluid of various substances formed during deep geological processes (for example, hydrogen), the conditions of migration of matter to the surface, and the reservoir conditions (Figure 1).

During catagenesis, kerogen components form the composition of the future crude oil, including hydrocarbons: alkanes, naphthenes, and arenes. These compounds arise from the thermal cleavage of complex organic kerogen molecules such as lipids, proteins, or carbohydrates. Chemically, the catagenesis process involves radical cleavage reactions, recombination, cyclization, and aromatization.

Nitrogen-containing oil components are generated by the breakdown of nitrogenous functional groups present in the kerogen structure (36). Nitrogen in kerogen exists in the form of amines, amides, pyrroles, pyridines, porphyrins, and other heterocyclic structures, which upon thermolysis convert into various nitrogen-containing oil compounds. The main classes of such compounds include amines, pyridines, pyrroles, quinolines, carbazoles, porphyrins, and their derivatives, as confirmed by high-resolution mass spectrometry data (36).

Sulfur-containing oil components are produced by the breakdown of sulfur-containing functional groups. Sulfur in kerogen is present in functional groups that, upon thermolysis and subsequent reactions, yield sulfur-containing oil compounds. These include thiols (mercaptans), sulfides, disulfides, thiophenes, and their derivatives.

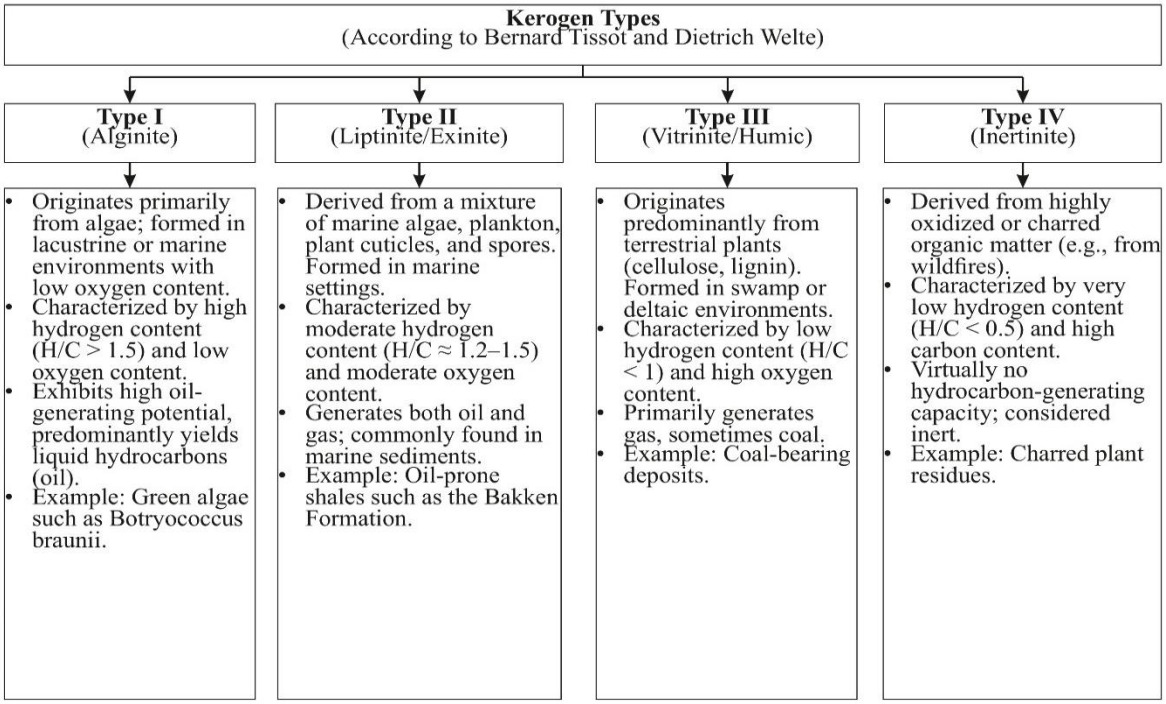


Figure 1. Classification of kerogen types according to Bernard Tissot and Dietrich Welte (35)

Oxygen-containing oil components also form during catagenesis through the breakdown of oxygen-containing functional groups in the kerogen structure. Kerogen contains oxygen in the form of carboxyl, hydroxyl, ether, carbonyl, and other groups, which upon thermolysis transform into various oxygen-containing oil compounds. The main classes of such compounds include carboxylic acids, phenols, alcohols, ethers, ketones, aldehydes, and furans.

Typical classes of heteroatom compounds are shown in Figure 2.

During the formation of liquid hydrocarbons from the parent kerogen, the following key reactions can be identified. Thermolysis of kerogen is accompanied by the cleavage of C–C, C–O, C–N, and C–S bonds, resulting in the formation of free radicals or unsaturated molecules.

The presence of sulfur-containing components in crude oil can serve as an indicator of the degree of kerogen transformation. As hydrocarbons are progressively formed, sulfur compounds derived from oil-prone kerogen can be detected, and their formation can be described by the following principal reactions:

Thermolysis of kerogen, involving the cleavage of C–S, S–S, and S–H bonds and the generation of reactive radicals or molecules containing sulfur atoms.

Radical reactions, including recombination of the formed radicals and their interaction with hydrocarbons or hydrogen, leading to the formation of new sulfur-containing compounds.

High-temperature cyclization and aromatization of sulfur-bearing molecules, resulting in the formation of heterocyclic compounds such as thiophenes.

Secondary reactions of recombination, oxidation, or reduction, leading to the generation of more complex sulfur-containing structures.

Thiols are formed via cleavage of weak S–S or C–S bonds in kerogen followed by hydrogen addition (Reaction 1):

$$\begin{aligned} &\text{R-S-S-R'} + H_2 \rightarrow \text{R-S-H} + \text{H-S-R'} \\ &C_2H_6S_2 + H_2 \rightarrow 2CH_4S \end{aligned} \quad (1)$$

Sulfides are formed by oxidation of thiols and interaction with hydrocarbons (Reaction 2) or by radical recombination (Reaction 3):

$$\begin{aligned} &\text{R}\bullet + \bullet\text{S-R'} \xrightarrow{[O]} \text{R-S-R'} + H_2O \\ &CH_4 + CH_3\text{-HS} + O_2 \xrightarrow{[O]} CH_3\text{-S-}CH_3 + H_2O \end{aligned} \quad (2)$$

$$\begin{aligned} &\text{R}\bullet + \bullet\text{S-R'} \rightarrow \text{R-S-R'} \\ &C_2H_6 + CH_3HS \rightarrow C_2H_5\text{-S-}CH_3 + H_2 \end{aligned} \quad (3)$$

Disulfides form via oxidation of thiols (Reaction 4) or recombination of thiyl radicals (Reaction 5):

$$\begin{aligned} &\text{R-S}\bullet + \bullet\text{S-R'} \xrightarrow{[O]} \text{R-S-S-R'} + H_2O \\ &CH_3\text{-HS} + CH_3\text{-HS} \xrightarrow{[O]} CH_3\text{-S-S-}CH_3 + H_2O \end{aligned} \quad (4)$$

$$\begin{aligned} &\text{R-S}\bullet + \bullet\text{S-R'} \rightarrow \text{R-S-S-R'} \\ &C_2H_5\text{-HS} + CH_3\text{-HS} \rightarrow C_2H_5\text{-S-S-}CH_3 + H_2 \end{aligned} \quad (5)$$

Thiophenes and their derivatives are generated through cyclization of sulfur-containing molecules followed by aromatization (37). The formation of thiophene (Reaction 6) is characteristic of more advanced stages of catagenesis.

$$CH_2{=}CH\text{-}CH{=}CH_2 + H_2S \rightarrow C_4H_4S + 2H_2 \quad (6)$$

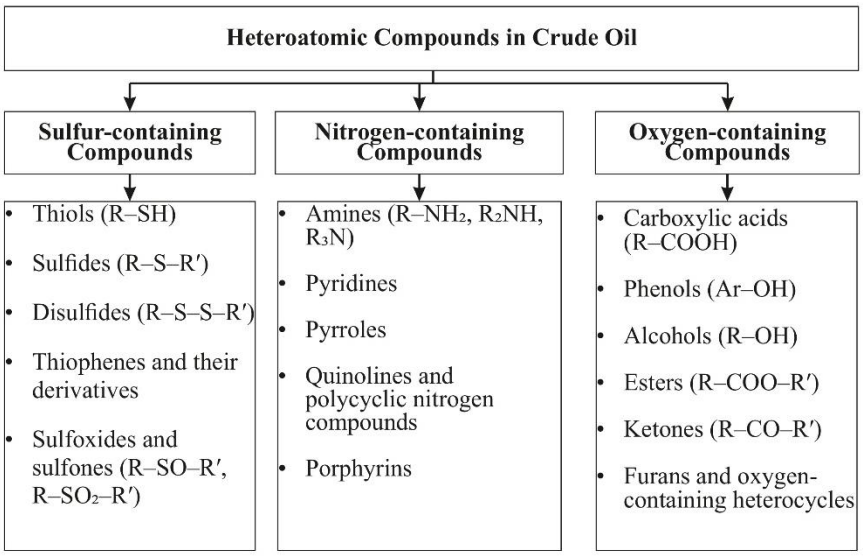


Figure 2. Classification of typical heteroatomic compounds included in formation fluids

More complex thiophenes, such as benzothiophene, arise from the condensation of aromatic hydrocarbons with sulfur-containing fragments.

Sulfoxides (Reaction 7) and sulfones (Reaction 8) form predominantly during the migration of the reservoir fluid to the surface via oxidation of sulfides in the presence of oxygen or other oxidants in pore waters.

$$\text{R-S-R'} \xrightarrow{[O]} \text{R-SO-R'}$$
$$\text{C}_2\text{H}_6\text{S} \xrightarrow{[O]} \text{C}_2\text{H}_6\text{OS} \quad (7)$$

$$\text{R-SO-R'} \xrightarrow{[O]} \text{R-SO}_2\text{-R'}$$
$$\text{C}_2\text{H}_6\text{OS} \xrightarrow{[O]} \text{C}_2\text{H}_6\text{O}_2\text{S} \quad (8)$$

For the above reactions, the Gibbs free energy changes as a function of temperature were calculated and are presented in Figure 3. The curves were calculated using standard thermodynamic data (ΔH°f, S°, Cp) from reference sources (15, 32) with extrapolation based on temperature-dependent heat capacity equations. The pressure was assumed to be 100–150 MPa, typical for depths of 5-7 km. The methodology used for these calculations is described by Cheremisina et al. (38).

From the obtained dependencies, it can be concluded that thiol formation is spontaneous throughout all stages of catagenesis. The formation of thiophene and its derivatives becomes feasible in high-temperature zones provided sufficient hydrogen sulfide is present. The formation of sulfides, disulfides, sulfoxides, and sulfones is only possible in the presence of oxidants and is most likely to occur during the migration of reservoir fluids toward the surface and their localization within reservoir traps.

Thus, the key factors governing the compositional profile of the reservoir fluid include the type of kerogen undergoing catagenesis and the pressure–temperature conditions under which catagenesis occurs. The effects of these parameters are summarized in Tables 2 and 3 (39).

3. 2. Interaction of Oil Components with Hydrogen Introduced into the Reservoir by Geological Processes

Earlier work, Prishchepa et al. (2) examined the mutual transformations of hydrocarbons in the presence of hydrogen under various reservoir conditions. In the present work, emphasis is placed on investigating similar transformations for a number of sulfur-containing compounds. Reactions of hydrodesulfurization of sulfur-containing compounds by hydrogen entering the reservoir have been observed in natural systems and confirmed by modeling (40, 41).

In these processes, thiols react with hydrogen to form the corresponding hydrocarbon and hydrogen sulfide. This reaction proceeds relatively easily due to the weak C–S bond (Reaction 9):

$$\text{R-SH} + \text{H}_2 \rightarrow \text{R-H} + \text{H}_2\text{S}$$
$$\text{CH}_3\text{-SH} + \text{H}_2 \rightarrow \text{CH}_3\text{-H} + \text{H}_2\text{S} \quad (9)$$

Sulfides are cleaved by hydrogen to form two hydrocarbon molecules and hydrogen sulfide. This reaction requires more severe conditions than for thiols (Reaction 10):

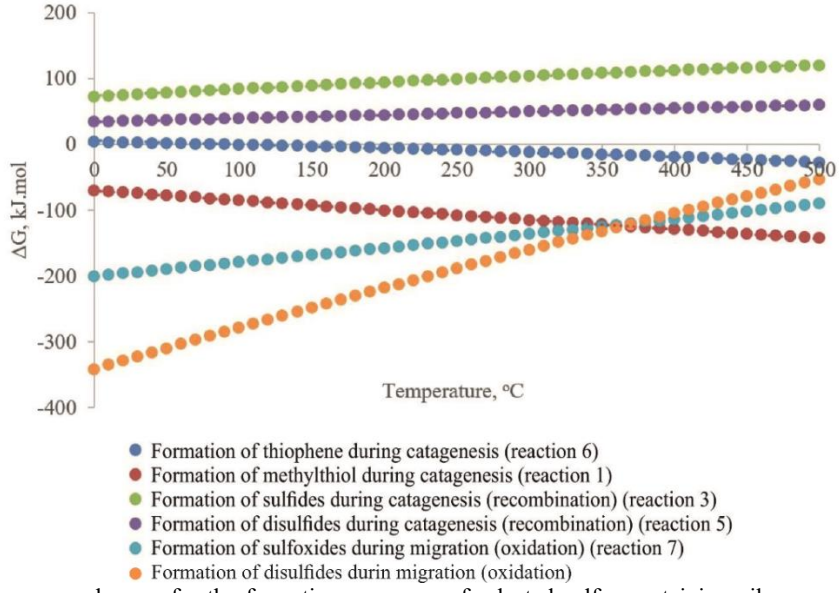


Figure 3. Gibbs free energy changes for the formation processes of selected sulfur-containing oil components as a function of temperature

TABLE 2. Effect of kerogen type on the formation of hydrocarbons and heteroatomic components of oil

Kerogen Type	Hydrocarbons	Sulfur-containing components	Oxygen-containing components	Nitrogen-containing components
Type I (algal)	Rich in lipids, generates a large amount of alkanes and naphthenes and a smaller quantity of arenes (aromatic hydrocarbons).	Low sulfur content; hydrocarbons predominate.	Low oxygen content; hydrocarbons predominate, but alcohols and carboxylic acids can form from lipids.	High nitrogen content (up to 2–3%) due to high protein and chlorophyll content, leading to the formation of porphyrins and amines.
Type II	Contains lipids and cyclic structures, forming a mixture of alkanes, naphthenes, and a moderate amount of arenes.	High sulfur content (up to 10–15%); generates a large amount of thiols, sulfides, and thiophenes.	Moderate oxygen content; carboxylic acids, esters, and furans are formed.	Moderate nitrogen content; pyrroles, pyridines, and quinolines predominate.
Type III	Rich in lignin, which promotes formation of arenes and yields fewer alkanes and naphthenes.	Low sulfur content and few sulfur-containing compounds.	High oxygen content (up to 20–30%) due to lignin and cellulose; phenols and furans predominate.	Low nitrogen content; nitrogen-containing compounds are minimal, mostly amines and simple heterocycles.

TABLE 3. Effect of temperature on the formation of hydrocarbons and heteroatomic components of oil

Kerogen transformation temperature (°C)	Hydrocarbons	Sulfur-containing components	Oxygen-containing components	Nitrogen-containing components
100-150	Predominant formation of alkanes and naphthenes via cracking and cyclization	Mainly thiols and sulfides form.	Primarily carboxylic acids, alcohols, and esters form.	Amines and simple heterocycles (pyrroles, pyridines) form
150-300	Increased proportion of arenes (aromatics) due to dehydrogenation and aromatization.	Growing proportion of thiophenes and derivatives due to aromatization.	Decarboxylation and dehydration reduce the content of oxygen-containing compounds, increasing the share of furans and phenols.	Formation of more complex structures (quinolines, carbazoles) through aromatization.

$$\begin{aligned} &R\text{-}S\text{-}R + H_2 \rightarrow 2R\text{-}H + H_2S \\ &CH_3\text{-}S\text{-}CH_3 + H_2 \rightarrow 2CH_4 + H_2S \end{aligned} \quad (10)$$

Disulfides, under hydrodesulfurization, split into two hydrocarbons and two molecules of hydrogen sulfide. The S–S bond is broken, and each sulfur is removed as H_2S (Reaction 11):

$$\begin{aligned} &R\text{-}S\text{-}S\text{-}R + 3H_2 \rightarrow 2R\text{-}H + 2H_2S \\ &CH_3\text{-}S\text{-}S\text{-}CH_3 + 3H_2 \rightarrow 2CH_4 + 2H_2S \end{aligned} \quad (11)$$

Thiophenes are compounds resistant to hydrodesulfurization because of the aromatic nature of the ring. The reaction requires higher temperatures and active catalysts. The thiophene ring is first hydrogenated, then cleaved, forming hydrocarbons and H_2S (Reaction 12):

$$CH_3\text{-}S\text{-}S\text{-}CH_3 + 3H_2 \rightarrow 2CH_4 + 2H_2 \quad (12)$$

Complex polycyclic sulfur-containing compounds (e.g., dibenzothiophene derivatives) are resistant to hydrodesulfurization due to their complex structure. These reactions require even more severe conditions, and often partial hydrogenation of rings occurs before sulfur removal. Products depend on the structure of the starting compound, but they are usually polyaromatic hydrocarbons and H_2S. For example, hydrogenation of 4-methyldibenzothiophene yields methylbiphenyl and hydrogen sulfide (Reaction 13):

$$C_{13}H_{10}S + 3H_2 \rightarrow C_{13}H_{12} + H_2S \quad (13)$$

It is shown that a formation fluid saturated with sulfur-containing compounds, when reacting with hydrogen generated by deep geochemical processes, reduces the concentration of thiols, sulfides, disulfides, thiophenes, sulfoxides, and sulfones, while the concentration of hydrocarbons increases and hydrogen sulfide (H_2S) is produced.

4. DISCUSSION

It has been demonstrated that kerogen type is a determining factor in the component composition of formation fluids. Algal-type kerogen predominantly generates liquid hydrocarbons with low sulfur content but high nitrogen content. Marine kerogen is characterized by high sulfur content, which contributes to the formation of significant amounts of thiols, sulfides, and thiophenes. Humic (terrestrial) kerogen, on the contrary, generates mainly gas and aromatic hydrocarbons, with a minimal amount of sulfur-containing compounds but a high content of oxygen-containing components.

Pressure–temperature (thermobaric) conditions during catagenesis have a decisive influence on the

chemical transformations of organic matter. At 100–150 °C, the formation of alkanes, naphthenes, thiols, and sulfides dominates, as well as linear oxygen- and nitrogen-containing compounds. As temperature rises to 150–300 °C, aromatization and dehydrogenation occur, leading to an increased fraction of arenes, thiophenes, and complex nitrogenous compounds, while the content of oxygen-containing components decreases due to decarboxylation.

The mechanisms of formation of sulfur-containing compounds include thermolysis, radical recombination, cyclization, and aromatization. Thiols and sulfides form in the early stages of catagenesis by breaking weak C–S and S–S bonds. Thiophenes and their derivatives form at higher temperatures via cyclization and aromatization of sulfur-bearing fragments, which requires the presence of H_2S. Sulfoxides and sulfones typically form not in the catagenesis zone but during fluid migration toward the surface by oxidation of sulfides in the presence of oxygen or other oxidants.

Thermodynamic analysis based on Gibbs free energy indicates that the formation of thiols is a spontaneous process at all stages of catagenesis. The formation of thiophenes is thermodynamically possible at high temperatures and sufficient H_2S concentrations. The formation of sulfides, disulfides, sulfoxides, and sulfones is possible only in the presence of oxidizers, which suggests their generation predominantly during migration and accumulation in traps.

Hydrodesulfurization of sulfur-containing compounds through interaction with hydrogen introduced into the reservoir by deep geochemical processes leads to a reduction in the concentration of thiols, sulfides, disulfides, and thiophenes, accompanied by an increase in hydrocarbon content and the formation of hydrogen sulfide (H_2S). These reactions proceed with varying degrees: thiols are easily desulfurized, whereas thiophenes and polycyclic sulfur compounds are resistant and require severe conditions.

The complex interplay of geochemical factors—kerogen type, temperature, pressure, availability of hydrogen and oxidants—determines the final composition of oil. This allows the use of data on heteroatomic compounds (S, N, O) as biomarkers to reconstruct the conditions of oil formation and migration, as well as to assess the degree of catagenetic maturity of organic matter.

Discoveries of oil reservoirs in zones of extreme temperatures and pressures indicate the need to revise long-established views on the formation of accumulations both in high-organic (domanik) strata and on their role in the petroleum potential of sedimentary basins. It is evident that to reconstruct the conditions of hydrocarbon formation in specific clayey, siliceous-clayey high-organic strata, one must consider not only the burial and heating history, but also the influence of pressure, tectono-geodynamic processes, and the interplay of organic transformation processes with deep fluid mass transfer and the energy input of deep gases (hydrogen in particular).

The research also findings highlight the need to reconsider the deep «oil window» limit: at high pressures, oil may persist at depths of up to 7–8 km, especially in the presence of hydrogen. For instance, the formation of thiophenes becomes thermodynamically favorable at $T > 220$ °C and $P > 120$ MPa, corresponding to a depth of ~6.5 km with an average geothermal gradient of 35 °C/km.

Another critical factor for hydrocarbon accumulation at great and ultra-great depths is the preservation of reservoirs and the reservoir properties of rocks, which requires additional scientific research (42-44).

There has also been a growing application of machine learning and hybrid models to optimize thermodynamic calculations in recent studies (45).

5. CONCLUSION

Despite significant progress in understanding the processes of petroleum generation, existing geochemical methods and basin modeling have limited applicability at great depths (>5–6 km) under extremely high temperatures and pressures.

The obtained results show that the formation of thiols is thermodynamically favorable throughout all stages of catagenesis, whereas thiophenes and more complex sulfur-bearing structures form only in high-temperature zones with sufficient H_2S. Natural hydrodesulfurization reactions, confirmed by modeling, indicate that interaction with hydrogen from deep geological processes reduces the concentration of sulfur-containing components and increases the proportion of hydrocarbons in deep reservoir fluids.

Comparison of kerogen types and their temperature-dependent transformation pathways also shows that the distribution of heteroatomic (S, N, O) compounds can serve as an indicator of catagenetic maturity and deep reservoir conditions.

The novelty of the study lies in integrating thermodynamic equilibrium calculations with geochemical reaction mechanisms of heteroatomic compounds under ultra-high P–T conditions, which allows reconsideration of hydrocarbon stability limits at depths beyond 5–6 km.

Current methods for predicting petroleum systems must be used very cautiously when moving into depth ranges beyond 5–6 km. Only a combination of different approaches – thermodynamic, kinetic, and geophysical – and their mutual calibration will gradually reduce uncertainty and increase the reliability of predictions of the presence of oil and gas at the greatest depths.

Further research directions include:

- Development of experimental setups to simulate conditions of 300 °C / 150 MPa.
- Deeper integration of the influence from deep-seated fluid flows (H_2, CH_4) into thermodynamic models.
- Combination of AI and physics-based models for extrapolating hydrocarbon system properties.
- Expansion of databases on heavy fractions and heteroatomic compounds under extreme P–T conditions.

Acknowledgements

The author appreciates Scientific Centre «Deep Metageology», Saint-Petersburg Mining University of Empress Ekaterina II for providing necessary facilities to conduct present research.

Funding

This article was prepared as part of State Assignment FSRW-2024-0008 “Investigation of the Earth’s thermodynamic processes from the perspective of hydrocarbon genesis at great depths”.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.
The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Authors Biosketches

O.M. Prischepa – is Professor and Head of the Department of Oil and Gas Geology at SPMU. He received his doctorate on "Strategy for Balanced Reproduction of Oil and Gas Reserves and Production in Northwest Russia" in the All-Russian Petroleum Research Geological Prospecting Institute (VNIGRI). His research interests include low-porosity shale reservoirs and the resource base of the Russian Federation and the world.
D. S. Lutskiy is an Assistant Professor in the Department of Physical chemistry at Saint Petersburg Mining University. He received his Ph.D. in "Extraction and separation of lanthanides by hydrometallurgical methods in the complex processing of low-concentration raw materials". His research interests include thermodynamic modeling of natural and man-made processes.
S. B. Kireev – Head of the Laboratory of Modern Technologies for Prospecting and Exploration of Mineral Deposits.
Y. V. Nefedov is an Assistant Professor in the Department of Oil and Gas Geology at SPMU. He received his Ph.D. in "patterns of nitrogen defect manifestation in Ural-type diamond crystals." His research interests include oil and gas generation potential, geochemistry, and low-porosity shale reservoirs.
V. S. Nikiforova is an Associate Professor in the Department of Oil and Gas Geology at SPMU.
N.V. Sinitsa – Assistant of the Department of Oil and Gas Geology at SPMU.
N. N. Vostrikov – PhD candidate of the Department of Oil and Gas Geology and engineer at the «Deep Metageology» problem-solving laboratory at SPMU.

REFERENCES

1. Dvoynikov MV, Sidorkin DI, Yurtaev SL, Grokhotov EI, Ulyanov DS. Drilling of deep and ultra-deep wells for the purpose of searching and exploration of new mineral deposits. Journal of Mining Institute, 2022; 258: 945-55. https://doi.org/10.31897/PMI.2022.55
2. Prishchepa OM, Lutsky DS, Kireev SB, Sinitsa NV. Thermodynamic modeling as the basis for predicting phase states of hydrocarbon fluids at great and ultra-great depths. Journal of Mining Institute, 2024; 269: 815-32.
3. Kontorovich AE, Burshtein LM, Gubin IA, Parfenova TM, Safronov PI. Deeply buried oil and gas systems of the Lower Paleozoic in the east of the Siberian Platform: Geological-geophysical characteristics and resource assessment of hydrocarbons. Journal of Mining Institute, 2024; 269: 721-37.
4. Kosenkova NN, Syngaevsky PE, Khafizov SF. Review of the modern ideas about the hydrocarbon accumulation formation processes at great depth. Oil Industry Journal, 2022; (5): 6-12. https://doi.org/10.24887/0028-2448-2022-5-6-12
5. Trushko VL, Rozanov AO, Saitgaleev MM, Petrov DN, Ilinov MD, Karmansky DA, Selikhov AA. Acoustic emission criteria for

analyzing rock destruction and evaluating fractured reservoir formation at great depths. Journal of Mining Institute, 2024; 269: 848-58.

6. Glazunov VV, Saitgaleev MM, Sekerina DD. Features of fracture formation in rock samples under comprehensive compression using acoustic emission. International Journal of Engineering Transactions C: Aspects, 2026; 39(6): 1382-89. https://doi.org/10.5829/ije.2026.39.06c.0
7. Sekerina DD, Saitgaleev MM, Senchina NP, Glazunov VV, Kalinin DF, Kozlov MP, Ismagilova EI. The role of strike-slip faults and graben-rifts in controlling the oil and gas potential of deep horizons of the Russko-Chaselsky Ridge (West Siberian Province). Mining Science and Technology, 2025; 10(2): 109-17. https://doi.org/10.17073/2500-0632-2025-2-399
8. Zou C, Du J, Xu C. Formation, distribution, resource potential, and discovery of Sinian–Cambrian giant gas field, Sichuan Basin, SW China. Petroleum Exploration and Development, 2014; 41(5): 543-54. https://doi.org/10.1016/S1876-3804(14)60063-X
9. Wang Z, Wang T, Wen L, Hua J, Zhang B. Basic geological characteristics and accumulation conditions of Anyue giant gas field, Sichuan Basin. China Offshore Oil and Gas, 2014; 28(2): 45-52.
10. Ma A. Kinetics of oil-cracking for different types of marine oils from Tahe Oilfield, Tarim Basin, NW China. Journal of Natural Gas Geoscience, 2016; 1(1): 35-43. https://doi.org/10.1016/j.jnggs.2016.03.001
11. Li J, Zhang Z, Zhu G, Zhao K, Chi L, Wang P, Chen Y. Geochemical characteristics and the origin of superdeep condensates in Tarim Basin, China. ACS Omega, 2021; 6(11): 7275-85. https://doi.org/10.1021/acsomega.0c04932
12. Cherdantsev GA, Zharkov AM. Prospects of oil and gas content of Upper Permian deposits in the southwestern Vilyuysk syneclise based on sedimentation environments and geochemical conditions. Journal of Mining Institute, 2021; 251: 698-711. https://doi.org/10.31897/PMI.2021.5.9
13. Fialkovsky IS, Litvinova TE, Lutsky DS, Alexeev AA. Determination of parameters of thermodynamic stability constants of bromide complexes of rare earth metals. Arab Journal of Basic and Applied Sciences, 2022; 29(1): 1-9. https://doi.org/10.1080/25765299.2021.2015897
14. Helgeson HC, Richard L, McKenzie WF, Norton DL, Schmitt A. A chemical and thermodynamic model of oil generation in hydrocarbon source rocks. Geochimica et Cosmochimica Acta, 2009; 73(3): 594-695. https://doi.org/10.1016/j.gca.2008.03.004
15. Holland TJB, Powell R. An internally consistent thermodynamic data set for phases of petrological interest. Journal of Metamorphic Geology, 1998; 16(3): 309-44. https://doi.org/10.1111/j.1525-1314.1998.00140.x
16. Nefedov Y, Gribanov D, Gasimov E, Peskov D, Han G, Vostrikov N, Pashayeva S. Development of Achimov deposits sedimentation model of a West Siberian field. Reliability: Theory & Applications, 2023; 18(S5): 441-48. https://doi.org/10.24412/1932-2321-2023-575-441-448
17. Zharkov AM, Peskov DV, Martynov AV. Identification of hydrocarbon accumulation zones on ancient platforms in the northern Eurasian lithospheric plate. International Journal of Engineering Transactions B: Applications, 2025; 38(11): 2713-27. https://doi.org/10.5829/ije.2025.38.11b.19
18. Peskov DV, Zharkov AM, Kalinin DF. Criteria for predicting the hydrocarbon potential of Riphean–Vendian deposits in the Mezen Syneclise. Russian Journal of Earth Sciences, 2025; 25(5): 1-14. https://doi.org/10.2205/2025ES001033
19. Mozhegova SV, Gerasimov RS, Paizanskaya IL, Alferova AA, Kravchenko EM. Kinetic features of thermal transformation of organic matter in Bazhenov and Domanik formations. Journal of Mining Institute, 2024; 269: 765-76.
20. Plotnikova IN, Ostroukhov SB, Pronin NV. Influence of oceanic anoxia on the conditions of Domanik deposit formation. Journal of Mining Institute, 2024; 269: 803-14.
21. Bolshakova MA, Sitar KA, Kozhanov DD. Features of composition and properties of ancient oil-and-gas source deposits. Journal of Mining Institute, 2024; 269: 700-07.
22. Prishchepa OM, Kireev SB, Nefedov YV, Martynov AV, Lutsky DS, Krykova TN, Sinitsa N, Xu R. Approaches to identifying deep accumulations of oil and gas in Russian basins. Frontiers in Earth Science, 2023; 11: 1-22. https://doi.org/10.3389/feart.2023.1192051
23. Liu J, Jin Z, Ma A. Hydrocarbon phase in the deep Cambrian of the Tarim Basin. Petroleum Geology & Experiment, 2015; 37(6): 681-88. https://doi.org/10.11781/sysydz201506681
24. Huang J, Deliao Y, Yu H. Petroleum geology features and accumulation controls for ultra-deep oil and gas reservoirs. Petroleum Geology & Experiment, 2016; 38(5): 635-40. https://doi.org/10.11781/sysydz201605635
25. Sokolov BA, Khain VE. Geofluidodynamic model of petroleum generation in sedimentary basins. Moscow: Nauka; 1997: 5-9.
26. Lukin AE. On the origin of oil and gas (geosynergetic concept). Geologicheskiy Zhurnal, 1999; 1: 30-42.
27. Chebanenko II, Klochko VP, Tokovenko VS, Evdoshuk NI. Sedimentary-inorganic theory of oil and gas field formation. Geology of Oil and Gas, 2000; 5: 50-52.
28. Iskaziev KO, Syngaevsky PE, Khafizov SF. Oil at great depths: Deposits of offshore fields in the Gulf of Mexico. Bulletin of the Oil and Gas Industry of Kazakhstan, 2017; 3(1): 3-26.
29. Guliyev IS, Kerimov VYu, Osipov AV, Mustaev RN. Generation and accumulation of hydrocarbons at great depths of the Earth's crust. Scientific Works of NIPI Neftegaz GNKAR, 2017; 1: 4-16.
30. Pepper AS, Corvi PJ. Simple kinetic models of petroleum formation: Part I. Marine and Petroleum Geology, 1995; 12(3): 291-319. https://doi.org/10.1016/0264-8172(95)98381-E
31. Gibbs JW. Thermodynamics. Statistical mechanics. Moscow: Nauka; 1982. 584 p.
32. Johnson JW, Oelkers EH, Helgeson HC. SUPCRT92: Software for calculating thermodynamic properties from 1 to 5000 bar and 0 to 1000°C. Computers & Geosciences, 1992; 18(7): 899-947. https://doi.org/10.1016/0098-3004(92)90029-Q
33. Gibbs JW. Thermodynamic works. Moscow: State Publishing House of Technical-Theoretical Literature; 1950. 492 p.
34. Marakushev SA, Belonogova OV. Thermodynamic factors of natural selection in autocatalytic chemical systems. 2012; 444(1): 92-97.
35. Tissot BP, Welte DH. Formation and distribution of oil. Moscow: Mir; 1981. 500 p.
36. Mitra-Kirtley S, Mullins O, Pomerantz A. Sulfur and nitrogen chemical speciation in crude oils. Applying Nanotechnology to the Desulfurization Process in Petroleum Engineering; 2016. 31 p. https://doi.org/10.4018/978-1-4666-9545-0.ch002
37. Wu J, Ma C, Zhang W, Chang W, Zhang Y, Shi Q. Molecular characterization of non-polar sulfur compounds in crude oil fractions. Fuel, 2023; 338: 127323. https://doi.org/10.1016/j.fuel.2022.127323
38. Cheremisina OV, Ponomareva MA, Mashukova YA, Nasonova NA, Burtseva MD. Thermodynamic characteristics of the ion-exchange process involving light rare earth metals. Separations, 2025; 12(7): 177. https://doi.org/10.3390/separations1207177
39. Craddock PR, Haecker A, Bake KD, Pomerantz AE. Universal curves describing chemical and physical evolution of Type II kerogen. Energy & Fuels, 2020; 34(12): 15217-33. https://doi.org/10.1021/acs.energyfuels.0c02376

40. Danilevskiy SA, Sklyarova ZP, Trifachev YM. Geofluidic systems of the Timan-Pechora region. Ukhta: Timan-Pechora Research Center; 2003. 298 p.
41. Huang H, Zhang S, Yang F, Zhao J, Zhang M, Li S. Natural hydrodesulfurization in deep petroleum systems. Geochimica et Cosmochimica Acta, 2023; 345: 125-42. https://doi.org/10.1016/j.gca.2023.01.015
42. Almenov T, Zhanakova R, Sarybayev M, Shabaz DM. Selecting rational supports for underground mining workings. Civil Engineering Journal, 2025; 11(3). https://doi.org/10.28991/CEJ-2025-011-03-022
43. Sloniec J, Kulisz M, Malecka-Dobrogowska M, Konurbayeva Z, Sobaszek L. Awareness of IT/AI impact on energy consumption in enterprises. Energies, 2025; 18(21): 5573.
44. Demin V, Kalinin A, Tomilova N, Tomilov A, Akpanbayeva A, Shokarev D, Popov A. Advanced digital modeling of stress-strain behavior in rock masses. Civil Engineering Journal, 2025; 11(3): 1072-87. https://doi.org/10.28991/cej-2025-011-03-014
45. Ma A. Kinetics of oil-cracking for different types of marine oils from Tahe Oilfield, Tarim Basin, NW China. Journal of Natural Gas Geoscience, 2016; 1(1): 35-43. https://doi.org/10.1016/j.jnggs.2016.03.001

Persian Abstract

چکیده

کشف مخازن نفتی در اعماق زیاد و بسیار زیاد، بازنگری در مفاهیم موجود در مورد امکان تولید و حفظ هیدروکربن در شرایط فشار–دمای بسیار بالا را ضروری ساخته است. این کار، رویکردهای ژئوشیمیایی و روش‌های مدل‌سازی حوضه مورد استفاده برای پیش‌بینی تولید هیدروکربن در اعماق زیاد را تجزیه و تحلیل می‌کند. این کار تأثیر دما و فشارهای بالا، سرعت دفن، انتقال فاز و سایر عوامل را بر تشکیل و تخریب نفت و گاز بررسی می‌کند. تجزیه و تحلیل سیستم‌های هیدروکربنی عمیق انجام می‌شود و پتانسیل مدل‌سازی ترمودینامیکی ترکیبات تعادل سیال و نتایج آزمایش‌های مربوط به تبدیل نفت در شرایط فشار–دمای بالا در نظر گرفته می‌شود. یک رویکرد مدرن برای ارزیابی یک سازند با کربن آلی بالا (Domanik) ارائه می‌شود که امکان ارزیابی پتانسیل تولید هیدروکربن را فراهم می‌کند. دلایلی که روش‌های ژئوشیمیایی موجود و مدل‌سازی حوضه می‌توانند نتایج بی‌نتیجه‌ای برای فرآیندهای مشابه در اعماق زیاد تحت شرایط فشار–دمای شدید ارائه دهند، شناسایی می‌شوند. به طور خاص، منابع عدم قطعیت در انتقال فاز هیدروکربن، مرتبط با فقدان داده‌های تجربی برای کالیبراسیون مدل و محدودیت‌های مدل‌های ترمودینامیکی فراتر از محدوده‌های معتبر آنها، در میان سایر عوامل، مورد بحث قرار می‌گیرند.

International Journal of Engineering

Journal Homepage: www.ije.ir

Investigation of the Biodegradability of Bio-Plastics Based on Polypropylene and Starch

A. Ganji[a], Z. Ghobadi Nejad[b], S. Yaghmaei[*a,b], R. Amini[c]

[a] *Department of Chemical and Petroleum Engineering, Sharif University of Technology, Tehran, Iran*
[b] *Biochemical and Bioenvironmental Research Center, Sharif University of Technology, Tehran, Iran*
[c] *Kimia Samaneh Sabz Co., Qom, Iran*

PAPER INFO

Paper history:
Received 13 September 2025
Received in revised form 17 December 2025
Accepted 08 January 2026

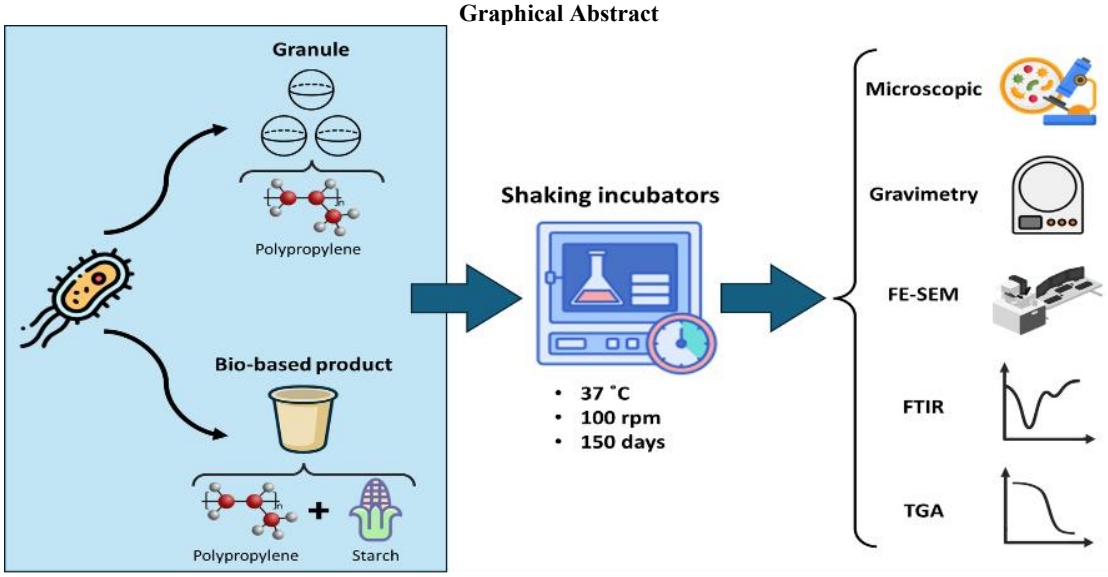
Keywords:
Bacillus sp.
Bacteria
Biodegradability
Bioplastic
Polypropylene
Starch

ABSTRACT

To address the environmental crisis caused by plastic accumulation, this study investigated the biodegradation potential of bio-based containers using indigenous microorganisms isolated and purified from plastic-contaminated sources. Two types of polymers were tested: pure polypropylene granules and starch-polypropylene based bioplastic. The biodegradation process was evaluated through visual observation, gravimetric analysis, field emission scanning electron microscope, Fourier transform infrared analysis, and thermogravimetric analysis. Results showed the indigenous bacterial strain Bacillus sp. achieved weight reductions of 4.73% for the granules and 7.1% for the bio-based product after 150 days of incubation. Fourier transform infrared analysis revealed significant surface oxidation, evidenced by the appearance of hydroxyl groups. The disappearance of methyl stretching and methylene asymmetrical stretching peaks indicated microbial carbon consumption. For the bio-based product, starch-related peaks at 713, 871, and 998 cm^{-1} disappeared, confirming the successful degradation of starch components. Thermogravimetric analysis results demonstrated a decrease in thermal stability for both polymer types, with weight loss beginning at 120°C for the granules and 150°C for the bio-based product. These findings highlight the effectiveness of indigenous microorganisms in biodegrading bioplastics, offering a targeted bioremediation approach to reduce persistent plastic waste.

***doi**: 10.5829/ije.2026.39.10a.17*

Graphical Abstract



* Corresponding Author Email: yaghmaei@sharif.edu (S. Yaghmaei)

Please cite this article as: Ganji A, Ghobadi-Nejad Z, Yaghmaei S, Amini R. Investigation of the Biodegradability of Bio-Plastics Based on Polypropylene and Starch. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2559-70.

1. INTRODUCTION

In the era of living in the plastic age, plastic production has significantly increased over recent decades (1). Synthetic plastics are recognized as highly significant materials in all modern societies. This is due to their outstanding characteristics acquired over time, including low cost, durability, and high resistance, which stem from their polymeric structure, making plastics an integral part of all aspects of our daily lives (2, 3). According to official reports, a large portion of waste consists of plastic waste, of which only 19.5% is recycled, 25.5% is incinerated, and the remaining 55% persists in the environment or landfills (4). Furthermore, according to a United Nations report, annual plastic production is expected to reach approximately 800 million tons by 2035 and 1600 million tons by 2050 (5).

More than 50% of plastic waste present in the environment may consist of polyolefins. The degradation of polyolefins in nature is a very slow process, leading to their annual accumulation in the environment (6). Polyethylene (PE) and polypropylene (PP) are considered the most common plastic wastes found (7). One of the factors contributing to the slow degradation of plastics is the weak formation of biofilm or attachment of microorganisms to the surface of polypropylene, which is related to its hydrophobic nature, causing difficulties in their degradation (6). In addition to high hydrophobicity, polypropylene is composed solely of carbon and hydrogen elements and does not provide the necessary nutrients for long-term survival and growth, such as nitrogen, phosphorus, sodium, potassium, trace elements, amino acids, and others (8).

Polypropylene, due to its excellent properties, including exceptional mechanical characteristics, ease of manufacturing, and cost-effectiveness, has been widely used in applications ranging from disposable products to long-lasting durable items. Improvements in the transparency, strength, and durability of polypropylene production have led to a significant increase in its consumption (9). Polypropylene is a cost-effective polymer derived from propylene, one of the derivatives of oil refining. The dense and hydrophobic nature of polypropylene, combined with its high crystallinity, makes it resistant to microbial attack, as it is not easily accessible to enzymes produced by microorganisms. Consequently, polypropylene is considered to have low biodegradability and can persist in the environment for extended periods, contributing to the accumulation of plastic waste (10). This polymer, with a low density of 0.90 to 0.92 g/cm³, possesses a range of superior properties, including ease of processability, desirable hardness and mechanical strength, chemical resistance, and moisture resistance, which enable its use in various industries (11). Only a limited number of microorganisms, such as *Bacillus* sp., *Brevibacillus* sp., *Phanerochaete chrysoporium*, and *Rhodococcus* sp., have been reported to degrade polypropylene (12-14). Experimental studies have shown that plastic degradation by microorganisms begins within a few days to several weeks (12). Microorganisms that grow on plastic polymers initiate the biodegradation process by adhering to the polymer surface and forming a biofilm (15, 16).

Bioplastics are among the most innovative materials, considered to be bio-based and biodegradable (17). According to the definition of bioplastics, a polymer is classified as a bioplastic if it is either bio-based, biodegradable, or possesses both characteristics. Therefore, the use of the prefix "bio" does not necessarily imply that a plastic is biodegradable, which may lead to confusion (18). The global market size for bioplastics is expanding daily and is projected to reach US$ 20 billion by 2026. Asia remains the primary hub for production, supplying 46% of the total bioplastics produced worldwide, followed by Europe with 26%, North America with 17%, and South America with 10% (19). Bioplastics can be synthesized from natural organic materials such as polysaccharides, proteins, and lipids, but starch-based bioplastics are considered the most promising type due to their abundance in nature (20). Starch is regarded as one of the most promising biopolymers for producing edible films, and its cost-effectiveness enhances its appeal. Starch-based bioplastics exhibit mechanical properties and transparency comparable to traditional plastics (21). Starch-based bioplastics offer several advantages, including biodegradability and ease of processability, such as those derived from potato starch, corn starch, wheat starch, jackfruit seed starch, and cassava plant starch (17). Starch is a natural polymer extracted from plants and, due to its inherent properties and low cost, can be utilized for producing biodegradable plastics (22).

The biodegradation process of plastics refers to a process in which microbes or microbial communities are used to degrade plastics (23). In this process, the conversion of organic compounds into biogas and residual biomass occurs as a result of microbial activity, with microorganisms utilizing the plastic as a carbon source (24-28). The biodegradation of polymers is a slow process influenced by multiple factors. Some factors are related to the polymer itself, such as its composition, structural arrangement, carbon content, chain mobility, crystallinity, molecular weight, and the presence of plasticizers or additives. Other factors depend on environmental conditions, such as the efficiency of degradation by microorganisms and the optimal activity of enzymes. These conditions can be affected by variables such as humidity, temperature, pH, nutrients, electron acceptors/donors, salinity, oxygen, sunlight, water, copolymers, and the composition of the environment which are the key factors influencing the efficiency of the biodegradation process. By optimizing

these conditions, the efficiency and rate of biodegradation can be enhanced, ideally leading to the mineralization of persistent pollutants (29-33). Microbial growth and enzyme activity are key to biological degradation and are shaped by environmental conditions. Accordingly, Ghobadi Nejad et al. (34) reported that optimizing temperature and pH enhanced bacterial growth and enzyme production.

One of the studies in this field was conducted by Wichithem et al., where a strain of the bacterium *Streptomyces ardesiacus* NBI0111 was isolated from plastic-contaminated soil in Thailand. This strain was selected based on its ability to grow on a Bushnell-Haas medium containing 1% polypropylene. The study was conducted in two phases: laboratory and natural. In the laboratory phase, sterilized 1×1 cm polypropylene pieces were added to the culture medium, and after 90 days of incubation, a 17.52% weight reduction of the plastic was observed. FESEM images confirmed the presence of cracks, surface roughness, and surface degradation. FTIR and GC-MS analyses showed a reduction in the intensity of characteristic peaks and the formation of new O-H bonds, indicating degradation of the polymeric structure. This study demonstrated that indigenous strains, through growth, surface colonization, and enzyme secretion, are capable of degrading resistant polymers such as PP (35). In another study by Lu et al., researchers investigated the role of deep-sea bacteria of the genus *Pseudoalteromonas* in degrading polypropylene and polystyrene. These bacteria were initially enriched with plastic pieces over one year at a depth of 4700 meters in the western Pacific Ocean and then cultured in the laboratory at 15°C. Two strains, *Pseudoalteromonas lipolytica* and *P. tetraodonis*, were able to reduce the weight of polypropylene films by 2.3% and 1.8%, respectively, after 80 days (36).

In recent decades, numerous efforts have been made to develop biodegradable materials. One promising approach is the combination of synthetic polymers, such as polypropylene, with bio-based compounds like starch to enhance their biodegradability. Starch, as a natural, biocompatible, and cost-effective material, can act as a dispersed phase in the polypropylene matrix, creating conditions favorable for microbial activity. However, these modified composites still exhibit relative resistance to degradation and require further investigation to assess their degradation performance by indigenous microorganisms. Given the limited information available regarding the precise performance of these microorganisms in the biodegradation of bioplastic composites, the present study was designed to evaluate the ability of indigenous microbial species to degrade disposable containers made from a combination of polypropylene and starch. The significance of this study lies in its potential to provide a biocompatible solution for managing plastic waste, contribute to the development of indigenous technologies in the field of biorecycling, and take a step toward reducing the environmental hazards caused by plastics.

2. MATERIALS AND METHODS

In this study, two different types of polymer samples were used to investigate biodegradability. One type of sample consists of pure polypropylene granules, referred to in this study as "granules." The granule samples, with an initial weight of 0.15 g, were isolated and examined. The other type is a composite of polypropylene and starch. These samples were divided into sheet-like pieces in the form of 1×1 cm squares. The initial mass of each square piece was measured as 0.04 g, and in this study, these samples are referred to as the "bio-based product."

Bacterial sampling was conducted at the site of a factory producing plant-based containers made from polypropylene and starch, which are the samples examined in this experiment. This factory, named "Kimia Samaneh Sabz Co. (with the Amelon brand)," specializes in the production of bio-based disposable containers based on polypropylene and starch. To study indigenous bacteria, two general culture media, Nutrient Agar and Nutrient Broth, and mineral salt medium (MSM) were used. The MSM culture medium is a type of medium containing inorganic compounds (mineral salts) necessary for microbial growth but lacking organic carbon sources. This is to ensure that the bacteria utilize the carbon from the structure of the material targeted for degradation instead of the carbon present in the culture medium. The components of the specialized culture medium are presented in Table 1. After performing the preparation and enrichment stages for potential bacteria, adapting the bacteria to the environment containing polymer samples, and acclimating the microorganisms to the specialized culture medium, suitable indigenous bacteria were isolated and purified. Following the

TABLE 1. Components of Mineral Salts Medium Culture medium

Component	Concentration [g. L^{-1}]
NH_4NO_3	1
K_2HPO_4	1
KCl	0.15
$CaCl_2.2H_2O$	0.1
$MgSO_4.7H_2O$	0.2
Yeast extract	0.1
$FeSO_4.7H_2O$	0.001
$ZnSO_4.7H_2O$	0.001
$MnSO_4·7H_2O$	0.001

isolation and purification of the bacteria, the polymer samples were added to erlenmeyer flasks containing the bacteria and the MSM culture medium, and they were incubated for 150 days in a shaker incubator at 37˚C and 100 rpm.

Additionally, polymer samples were placed in Erlenmeyer flasks without bacteria to serve as control samples, allowing consideration of physical changes unrelated to bacterial activity, such as sample collisions with the flask walls, sample-to-sample collisions, and temperature effects. After 150 days, the samples were removed from the flasks. The samples were then placed in an oven at 80˚C for 24 hours to eliminate moisture and ensure complete drying. To thoroughly investigate the changes in the samples and the performance of the bacteria, macroscopic observations and various analyses, including gravimetric analysis, FESEM, FTIR, and TGA, were conducted on the granule and bio-based product samples. For FESEM analysis, after coating the samples with gold and placing them under high vacuum conditions at ambient temperature, imaging was performed at 1000x magnification and a 50 µm scale. For FTIR analysis, the samples were analyzed in the wavenumber range of 600 to 4000 cm^{-1}. TGA analysis was performed on both the granule and bio-based product sheet samples, subjected to heating at a rate of 20˚C per minute, with the temperature ranging from 40˚C to 800˚C.

3. RESULTS AND DISCUSSION

After performing the isolation and purification stages of the sample collected from the site, it was observed that only one bacterial strain was present in the initial sample. Following Gram staining, it was determined that this bacterium is a Gram-positive *bacillus*, with its microscopic image visible in Figure 1 and its characteristics presented in Table 2.

Bacterial biodegradation of bioplastics based on polypropylene and starch involves distinct mechanisms due to their differing chemical structures, where starch-based bioplastics are generally more readily biodegraded by bacteria, while PP requires specific microbial adaptations and often pretreatment to enhance degradation. For polypropylene (PP), initial oxidation occurs as bacteria such as *Bacillus cereus*, *Bacillus pasteurii*, and *Gordonia spp.* initiate PP degradation by oxidizing the polymer surface, introducing carbonyl, hydroxyl, and ester groups, with this often being the rate-limiting step as PP's hydrophobic, high-molecular-weight structure resists enzymatic attack (37-42). Enzymes are natural, green catalysts produced from sources such as fungi and bacteria, enabling more sustainable and cost-effective processes (43). Enzymatic action involves key enzymes including hydrolases, alkane hydroxylases, alcohol dehydrogenases, and cytochrome P450s, which break down oxidized PP into smaller oligomers (40, 44-46). Biofilm formation happens as bacteria form biofilms on PP surfaces, facilitating localized enzyme secretion and further polymer breakdown (37, 40, 41). Biofragmentation and mineralization follow as the fragmented oligomers are assimilated and mineralized to CO_2 and H_2O via microbial metabolic pathways (37, 44, 45). For starch-based bioplastics, rapid hydrolysis takes place as starch's hydrophilic and amorphous structure allows bacterial amylases and glycoside hydrolases to quickly hydrolyze glycosidic bonds, producing sugars that are readily metabolized (40, 47, 48). Sequential degradation occurs in blends (e.g., starch with PBAT or PLA), where starch is degraded first, followed by slower breakdown of the synthetic component (48, 49). Environmental factors such as temperature, moisture, and microbial diversity significantly influence degradation rates, with thermophilic conditions accelerating the process (48-50). In conclusion, bacterial biodegradation of starch-based bioplastics is efficient due to accessible hydrolytic pathways, while pure PP degradation is slower, requiring initial oxidation and specialized enzymes, and blending starch with PP or other polymers can enhance overall biodegradability, but the synthetic fraction remains the bottleneck, with environmental conditions and microbial

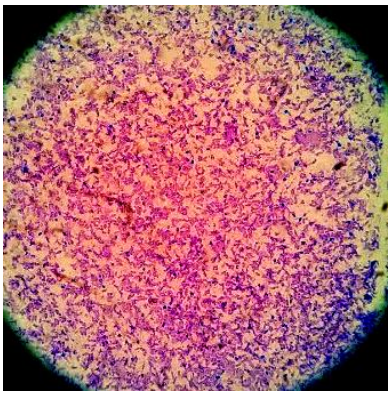


Figure 1. Microscopic image of purified indigenous bacteria

TABLE 2. Scientific classification of purified indigenous bacteria

Classification	***Bacillus* sp.**
Domain	*Bacteria*
Phylum	*Bacillota*
Class	*Bacilli*
Order	*Bacillales*
Family	*Bacillaceae*
Genus	*Bacillus*
Species	*Bacillus* sp.

community composition being critical for optimizing degradation.

3. 1. Macroscopic One of the initial indications of biodegradation of the samples by bacteria was the observation of microbial growth in the culture medium containing the polymer samples and bacteria, indicating that the microorganisms utilized the granules and the bio-based product as a carbon source. In this study, after the designated time intervals, the appearance of the samples visibly changed. Examinations revealed that due to the formation of a biofilm on and around the surface of the sample particles, these samples had settled at the bottom of the Erlenmeyer flasks. Notably, in the control samples, i.e., in the absence of bacteria, no sedimentation or biofilm formation was observed; thus, the presence of bacteria was a significant factor in the observed changes and biodegradation of the samples in this experiment. Another macroscopic observation was the high turbidity in the flasks containing bacteria compared to the control flasks. This turbidity can be attributed to two factors. The first is that the samples underwent degradation, and small fragments detached from them caused turbidity in the liquid medium. The second factor indicates bacterial growth and an increase in their population, which stems from their access to the necessary nutritional resources. Given the absence of carbon in the culture medium, this suggests that the bacteria utilized the carbon present in the polymeric bonds of the samples, leading to their degradation.

3. 2. Gravimetric One of the methods used to assess the extent of polymer degradation is to examine the weight changes of the samples compared to the initial samples. For this purpose, the weight of the granule and bio-based product samples, from which the biofilm had been removed and moisture eliminated, was measured using a balance. The dry weight of the samples and the percentage of weight reduction for each sample over the 150-day period are presented in Table 3. According to Table 3, the results indicate that the samples placed in the control environment, as expected, exhibited the least weight reduction, highlighting the significance of the presence of bacteria and their role in biodegradation. The slight weight reduction in the control environment may be attributed to physical factors such as abrasion of the samples against each other or against the flask walls, leading to the detachment of a small portion of the sample surface and consequently a reduction in their weight.

In contrast, the greatest weight reduction was observed in the bio-based product, which, from the perspective of gravimetric analysis, indicates higher biodegradability of this sample. This difference can be attributed to the presence of starch in the composition of the bio-based product, which increases the hydrophilicity of the sample, thereby enhancing its degradation.

TABLE 3. Gravimetric analysis results

Sample	Condition	Weight (g)	Weight loss (%)
Granule	Sample	0.15000	N/A
	Control	0.14950	0.31
	Indigenous bacteria	0.14440	4.73
Bio-based product	Sample	0.04000	N/A
	Control	0.03991	0.22
	Indigenous bacteria	0.03716	7.10

3. 3. Field Emission Scanning Electron Microscope (FESEM) In Figures 2 and 3, images of the surface of the granule and bio-based product samples, respectively, are presented. Each figure depicts the sample in the initial environment, control environment, and under biological degradation by the indigenous bacterial strain. Upon examination of the control samples, it is observed that in the control environment, which lacks bacteria, very minor changes, including the formation of some irregularities and small cavities in the polymer structure, have occurred. These changes may be due to the abrasion of samples against each other or their collision with the flask walls.

Figures 2.a.3 and 3.b.3 indicate that the samples subjected to biological degradation exhibit more significant surface structural changes compared to the initial and control samples, demonstrating an effective biodegradation process. Specifically, more and larger cavities are observed in the biologically degraded samples compared to the control and initial samples. When comparing the different types of samples, it can be claimed that the granule samples underwent less degradation than the bio-based product samples. This observation can be attributed to the presence of starch in the composition of the bio-based product samples compared to the granules. A similar phenomenon was also reported in a previous study by Wang et al., and Jeon et al. observed comparable results, noting that the PP microplastics treated with *Lysinibacillus* sp. showed similar outcomes (23, 51). Also, in the study by Anggiani et al., which investigated the degradation of polypropylene in contact with a bacterial consortium, it was observed that the control sample had minor surface changes, but cracks and holes were observed in the samples exposed to the bacterial consortium (52).

3. 4. Fourier Transform Infrared Spectroscopy (FTIR) The FTIR spectra for the initial granule samples, control samples, and those exposed to indigenous bacteria reveal key structural changes due to the influence of different environments. As shown in Figure 4, in the initial granule sample, characteristic peaks at 808 cm^{-1} (C-C stretching), 971 cm^{-1} (CH_3 rocking vibration), 1159 cm^{-1} (C-H wagging vibration),

1374 cm^{-1} and 1454 cm^{-1} (CH_3 symmetrical bending vibration), 2838 cm^{-1} and 2871 cm^{-1} (CH_3 stretching), 2913 cm^{-1}, and 2950 cm^{-1} (CH_2 asymmetrical stretching) indicate the primary structure of polypropylene. These findings are consistent with those reported in several other studies, , which also identify these peaks as characteristic of polypropylene structures (53). When comparing the control sample to the initial sample, slight

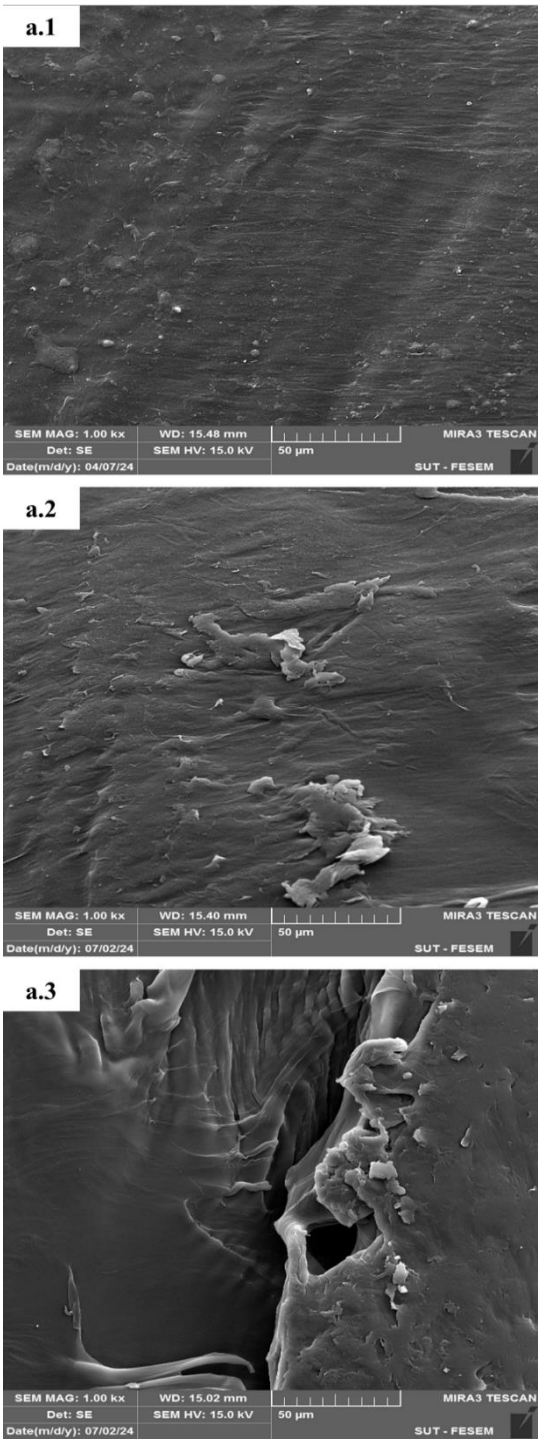


Figure 2. Field emission scanning electron microscope images of the granule at 1000x magnification: a.1, a.2, and a.3 sequentially represent the initial, control, and subjected to biodegradation by indigenous bacteria samples

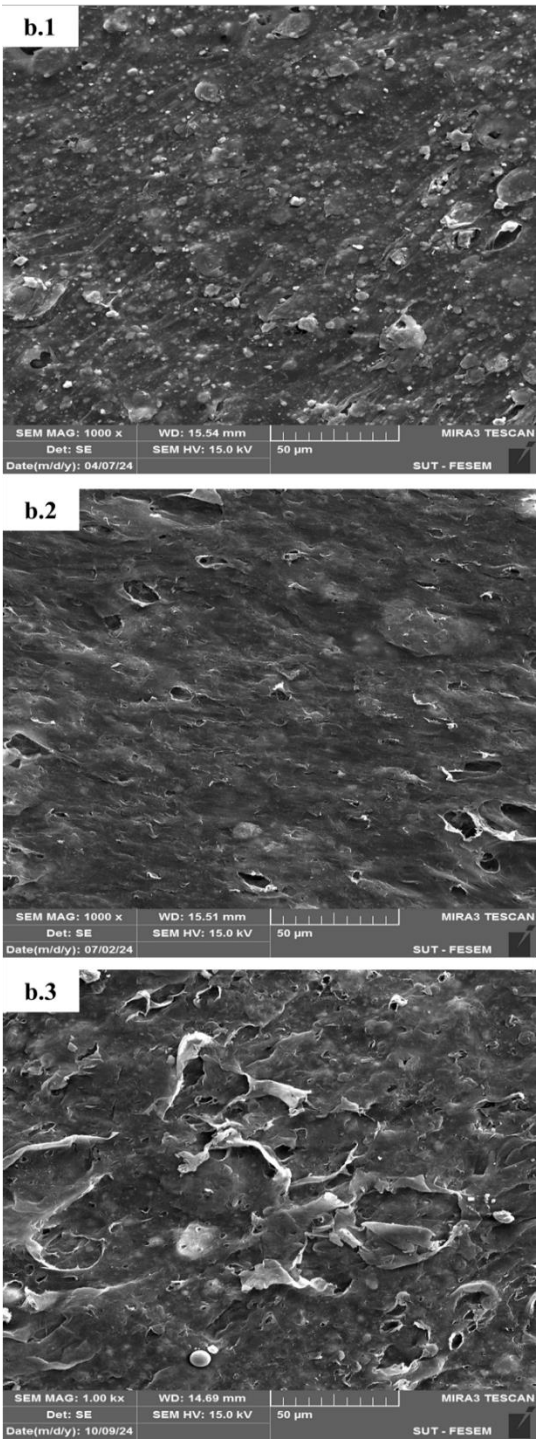


Figure 3. Field emission scanning electron microscope images of the bio-based product at 1000x magnification: b.1, b.2, and b.3 sequentially represent the initial, control, and subjected to biodegradation by indigenous bacteria samples

shifts and changes in peak positions are observed, suggesting that the control environment did not cause significant structural degradation in polypropylene. In the environment with indigenous bacteria, new intense peaks at 3425, 3521, 3637, and 3729 cm^{-1} emerged, indicating the presence of the hydroxyl (OH) functional group, which signifies strong oxidation of the samples and the progression of oxidative processes. This pattern aligns with previous studies, which also observed OH group formation at similar wavenumbers (35). These changes also report a shift in the sample surface from hydrophobic to hydrophilic, facilitating greater biofilm formation and, consequently, enhancing the biodegradability of the sample. Additionally, the previous peaks at 2834 to 2954 cm^{-1}, which represent CH_3 stretching and CH_2 asymmetrical stretching, disappeared, indicating bacterial activity and likely the consumption of carbon from these functional groups. Furthermore, the peaks at 1374 cm^{-1} (CH_2 methylene bending vibration) and 1454 cm^{-1} (CH_3 methyl bending vibration) were significantly weakened and disappeared after exposure to bacteria (54, 55). The observed changes indicated a progressive establishment and firm adherence of the microbial population onto the surface of the PP plastic, signifying increased colonization and biofilm formation. This trend, wherein the plastic-degrading community becomes more prominent over time, corroborates the findings reported by previous studies (56).

As shown in Figure 5, for the bio-based product samples, including those placed in control and indigenous environments, the control group exhibits no significant changes in peak intensity, as expected in a bacteria-free environment. Peaks around 713 cm^{-1} (CH out-of-plane vibration in starch), 871 cm^{-1} (CH rocking vibration), 998 cm^{-1} (C-O-C and OH bending in starch), 1157 cm^{-1} (C-H wagging vibration), and 1411 cm^{-1} (CH bending in starch or CH_3 symmetrical bending vibration) remain stable, indicating that the structure of the bio-based product is preserved in the absence of bacteria. The C-H stretching peaks at 2842, 2915, and 2950 cm^{-1}, corresponding to CH_3 stretching, CH_2 asymmetrical stretching, and CH_3 asymmetrical stretching in polypropylene, respectively, remain unchanged, confirming the material's stability. However, in the indigenous environment, most peaks related to starch have disappeared, indicating partial or complete consumption of this material during the process. The disappearance of functional groups at peaks 2842, 2915, and 2950 cm^{-1}, corresponding to CH_3 stretching, CH_2 asymmetrical stretching, and CH_3 asymmetrical stretching, respectively, suggests their consumption by bacteria (54, 55). Additionally, peaks at 3417, 3517, 3633, and 3722 cm^{-1} indicate the formation of hydroxyl (OH) groups, with peak intensity reflecting high oxidation in these environments and the transition of the sample surface from hydrophobic to hydrophilic (57). Regarding the FTIR peak in the range of 2850 to 3000 cm^{-1}, which corresponds to aliphatic CH stretching, it was observed that this peak disappeared after the samples were exposed to bacteria. This is consistent with findings from other investigations, where similar peak formation was observed after degradation. The biodegradation of starch-based nanocomposites showed comparable FTIR peaks, especially in the OH groups, indicating microbial activity. In these studies, the observed peaks are as follows: 1455 cm^{-1} due to the symmetric bending of CH_3, 2919 and 2953 cm^{-1} due to the asymmetric stretching of CH_2 and CH_3, and a broad band between 3000 and 3600 cm^{-1} due to OH groups linked by hydrogen bonding (58-60).

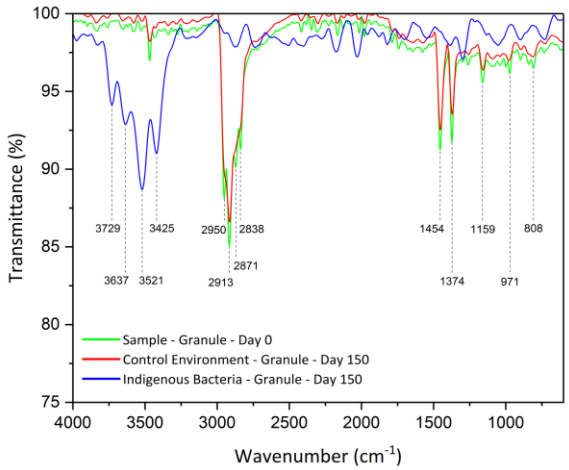


Figure 4. Fourier transform infrared analysis for the granule sample in initial, control, and indigenous bacterial biodegradation conditions

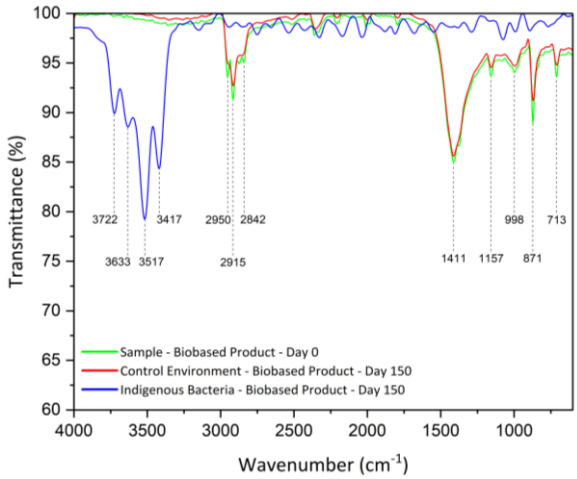


Figure 5. Fourier transform infrared analysis for the bio-based product sample in initial, control, and indigenous bacterial biodegradation conditions

3. 5. Thermogravimetric Analysis (TGA) TGA data for the granule sample reveal a clear difference in thermal stability between the control and indigenous bacterial environments. As shown in Figure 6, in the control environment, the granule sample exhibits excellent thermal stability, with negligible weight loss up to approximately 355°C. This indicates that the polymer maintains its structure in the absence of bacterial influence, retaining over 98% of its original weight up to this temperature. In the indigenous environment, the granule sample begins to lose weight at a lower temperature, suggesting that exposure to bacteria weakens the polymer structure. In the bacterial environment, the onset of thermal degradation is observed at approximately 120°C, earlier than in the control sample. At 460°C, the sample exposed to indigenous bacteria loses nearly all its weight, highlighting the impact of bacterial degradation on the polymer's thermal stability. It is likely that indigenous bacteria initiate hydrolysis or oxidation processes that destabilize the polymer chains, leading to earlier and more severe thermal degradation.

As depicted in Figure 7, in the control environment, the bio-based product remains stable up to approximately 280°C, where less than 3% weight loss is observed. The main degradation stage begins at 300°C, and by 800°C, approximately 70% of the material has degraded. In the indigenous environment, the bio-based product shows earlier degradation compared to the control sample, with weight loss starting at approximately 150°C. At 450°C, about 35% of the polymer has degraded, indicating that exposure to bacteria triggers the onset of polymer degradation, and by 800°C, approximately 80% of its weight is lost. A similar pattern of results was reported by Shiwei et al. (36).

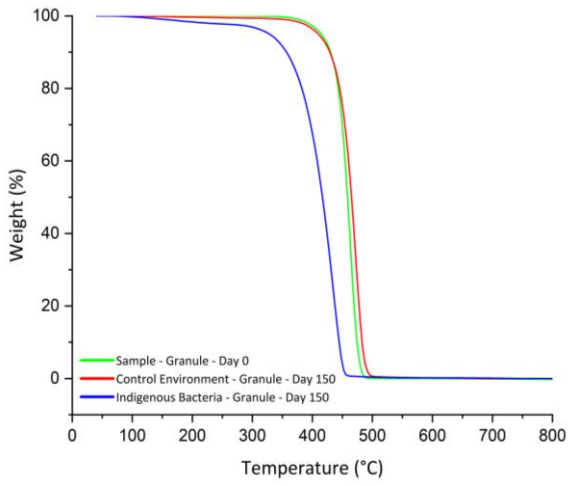


Figure 6. Thermogravimetric analysis for the granule sample in initial, control, and indigenous bacterial biodegradation conditions

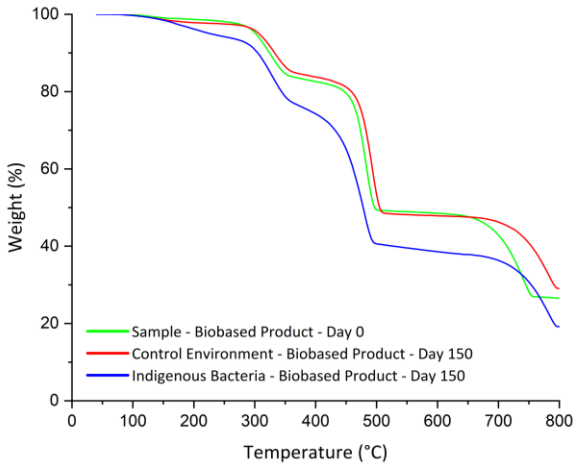


Figure 7. Thermogravimetric analysis for the bio-based product sample in initial, control, and indigenous bacterial biodegradation conditions

To compare the TGA analysis results from the previous studies, this evaluation can be approached from two distinct perspectives. From the first perspective, the analysis of each sample can be examined over time and under the influence of biological degradation. In this regard, the TGA analysis obtained exhibit complete concordance with the investigations by Jain et al. (61) and Pires et al. (62), wherein samples subjected to biological degradation display significantly weaker structural integrity and initiate weight loss at lower temperatures than the initial and control samples. This pattern continues throughout the analysis, and consequently, samples in contact with bacteria, achieve complete weight loss at lower temperatures relative to the initial and control samples. From another perspective, studies indicates that the weight loss profile for pure polypropylene proceeds uniformly (61, 62). In contrast, for bioplastic based on polypropylene and starch, the weight loss profile is non-uniform, featuring alterations during the degradation process. This variability arises because the bioplastic is not constituted from a single pure substance but rather from multiple components (polypropylene and starch), which commence thermal decomposition and weight loss at different temperatures and exhibit differential responses to thermal variations. These findings are fully consistent with the experimental results of the present study, and similar trends are observed with previous studies for the two different samples tested (53).

4. CONCLUSION

Analyses of bio-based product and granule samples showed high similarity between control and initial conditions, while samples in the control environment

differed from those exposed to indigenous bacteria, indicating no structural change in the control environment and bacterial degradation in the presence of bacteria. This study investigated the biodegradability of pure polypropylene and starch-polypropylene based bioplastic samples using control and indigenous bacterial environments. Samples were collected from a factory producing bio-based disposable containers, and a *Bacillus* sp. strain capable of degrading polypropylene was isolated. Over 150 days, bio-based product, due to their starch content, exhibited better performance in biological degradation, decomposing faster and reducing environmental pollution compared to pure polypropylene, thereby contributing to mitigating plastic accumulation in ecosystems, lowering the carbon footprint associated with persistent plastic waste, and supporting broader sustainability goals by decreasing reliance on non-degradable materials in everyday applications. FESEM analysis revealed more cracks, pits, and erosions on bio-based product surfaces than on granules, with bacterial environments showing significantly more severe surface changes than the control, which remained largely unchanged. FTIR analysis indicated reduced methyl bond peak intensity and hydroxyl bond formation in bacterially degraded samples, suggesting oxidation. TGA analysis showed lower thermal stability in bacteria-exposed samples, confirming structural degradation. These results highlight the potential for industrial scaling of microbial biodegradation processes, such as through the integration of bacterial consortia in bioreactors for efficient large-scale treatment of plastic waste in manufacturing and waste management facilities, potentially enabling cost-effective implementation in recycling plants and promoting circular economy practices. Future studies could explore *Bacillus* sp. metabolic pathways for polymer degradation and the impact of environmental factors like pH and temperature on degradation efficiency.

Acknowledgements

The author appreciates Biochemical and Bioenvironmental Research Center, Sharif University of Technology for supplying the necessary facilities and chemicals to conduct present research.

Funding

This work was supported by Kimia Samaneh Sabz Company (Amolon), Qom, Iran.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares that there are no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT exclusively for minor language editing and stylistic refinement to improve clarity and readability. The author carefully reviewed, revised, and approved the final content and takes full responsibility for the accuracy, integrity, and originality of the work.

Authors Biosketch

Alireza Ganji received his MSc in Chemical Engineering (Environmental Engineering) from Sharif University of Technology.
Zahra Ghobadi Nejad is a Senior Researcher at the Biochemical and Bioenvironmental Research Center, Sharif University of Technology.
Soheila Yaghmaei is a Professor in the Department of Chemical and Petroleum Engineering, Sharif University of Technology.
Reza Amini is an Engineer at Kimia Samane Sabz Company (Amelon).

REFERENCES

1. Silva AL, Prata JC, Walker TR, Campos D, Duarte AC, Soares AM, Barcelo D, Rocha-Santos T. Rethinking and optimising plastic waste management under COVID-19 pandemic: Policy solutions based on redesign and reduction of single-use plastics and personal protective equipment. Science of the Total Environment, 2020;742:140565. https://doi.org/10.1016/j.scitotenv.2020.140565.
2. Ali SS, Elsamahy T, Koutra E, Kornaros M, El-Sheekh M, Abdelkarim EA, Zhu D, Sun J. Degradation of conventional plastic wastes in the environment: A review on current status of knowledge and future perspectives of disposal. Science of the Total Environment, 2021;771:144719. https://doi.org/10.1016/j.scitotenv.2020.144719.
3. Ru J, Huo Y, Yang Y. Microbial degradation and valorization of plastic wastes. Frontiers in Microbiology, 2020;11:442. https://doi.org/10.3389/fmicb.2020.00442.

4. Osman AI, Nasr M, Aniagor CO, Farghali M, Huang MM, Chin BL, Sun Z, Lock SS, Lopez-Maldonado EA, Yiin CL, Chinyelu CE. Synergistic technologies for a circular economy: upcycling waste plastics and biomass. Frontiers of Chemical Science and Engineering, 2025;19(1):2-15. https://doi.org/10.1007/s11705-024-2507-0.
5. Ali SS, Elsamahy T, Al-Tohamy R, Zhu D, Mahmoud YA, Koutra E, Metwally MA, Kornaros M, Sun J. Plastic wastes biodegradation: mechanisms, challenges and future prospects. Science of the Total Environment, 2021;780:146590. https://doi.org/10.1016/j.scitotenv.2021.146590.
6. Jeyakumar D, Chirsteen J, Doble M. Synergistic effects of pretreatment and blending on fungi mediated biodegradation of polypropylenes. Bioresource Technology, 2013;148:78-85. https://doi.org/10.1016/j.biortech.2013.08.074.
7. Pabortsava K, Lampitt RS. High concentrations of plastic hidden beneath the surface of the Atlantic Ocean. Nature Communications, 2020;11(1):4073. https://doi.org/10.1038/s41467-020-17932-9.
8. Yang SS, Ding MQ, He L, Zhang CH, Li QX, Xing DF, Cao GL, Zhao L, Ding J, Ren NQ, Wu WM. Biodegradation of polypropylene by yellow mealworms (Tenebrio molitor) and superworms (Zophobas atratus) via gut-microbe-dependent depolymerization. Science of the Total Environment, 2021;756:144087. https://doi.org/10.1016/j.scitotenv.2020.144087.
9. Jeon HJ, Kim MN. Isolation of mesophilic bacterium for biodegradation of polypropylene. International Biodeterioration and Biodegradation, 2016;115:244-49. https://doi.org/10.1016/j.ibiod.2016.08.025.
10. Bautista-Zamudio PA, Florez-Restrepo MA, Lopez-Legarda X, Monroy-Giraldo LC, Segura-Sanchez F. Biodegradation of plastics by white-rot fungi: A review. Science of the Total Environment, 2023;858:165950. https://doi.org/10.1016/j.scitotenv.2023.165950.
11. Wang S, Muiruri JK, Soo XY, Liu S, Thitsartarn W, Tan BH, Suwardi A, Li Z, Zhu Q, Loh XJ. Bio-polypropylene and polypropylene-based biocomposites: solutions for a sustainable future. Chemistry–An Asian Journal, 2023;18(2):e202200972. https://doi.org/10.1002/asia.202200972.
12. Auta HS, Emenike CU, Jayanthi B, Fauziah SH. Growth kinetics and biodeterioration of polypropylene microplastics by Bacillus sp. and Rhodococcus sp. isolated from mangrove sediment. Marine Pollution Bulletin, 2018;127:15-21. https://doi.org/10.1016/j.marpolbul.2017.11.036.
13. Skariyachan S, Patil AA, Shankar A, Manjunath M, Bachappanavar N, Kiran S. Enhanced polymer degradation of polyethylene and polypropylene by novel thermophilic consortia of Brevibacillus spp. and Aneurinibacillus sp. Polymer Degradation and Stability, 2018;149:52-68. https://doi.org/10.1016/j.polymdegradstab.2018.01.018.
14. Shimpi N, Borane M, Mishra S, Kadam M, Sonawane SS. Biodegradation of isotactic polypropylene/poly(lactic acid) and iPP/PLA/nano calcium carbonates using Phanerochaete chrysosporium. Advances in Polymer Technology, 2018;37(2):522-30. https://doi.org/10.1002/adv.21691.
15. Eich A, Mildenberger T, Laforsch C, Weber M. Biofilm and diatom succession on polyethylene and biodegradable plastic bags in two marine habitats. PLoS One, 2015;10(9):e0137201. https://doi.org/10.1371/journal.pone.0137201.
16. Sekhar VC, Nampoothiri KM, Mohan AJ, Nair NR, Bhaskar T, Pandey A. Microbial degradation of high impact polystyrene, an e-plastic with decabromodiphenyl oxide and antimony trioxide. Journal of Hazardous Materials, 2016;318:347-54. https://doi.org/10.1016/j.jhazmat.2016.07.008.
17. Anugrahwidya R, Armynah B, Tahir D. Bioplastics starch-based with additional fiber and nanoparticle: characteristics and biodegradation performance. Journal of Polymers and the Environment, 2021;29:3459-76. https://doi.org/10.1007/s10924-021-02152-z.
18. Lavagnolo MC, Poli V, Zampini AM, Grossule V. Biodegradability of bioplastics in different aquatic environments: a systematic review. Journal of Environmental Sciences, 2024;134:169-81. https://doi.org/10.1016/j.jes.2023.06.013.
19. Nanda S, Patra BR, Patel R, Bakos J, Dalai AK. Innovations in applications and prospects of bioplastics and biopolymers: a review. Environmental Chemistry Letters, 2022;20:379-95. https://doi.org/10.1007/s10311-021-01334-4.
20. Shafqat A, Tahir A, Mahmood A, Tabinda AB, Yasar A, Pugazhendhi A. Environmental significance of carbon footprints of starch-based bioplastics. Biocatalysis and Agricultural Biotechnology, 2020;27:101540. https://doi.org/10.1016/j.bcab.2020.101540.
21. Shahabi-Ghahfarrokhi I, Goudarzi V, Babaei-Ghazvini A. Production of starch-based biopolymer by green photochemical reaction for food packaging applications. International Journal of Biological Macromolecules, 2019;122:201-09. https://doi.org/10.1016/j.ijbiomac.2018.10.154.
22. Wahyuningtyas N, Suryanto H. Analysis of biodegradation of bioplastics made of cassava starch. Journal of Mechanical Engineering Science and Technology, 2017;1(1):24-31. https://doi.org/10.17977/um016v1i12017p024.
23. Jeon JM, Park SJ, Choi TR, Park JH, Yang YH, Yoon JJ. Biodegradation of polyethylene and polypropylene by Lysinibacillus species JJY0216. Polymer Degradation and Stability, 2021;191:109662. https://doi.org/10.1016/j.polymdegradstab.2021.109662.
24. Ali SS, Al-Tohamy R, Manni A, Luz FC, Elsamahy T, Sun J. Enhanced digestion of bio-pretreated sawdust using a novel bacterial consortium. Fuel, 2019;254:115597. https://doi.org/10.1016/j.fuel.2019.06.012.
25. Ali SS, Nessem AA, Sun J, Li X. Effects of water hyacinth pretreated digestate on Lupinus termis seedlings under salinity stress. Journal of Environmental Chemical Engineering, 2019;7(3):103159. https://doi.org/10.1016/j.jece.2019.103159.
26. Shah AA, Hasan F, Hameed A, Ahmed S. Biological degradation of plastics: a comprehensive review. Biotechnology Advances, 2008;26:246-65. https://doi.org/10.1016/j.biotechadv.2007.12.005.
27. Ali SS, Sun J. Effective thermal pretreatment of water hyacinth for enhancement of biomethanation. Journal of Environmental Chemical Engineering, 2019;7(1):102853. https://doi.org/10.1016/j.jece.2018.102853.
28. Magnin A, Pollet E, Phalip V, Averous L. Evaluation of biological degradation of polyurethanes. Biotechnology Advances, 2020;39:107457. https://doi.org/10.1016/j.biotechadv.2019.107457.
29. Priya A, Dutta K, Daverey A. Biotechnological and molecular insight into plastic degradation by microbial communities. Journal of Chemical Technology and Biotechnology, 2022;97:381-90. https://doi.org/10.1002/jctb.6675.
30. Siracusa V. Microbial degradation of synthetic biopolymers waste. Polymers, 2019;11(6):1066. https://doi.org/10.3390/polym11061066.
31. Li S, Yang Y, Yang S, Zheng H, Zheng Y, Nagarajan D, Varjani S, Chang JS. Recent advances in biodegradation of microplastics. Chemosphere, 2023;331:138776. https://doi.org/10.1016/j.chemosphere.2023.138776.
32. Stoleru E, Hitruc EG, Vasile C, Oprica L. Biodegradation of poly(lactic acid)/chitosan stratified composites by Phanerochaete chrysosporium. Polymer Degradation and Stability, 2017;143:118-29. https://doi.org/10.1016/j.polymdegradstab.2017.06.023.

33. Samadi A, Sharifi H, Ghobadi Nejad Z, Hasan-Zadeh A, Yaghmaei S. Biodegradation of 4-chlorobenzoic acid by Lysinibacillus macrolides DSM54T. International Journal of Environmental Research, 2020;14(2):145-54. https://doi.org/10.1007/s41742-020-00247-4.
34. Ghobadi Nejad Z, Yaghmaei S, Hajhosseini R. Production of extracellular protease by Bacillus licheniformis. Chemical Engineering Transactions, 2009;21:242-47. https://doi.org/10.3303/CET1021242.
35. Wichatham K, Piyaviriyakul P, Boontanon N, Surinkul N, Visvanathan C, Fujii S, Boontanon SK. Biodegradation of polypropylene plastics by Streptomyces sp. Environmental Technology & Innovation, 2024;35:103681. https://doi.org/10.1016/j.eti.2024.103681.
36. Lv S, Wang Q, Li Y, Gu L, Hu R, Chen Z, Shao Z. Biodegradation of polystyrene and polypropylene by deep-sea psychrophilic bacteria. Science of the Total Environment, 2024;948:174857. https://doi.org/10.1016/j.scitotenv.2024.174857.
37. Nyamjav I, Jang Y, Park N, Lee YE, Lee S. Physicochemical evidence of polypropylene biodegradation by Bacillus cereus from waxworm gut. Journal of Polymers and the Environment, 2023;31(10):4274-87. https://doi.org/10.1007/s10924-023-02878-y.
38. Wang Z, Li Y, Bai L, Hou C, Zheng C, Li W. Biodegradation of polypropylene microplastics by Bacillus pasteurii. Emerging Contaminants, 2025;11(1):100397. https://doi.org/10.1016/j.emcon.2024.100397.
39. He L, Ding J, Yang SS, Zang YN, Pang JW, Xing D, Zhang LY, Ren N, Wu WM. Molecular-weight-dependent degradation of polypropylene microplastics by mealworms. Environmental Science & Technology, 2024;58(15):6647-58. https://doi.org/10.1021/acs.est.3c06954.
40. Pawano O, Jenpuntarat N, Streit WR, Perez-Garcia P, Pongtharangkul T, Phinyocheep P, Thayanukul P, Euanorasetr J, Intra B. Metagenomic exploration of polypropylene-degrading enzymes from mangrove sediment. Frontiers in Microbiology, 2024;15:1347119. https://doi.org/10.3389/fmicb.2024.1347119.
41. Jeyavani J, Al-Ghanim KA, Govindarajan M, Nicoletti M, Malafaia G, Vaseeharan B. Bacterial screening for polypropylene microplastics biodegradation. Science of the Total Environment, 2024;918:170499. https://doi.org/10.1016/j.scitotenv.2024.170499.
42. Rana AK, Thakur MK, Saini AK, Mokhta SK, Moradi O, Rydzkowski T, Alsanie WF, Wang Q, Grammatikos S, Thakur VK. Recent developments in microbial degradation of polypropylene. Science of the Total Environment, 2022;826:154056. https://doi.org/10.1016/j.scitotenv.2022.154056.
43. Sadeghzadeh S, Ghazvini S, Hejazi S, Yaghmaei S, Ghobadi NZ. Immobilization of laccase from Trametes hirsuta onto CMC-coated magnetic nanoparticles. International Journal of Engineering, 2020;33(4):601-08. https://doi.org/10.5829/ije.2020.33.04a.01.
44. da Silva MR, Souza KS, Motteran F, de Araujo LC, Singh R, Bhadouria R, de Oliveira MB. Microplastic-degrading bacteria: a systematic review. Frontiers in Microbiology, 2024;15:1360844. https://doi.org/10.3389/fmicb.2024.1360844.
45. Lv S, Li Y, Zhao S, Shao Z. Biodegradation of typical plastics: microbial diversity and metabolic mechanisms. International Journal of Molecular Sciences, 2024;25(1):593. https://doi.org/10.3390/ijms25010593.
46. Zhu Y, Wang H, Bai J, Qi Y, Han D. Biodegradation potential of Gordonia spp. on polypropylene and polystyrene. Frontiers in Microbiology, 2025;16:1621498. https://doi.org/10.3389/fmicb.2025.1621498.
47. Eronen-Rasimus EL, Nakki PP, Kaartokallio HP. Degradation rates and bacterial communities of bioplastics in brackish marine environments. Environmental Science & Technology, 2022;56(22):15760-69. https://doi.org/10.1021/acs.est.2c06280.
48. Shafana Farveen M, Munoz R, Narayanan R, Garcia-Depraect O. Enhancing bioplastic degradation in anaerobic digestion. Polymers, 2025;17(13):1756. https://doi.org/10.3390/polym17131756.
49. Yu C, Dongsu B, Tao Z, Xintong J, Ming C, Siqi W, Zheng S, Yalei Z. Anaerobic co-digestion of commercial bioplastic bags with food waste. Science of the Total Environment, 2023;859:159967. https://doi.org/10.1016/j.scitotenv.2022.159967.
50. El Feky AR, Ismaiel M, Yilmaz M, Madkour FM, El Nemr A, Ibrahim HAH. Biodegradable plastic from chitosan with castor oil and starch. Scientific Reports, 2024;14(1):11161. https://doi.org/10.1038/s41598-024-61377-9.
51. Wang P, Zhao J, Ruan Y, Cai X, Li J, Zhang L, Huang H. Degradation of polypropylene by Pseudomonas aeruginosa strains. Journal of Polymers and the Environment, 2022;30(9):3949-58. https://doi.org/10.1007/s10924-022-02480-8.
52. Anggiani M, Kristanti RA, Hadibarata T, Kurniati TH, Shiddiq MA. Degradation of polypropylene microplastics by bacterial consortium. Water, Air, & Soil Pollution, 2024;235(5):308. https://doi.org/10.1007/s11270-024-07113-5.
53. Raee E, Kaffashi B. Biodegradable polypropylene/thermoplastic starch nanocomposites. Journal of Applied Polymer Science, 2018;135(4):45740. https://doi.org/10.1002/app.45740.
54. Fang J, Zhang L, Sutton D, Wang X, Lin T. Needleless melt-electrospinning of polypropylene nanofibres. Journal of Nanomaterials, 2012;2012:382639. https://doi.org/10.1155/2012/382639.
55. Abdullah AHD, Chalimah S, Primadona I, Hanantyo MHG. Physical and chemical properties of corn, cassava, and potato starches. IOP Conference Series: Earth and Environmental Science, 2018;160(1):012003. https://doi.org/10.1088/1755-1315/160/1/012003.
56. Park SY, Kim CG. Biodegradation of micro-polyethylene particles by bacterial colonization. Chemosphere, 2019;222:527-33. https://doi.org/10.1016/j.chemosphere.2019.01.159.
57. Sun Y, Zhang Y, Hao X, Zhang X, Ma Y, Niu Z. Degradation of additive-free polypropylene by Exiguobacterium marinum. Environmental Pollution, 2023;336:122390. https://doi.org/10.1016/j.envpol.2023.122390.
58. Schmitt H, Prashantha K, Soulestin J, Lacrampe MF, Krawczak P. Halloysite nanotubes/wheat starch nanocomposites. Carbohydrate Polymers, 2012;89(3):920-27. https://doi.org/10.1016/j.carbpol.2012.04.037.
59. Hanifi S, Oromiehie A, Ahmadi S, Farhadnejad H. Corn starch and montmorillonite nanocomposite-reinforced polypropylene. Journal of Vinyl and Additive Technology, 2014;20(1):16-23. https://doi.org/10.1002/vnl.21333.
60. Orue A, Corcuera MA, Pena C, Eceiza A, Arbelaiz A. Bionanocomposites based on thermoplastic starch and cellulose nanofibers. Journal of Thermoplastic Composite Materials, 2016;29(6):817-32. https://doi.org/10.1177/0892705714536424.
61. Jain K, Bhunia H, Sudhakara Reddy M. Degradation of polypropylene–poly-L-lactide blend by bacteria. Bioremediation Journal, 2018;22(3-4):73-90. https://doi.org/10.1080/10889868.2018.1516620.
62. Pires JP, Miranda GM, de Souza GL, Fraga F, da Silva Ramos A, de Araujo GE, Ligabue RA, Azevedo CM, Lourega RV, de Lima JE. Degradation of polypropylene in soil using an enzymatic additive. Iranian Polymer Journal, 2019;28(12):1045-55. https://doi.org/10.1007/s13726-019-00766-8.

Persian Abstract

چکیده

برای مقابله با بحران زیست‌محیطی ناشی از تجمع پلاستیک‌ها، در این مطالعه پتانسیل تجزیه‌زیستی ظروف زیست‌پایه با استفاده از میکروارگانیسم‌های بومی جداسازی و خالص‌سازی‌شده از منابع آلوده به پلاستیک مورد بررسی قرار گرفت؛ بدین منظور دو نوع پلیمر شامل گرانول پلی‌پروپیلن خالص و بیوپلاستیک بر پایه نشاسته–پلی‌پروپیلن انتخاب شدند و فرآیند تجزیه‌زیستی از طریق مشاهده ظاهری، آنالیز وزنی، میکروسکوپ الکترونی روبشی گسیل میدانی، طیف‌سنجی مادون قرمز تبدیل فوریه و آنالیز گرماوزنی ارزیابی شد. نتایج نشان داد سویه باکتریایی بومی *Bacillus* sp. پس از ۱۵۰ روز انکوباسیون، کاهش وزن معادل٤/٧٣٪ برای نمونه‌های گرانولی و ۷/۱٪ برای محصول زیست‌پایه ایجاد کرده است. طیف‌سنجی مادون قرمز تبدیل فوریه بیانگر اکسیداسیون قابل‌توجه سطحی با ظهور گروه‌های هیدروکسیل بود و حذف پیک‌های کشش متیل و کشش نامتقارن متیلن به مصرف کربن پلیمر توسط میکروارگانیسم‌ها نسبت داده شد؛ همچنین در محصول زیست‌پایه، ناپدید شدن پیک‌های مرتبط با نشاسته در ۷۱۳، ۸۷۱ و ۹۹۸ cm^{-1}، تجزیه موفق اجزای نشاسته‌ای را تأیید کرد. نتایج آنالیز گرماوزنی نیز کاهش پایداری حرارتی هر دو نوع پلیمر را نشان داد، به‌طوری‌که آغاز کاهش وزن برای نمونه‌های گرانولی در ۱۲۰ درجه سانتی‌گراد و برای محصول زیست‌پایه در ۱۵۰ درجه سانتی‌گراد مشاهده شد که در مجموع کارایی میکروارگانیسم بومی را در تجزیه‌زیستی بیوپلاستیک‌ها و ارائه رویکردی هدفمند برای زیست‌پالایی و کاهش پسماندهای پلاستیکی پایدار برجسته می‌سازد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2571-2582

International Journal of Engineering

Journal Homepage: www.ije.ir

Multi-Stream Nozzle Apparatus for Advanced Energy-Efficient Jet Devices and Systems

Y. A. Sazonov[a], M. A. Mokhov[a], I. V. Gryaznova[a], V. V. Voronova[a], H. A. Tumanyan[a], E. I. Konyushkov[a], A. V. Dengaev[a], L. R. Sagirova*[b], S. S. Zamaraev[b]

[a] *Gubkin Russian State University of Oil and Gas (National Research University), Moscow, Russia*
[b] *Empress Catherine II St. Petersburg Mining University, Saint Petersburg, Russia*

PAPER INFO

Paper history:
Received 09 October 2025
Received in revised form 09 December 2025
Accepted 29 December 2025

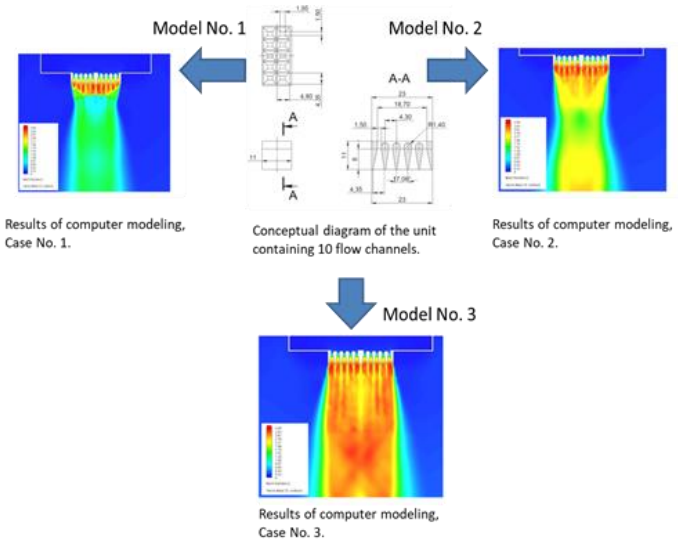
Keywords:
Nozzle Apparatus
Ejector
Thrust Vector
Computer Modeling
Euler's Methodology
Jet Systems

ABSTRACT

This paper aims to carry out interdisciplinary research within an emerging scientific field focused on thrust vector control over a full spherical geometry, where the thrust vector can be deflected continuously through a complete angular range of ±180° in all spatial directions. A patented variant of a multi-stream nozzle apparatus is analyzed. The distribution of energy of the working gas is analyzed for cases with different working pressures. A critical cross-section of rectangular geometry has been examined. The article presents 3 simulation results with a range of values for gas mass flow rate from 0.296 kg/s to 1.173 kg/s and for thrust from 396.53 N to 2140.15 N. Prospects for the practical application of the obtained results are discussed, including the creation of digital twins for use in the educational process in training engineers. For conceptual design, Euler's methodology in combination with CFD technologies is proposed, consistent with contemporary approaches to handling large datasets. A scientific foundation has been established for the development of lattice-type multistream jet systems based on Euler's methodology.Three main directions for further scientific research are identified: energy-efficient power systems; effective development of oil and gas fields; and transport robotic systems for operation on land, at sea, and in the air.

***doi**: 10.5829/ije.2026.39.10a.18*

Graphical Abstract



*Corresponding Author Institutional Email: gumersan@mail.ru (L. R. Sagirova)

Please cite this article as: Sazonov YA, Mokhov MA, Gryaznova IV, Voronova VV, Tumanyan HA, Konyushkov EI, Dengaev AV, Sagirova LR, Zamaraev SS. Multi-Stream Nozzle Apparatus for Advanced Energy-Efficient Jet Devices and Systems. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2571-82.

1. INTRODUCTION

To tackle the existing challenges in lowering energy costs in technological processes, it is essential to explore new directions of technological development, including those related to the extraction and processing of liquid and gaseous hydrocarbons.

Contemporary CFD technologies provide unique opportunities for studying complex gas-dynamic and hydrodynamic phenomena occurring within the channels of multiphase ejectors. Drawing upon Euler's foundational concepts, the authors have formulated an innovative scientific methodology in the field of jet propulsion systems through fundamental theoretical and experimental research. The proposed systems facilitate thrust vector control over the entire geometric sphere, enabling vector deflection angles spanning ±180 degrees in all spatial orientations (1-3). Several technical solutions have been patented, including a novel class of mesh turbomachinery equipped with mesh-based jet control systems. Prospective areas of application for the results of these studies include energy, hydrocarbon production, transportation, and robotics. This article serves as a logical continuation of a series of scientific papers published in international peer-reviewed journals for the years 2020 and 2025.

In this context, a comprehensive analysis of scientific and technical information was conducted taking into account the well-established big data fusion technology "MSIF," which integrates inputs from numerous heterogeneous sources (4-6). In this case, the conditions in the energy resource market (7-9), as well as various difficulties in oil and gas production (10-12) have been taken into account. By analogy with reversible turbomachinery (13), reversible jet pumps and ejectors (1-3) are considered and studied. New possibilities of flow control in the main and additional channels of hydraulic machines are being studied (14-16). More complex hybrid systems are being increasingly considered (17, 18). Issues related to the optimization of nozzle assembly design are being addressed (19-22). Electronic systems for controlling the ejector nozzle are being considered in the field of hydrogen energy (23-26). Ejectors with annular mixing chambers are being considered separately (27-29). In aviation technologies, ejectors for enhancing engine thrust are being actively studied (30, 31). Jet control systems are recognized as revolutionary advancements in the aviation industry and are viewed as one of the most promising fields of development, particularly in the research carried out by DARPA and Boeing (32-34). Utilizing jet control and monitoring systems can notably enhance the maneuverability of aircraft and optimize their energy consumption, reduce their radar visibility, and make them less detectable. However, it should be noted that there is not enough information about such technologies in scientific publications, which underlines the importance of conducting scientific research and creating new technical concepts in the field of jet technology. The potential to expand the range of ejector applications through the use of curved mixing chambers are currently being studied (35-37). When developing ejectors featuring a curved mixing chamber, the results from studies on S-shaped channels are taken into account (38-40) and related systems (41-43) should be considered.

In summary, it is worth noting that modern scientific and technical progress is largely driven by a set of interrelated areas and technologies. These include automation, artificial intelligence technologies, shape modification systems, as well as groups of autonomous vehicles or robotic devices. The advancement of energy-efficient and environmentally safe industrial technologies relies heavily on interdisciplinary research efforts, which are essential for addressing complex engineering and sustainability challenges.

This research focuses on the computational modeling and performance analysis of a multiphase nozzle system intended for use in energy-efficient jet technologies. The broader aim is to enable the development of adaptive jet control systems characterized by high responsiveness, structural simplicity, and broad applicability. In parallel, the study includes the development of digital twin models to aid in the training and professional growth of design engineers, based on methodological principles derived from Eulerian mechanics (1, 2). The switching time between nozzle operating modes can be reduced by employing a multi-stream nozzle, where a single large nozzle is replaced by a cluster of smaller, geometrically similar ones.

2. MATERIALS AND METHODS

To provide a clear representation of the research methodology employed, a flowchart has been constructed and is presented in Figure 1. This methodology encompasses the full scope of activities contributing to the development of new technological solutions in the field of energy-efficient jet installations. First of all, it is necessary to prepare theoretical hypotheses which are then translated into schematic representations of various devices through comprehensive analysis of scientific and technical literature. Based on these schematics, computer modeling is carried out using CFD technology. Following the analysis of modeling results, three-dimensional prototypes are created using 3D printing. The subsequent stages include assessing the potential for publishing scientific papers and patents, as well as incorporating the obtained research data into a unified database to support further fundamental and applied studies. The outcomes of this work contribute to educating students and training

intelligent systems in new developments and recommendations within the research domain.

The formulation of hypotheses is carried out within the framework of multiparametric (multidimensional) problems, drawing on Euler's ideas. Consideration is also given to the prospect of more active practical application of artificial intelligence tools. A fundamental research direction has been selected, focused on the study of gas-dynamic and hydrodynamic processes in thrust vector control under extreme conditions, thereby opening new opportunities to transcend established boundaries and expand the existing body of knowledge (1, 2).

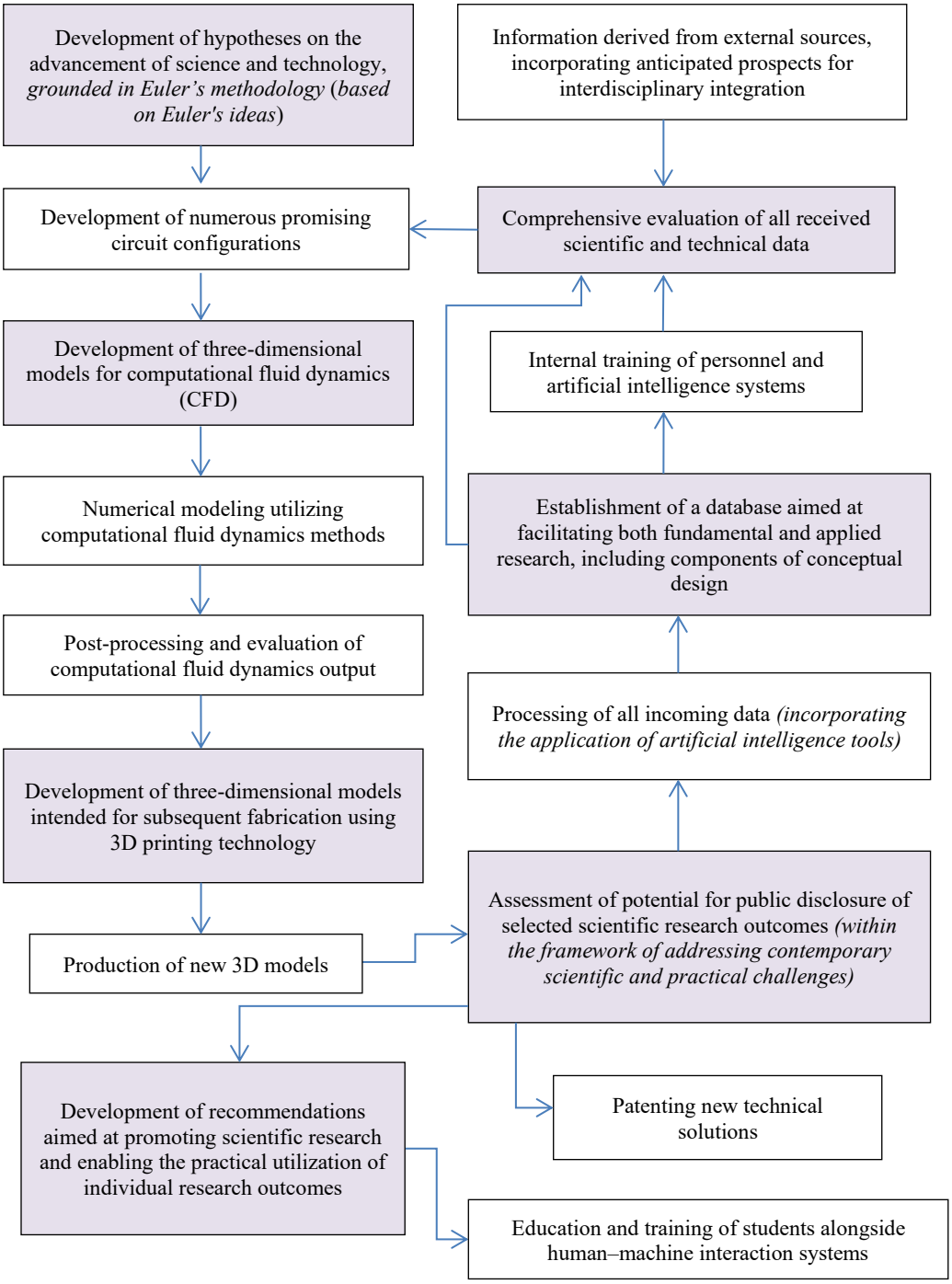


Figure 1. Block diagram of the methodology employed in the course of scientific research

Within fundamental research, the objective is to broaden the range of possibilities for achieving competitive advantages in the advancement of science and technology. Within applied research, the objective is to develop recommendations for the practical implementation of new competitive advantages from the aforementioned set of possibilities.

The scientific and engineering research was carried out with reference to the theoretical works of Euler. As demonstrated by the results obtained, Euler's legacy has not yet been fully explored. Many of his ideas were far ahead of their time and, in fact, continue to surpass our present understanding. It is appropriate here to recall the remark made by Feyerabend (44): "There is no idea, however ancient or absurd, that is not capable of improving our knowledge. The whole history of thought is condensed in science and is used for the improvement of every single theory. No idea has ever been studied in all its implications, and no concept has ever been given all the chances it deserves to succeed. Theories are eliminated and replaced by more fashionable ones long before they have had the opportunity to display their full worth." One of Euler's fundamental works, originally written in 1754, was reintroduced by modern historians in digital format in 2021 (45). In this work, Euler described his new turbine and presented a mathematical model for calculating the Segner turbine. Building on this foundation, it remains possible today to formulate new directions for the advancement of science and engineering.

According to Euler's methodology, scientific, inventive, and engineering tasks should be addressed as part of an integrated whole within the framework of interdisciplinary studies. In such interdisciplinary research, mathematics and computer technologies (CFD) currently serve as the principal unifying elements.

3. RESULTS

3. 1. Development of a conceptual design of a Multi-stream Nozzle Apparatus Building upon Euler's work (45), the present study extends the investigation to a variant of the patented technical solution (1-3). Figure 2 presents a schematic diagram of a jet device enabling angular displacement of the nozzle 1.

In the general case, when Euler's work is interpreted more broadly, various modes of fluid flow through the nozzle 1 can be considered. These may involve a liquid, a gas, or a multiphase multicomponent medium. The flow velocity at the nozzle exit is denoted by v_c. The nozzle may move with velocity v_r or remain stationary, depending on the boundary conditions of the problem. The flow direction can be varied by controlling the thrust vector within the limits of a complete geometric sphere, with thrust vector control considered in terms of its magnitude, direction, and point of application.

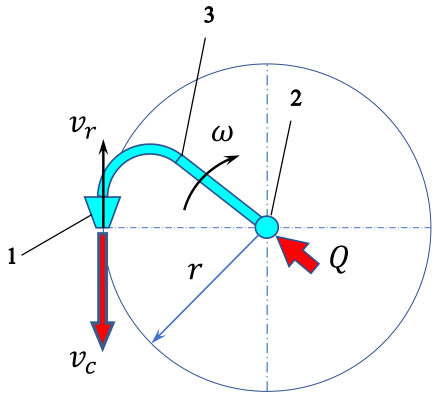


Figure 2. Schematic of the jet device variant (1-3): 1 – nozzle, 2 – energy supply point, 3 – energy transmission channel

In a specific case, Euler employed this schematic when developing the theoretical model for calculating the Segner turbine (45). It is therefore reasonable to reproduce Euler's mathematical reasoning. The reactive force (thrust) F generated by the flow of a fluid with density ρ and volumetric flow rate Q is expressed as:

$$F = Q\rho(v_c - v_r) \tag{1}$$

$$v_r = \omega r \tag{2}$$

where r is the radius of the rotor in the Segner turbine, and ω is the angular velocity of the rotor.

Torque of the rotor:

$$M = Fr \tag{3}$$

Effective power of the turbine:

$$N_i = M\omega = Fr * \frac{v_r}{r} = Fv_r = Q\rho(v_c - v_r)v_r \tag{4}$$

The power of the flow at the turbine inlet, given the inlet pressure P_0, and the outlet pressure P_r , is expressed as:

$$N_0 = Q(P_0 - P_r) = Q\rho * \frac{v_0^2}{2} \tag{5}$$

The efficiency of the turbine (for the idealized turbine model) is given by:

$$\eta = \frac{N_i}{N_0} = \frac{2(v_c - v_r)v_r}{v_0^2} \tag{6}$$

$$\frac{v_c^2}{2}\rho = \left(P_0 + \frac{v_r^2}{2}\rho\right) - P_r = \frac{v_r^2}{2}\rho + \frac{v_0^2}{2}\rho \tag{7}$$

$$v_c = \sqrt{v_r^2 + v_0^2} \tag{8}$$

Relative velocity:

$$v = \frac{v_r}{v_0} \tag{9}$$

$$\eta = \frac{2\left(\sqrt{v_r^2+v_0^2}-v_r\right)v_r}{v_0^2} \tag{10}$$

A representation of Euler's formula for the idealized Segner turbine (45) is given as:

$$\eta = 2\left(\sqrt{v^2+1}-v\right)v \tag{11}$$

An alternative formulation of Euler's equation – adapted for the actual Segner turbine and its possible modifications, and incorporating the nozzle discharge coefficient $\mu = \varepsilon\varphi$, where ε is the jet (flow) contraction coefficient and φ is the velocity coefficient of the nozzle assembly and the flow passages as a whole, introduced to account for hydraulic losses – is given as:

$$\eta = 2\left(\mu\sqrt{v^2+1}-v\right)v \tag{12}$$

Base on Bistafa (45) and Equations 1-11 partially reflect the scientific foundation laid by Euler for the development of modern turbomachinery and jet technology as a whole. The potential of Euler's ideas has not yet been fully revealed, and future generations of researchers will need to continue advancing this line of inquiry.

Based on Euler's Formulas 1-11, the parameters of a reactive jet with cross-sectional area f can be separately described for the case when $\omega = 0$. In this case, the thrust of a stationary reactive nozzle, F_0, can be defined as follows:

$$F_0 = Q\rho v_c = Q\rho v_0 = f\rho v_0^2 = f\rho(2\frac{N_0}{F_0})^2 \tag{13}$$

According to Euler's static theory, the principal parameters of an ideal jet are related by the following expression:

$$4\rho f\frac{N_0^2}{F_0^3} = 1 \tag{14}$$

This form of Euler's Equation 14 is applicable to the calculation of jet systems with arbitrary cross-sectional shapes for an idealized jet. In the specific case of a jet with a circular cross section – for instance, the jet at the exit of an air propeller of diameter D, Equation 14 can be transformed almost entirely into the Wellner–Zhukovsky equation (46), where the jet power N_0 (in watts) is expressed as the product of the input power supplied to the air propeller and its efficiency (units of measure: thrust F_0 in newtons; propeller diameter D in meters).

$${F_0}^3 = (1{,}39 * D * N_0)^2 = (1{,}39 * D * N_D * \eta_D)^2 \tag{15}$$

In the Wellner–Zhukovsky equation (46), the coefficient takes a value of 1.37. As shown by the derived formula based on Euler's theory 15, a discrepancy arises in the third significant digit of this coefficient. However, given the universality of Equation 14, it can be applied to a broad range of nozzle configurations and jet geometries – circular, rectangular, triangular, annular, and others, as well as to media of different densities, thereby enabling adaptation to varying flight altitudes in aeronautical applications.

Summarizing the intermediate results, the following conclusions can be drawn: 1 – one of the patented variants of a jet apparatus with distributed energy supply has been analyzed within the framework of Euler's theory (1-3), demonstrating the capability for thrust vector control across the entire geometric sphere; 2 – numerical experiments have confirmed the operability of a multi-stream jet apparatus with distributed energy supply, intended for the development of compact, high-response, and ergonomically efficient control systems capable of addressing a wide range of applied engineering tasks; 3 – based on the obtained results, a preliminary conclusion can be drawn regarding the applicability of such jet systems for processes involving the transfer of gases, liquids, and gas–liquid mixtures.

Figure 3 presents the conceptual design developed in the course of investigating advanced jet systems (1, 2). These systems are intended to address a range of practical tasks, including hydrocarbon extraction and processing, as well as thrust vector control within the limits of a complete geometric sphere.

The block shown contains flow channels. At the critical cross-section, the channel is rectangular, measuring 1.5 mm by 1.95 mm. The block height is 11 mm.

The nozzle apparatus shown in Figure 3 is assembled from 8 units. The height of the nozzle device can be decreased by an order of magnitude through the parallel integration of 80 channels. A geometrically similar nozzle apparatus with only a single flow channel would have a height of 110 mm instead of 11 mm.

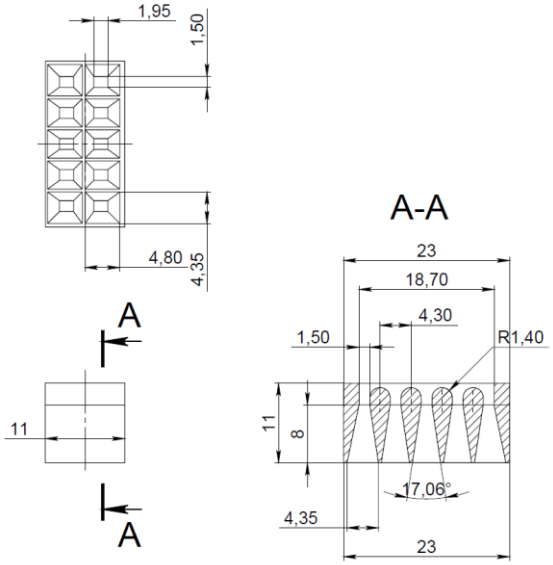


Figure 3. Conceptual diagram of the unit containing 10 flow channels

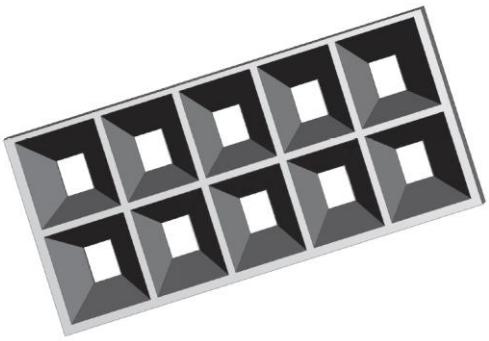


Figure 4 (a). Three-dimensional computer model of the nozzle apparatus, which includes 10 flow channels, was created for numerical experiments

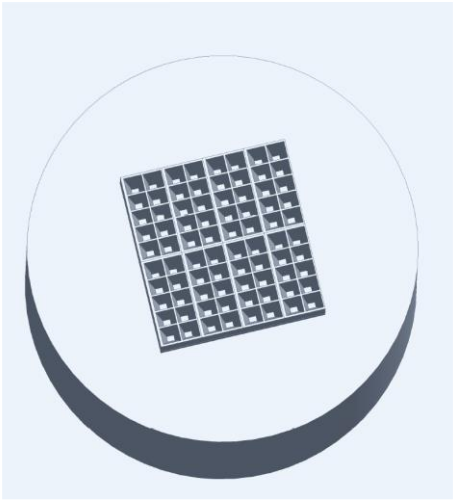


Figure 4 (b). Three-dimensional computer model of the nozzle apparatus, which includes 8 units and a total of 80 flow channels, was created for numerical experiments

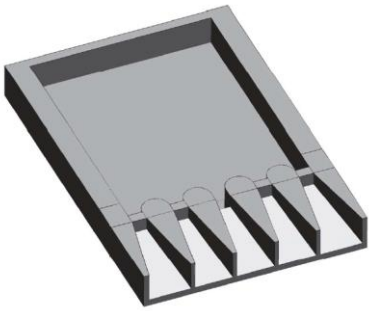


Figure 4 (c). A three-dimensional computational model of the nozzle apparatus was developed to perform numerical experiments. The model represents a block comprising five adjustable flow channels, configured with one open central channel and four closed peripheral channels

3. 2. Results of Computer Modeling In the numerical experiments, air served as the model working medium. The baseline parameters used for the calculations in this case study included gas inlet pressures of 1.5 MPa (Case No. 1), 3 MPa (Case No. 2), and 6 MPa (Case No. 3). The temperature of the gas entering the nozzle was maintained at 2000 °C.

Gas pressure at the nozzle inlet: P=3 MPa (Cases No. 4 and No. 5). Gas temperature at the nozzle inlet: 1000 ºC.

The outlet pressure was set to atmospheric conditions, specifically 101,325 Pa, while the surrounding ambient temperature was 20 °C.

- Computer parameters used for CFD:
- Software Product (SW for calculations): Flow Simulation 2018 SP5.0
- CPU type (processor): Intel(R) Core (TM) i5-6200U CPU @ 2.30GHz
- CPU speed: 2401 MHz
- RAM: 8065 MB
- Operating system: Windows 10

The k–ε turbulence model was used; the computational mesh was generated automatically, A grid consisting of more than 700,000 cells was employed, with the solver performing 1,500 iterations.

In agreement with the client, the parameters of the computer system used for performing CFD simulations were determined at the conceptual design stage. According to the current state of the art, similar simulation conditions are commonly applied in the study of nozzle devices for aerospace and rocket applications. However, based on expert recommendations, more precise and higher-performance computational equipment and technologies will be required to conduct CFD analyses during the transition to optimization tasks aimed at the development of specific products and production series.

Figures 5 through 9 offer a detailed graphical representation of specific results obtained from the computational modeling process, highlighting significant trends and observations.

In Case No. 1, the following data were obtained: mass flow rate of gas – 0.296 kg/s; thrust – 396.53 N. Within

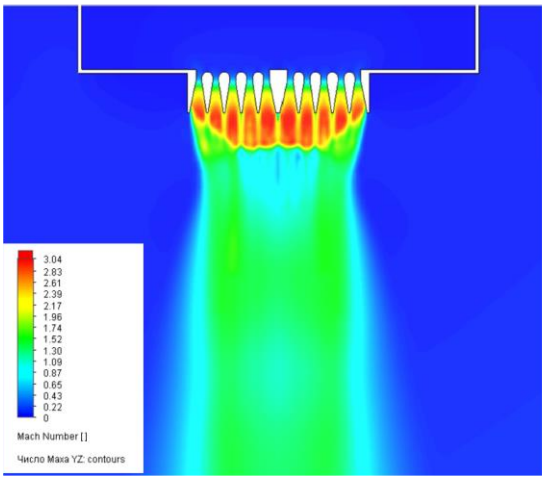


Figure 5. Results of computer modeling, Case No. 1

the framework of digital twin development, additional rapid calculations were also performed using the methodology described in (43). The results were as follows: mass flow rate of gas – 0.302 kg/s; thrust – 472.90 N.

In Case No. 2, the following data were obtained: mass flow rate of gas – 0.578 kg/s; thrust – 950.42 N. Using the methodology described in (43), the corresponding results were: mass flow rate of gas – 0.603 kg/s; thrust – 1,015.95 N.

In Case No. 3, the following data were obtained: mass flow rate of gas – 1.173 kg/s; thrust – 2,144.04 N. Using the methodology described in (43), the corresponding results were: mass flow rate of gas – 1.206 kg/s; thrust – 2,140.15 N.

In Case No. 4, the following data were obtained at a working gas temperature of 1000 ºC: gas mass flow rate: 0.00905 kg/s; thrust: 11.20 N. Specific thrust (defined as the ratio of thrust to gas mass flow rate): 1237.7 N/(kg/s).

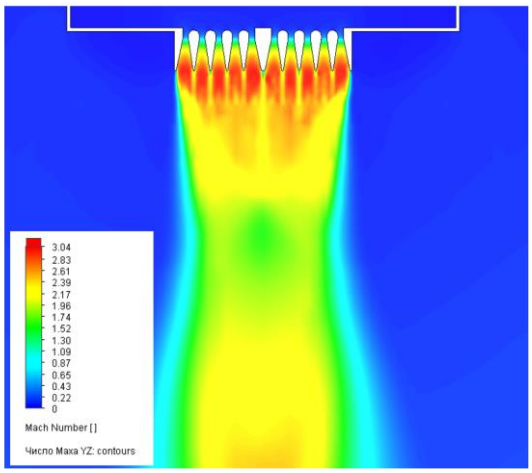


Figure 6. Results of computer modeling, Case No. 2

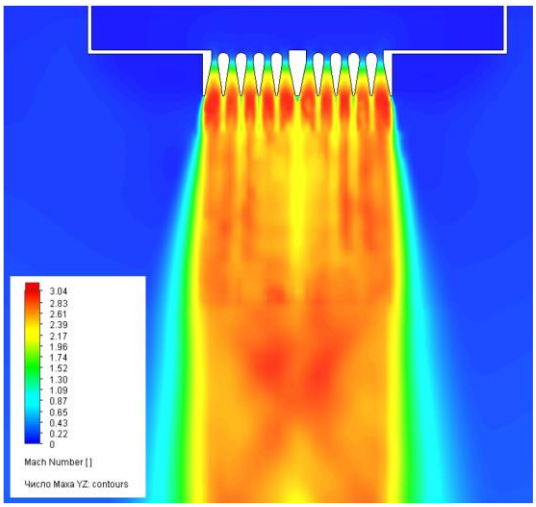


Figure 7. Results of computer modeling, Case No. 3

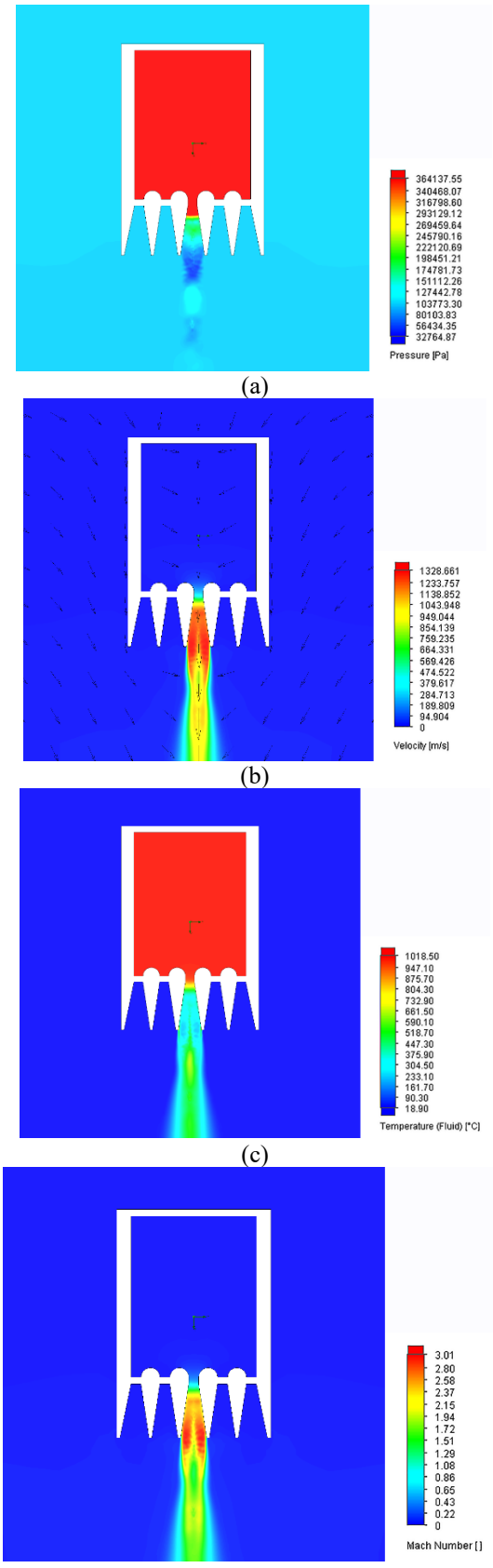
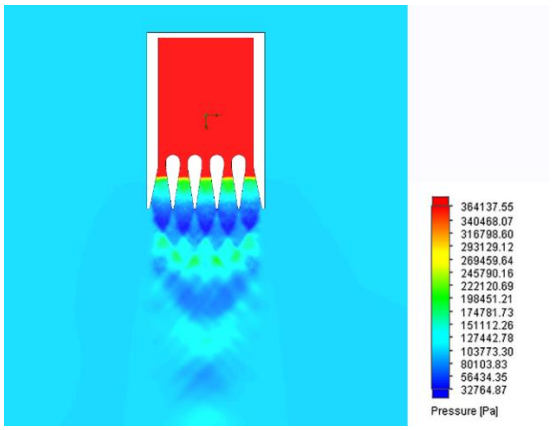


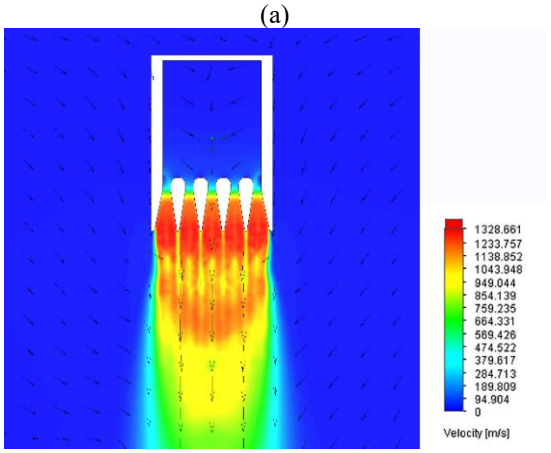
Figure 8. Results of computer simulation, Case No. 4

As part of the development of digital twins for rapid computational analysis, additional calculations were carried out using the methodology described by Sokolov and Singer (43). The corresponding results were as follows: gas mass flow rate: 0.01007 kg/s; thrust: 12.70 N. Specific thrust (the ratio of thrust to gas mass flow rate): 1260.8 N/(kg/s).

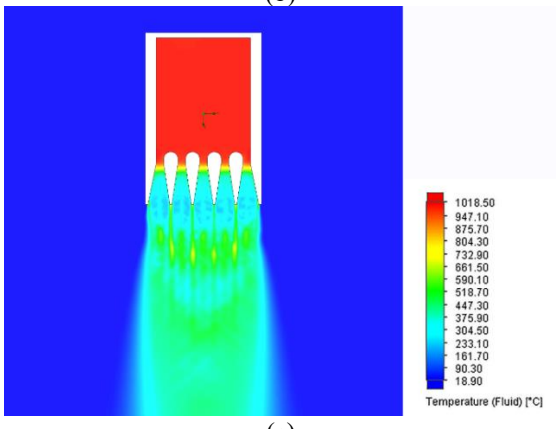
In Case No. 5, the following data were obtained at a working gas temperature of 1000 ºC: gas mass flow rate: 0.04807 kg/s; thrust: 60.125 N. Specific thrust (defined as the ratio of thrust to gas mass flow rate): 1250.7 N/(kg/s).



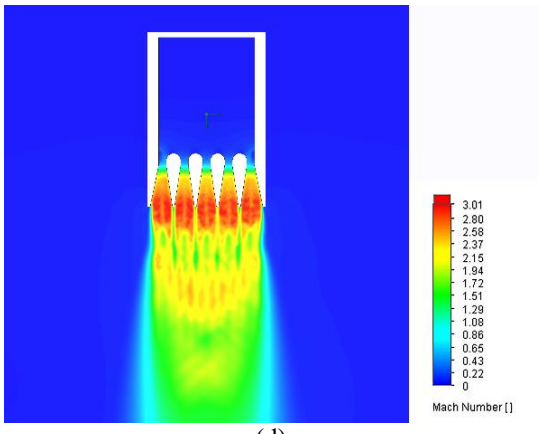
(a)



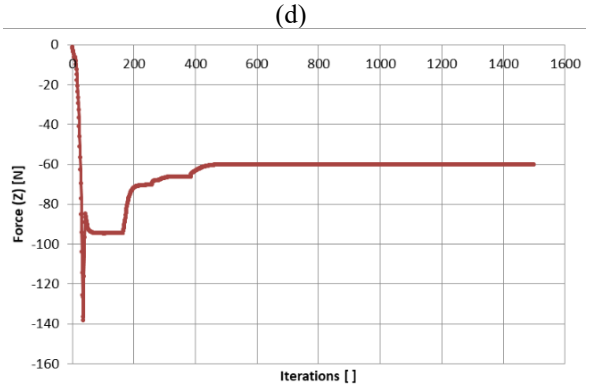
(b)



(c)



(d)



(e)

Figure 9. Results of computer simulation, Case No. 5: assessment of convergence in nozzle thrust calculation

As part of the development of digital twins for rapid computational analysis, additional calculations were carried out using the methodology described by Sokolov and Singer (43). The corresponding results were as follows: gas mass flow rate: 0.0504 kg/s; thrust: 63.50 N. Specific thrust (the ratio of thrust to gas mass flow rate): 1260.8 N/(kg/s).

Similar calculations were carried out for the operating conditions of the nozzle apparatus in a cold-gas environment and in liquid. In all cases, the operability of the new multi-stream nozzle apparatus was confirmed. The results of the research will be used in creating advanced energy-efficient jet devices and systems.

4. DISCUSSION

Within the framework of conceptual design, digital twins are required that provide capabilities for rapid calculations and for verifying the operability of new technical solutions. In regions of high velocities and supersonic flows, such tasks are particularly challenging. Modern CFD technologies provide unique opportunities to investigate complex gas-dynamic and hydrodynamic processes within the channels of multi-stream ejectors and nozzle apparatuses.

The jet system shown in Figure 2, under conditions where $\omega = 0$, allows the nozzle 1 to be oriented in a specified direction by means of angular displacement. Once positioned, the nozzle 1 can be fixed in place to provide the required thrust vector, enabling control over its direction. Thrust vector control by absolute value can be achieved by regulating the mass flow rate of the working medium at the nozzle 1 outlet. This simple configuration, originally presented in Euler's work (45), can be – and already is – applied in the development of more advanced and sophisticated systems for thrust vector control with respect to absolute value, direction, and the spatial coordinates of the point of force application.

For Cases No. 4 and No. 5, the gas temperature used in the calculations was 1000 ºC, which corresponds to the current state of the art in material selection. For Cases No. 1, No. 2, and No. 3, the gas temperature was set to 2000 ºC, representing near-term technological prospects achievable through the use of advanced materials.

Euler's mathematical model, originally developed for the Segner turbine, together with formulas 11-15, can be applied to describe jet systems designed for thrust vector control (1-3). In the particular case of a static propeller, the Wellner–Zhukovsky equation (46) can be directly derived from Euler's theory. It is therefore appropriate to extend the analysis of Euler's seminal work (45) with the goal of developing new computational algorithms and digital twin models for the design of jet propulsion systems – both for applied engineering applications and for educational purposes in the training of design engineers.

Summarizing the intermediate results, the following conclusions can be drawn:

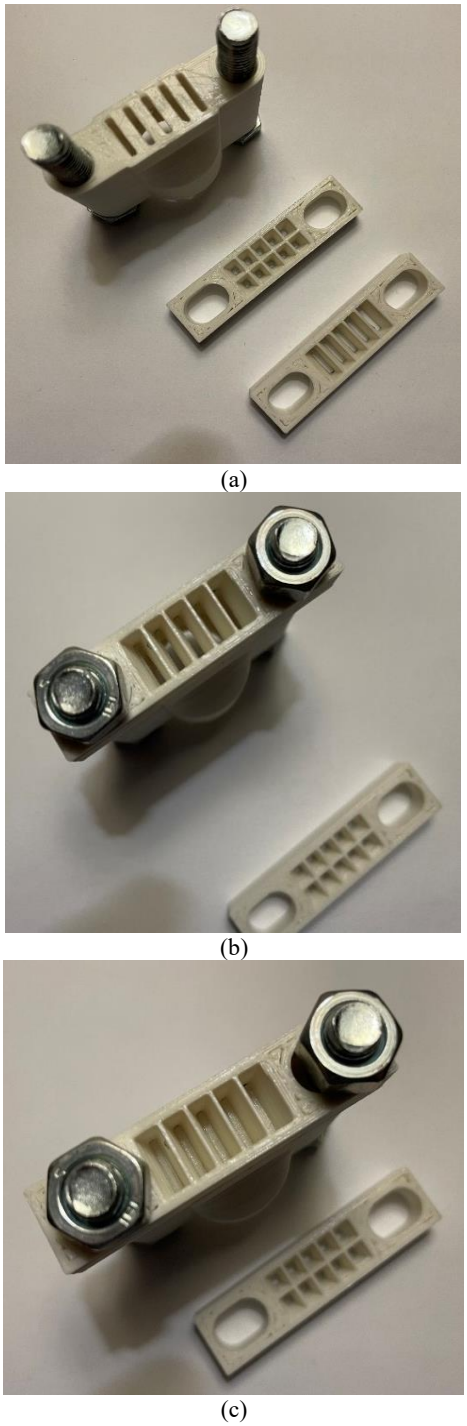
One of the patented variants of the jet system (3), designed for energy distribution with the capability of thrust vector control within the limits of a complete geometric sphere, has been investigated.

Numerical experiments confirmed the operability of the multi-stream lattice-type nozzle apparatus, also oriented toward the development of compact, high-speed, and ergonomic control systems for a wide range of applied tasks.

Based on the results of computer modeling, a preliminary conclusion can be made regarding the suitability of such lattice-type jet systems for operation under conditions of liquid pumping and gas–liquid mixture handling, including applications with hybrid blade machines.

The materials presented in this article constitute a logical continuation of a series of fifteen publications in international scientific journals prepared by the authors between 2020 and 2025 (including journals ranked in Q1-2). The scientific novelty of the developed technological solutions is confirmed by a number of Russian patents for inventions: Nos. 2847614, 2847612, 2839870, 2819487, 2813562, 2802351, 2781534, 2781455, 2778961, 2750833, 2726511, 2714989.

Figure 10 shows photographs of the fabricated micromodel of the multi-stream nozzle.



(a) (b) (c)

Figure 10. Micromodel of the multistream nozzle: a) Nozzle body and two movable blocks with diffusers (variants); b) Configuration with open nozzle flow channels; c) Configuration with closed nozzle flow channels

The micromodel shown in Figure 10 was fabricated according to the schematic presented in Figure 3. The design allows the diffuser block to shift in a plane perpendicular to the working medium flow passing through the multi-stream nozzle. Two operational positions are illustrated in Figure 10: 1 – nozzle open, 2 – nozzle closed. Intermediate positions of the diffuser block are also possible, resulting in a partially open critical section of the nozzle. Additional nozzle adjustment can be achieved via a movable perforated diaphragm placed between the converging section and the diffuser section. In this configuration, the diffuser block can be rigidly attached to the nozzle body, representing one of many possible arrangements. One such configuration is illustrated in Figures 4 and 8 for Case No. 4, where the central channel remains open while the four peripheral channels are closed using diaphragms at the critical cross-sections.

A lightweight movable perforated diaphragm can serve as a shut-off regulating device capable of simultaneously closing the critical sections of all channels in the multi-stream nozzle (Figure 4b). In the cases corresponding to Figures 3 and 4b, a diaphragm displacement of 1.5 mm is sufficient for complete nozzle closure. By comparison, a conventional nozzle of equivalent design speed would require a diaphragm displacement of approximately 15 mm at the critical section, assuming equal diaphragm movement speed.

Calculations confirm that the time required to switch the nozzle operating mode can be significantly reduced through the use of a multi-stream nozzle, where a single large nozzle is replaced by a group of smaller, geometrically similar nozzles. Examples demonstrate that, using this technology, actuation time can be reduced by nearly a factor of ten.

Further research into multi-stream nozzle apparatuses is warranted. These nozzles are more compact and can be adapted to complex fuselage contours of aircraft. Linear channel arrangements (Cases No. 4 and No. 5) are preferable, as they maintain specific thrust comparable to that of a single conventional nozzle. Grouping multiple channels into a single block (Case No. 3) is estimated to reduce specific thrust by approximately 15% relative to a single conventional nozzle.

A patented modification of a multi-flow nozzle assembly (3) has been developed and tested, allowing for energy distribution during thrust vector control within a full geometric sphere. This aspect was partially addressed in works (1, 2). Calculations confirm the possibility of reducing the time required to switch the nozzle operating mode, which is achieved by using a multi-stream nozzle, where a single large nozzle is replaced by a group of smaller, geometrically similar nozzles. Examples demonstrate that, using this technology, the actuation time can be reduced by nearly a factor of ten.

New methodological approaches for the design of complex jet systems with the ability to work with supersonic flows are proposed as a continuation of previously published works (1, 2).

New possibilities for the control of extreme gas flows are proposed. The findings of this study have potential applications across energy sectors, oil and gas extraction, and multiple domains within robotics. Additionally, an extension of Euler's methodological framework is proposed, aimed at addressing practical challenges in higher education and enhancing the training process for contemporary design engineers.

Computer modeling was conducted, yielding the following data:

1. In example No. 1: gas mass flow rate is 0.296 kg/s; thrust is 396.53 N. As part of the creation of digital twins, for fast calculations, calculations were also performed using the method (43), respectively, the gas mass flow rate is 0.302 kg/s; thrust is 472.90 N.
2. In example No. 2: gas mass flow rate 0.578 kg/s; thrust 950.42 N. Using the method (43), the following values were obtained: gas mass flow rate 0.603 kg/s; thrust 1015.95 N.
3. In example No. 3: gas mass flow rate 1.173 kg/s; thrust 2144.04 N. Using the method (43), the following values were obtained: gas mass flow rate 1.206 kg/s; thrust 2140.15 N.

What makes this device suitable for effective use in pumping and compressor units within the oil and gas industry.

5. CONCLUSIONS

The application of artificial intelligence to solving scientific and inventive problems with multiple parameters remains a major and unresolved challenge. A clear interpretation of the results of artificial intelligence work is still lacking. Scientific groundwork has been laid for addressing extreme maneuvering problems with thrust vector control within a full geometric sphere. When summarizing the intermediate results (for the period 2020–2025), it should be noted that with the transition from the conceptual design stage to the preliminary design stage, including prototype development, a significant increase in funding will be required to carry out more precise and higher-cost computational and design work, including CFD simulations and additive manufacturing technologies.

Acknowledgements

The experimental investigations reported in this study were conducted at the laboratories of Empress Catherine II Saint Petersburg Mining University. The authors gratefully acknowledge Mining University for providing

the facilities, materials, chemical reagents, and equipment required to carry out this research.

Funding

This research received no external funding.

Ethics Approval and Consent to Participate

This study did not involve human participants or animals. Therefore, ethical approval and informed consent were not required.

Competing Interests

The authors declare that they have no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Data Availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

The authors used artificial intelligence–based tools to assist with the accurate scientific translation of selected phrases from Russian into English. All translated content was carefully reviewed, edited, and validated by the authors, who take full responsibility for the integrity, originality, and accuracy of the final manuscript.

REFERENCES

1. Sazonov YA, Mokhov MA, Bondarenko AV, Voronova VV, Tumanyan KA, Konyushkov EI. Investigation of a multiflow ejector equipped with variable-length links for thrust vector control using Euler's methodology. Engineering. 2024;5(4):2999-3022. https://doi.org/10.3390/eng5040156
2. Sazonov YA, Mokhov MA, Bondarenko AV, Voronova VV, Tumanyan KA, Konyushkov EI. Study of reversible nozzle apparatuses using Euler methodology and CFD technologies. Civil Engineering Journal. 2024;10(11):3640-71. https://doi.org/10.28991/CEJ-2024-010-11-013
3. Sazonov YA, Mokhov MA, Bondarenko AV, Voronova VV, Tumanyan KA, Konyushkov EI. Jet apparatus. Russian Federation Patent 2839870 C1. 2025.
4. Li X, Dunkin F, Dezert J. Multi-source information fusion: Progress and future. Chinese Journal of Aeronautics. 2024;37(7):24-58. https://doi.org/10.1016/j.cja.2023.12.009
5. Schafer Y, Stossel M, Barnique A, Kozulovic D. Effects of tip injection on a turbofan engine with non-invasive high-speed actuators. International Journal of Turbomachinery, Propulsion and Power. 2025;10(2):9. https://doi.org/10.3390/ijtpp10020009
6. Luo Y, Wu Z, Li Z, Cheng L, Qi P, Mou J. Design of a novel pump cavitation valve and study of its cavitation characteristics. Water. 2025;17:1503. https://doi.org/10.3390/w17101503
7. Ulger Danaci O. The effect of renewable and non-renewable energy on economic growth: A panel cointegration analysis for the top renewable energy consumers (1970-2023). Energies. 2025;18:4745. https://doi.org/10.3390/en18174745
8. Hesami A, Kazempoor P, Acacio VJ, Van Dam JD. Systems and methods for gas pulse jet pump. United States Patent 10837463. 2020.
9. Teneva EV. Emotionalization of the 2021-2022 global energy crisis coverage: Analyzing the rhetorical appeals as manipulation means in the mainstream media. Journalism and Mass Media. 2025;6(1):14. https://doi.org/10.3390/journalism6010014
10. Legrand B, Gaillard A, Bouquain D. Comparative study of hybrid electric distributed propulsion aircraft through multiple powertrain component modeling approaches. Aerospace. 2025;12:732. https://doi.org/10.3390/aerospace12080732
11. Verbitsky VS, Dengaev AV, Safiullina EU, et al. Analyzing efficiency of joint operation of gas separator and screw pump on high-viscosity gas-liquid mixture. International Journal of Engineering, Transactions A: Basics. 2024;37(10):1901-13. https://doi.org/10.5829/ije.2024.37.10a.04
12. Sazonov YA, Mokhov MA, Gryaznova IV, et al. Development of a scientific approach to the advancement of multi-stream jet systems based on Euler's methodology. Neftyanoye Khozyaystvo. 2025;(9):30-35. https://doi.org/10.24887/0028-2448-2025-9-30-35
13. Rubison WC. Reversible flow centrifugal pump. United States Patent 2442512. 1948.
14. Hou Y, Xue R. Hydraulic performance and flow characteristics of a high-speed centrifugal pump based on multi-objective optimization. Fluids. 2025;10:174. https://doi.org/10.3390/fluids10070174
15. Shim HS. Influence of impeller-diffuser side-gap flow with a simplified leakage model on the performance and internal flow of a centrifugal pump. Energies. 2025;18:1278. https://doi.org/10.3390/en18051278
16. Koteleva N, Buslaev G, Valnev V, Kunshin A. Augmented reality system and maintenance of oil pumps. International Journal of Engineering, Transactions B: Applications. 2020;33(8):1620-28. https://doi.org/10.5829/ije.2020.33.08b.20
17. Zhang L, Li Y, Ma M, Kong L, Huang Z, Xu L, Wang B. Numerical investigation of clocking effects on the hydraulic performance of pump-turbine in pump mode. Energies. 2025;18:4317. https://doi.org/10.3390/en18164317
18. Tsentis SE, Gkoutzamanis VG, Gaitanis AD, Kalfas AI. Multi-platform app-embedded model for hybrid air-breathing rocket-cycle engine in hypersonic atmospheric ascent. Aeronautical Journal. 2021;125(1291):1631-53. https://doi.org/10.1017/aer.2021.3
19. Cardenas IR, Lain S, Lopez OD. A review of aerospike nozzles: Current trends in aerospace applications. Aerospace. 2025;12:519. https://doi.org/10.3390/aerospace12060519
20. Gonzalez-Viana A, Sastre F, Martin E, Velazquez A. Multi-objective optimization of rocket-type pulse detonation engine nozzles. Aerospace. 2025;12:502. https://doi.org/10.3390/aerospace12060502
21. Shumeiko A, Telekh V, Ryzhkov S. Thrust-vectoring schemes for electric propulsion systems: A review. Chinese Journal of Aeronautics. 2025;38(6):103401. https://doi.org/10.1016/j.cja.2025.103401
22. Cheng Z, Yang W, Zeng L, Wu L. Pressure characteristics analysis of the deflector jet pilot stage under dynamic skewed velocity distribution. Aerospace. 2025;12:638. https://doi.org/10.3390/aerospace12070638
23. Xu Z, Liu B, Tong Y, Dong Z, Feng Y. Modeling and control of ejector-based hydrogen circulation system for proton exchange membrane fuel cell systems. Energies. 2024;17:2460. https://doi.org/10.3390/en17112460

24. Li J, Wu T, Cheng C, Li J, Zhou K. A review of the research progress and application of key components in the hydrogen fuel cell system. Processes. 2024;12:249. https://doi.org/10.3390/pr12020249
25. Li C, Sun B, Bao L. Coupling global parameters and local flow optimization of a pulsed ejector for proton exchange membrane fuel cells. Sustainability. 2024;16:4170. https://doi.org/10.3390/su16104170
26. Brunner DA, Marcks S, Bajpai M, Prasad AK, Advani SG. Design and characterization of an electronically controlled variable flow rate ejector for fuel cell applications. International Journal of Hydrogen Energy. 2012;37(5):4457-66. https://doi.org/10.1016/j.ijhydene.2011.11.116
27. Lysak IA, Lysak GV, Konyukhov VY, Stupina AA, Gozbenko VE, Yamshchikov AS. Efficiency optimization of an annular-nozzle air ejector under the influence of structural and operating parameters. Mathematics. 2023;11:3039. https://doi.org/10.3390/math11143039
28. Gu R, Sun M, Cai Z, Chen J, Li P, Dong Z, Huang Y. Experimental study on the rocket-ejector system with a throat in the secondary stream. Aerospace Science and Technology. 2021;113:106697. https://doi.org/10.1016/j.ast.2021.106697
29. Riaz S, Aaltonen J, Pinkse T, Koskinen K. Numerical investigation and validation of multiphase flow in annular jet pump: A mixture model approach. Engineering Science and Technology, an International Journal. 2025;69:102100. https://doi.org/10.1016/j.jestch.2025.102100
30. Cican G, Frigioescu TF, Crunteanu DE, Cristea L. Micro turbojet engine nozzle ejector impact on acoustic emission, thrust force and fuel consumption analysis. Aerospace. 2023;10:162. https://doi.org/10.3390/aerospace10020162
31. Kim SW. Traction booster. Japanese Patent JP2024516044. 2024.
32. Wlezien R. Control of revolutionary aircraft with novel effectors (CRANE). Defense Advanced Research Projects Agency. 2023.
33. Smith C. Aurora begins building full-scale active flow control X-plane. Aurora Flight Sciences. 2024.
34. Xu B, Feng L. Experimental research on three-axis control of flying-wing aircraft based on active flow control. Chinese Journal of Aeronautics. 2025;38(8):103443. https://doi.org/10.1016/j.cja.2025.103443
35. Suction jet pump. United States Patent 10072674. 2018.
36. Bolobov V, Martynenko YV, Voronov V, Latipov I, Popov G. Improvement of the liquefied natural gas vapor utilization system using a gas ejector. Inventions. 2022;7(1):14. https://doi.org/10.3390/inventions7010014
37. Korshak AA, Pshenin VV. Determination of parameters of non-pump ejector gasoline vapor recovery unit. Science and Technologies: Oil and Oil Products Pipeline Transportation. 2023;13(1):25-31. https://doi.org/10.28999/2541-9595-2023-13-1-25-31
38. Wang C, Lu H, Kong X, Wang S, Ren D, Huang T. Effects of pulsed jet intensities on the performance of the S-duct. Aerospace. 2023;10:184. https://doi.org/10.3390/aerospace10020184
39. Ahmed F, Eames I, Moeendarbary E, Azarbadegan A. High-Strouhal-number pulsatile flow in a curved pipe. Journal of Fluid Mechanics. 2021;923:A15. https://doi.org/10.1017/jfm.2021.553
40. Brethouwer G. Turbulent flow in curved channels. Journal of Fluid Mechanics. 2022;931:A21. https://doi.org/10.1017/jfm.2021.953
41. Zhang X, Shan Y, Zhang J. Influence of vertical/spanwise offsets on aerodynamic performance of double serpentine nozzles. Aerospace. 2025;12:171. https://doi.org/10.3390/aerospace12030171
42. Liu J, Wang D, Huang X. Design methodology for fishtailed pipe diffusers and its application to a high-pressure ratio centrifugal compressor. Aerospace. 2025;12:746. https://doi.org/10.3390/aerospace12080746
43. Sokolov EYa, Singer NM. Jet devices. Moscow: Energoatomizdat; 1989.
44. Feyerabend PK. Against method: Outline of an anarchistic theory of knowledge. London; 1975.
45. Bistafa SR. Investigation of a water turbine built according to Euler's proposals (1754). arXiv. 2021. https://arxiv.org/abs/2108.12048
46. Redkin AV, Yaloza YA, Kovalev IE. Reliability assessment of a convertible aircraft with a hybrid powerplant and a multi-rotor lifting system. Scientific Bulletin of MSTU GA. 2020;23(5):76-96. https://doi.org/10.26467/2079-0619-2020-23-5-76-96

Persian Abstract

چکیده

هدف این مقاله انجام تحقیقات میان‌رشته‌ای در یک حوزه علمی نوظهور است که بر کنترل بردار رانش بر روی یک هندسه کروی کامل متمرکز است، که در آن بردار رانش می‌تواند به طور مداوم در یک محدوده زاویه‌ای کامل ±۱۸۰ درجه در تمام جهات فضایی منحرف شود. یک نوع ثبت اختراع شده از دستگاه نازل چند جریانی مورد تجزیه و تحلیل قرار می‌گیرد. توزیع انرژی گاز عامل برای مواردی با فشارهای کاری مختلف تجزیه و تحلیل می‌شود. یک مقطع بحرانی از هندسه مستطیلی بررسی شده است. این مقاله ۳ نتیجه شبیه‌سازی را با طیف وسیعی از مقادیر برای نرخ جریان جرمی گاز از ۰.۲۹۶ کیلوگرم بر ثانیه تا ۱.۱۷۳ کیلوگرم بر ثانیه و برای رانش از ۳۹۶.۵۳ نیوتن تا ۲۱۴۰.۱۵ نیوتن ارائه می‌دهد. چشم‌اندازهای کاربرد عملی نتایج به‌دست‌آمده، از جمله ایجاد دوقلوهای دیجیتال برای استفاده در فرآیند آموزشی در آموزش مهندسان، مورد بحث قرار گرفته است. برای طراحی مفهومی، روش اویلر در ترکیب با فناوری‌های CFD پیشنهاد شده است که با رویکردهای معاصر برای مدیریت مجموعه داده‌های بزرگ سازگار است. یک پایه علمی برای توسعه سیستم‌های جت چندجریانی از نوع شبکه‌ای بر اساس روش اویلر ایجاد شده است. سه جهت اصلی برای تحقیقات علمی بیشتر شناسایی شده است: سیستم‌های قدرت با بهره‌وری انرژی؛ توسعه مؤثر میدان‌های نفت و گاز؛ و سیستم‌های رباتیک حمل و نقل برای عملیات در زمین، دریا و هوا.

International Journal of Engineering

Journal Homepage: www.ije.ir

An Ensemble Deep Learning Approach for Automated Bone Fracture Detection in Medical Imaging

M. Mahtabi[a], S. Asadi Amiri*[a], S. Mavaddati[b]

[a] *Department of Computer Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*
[b] *Department of Electronic Engineering, Faculty of Engineering and Technology, University of Mazandaran, Babolsar, Iran*

PAPER INFO

Paper history:
Received 02 October 2025
Received in revised form 02 November 2025
Accepted 31 Januray 2026

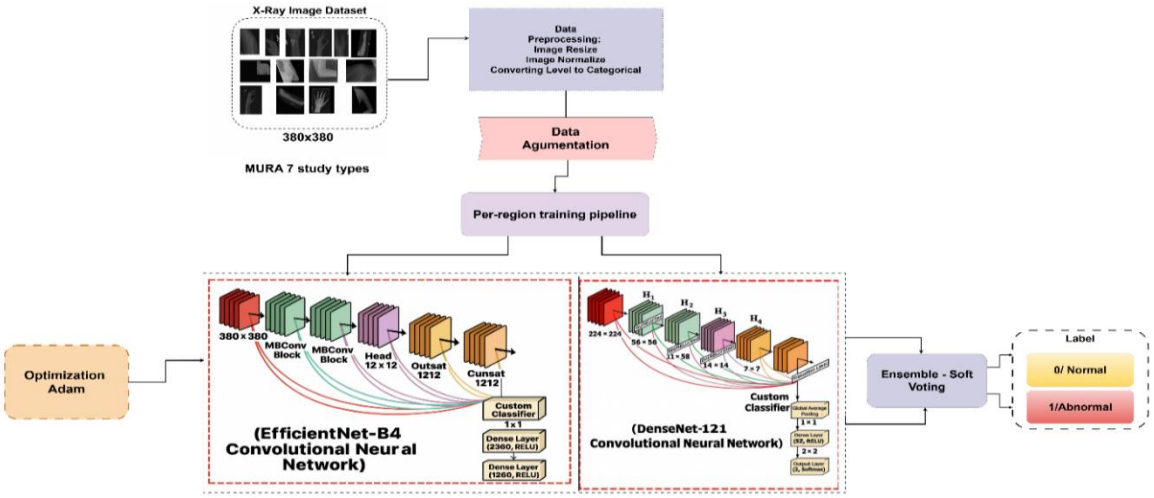
Keywords:
Deep Learning
Bone Abnormality Detection
Medical Image Analysis
Radiographic Images
EfficientNet-B4
DenseNet-121
Ensemble Learning

ABSTRACT

Recent advances in deep neural networks have significantly improved medical image analysis; however, detecting bone fractures in radiographic images remains challenging due to the complex skeletal structure, subtle fracture patterns, and limited annotated data. This study proposes an advanced ensemble deep learning framework that integrates two optimized architectures, EfficientNet-B4 and DenseNet-121, through a soft-voting fusion strategy to enable automated and precise bone fracture detection. The hybrid framework introduces adaptive weighting and optimized dense layers, which enhance feature discrimination and strengthen the network's capacity to distinguish fine-grained fracture details. Moreover, transfer learning and fine-tuning techniques are employed to address data imbalance and improve model generalization across multiple anatomical regions. Comprehensive experiments conducted on the MURA dataset, consisting of radiographs from seven distinct anatomical regions, demonstrate that the proposed model achieves superior performance with 83.52% accuracy and 90.76% sensitivity, outperforming each individual baseline. The model's robustness under different training configurations confirms its reliability and stability for clinical deployment. Overall, this research presents a novel ensemble-based diagnostic system that leverages complementary architectural strengths and adaptive feature fusion to achieve high diagnostic precision. The proposed method contributes not only to improving classification accuracy but also to establishing a scalable and interpretable framework for computer-aided fracture diagnosis, offering a practical step toward intelligent and reliable radiological decision support.

***doi**: 10.5829/ije.2026.39.10a.19*

Graphical Abstract



*Corresponding Author Email: s.asadi@umz.ac.ir (S. Asadi Amiri)

Please cite this article as: Mahtabi M, Asadi Amiri S, Mavaddati S. An Ensemble Deep Learning Approach for Automated Bone Fracture Detection in Medical Imaging. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2583-601.

1. INTRODUCTION

Bone fractures are among the most common orthopedic injuries, affecting millions of people worldwide each year. Increasing life expectancy and changing lifestyle patterns in industrialized societies have led to a considerable rise in the incidence of these injuries. Fractures typically occur when forces exceed the load-bearing capacity of bone and may result in from causes such as traffic accidents, falls, sports injuries, and underlying conditions like osteoporosis (1). Timely diagnosis is vital because delays in treatment can lead to serious consequences such as infection, permanent bone deformity, and loss of function (1).

Medical imaging modalities play a key role in the identification and assessment of fractures. X-ray radiography, computed tomography (CT), and magnetic resonance imaging (MRI) are among the primary options, chosen based on the type and location of the injury (1-3). Due to its speed, accuracy, and ability to reveal fine structural details, X-ray imaging is widely used to diagnose osseous problems and certain soft-tissue abnormalities. Figure 1 shows examples of radiographic images that illustrate bone abnormalities and fractures. Beyond its role in initial diagnosis, this modality is an important tool for the treatment planning and monitoring of recovery (2). Although X-rays are relatively safe, adherence to radiation safety principles remains necessary to avoid unnecessary exposure.

Musculoskeletal disorders, as a prevalent group of conditions, have a substantial impact on quality of life (4). These disorders, including compression fractures, carpal tunnel syndrome, and ligament tears, often arise from occupational incidents, repetitive movements, sports injuries, and poor ergonomic conditions. Occupations such as banking, software development, or hairdressing, which involve prolonged non-neutral postures, are associated with increased prevalence, particularly in the neck and lower back. Research indicates that both physical and psychosocial workplace factors play important roles in the emergence of these disorders (3, 4). Early diagnosis is especially important, since timely interventions such as physiotherapy and shock-wave therapy can effectively relieve mild-to-moderate pain and prevent disease progression. Nevertheless, accurate and timely diagnosis is often challenging and demands precise, sensitive diagnostic approaches.

To date, the detection of musculoskeletal abnormalities has largely been performed manually by radiologists. However, the complexity of shapes and the variability in lesion size, especially when patterns remain hidden in images, make the diagnostic process difficult and error-prone (5). In addition, the large volume of imaging data, together with factors such as fatigue and human error, can negatively affect accuracy and increase the likelihood of delays or mistakes in diagnosis (6, 7). Consequently, conventional approaches face serious limitations, and the need for modern, automated solutions is increasingly felt. In this context, advances in image processing and computer vision have provided powerful tools for automatic abnormality detection, which, by improving diagnostic speed and accuracy, can enhance clinical outcomes and reduce healthcare costs (8, 9).

In recent years, deep neural networks have been widely applied to bone-fracture detection. By learning complex patterns, these models can automatically highlight suspicious regions and, while reducing human error, improve both the accuracy and efficiency of diagnosis. They can also detect subtle fractures and provide quantitative analyses related to severity and displacement. Despite challenges such as the need for high-quality data and the risk of false results, for example, false positives, the application of deep learning in this area holds great potential for improving the quality of medical services and accelerating diagnosis.

In this paper, we present an innovative approach for more accurate detection of musculoskeletal abnormalities in radiographic images by combining two advanced models, EfficientNet-B4 and DenseNet-121, within an ensemble framework based on soft voting. Manual detection of these abnormalities is consistently challenging due to structural complexity, variability in lesion morphology, and fluctuating image quality, and thus requires approaches with higher accuracy and generalizability. Accordingly, our method leverages the strengths of deep models and intelligently fuses their outputs via soft voting to yield decisions that are more accurate, more stable, and more robust to noise and imaging variability.

To comprehensively evaluate performance, the models were trained separately on the seven anatomical regions of the MURA dataset. To address data imbalance, between 7,000 and 10,000 augmented images were generated per region, depending on its size. The outputs of the individual models for each region were then combined using soft voting, and the final accuracy for that region was obtained. In the final step, aggregating the results across all seven regions yielded an overall evaluation of the ensemble's performance, showing that our approach outperforms previous methods in terms of accuracy. These results corroborate the high efficacy of our hybrid model for automatic skeletal-lesion detection and its suitability for use in intelligent computer-aided diagnostic systems (10, 11). The main findings of this study are as follows:

- Improved performance over individual models: The proposed ensemble of EfficientNet-B4 and DenseNet-121 with soft voting outperformed either model used in isolation in detecting bone abnormalities.
- Notable accuracy: The model achieved an accuracy of 83.52%, demonstrating the strong capability of the

proposed framework in correctly classifying radiographs.

- High sensitivity: The sensitivity was 90.76%, indicating strong ability to detect positive cases (presence of abnormality) with a low miss rate.
- Enhanced high-level feature extraction: Adding customized dense layers at the end of each network enabled extraction of higher-level, more discriminative features, increasing classification power.
- Strong performance on MURA: Evaluation on the large MURA dataset, radiographs from seven anatomical regions, showed that the proposed model effectively detected skeletal abnormalities across regions.
- Clinical potential: The findings suggest that this framework can serve as a clinical decision-support tool, helping radiologists improve diagnostic accuracy and speed.

The novelty of this work lies in designing a hybrid ensemble framework that combines EfficientNet-B4 and DenseNet-121 through a soft-voting strategy to leverage their complementary feature extraction capabilities. In addition, customized dense layers were incorporated at the end of each network to improve high-level feature discrimination. This design provides a more generalizable and accurate model for bone abnormality detection than existing single-network approaches.

In the remainder of this paper, Section 2 reviews the literature and prior work on bone-abnormality detection using machine-learning and deep-learning methods. Section 3 details the proposed method, including the ensemble architecture based on EfficientNet-B4 and DenseNet-121, the customized classification layers, the dataset description, and the data-augmentation strategies. Section 4 presents experimental results and performance metrics and compares the proposed method with individual baselines. Finally, Section 5 offers discussion and conclusions, summarizing the findings and suggesting directions for future research.

2. LITERATURE REVIEW

Image processing and computer vision are among the most important branches of computer science in the analysis of medical images. The application of machine learning algorithms, intensive learning, enables the recognition of complex patterns in radiological images and the automatic detection of abnormalities. Research has shown that these approaches can increase diagnostic accuracy and speed, while supporting radiologists in clinical decision-making. Among modern techniques, convolutional neural networks (CNNs) hold a special place in medical image analysis due to their ability to extract local and complex features from images (12). Different CNN architectures, such as AlexNet, VGG, and ResNet, have been applied to various medical tasks.

Nevertheless, the lack of annotated datasets remains a major challenge in developing these models. To overcome this limitation, techniques such as data augmentation (rotation, cropping, color transformation), transfer learning from pre-trained models, and semi-supervised approaches are widely used. Furthermore, issues such as class imbalance, image noise, and the difficulty of result interpretation remain significant challenges. Despite these obstacles, CNNs have demonstrated remarkable performance in diagnosing diseases such as cancer, cardiovascular conditions, and neurological disorders. Karthik et al. (4) and Krizhevsky et al. (13) by using a deep CNN achieved groundbreaking success in the ImageNet competition, which paved the way for widespread CNN applications across domains, including medicine.

Ebsim et al. (14) introduced an innovative approach was that combined feature fusion with ensemble random forests for wrist fracture detection. By integrating diverse image features, this method significantly improved classification accuracy.

Chada (15) applied transfer learning with DenseNet and InceptionResNetV2 on the MURA dataset (finger and humerus, binary image-level). The best result on the humerus reached 88.2% accuracy, while on the finger, the performance was lower at only 77.7%. This shows the variability of CNN performance across different body parts.

Yahalom. et al. (16) used the Faster R-CNN model to detect distal radius fractures in posteroanterior wrist radiographs. Optimized with ImageNet pre-trained weights, the model achieved performance superior to that of radiologists in identifying fracture regions, despite being trained on a relatively small dataset.

The MSDNet model was introduced by Karthik et al. (4) combines multiple CNNs to enhance the detection and classification accuracy of radiographic abnormalities. In addition to identifying the presence or absence of abnormalities, it was also capable of localizing their exact position in the image and generating automated diagnostic reports with high accuracy.

A deep neural network was developed by Yadav and Rathor (17) to classify healthy and fractured bones. To improve performance and prevent overfitting, data augmentation techniques were employed. The results showed that this model achieved an accuracy above 92%, outperforming earlier approaches and demonstrating high capability in fracture detection.

Several methods for fracture detection and classification were reviewed by Meena and Roy (18), concluding that CNN-based architectures, particularly InceptionNet and XceptionNet, perform strongly in fracture detection. However, the lack of sufficiently large annotated datasets continues to be a barrier to the development of high-performance automated algorithms.

Spahr et al. (19) compared interpretability techniques such as Grad-CAM, attention maps, and Chefer across four medical datasets. The results showed that Chefer provided more accurate explanations, underscoring the importance of selecting appropriate interpretability methods in sensitive medical applications. Dahal et al. (20) proposed a hybrid model combining VGG-16 and Vision Transformer (ViT) for musculoskeletal abnormality detection on the full MURA dataset, achieving 82.88% accuracy and 0.88 sensitivity. The study highlights that integrating local and global feature extractors can improve automated radiographic diagnosis, though performance is still constrained by dataset limitations. Hardalaç et al. (21) used models such as Faster R-CNN, RetinaNet, and ResNet for automatic wrist fracture detection in X-ray images. By combining five models, the ensemble model WFD-C was introduced, which achieved strong performance with AP50 = 0.8639. Such models can serve as supportive tools in clinical diagnosis.

Beyaz et al. (22) combined three models, Xception, EfficientNet, and NfNet, using majority voting for hip fracture detection. Additionally, DenseNet-169, DenseNet-201, and InceptionResNetV2 were tested on the MURA dataset for detecting arm and finger abnormalities. The accuracy for fingers was lower due to disagreement among radiologists. Ghosh et al. (23) developed shallow neural networks and convolutional neural networks within an AdaBoost framework for humerus data, demonstrating that with an appropriate number of epochs and the use of ensemble models, performance close to that of radiologists and the 169-layer deep networks can be achieved, while significantly reducing training time and computational resources. In recent studies Lu et al. (24) have focused on CNN-based and hybrid models for fracture detection, with Inception/Xception showing strong performance despite data limitations. Hybrid CNN–Transformer models improved accuracy on MURA by combining local and global features. Building on this, HAMIL-Net used hierarchical attention and multiple-instance learning, achieving robust study-level results (AUC ≈ 0.87) for foot and ankle radiographs.

The main objective of the present study is to develop a CNN-based model for the automatic detection of musculoskeletal radiographic abnormalities with improved performance. To achieve this, an ensemble framework was designed, combining two advanced architectures, EfficientNet-B4 and DenseNet-121, in a jointly optimized manner to enhance both feature extraction and classification accuracy. Furthermore, the use of a soft-voting mechanism to integrate the results of individual models increases system stability and improves final detection accuracy. Customized dense layers were added to the end of each network to enhance learning capacity and extraction of high-level, discriminative features, thereby boosting model performance in abnormality detection. A comprehensive evaluation of the proposed framework was performed on the MURA dataset, which covers seven different anatomical regions, demonstrating the model's applicability in real-world, multi-purpose medical contexts. Ultimately, this approach not only maintains high accuracy but also offers the capability to be integrated as a clinical decision-support tool to accelerate diagnostic processes in radiology.

In recent years, several studies have explored the application of deep learning models for detecting bone fractures and abnormalities. A summary of the most relevant works is presented in Table 1.

3. DETECTION OF BONE FRACTURES AND ABNORMALITIES

In this section, the detailed procedure of the proposed system for detecting musculoskeletal abnormalities in radiographic images is described. The overall framework is based on ensemble learning, and by combining two powerful deep models, EfficientNet-B4 and DenseNet-121, it seeks to simultaneously leverage their high accuracy and generalization capability.

The overall process consists of the following stages: data preparation, design and customization of the model architectures, independent training of the models, and, finally, fusion of the models' outputs via soft voting to increase the accuracy of the final prediction.

3. 1. Dataset In studies related to detecting musculoskeletal disorders through image processing and computer vision, the use of benchmark datasets is very common and practical (25). The MURA dataset was developed specifically for training and evaluating deep learning models, particularly convolutional neural networks, for detecting musculoskeletal disorders from radiographic images. This dataset is one of the largest radiographic image datasets for musculoskeletal disorder detection. It was prepared by Stanford University in the United States and is widely used in computer vision and deep learning research on medical image analysis. The dataset contains more than 40,000 radiographic images from approximately 14,000 studies. These studies include radiographs from different body regions such as wrist, arm, elbow, shoulder, finger, forearm, and others. All images in this dataset are radiographs collected to investigate musculoskeletal problems and disorders. Each image in the dataset is categorized into one of two classes: "normal" or "abnormal" (presence of an abnormality). The labeling was performed by specialist radiologists, and these labels serve as the main basis for training machine learning models.

The dataset is publicly and freely available to researchers and can be downloaded from reputable platforms such as GitHub or websites related to Stanford University (25). Using this large dataset, together with the accuracy of deep learning models, improves the detection of musculoskeletal abnormalities and enables comparison of results across studies. The dataset includes 9,045 normal images and 5,818 abnormal images labeled by radiologists, and with diverse images from different body parts, poses challenges for automatic detection algorithms. The distribution of studies in the MURA dataset for normal and abnormal classes is summarized in Table 2.

3. 2. Preprocessing and Data Augmentation in the Proposed Method Deep learning models require diverse data for accurate detection of bone fractures. Although the MURA dataset is large, it does not provide sufficient diversity to cover all fracture cases and inter-patient variations. Incorporating additional data can prevent overfitting and improve the generalizability of the model (26-30). This is particularly important due to the inherent imbalance between "Normal" (healthy) and "Abnormal" or fractured samples in medical datasets. In fact, when the number of samples across classes is imbalanced, the model may become biased toward the majority class and show weakness in correctly

TABLE 1. Recent studies on bone fracture and abnormality detection using deep learning models.

Reference	Classes	Dataset	Classifier	Year
(13)	1000 classes (animals, objects, scenes, etc.)	ImageNet (ILSVRC)	AlexNet, ResNet, VGG	2012
(12)	Normal vs. abnormal (2 classes)	MURA V1.1	DenseNet-169	2018
(14)	Normal vs. abnormal (2 classes)	Wrist dataset	RF-CLM	2018
[15]	Normal vs. abnormal (2 classes)	MURA V1.1	DenseNet-169 ،DenseNet-201 ، InceptionResNetV2	2019
(16)	Normal vs. abnormal (2 classes)	Distal Radius Wrist X-ray dataset	MSDNet (Ensemble CNN), Faster R-CNN	2020
(23)	Normal vs. abnormal (2 classes)	MURA V1.1	Shallow NN, CNN + AdaBoost Ensemble	2021
(4)	Normal vs. abnormal (2 classes)	MURA + Indiana	MSDNet (Ensemble CNN)	2022
(24)	Normal vs. abnormal (2 classes)	Foot & Ankle Radiographs	HAMIL-Net (Hierarchical Attention + MIL)	2023
(22)	Normal vs. abnormal (2 classes: hip fracture vs. healthy)	Hip Fracture X-ray dataset (10,849 images, 5 hospitals)	Xception, EfficientNet, NfNet + Majority Voting	2023
(20)	Normal vs. abnormal (2 classes)	MURA V1.1	VGG-16 and Vision Transformer (ViT)	2024
(18)	Normal vs. abnormal (2 classes)	Hip Fracture X-ray (10,849 images), MURA, CP-Child, Duke Breast Cancer, Kvasir GI Disease (Z-line subset)	Vision Transformers	2025
Proposed Method	Normal vs. abnormal (2 classes)	MURA V1.1	EfficientNet-B4, DenseNet-121	–

TABLE 2. Distribution of studies in the MURA dataset with 9045 cases in the normal class and 5818 cases in the abnormal class

Region (Body Part)	Total Images	Validation (Normal)	Validation (Abnormal)	Training (Normal)	Training (Abnormal)
Elbow	1,912	92	66	1094	660
Finger	2,110	92	83	1280	655
Hand	2,185	101	66	1,497	521
Arm (Humerus)	727	68	67	321	271
Forearm	1,010	69	64	590	267
Shoulder	3,015	99	95	1364	1457
Wrist	3,697	140	97	2134	1326
Total	14,656	661	538	8280	5177

Figure 1. Examples of radiographic images from the MURA dataset illustrate the variety of abnormalities in different regions of the upper limb. Each image pair represents a single abnormal case. These images highlight challenges such as fractures and other anomalies, as well as the presence of medical devices that the model needs to handle

TABLE 3. Number of images in the dataset before and after augmentation

Anatomical Region	# Before Augmentation	Target per Class	# After Augmentation
Elbow	4437	4000	8062
Finger	4595	20000	39319
Forearm	1641	5000	10060
Hand	4988	10000	19591
Arm (Humerus)	1144	10000	19382
Shoulder	7540	10000	20030
Wrist	9441	7000	14057
Total	**33786**	–	**130501**

identifying the minority class (which is often the more clinically significant one).

In this study, data augmentation was applied online and simultaneously during training. The augmentation operations included a set of geometric and standard compression transformations, described as follows:

- Random rotation: images were rotated at random angles up to ±180°.
- Horizontal and vertical translation: up to 10% of the image width and height.
- Zooming: up to 10% to simulate scale variations.
- Flipping: both horizontally and vertically to increase orientation diversity.

This procedure was performed independently for each of the seven anatomical regions, increasing the number of training samples per class in each region to approximately 5,000 to 20,000 images. This process not only helped balance the classes but also forced the model to learn invariant features that are robust against these transformations. The number of images in the dataset before and after data augmentation is reported in Table 3. After augmentation, two main preprocessing steps were carried out to match the requirements of pre-trained neural networks:

- Pixel normalization using the standard functions of DenseNet or EfficientNet.
- Image resizing to 380×380 pixels to fit the architecture input.

These preprocessing steps enabled the model to learn features independent of position and shape, providing more stable and accurate performance when dealing with diverse real-world radiographs.

3. 3. Designed Architecture of EfficientNet-B4

The EfficientNet-B4 model is one of the advanced members of the EfficientNet family, a series of convolutional neural network architectures designed using the compound scaling approach. Unlike traditional methods that scale only one dimension of the network, for example, depth, width, or resolution, compound scaling expands all three dimensions simultaneously and in a balanced manner. This strategy increases efficiency and enhances model performance in visual classification tasks. EfficientNet-B4 is pre-trained on the large-scale ImageNet dataset, providing substantial transfer learning capability for visual pattern recognition. Its architecture is composed of hierarchical MBConv blocks, which reduce computational complexity and the number of parameters while maintaining the ability to extract high-level features. In this study, to adapt EfficientNet-B4 to the specific task of bone abnormality detection, the base structure of the model was modified in the classifier head. As shown in Figure 2, after the global average pooling layer, two dense layers with 2,560 and 1,200 nodes were added. The inclusion of these layers allows the extraction of more complex and specialized patterns from high-level features.

This structural adaptation enables the model to more effectively distinguish subtle differences between “Normal” and “Abnormal” images, thereby improving the overall classification performance (26).

3. 4. Designed Architecture of DenseNet-121

DenseNet-121 is one of the prominent models in the family of Dense Convolutional Networks, which revolutionized neural network design by introducing direct connections among all layers within a block. In the original DenseNet architecture, each layer passes its output not only to the subsequent layer but also to all the following layers. This dense connectivity mechanism enhances feature reuse, strengthens gradient flow during training, and significantly reduces the number of parameters compared to conventional architectures such as ResNet. The structure of DenseNet consists of dense blocks and transition layers. Within each dense block, the feature maps of the layers are concatenated along the channel dimension, a method that, unlike the additive operation in ResNet, preserves information and expands feature diversity. Transition layers, through convolution

and down-sampling operations, compress the feature map volume and prevent excessive growth of dimensions. In this study, the DenseNet-121 model, pre-trained on ImageNet, has been purposefully customized for optimizing radiographic image classification. As illustrated in Figure 3, following the global average pooling layer, two fully connected layers with 1024 and 512 nodes were added, serving as intermediaries between the feature extractor and the final output. These layers enhance the model's capability to learn complex nonlinear relationships between image features and classification labels (27).

This modified design is not only aligned with the specific objectives of the research but also preserves the ability of DenseNet's deep and rich structure to extract multi-level features from bone structures. Consequently, the optimized DenseNet-121, as part of the proposed hybrid framework, makes a substantial contribution to improving the accuracy of abnormality detection (27).

As mentioned, Figure 3 presents the detailed architecture of the customized DenseNet-121 model used in the proposed ensemble framework. In this design, the standard DenseNet-121 backbone is enhanced with additional dense layers at the end to improve the extraction of high-level and discriminative features from radiographic images.

The model processes images from various anatomical regions of the upper limb and outputs class probabilities for Normal and Abnormal categories, which are later combined in the ensemble model. This Figure provides a clear visualization of how the DenseNet-121 architecture is adapted for musculoskeletal abnormality detection.

Table 4 illustrates the stage-wise process of image processing in both architectures. Both models start with the same input layer and, through feature extraction blocks, progressively reduce the spatial dimensions of the feature maps while increasing their depth. The main difference lies in the structure and number of blocks in each architecture. The critical stage is the customized classifier, which includes a global average pooling layer and two fully connected layers for extracting high-level features. In EfficientNet-B4, the sizes of these layers are 2560 and 1200, whereas in DenseNet-121, they are 1024 and 512, respectively. Finally, the output layer with Softmax predicts the probability of an image belonging to the "normal" or "abnormal" class, and both models are optimized for bone abnormality detection.

3. 5. Soft Voting in the Proposed Method

Although EfficientNet-B4 and DenseNet-121 are well-established pretrained architectures, in this work, they were individually fine-tuned on the MURA dataset and enhanced with customized dense layers to improve high-level feature extraction. The outputs of the two optimized models were then combined through a soft-voting ensemble strategy to leverage their complementary strengths and achieve higher detection accuracy.

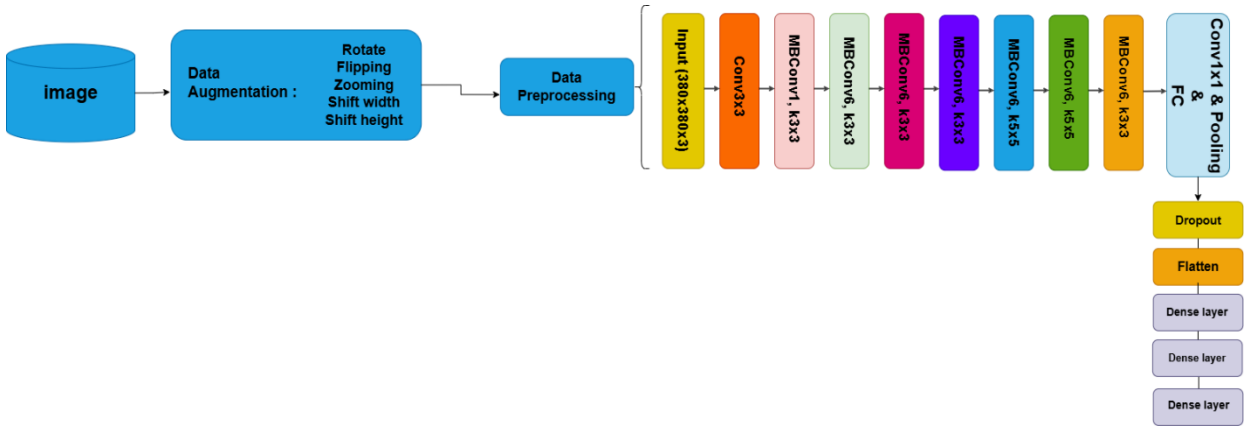


Figure 2. Schematic of the EfficientNet-B4 architecture as designed in the proposed method

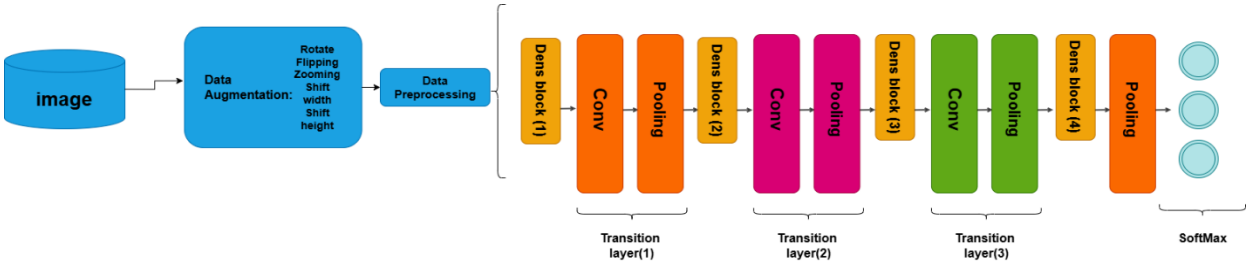


Figure 3. Overview of the architecture of the customized DenseNet-121 model in the proposed method.

TABLE 4. Comparative overview of the designed architectures of EfficientNet-B4 and DenseNet-121. This table details the stage-wise structure of each model, from the input layer to the final classification head, which has been optimized with customized fully connected layers

Stage	Designed EfficientNet-B4 Architecture	Designed DenseNet-121 Architecture
Input	Size: 380×380 Details: Image input	Size: 380×380 Details: Image input
Initial Conv/Pool	Size: 190×190 Details: Stem (Conv, BN, Swish)	Size: 95×95 Details: Conv1, Pool1
Feature Extraction 1	Size: 95×95 Details: 4 × MBConv Block	Size: 47×47 Details: Dense Block (1), Transition Layer (1)
Feature Extraction 2	Size: 48×48 Details: 4 × MBConv Block	Size: 23×23 Details: Dense Block (2), Transition Layer (2)
Feature Extraction 3	Size: 24×24 Details: 6 × MBConv Block	Size: 11×11 Details: Dense Block (3), Transition Layer (3)
Feature Extraction 4	Size: 12×12 Details: 10 × MBConv Block + Head	Size: 11×11 Details: Dense Block (4) + Classification Layer
Custom Classifier	Size: 1×1 Details: Global Average Pooling Dense Layer (2560, ReLU) Dense Layer (1200, ReLU) Output Layer (2, Softmax)	Size: 1×1 Details: Global Average Pooling Dense Layer (1024, ReLU) Dense Layer (512, ReLU) Output Layer (2, Softmax)

Soft voting is one of the common techniques in the field of ensemble learning, where multiple machine learning models, known as base learners, are trained in parallel, and their outputs are combined to improve the overall performance of the system (29).

Unlike individual models that make decisions solely based on their internal structure, ensemble models benefit from the diversity of architectures, perspectives, and capabilities of each model to enhance prediction accuracy (30).

Among the methods for combining model outputs, hard voting and soft voting are the two main approaches. Hard voting operates purely based on the majority vote (the class selected by each model) and assigns equal importance to all models. However, this approach has limitations, including the disregard for the confidence level of the models in their predictions. In contrast, soft voting takes into account the probability distributions produced by each model, providing a more precise and sensitive approach. In this method, the final prediction is determined by averaging the predicted probabilities for each class across the base models, which leads to more informed decision-making. This approach reduces the likelihood of errors when handling inconsistent outputs among models and results in greater stability and higher classification accuracy (31).

In the proposed framework of this study, the EfficientNet-B4 and DenseNet-121 models were independently trained on the dataset, and their outputs were extracted as probability vectors corresponding to the two target classes (normal and abnormal). By performing element-wise averaging of these vectors, the final output of the combined system was generated. This ensemble strategy, based on soft voting, leverages the distinctive features of each model to enhance overall performance while increasing robustness against noise, data inconsistencies, and structural variability in radiographic images. Ultimately, the application of soft voting in this research not only improved the accuracy of the hybrid model but also demonstrated its potential as a practical and reliable approach for integration into intelligent computer-aided medical diagnosis systems.

The block diagram of the proposed method is shown in the Graphical Abstract Section for automated bone fracture and abnormality detection. This diagram shows the step-by-step process starting from input radiographic images, followed by preprocessing, augmentation, and feature extraction using the customized DenseNet-121 and EfficientNet-B4 models. The outputs of these base models are then combined through a soft-voting ensemble mechanism to generate the final classification for each anatomical region as Normal or Abnormal. This Figure provides a concise visual summary of the methodology, highlighting how the ensemble framework integrates multiple deep learning architectures for robust detection of musculoskeletal abnormalities.

Algorithm 1 illustrates the proposed ensemble deep learning framework for the automatic detection of musculoskeletal abnormalities in radiographic images. The algorithm takes as input radiographs from different anatomical regions, including the elbow, finger, hand, arm, forearm, shoulder, and wrist. Each image is first preprocessed by resizing, normalization, and optional augmentation to enhance model generalization. The dataset is then divided into training and validation sets for each body part. Two optimized deep learning

architectures, EfficientNet-B4 and DenseNet-121, serve as base models. Their final layers are customized with dense layers to improve feature extraction and discrimination. Both models are trained separately on the training set and evaluated on the validation set. The ensemble model is then constructed using a soft-voting mechanism that combines the predictions of the base models to produce a final probability score for each class (Normal or Abnormal). For each validation image, the ensemble model predicts the likelihood of abnormality, which is then thresholded to assign a class label. The performance metrics, including accuracy, sensitivity, precision, and F1-score, are computed separately for each anatomical region, allowing detailed evaluation of the model's effectiveness across different parts of the upper limb.

3. 6. Relations in the Proposed Method This framework provides a robust and clinically relevant approach for automated musculoskeletal abnormality detection, facilitating faster and more accurate radiological assessments.

In the training process of the proposed model for classifying shoulder X-ray images, the categorical cross-entropy loss function was employed.

This function is one of the most common and effective loss functions in deep learning for binary or multi-class classification tasks, particularly when labels are represented using one-hot encoding. Its primary purpose is to measure the discrepancy between the probability distribution predicted by the model and the actual label distribution. In other words, it evaluates how well the model's predictions match the ground truth. The categorical cross-entropy loss is mathematically formulated as:

$$L = -\sum_{i=1}^{C} t_i \log(P_i) \quad (1)$$

where t_i denotes the true label for class i, P_i represents the probability predicted by the model's Softmax output, and C is the number of classes. In this study, given the two classes "Normal" and "Abnormal," the loss function encourages the model to assign the highest probability to the correct class. Incorrect predictions with high confidence result in higher loss, guiding the model toward improved accuracy.

For optimization, the Adaptive Moment Estimation (ADAM) algorithm was utilized. By leveraging both momentum and adaptive learning rates, ADAM updates the network weights to minimize the loss efficiently (28).

Let $\tilde{X}_i$ denote the preprocessed input image. In the proposed method, two optimized deep learning architectures, DenseNet-121 and EfficientNet-B4, are employed as base models to extract discriminative features from the input image. The outputs of the two base models, DenseNet-121 and EfficientNet-B4, for image i are given by:

$$\begin{aligned} y_i^{(1)} &= f_{DensNet}(\tilde{X}_i; \Theta_1), \\ y_i^{(2)} &= f_{EfficientNet}(\tilde{X}_i; \Theta_2) \end{aligned} \quad (2)$$

where Θ_1 and Θ_2 represent the learnable parameters of DenseNet-121 and EfficientNet-B4, respectively. These outputs correspond to the predicted probabilities for the "Normal" and "Abnormal" classes and form the basis for the subsequent ensemble prediction. The ensemble output combines the predictions of the base models using a soft-voting mechanism:

$$y_i^{Ensemble} = \alpha y_i^{(1)} + (1-\alpha) y_i^{(2)} \quad (3)$$

where $\alpha \in [0,1]$ is the weight assigned to DenseNet-121 and $(1-\alpha)$ is the weight for EfficientNet-B4.

In this paper, the weight parameter α in the soft-voting ensemble was empirically determined using a validation subset of the MURA dataset. Several candidate values were tested, and the value that maximized overall accuracy and sensitivity was selected to optimally balance the contributions of DenseNet-121 and EfficientNet-B4.

The predicted class for image iii is determined by selecting the class with the highest probability:

$$\hat{C}_i = arg \max_{c \in \{\text{Normal,Abnormal}\}} y_i^{Ensemble}[c] \quad (4)$$

The weights of both models are updated during training using the ADAM optimizer, which combines momentum

Algorithm 1: Ensemble Deep Learning Framework for Musculoskeletal Abnormality Detection

Input: Radiographic images from MURA dataset (Elbow, Finger, Hand, Arm, Forearm, Shoulder, Wrist)
Output: Classification result for each region (Normal / Abnormal)

1: Load MURA dataset
2: For each image I in the dataset do
3: Preprocess I:
- Resize to fixed dimensions
- Normalize pixel intensities
- Apply augmentation (rotation, flipping, zoom, etc.)

4: Divide the dataset into a Training set and a Validation set for each body part
5: Initialize base models: EfficientNet-B4 and DenseNet-121
6: For each base model M do
7: Replace final layers with customized dense layers
8: Train M on the Training set of each body part
9: Validate M on Validation set
10: End For
11: Construct an Ensemble model using a soft-voting mechanism:
- Input: Predictions from EfficientNet-B4 and DenseNet-121
- Output: Weighted probability scores for each class (Normal / Abnormal)

12: For each image I in the Validation set do
13: Predict abnormality probability using an Ensemble model
14: Assign class label:
If probability ≥ threshold, then Abnormal
Else Normal
15: End For
16: Compute performance metrics (Accuracy, Sensitivity, Precision, F1-score) separately for each body part
17: Return classification results for all anatomical regions

and adaptive learning rates to minimize the categorical cross-entropy loss efficiently.

4. EXPERIMENTS AND RESULTS

The effectiveness of the proposed method was evaluated through a comprehensive experimental study conducted on the MURA dataset. In the following sections, we first describe the performance metrics used for evaluating the method. Next, a detailed explanation of the dataset employed in this study is provided. Finally, we present an in-depth analysis of the experimental results, offering insights into the performance and potential implications of the proposed approach.

4. 1. Performance Metrics Metric rationale. Given the class imbalance and the clinical cost of missed fractures, we report Accuracy together with Sensitivity (Recall), Precision, F1, Specificity, and Cohen's κ. Sensitivity captures clinical safety, F1 balances Precision–Recall under imbalance, and κ reflects agreement beyond chance across regions with varying prevalence. We also report ROC-AUC (and scope-matched PR-AUC where noted) to assess threshold-free discrimination. These metrics are derived from the confusion matrix and provide a comprehensive view of the model's reliability and correctness in image classification (32, 33). A brief description of each metric is provided below:

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (5)$$

$$Spe = TN/(TN+FP) \quad (6)$$

$$Ppr = TP/(TP+FP) \quad (7)$$

$$Sen = TP/(TP+FN) \quad (8)$$

$$F\text{-}Measure=2\times(Ppr\times Sen)/(Ppr+Sen) \quad (9)$$

In these relations, TP denotes the number of positive samples correctly classified as positive, FN represents the number of positive samples incorrectly classified as negative, FP indicates the number of negative samples incorrectly classified as positive, and TN refers to the number of negative samples correctly classified as negative.

Accuracy reflects the proportion of samples correctly classified (both positive and negative) out of the total number of samples. This metric represents the overall percentage of correct predictions made by the model compared to the ground truth labels.

Sensitivity (or recall) measures the percentage of actual positive samples correctly identified as positive. This metric indicates the model's ability to detect positive instances, such as abnormal images, correctly.

Specificity measures the percentage of actual negative samples correctly identified as negative. This metric reflects the model's capability to avoid misclassifying negative instances as positive (34, 35).

Additionally, Cohen's Kappa coefficient is a measure of agreement between the model's predictions and the actual results, taking into account the effect of random agreement. This metric is particularly important in binary classification problems, such as "normal" versus "abnormal" cases. The Kappa value ranges from -1 to +1, where 1 indicates perfect agreement, 0 corresponds to agreement by chance, and negative values signify agreement worse than chance (30).

$$Kappa = \frac{\Pr(a)-\Pr(c)}{1-\Pr(c)} \quad (7)$$

Here, $\Pr(a)$ represents the observed agreement between the model's predictions and the ground truth labels, while $\Pr(c)$ denotes the expected agreement purely by chance.

In this study, in addition to the aforementioned metrics, the confusion matrix was employed for a more detailed analysis of the model's performance. This matrix provides precise information regarding the number of correctly and incorrectly classified samples, offering a deeper understanding of the model's strengths and weaknesses in image classification.

Furthermore, the ROC curve was introduced to illustrate the model's capability to distinguish between positive and negative samples across different thresholds. The area under the curve (AUC) serves as a comprehensive metric for evaluating the overall performance of the model in musculoskeletal radiographs classification.

These metrics collectively provide a precise and thorough assessment of the system's ability to identify and differentiate various types of bone fractures, offering an overall view of the model's effectiveness in managing the complexity and inherent variability of bone fracture images.

4. 2. Experimental Results of the Proposed Method In this section, the results of evaluating the proposed method on the MURA dataset are presented. The proposed model, designed for detecting abnormalities in bone radiographs, was developed based on the combination of two advanced and customized neural networks: EfficientNet-B4 and DenseNet-121. The models were trained using the Adam optimizer and the categorical cross-entropy loss function. The training process was conducted independently for each of the seven anatomical regions, and the best weights were stored according to validation performance.

The final architecture employed an ensemble model, in which the predictions of the two networks were integrated through a soft-voting mechanism. In this approach, for each anatomical region, the final output

was calculated as the weighted average of the class probabilities predicted by both models. Table 5 summarizes the hyperparameters applied for training the EfficientNet-B4 and DenseNet-121 models on musculoskeletal radiographs from the MURA dataset. This includes settings such as learning rate, batch size, optimizer choice, number of epochs, and any regularization techniques, providing a clear overview of the configurations used to achieve optimal model performance. The hyperparameters used for training EfficientNet-B4 and DenseNet-121 on the MURA musculoskeletal radiographs are reported in Table 6. Both models were trained with a learning rate of 0.001, a batch size of 8, and for 15 epochs. The Adam optimizer with an epsilon of 1.0 was employed, and transfer learning from ImageNet was used for weight initialization. Softmax was chosen as the activation function, and categorical cross-entropy served as the loss function.

These hyperparameters were selected empirically based on stability, convergence behavior, and prior experience with similar medical imaging tasks, ensuring balanced learning and optimal performance across the models. Table 6 presents the classification performance of the proposed ensemble model with soft voting across different anatomical regions in the MURA dataset. The results show that the model achieves robust overall accuracy (83.52%) with a balanced precision (75.62%), high sensitivity (90.76%), and reasonable specificity (78.66%). Among individual regions, the elbow and humerus (arm) achieved the highest accuracies of 87.31% and 87.50%, respectively, indicating strong reliability in these categories. In contrast, hand and finger regions obtained relatively lower precision values, which may reflect the higher structural variability and complexity of these areas. Notably, the consistently high sensitivity across all regions highlights the model's strong capability in correctly identifying abnormal cases, while specificity values suggest some challenges in distinguishing normal samples, especially in smaller anatomical regions.

These findings confirm the effectiveness of the ensemble model while also pointing to potential areas for further refinement in clinical applications.

TABLE 5. Hyperparameters used for training EfficientNet-B4 and DenseNet-121 on musculoskeletal radiographs from the MURA dataset.

Hyperparameters	EfficientNet-B4	DenseNet-121
Learning rate	0.001	0.001
Batch size	8	8
Epochs	15	15
Optimizer	Adam(epsilon = 1.0)	Adam(epsilon = 1.0)
Weight initialization	Transfer learning (ImageNet)	Transfer learning (ImageNet)
Activation function	Softmax	Softmax
Loss function	Categorical Cross Entropy	Categorical Cross Entropy

TABLE 6. Classification results of the ensemble model (with soft voting) for each region in the MURA dataset.

Anatomical Region	Accuracy	Precision	Sensitivity	Specificity
Elbow	87.31	80.87	93.62	74.59
Finger	80.69	77.33	84.58	61.47
Forearm	82.06	72.19	92.00	64.14
Hand	80.65	61.90	93.73	58.19
Arm (Humerus)	87.50	88.57	86.49	75.00
Shoulder	81.17	76.26	85.96	62.30
Wrist	85.74	73.90	95.33	70.59
Overall Average	**83.52**	**75.62**	**90.76**	**78.66**

4. 3. Analysis of Results by Anatomical Region

In this chart, the performance results of the proposed ensemble model for each of the seven anatomical regions are presented. These regions include the elbow, finger, forearm, hand, arm, shoulder, and wrist. Each region was evaluated separately using the metrics of accuracy, sensitivity, and F1-score.

The overall performance of the model indicates that the proposed ensemble demonstrated stable and strong results across most regions. In particular, in six out of the seven regions, the F1-score was higher than 0.80, which reflects the high reliability and generalizability of the model.

The best performance was achieved in the elbow and arm regions, with F1-scores of 0.87 and 0.86, respectively. This result highlights the strong ability of the model to detect abnormalities in these regions, which may be due to the simpler bone structures and clearer fracture patterns found there.

The hand region, however, proved to be the most challenging, with an F1-score of 0.72. This performance resulted from a significant imbalance between high precision (0.87) and low sensitivity (0.62). This finding suggests that the model acts very cautiously when identifying fractures in the hand region. The likely reason is the structural complexity of the hand, with its multiple small bones and subtle fractures, which makes accurate detection more difficult.

In some cases, misclassifications occurred due to inherent challenges in the dataset. These include the complexity of small bone structures, subtle fractures that are visually similar to normal anatomy, and image

artifacts or noise present in radiographs. Understanding these factors can guide future improvements in model design and data preprocessing to further enhance classification accuracy.

4. 4. Experimental Results and Performance Analysis In this section, the performance of the proposed and baseline models on the MURA dataset is thoroughly analyzed and reported. This analysis aims to carefully examine the effectiveness of the proposed framework in detecting skeletal abnormalities in radiographic images across seven anatomical regions. As explained earlier, the two base architectures, EfficientNet-B4 and DenseNet-121, were trained independently, and their predictions were then combined into an ensemble model using the soft-voting method.

While some prior papers report higher Cohen's κ, cross-study κ comparisons are fragile because κ is sensitive to class prevalence, labeling variability, and evaluation scope (image- vs study-level) and subset coverage (single region vs all seven). Under our all-regions, image-level protocol, $\kappa = 0.68$ coexists with high Sensitivity (90.76%) and robust AUC (0.89), reflecting a recall-oriented operating point desirable in screening. Regions with small, complex anatomy (e.g., hand/finger) naturally depress κ despite strong discrimination, whereas simpler regions (e.g., elbow, humerus) yield higher per-region scores, consistent with our accuracy peaks (87.31% elbow; 87.50% humerus).

To evaluate performance, the following metrics were employed: accuracy, precision, sensitivity (recall), F1-score, and area under the ROC curve (AUC). In the next step, the overall performance of the models was assessed based on all test data from the seven anatomical regions. The comparison of precision, recall, and F1-score for the ensemble model across different upper-limb regions is presented in Figure 4.

Figure 5 presents the confusion matrices for the three models. Analysis of these matrices reveals the following:

- The DenseNet-121 model shows higher accuracy in predicting the *normal* class compared to the others, but` it is weaker in detecting the *abnormal* class.
- The EfficientNet-B4 model performs better in detecting abnormal samples, though it has a slightly higher error rate in the normal class.
- The ensemble model achieves the best balance between true positives (TP) and true negatives (TN), resulting in the lowest error rate across both classes.

In this section, the ROC curves for the ensemble model across each anatomical region are presented. As shown in Figure 6, the ROC curves illustrate the relationship between true positives (TP) and false positives (FP) at different classification thresholds.

A strong discriminative power is observed: as the chart indicates, the ROC curves for all regions lie significantly above the diagonal chance line (AUC = 0.5). This demonstrates the high capability of the model in distinguishing between normal and abnormal classes across all anatomical regions.

As shown in Figure 7, the ROC curves comparing DenseNet-121, EfficientNet-B4, and the Soft Voting Ensemble further highlight that while each model performs well, the ensemble consistently achieves a more balanced and superior discriminative power across test data.

AUC analysis: The area under the curve values for different regions show that the elbow achieved the highest AUC of 0.93, representing the best performance, whereas the hand region, with an AUC of 0.85, was identified as the most challenging. These results align well with the F1-score outcomes, confirming that simpler regions, such as the elbow, yield considerably better performance.

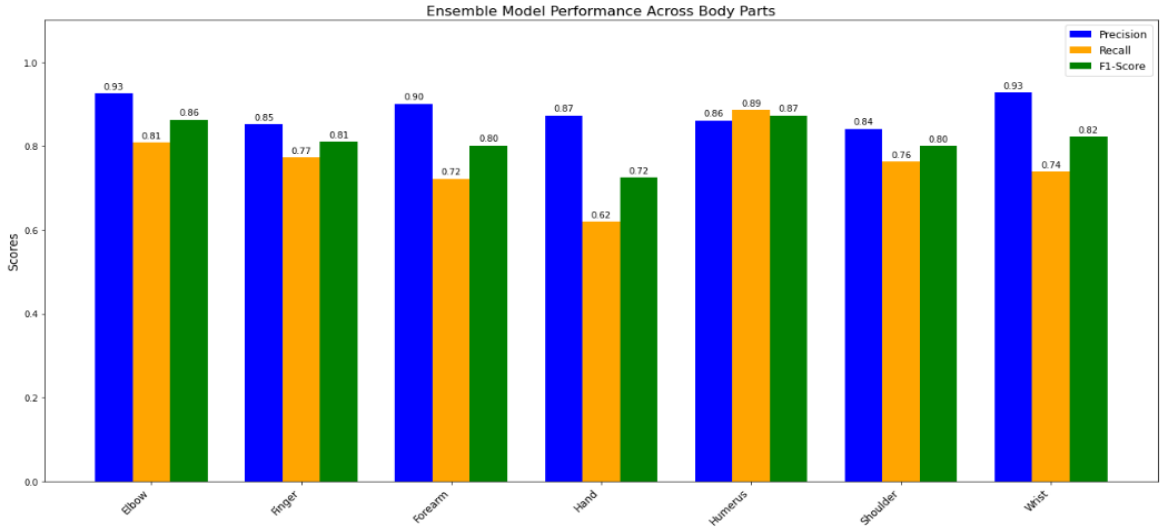
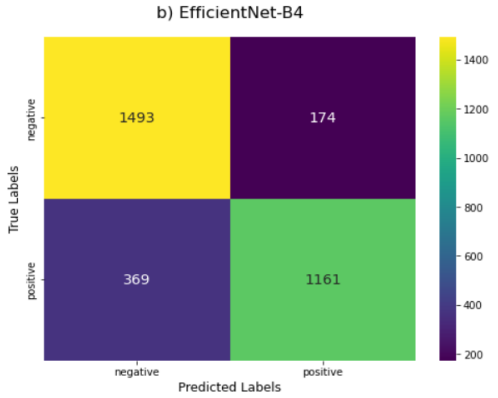
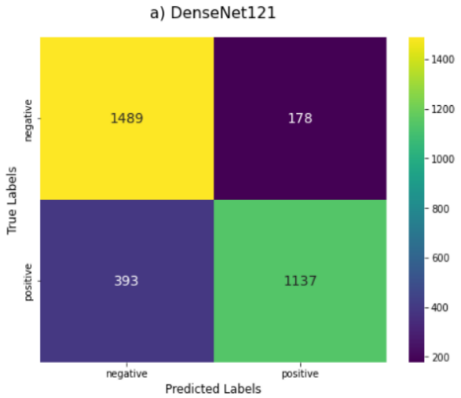


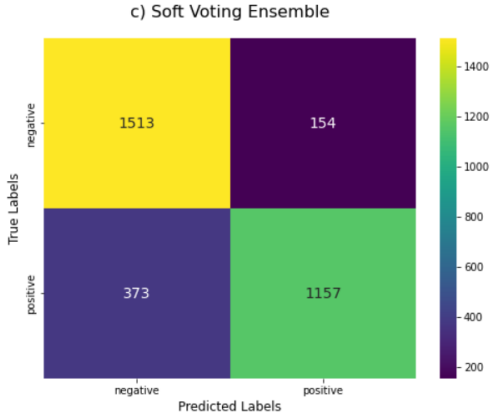
Figure 4. Comparison of precision, recall, and F1-score for the ensemble model across different upper-limb regions



a) Performance of the EfficientNet-B4 model



b) Performance of the DenseNet-121 model



c) Performance of the combined Soft Voting Ensemble method

Figure 5. Overall confusion matrices for comparing the performance of methods on the test dataset

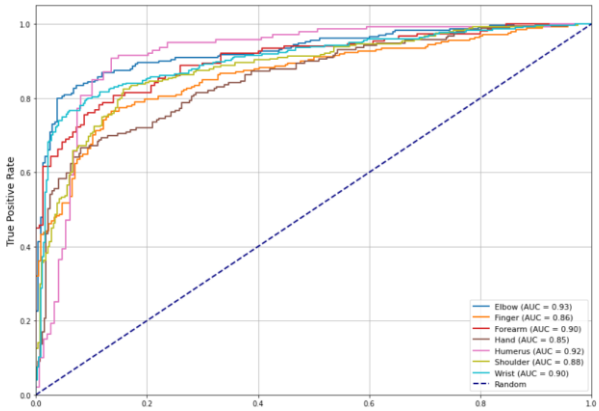


Figure 6. Comparison of ROC curves and AUC values for the Ensemble model across different parts of the upper limb.

Superiority of the ensemble model: By combining predictions from the two models, DenseNet-121 and EfficientNet-B4, the ensemble improved its overall performance, raising the AUC to 0.89. This improvement highlights the strong potential of model combination in building a more robust and stable system. This increase in AUC for the ensemble model demonstrates enhanced discriminative power and greater confidence in the model's diagnostic decision-making. To further assess the effectiveness of the proposed approach, its results were compared with several influential prior studies on fracture detection in radiographic images, particularly those based on the MURA dataset. The DenseNet-121 model was employed for binary classification of abnormalities, and its performance was reported at a radiologist-level standard.

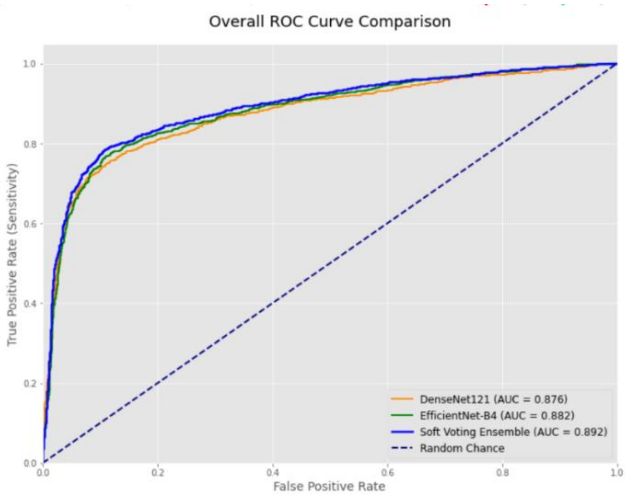


Figure 7. ROC curves for comparing the performance of methods on the test dataset

TABLE 7. Comparison of the proposed method with previous studies on MURA. (*Starred papers indicate that only part or 50% of the dataset was used.*)

Reference	Year	Used Model	Accuracy	AUC	Dataset
(15)*	2019	DenseNet-201, InceptionResNetV2	82	0.88	Humerus finger
(19)	2021	ST-SSAD	-	0.78	All 7 study types
(20)	2024	VGG-16 + ViT	82.88	-	All 7 study types
(4)	2022	Ensemble AlexNet + ResNet18	82.69	0.90	All 7 study types
Proposed Method	-	EfficientNet-B4 + DenseNet-121 (Soft Voting)	**84.1**	**0.90**	Humerus finger
Proposed Method	-	EfficientNet-B4 + DenseNet-121 (Soft Voting)	**83.52**	**0.89**	All 7 study types

However, this approach could not localize the fracture region and was limited only to overall decision-making. Subsequently, studies utilizing more advanced architectures such as MSDNet and DenseNet-169 attempted to enhance diagnostic performance.

While these models demonstrated high accuracy in certain metrics, they also faced challenges such as high computational requirements, complexity in implementation, and limited generalizability (4, 15).

To enable scope-matched comparison, we report results on both the full seven-region setting (Overall Accuracy 83.52%, Sensitivity 90.76%, Specificity 78.66%) and the restricted "finger + humerus" subset used by prior work (15). In the latter, our ensemble attains 84.1% Accuracy and 0.90 AUC, exceeding the 80% Accuracy and 0.84 AUC reported in (15). These results indicate that differences in anatomical coverage and granularity (image- vs study-level) can inflate metrics in narrower settings, and that our method remains competitive even under matched conditions.

Moreover, to enable a fair comparison with Chada (15), who reported results only on the finger and humerus subsets of the MURA dataset, we also evaluated our proposed method under the same restricted setting. In this case, the ensemble model achieved an average accuracy of 84.1% and an AUC of 0.90, outperforming the accuracy (80%) and AUC (0.84) reported by Chada (15). This demonstrates that even when constrained to a limited portion of the dataset, our approach maintains superior discriminative ability and robustness compared to previous methods. In contrast, the proposed model in this study, by combining the two advanced architectures DenseNet-121 and EfficientNet-B4 within an ensemble framework using a soft-voting mechanism, achieved competitive performance and, in some cases, outperformed previous works. To assess the position of the proposed method relative to prior advanced approaches, Table 6 compares its overall performance with several selected studies conducted on the MURA dataset. These results show that, in terms of accuracy,

balanced performance across different regions, and ease of implementation, the proposed method offers a significant advantage over many earlier techniques.

Ultimately, the presented model, with its balanced approach, relative simplicity, and generalizability in real clinical settings, holds strong potential for integration into computer-aided diagnostic systems. Our proposed framework has demonstrated stable and reliable performance by maintaining a balance among different evaluation metrics. In particular, by achieving an overall accuracy of 83.52% and a meaningful AUC of 0.89, the proposed model shows considerable promise as an auxiliary diagnostic tool in real clinical environments.

Table 7 presents a comparative analysis of the proposed ensemble method against several state-of-the-art approaches previously applied to the MURA dataset. As shown, earlier studies such as Chada (15) and others (4, 19, 20) employed models including DenseNet-201, InceptionResNetV2, ST-SSAD, VGG-16 combined with ViT, and ensemble variants of AlexNet and ResNet18, reporting accuracies in the range of 77–83%. In contrast, our proposed method, based on the integration of EfficientNet-B4 and DenseNet-121 through a soft voting strategy, achieved an accuracy of 83.52% and an AUC of 0.89 across all seven study types, thereby outperforming most of the prior works. Moreover, to enable a fair comparison with Chada (15), who reported results only on the finger and humerus subsets of the MURA dataset, we also evaluated our method under the same restricted setting. In this case, the ensemble model achieved an average accuracy of 84.1% and an AUC of 0.90, surpassing the accuracy (80%) and AUC (0.84) reported by Chada (15).

These findings demonstrate that even under limited data conditions, the proposed approach maintains superior discriminative ability and robustness compared to previous methods. This table summarizes the models used, evaluation metrics, and reported results from prior work. Starred papers indicate that only a portion of the dataset, or 50% of it, was utilized. As shown, previous approaches have employed various architectures, including DenseNet, InceptionResNetV2, ST-SSAD, VGG-16 combined with ViT, and ensemble models such as AlexNet and ResNet18. The proposed method integrates EfficientNet-B4 and DenseNet-121 through a soft-voting mechanism and demonstrates competitive performance, achieving higher accuracy compared to most previous studies while maintaining a strong AUC score.

To qualitatively assess the performance of the proposed model, several random predictions on the MURA dataset are presented. Figure 8 shows the ensemble model outputs across different skeletal regions, including the elbow, forearm, wrist, finger, hand, and shoulder. As can be observed, the model successfully identified many cases correctly (e.g., detecting fractures in the shoulder and finger regions), while some misclassifications still occurred (e.g., false negatives in the forearm region). These examples highlight the inherent challenges in automated fracture detection, where subtle abnormalities and structural similarities to normal anatomy may lead to errors. Overall, the qualitative results in Figure 8 complement the quantitative evaluation and confirm that the proposed model demonstrates robust predictive ability, making it a valuable clinical decision support tool in radiological workflows.

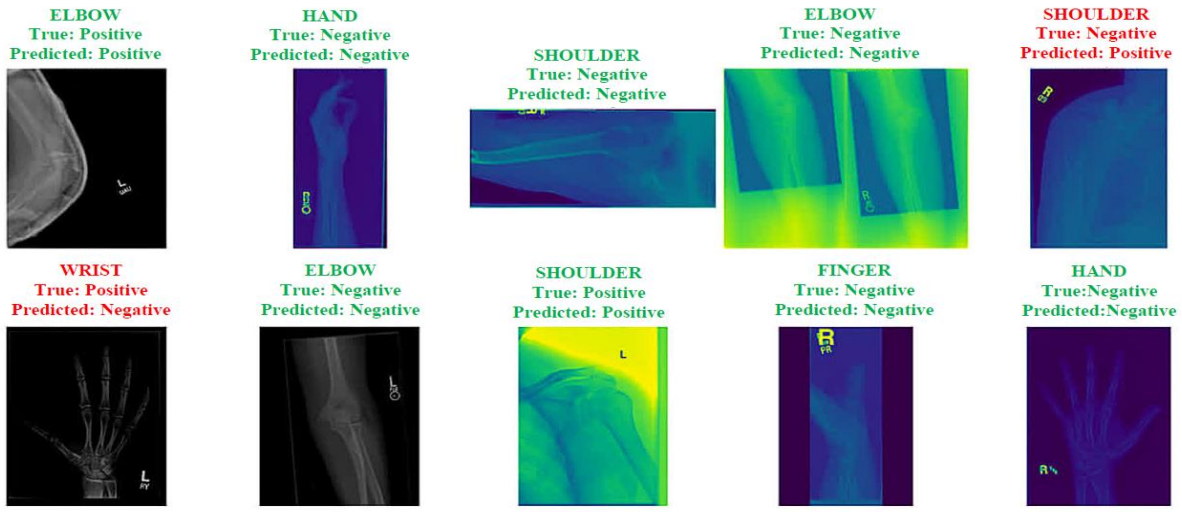


Figure 8. Random sample predictions across different skeletal regions using the proposed ensemble model. The figure presents true positive, true negative, false positive, and false negative predictions on radiographic images of various anatomical regions (elbow, forearm, wrist, finger, hand, and shoulder). The results highlight both the model's reliable performance and the challenges of detecting subtle bone abnormalities in medical imaging

4. 4. 5. Computational Considerations For clinical integration, computational efficiency is as important as accuracy. Our ensemble comprises two pretrained backbones plus customized dense heads. The approximate parameter counts are: EfficientNet-B4 ≈ 19M, DenseNet-121 ≈ 8M, with the added dense layers contributing ≈ 7.7M (2560→1200→2 on the EfficientNet branch) and ≈ 1.6M (1024→512→2 on the DenseNet branch). In total, the ensemble has ≈ 36–37M trainable parameters. At 32-bit precision, the parameter memory footprint is on the order of ≈ 140 MB, which roughly halves under FP16.

We measured end-to-end inference in a Kaggle GPU environment equipped with an NVIDIA Tesla T4 (batch size = 8, input size 380×380, FP32). Including preprocessing and both forward passes, the average latency was ≈ 4.0 s per image. This value serves as our deployment baseline. In settings where lower latency is required (e.g., interactive triage), several standard accelerations are available without retraining the ensemble: mixed-precision (FP16) inference, export to ONNX/TensorRT, and layer-level pruning of the added dense heads. For resource-constrained deployments, a knowledge-distilled single-backbone student network is a practical alternative that preserves most of the ensemble's accuracy while reducing latency and memory. Finally, a simple confidence-gated protocol can route easy cases through a single backbone and reserve the full ensemble for ambiguous cases, trading a small accuracy drop for significant average-latency reductions.

We report latency as time per image at batch size 8 because this setting best matches point-of-care use. Throughput under mini-batching scales predictably on GPUs and can be reported for site-specific hardware if required.

5. CONCLUSION

In this study, an innovative deep learning–based framework was proposed for the detection and classification of bone abnormalities in radiographic images. To this end, two advanced models, EfficientNet-B4 and DenseNet-121, were redesigned and optimized by adding customized dense layers at the output, enabling the extraction of precise and specialized features. To further improve classification accuracy and stability, the predictions of these two models were integrated into an ensemble structure through a soft-voting mechanism. The proposed model was evaluated on the complex and widely recognized MURA dataset, which includes seven distinct anatomical regions. Experimental results demonstrated that combining these two architectures produced reliable, stable, and accurate performance, with the ensemble model achieving an overall accuracy of 83.52%, thus offering competitive results compared to existing methods.

The proposed ensemble model demonstrated robust and consistent performance across all seven anatomical regions in the MURA dataset, highlighting its ability to generalize across diverse skeletal structures with minimal bias toward specific bone types.

The proposed framework contributes to the field by demonstrating that an ensemble of heterogeneous CNN architectures, enhanced with customized dense layers, can effectively capture subtle fracture patterns and improve both accuracy and sensitivity. This approach offers a robust and practical foundation for clinical decision-support systems in medical imaging.

Taken together, the all-regions, image-level evaluation explains why κ may be lower than reports limited to specific regions, while our Sensitivity (90.76%), AUC (0.89), and scope-matched improvements over support the clinical utility and robustness of the proposed ensemble.

The proposed models were implemented in Python 3.7.12 using TensorFlow 2.x with the built-in tf.keras API, and were trained and evaluated in a Kaggle GPU environment equipped with an NVIDIA Tesla T4 accelerator, ensuring reproducible results.

6. FUTURE WORK

Despite the promising results of the proposed method, there remain several opportunities for further improvement and extension in future research. While the current study focuses on achieving high classification accuracy, it does not explicitly address model interpretability. Incorporating methods such as CAM or Grad-CAM in future work can help highlight the most influential regions in the images, providing clinicians with more trustworthy and interpretable decisions. One potential direction is the use of CAM or Grad-CAM to generate heatmaps that highlight the region's most influential in the model's decisions, thereby improving the clinical interpretability and trustworthiness of the system. Another avenue is the incorporation of Vision Transformers (ViT) or hybrid CNN-Transformer architectures, which could enhance the model's ability to capture global contextual relationships within images and improve accuracy in detecting more complex fractures. Additionally, the integration of attention mechanisms would enable the model to selectively focus on the most relevant areas of an image while ignoring non-essential information, ultimately leading to greater precision, robustness, and reliability in diagnostic performance. Furthermore, employing k-fold cross-validation in future studies could provide a more robust and reliable estimate of model performance and generalization across different subsets of the dataset.

Acknowledgements

The authors would like to thank the Faculty of Engineering and Technology at the University of Mazandaran for providing the computational resources necessary for this study.

Funding

This research received no external funding or financial support.

Ethics Approval and Consent to Participate

This study used the publicly available MURA dataset and did not involve experiments on human participants or animals conducted by the authors; therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The authors declare that they have no known financial or organizational conflicts of interest that could have influenced the work reported in this paper.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT solely for minor language editing and clarity improvement. All scientific content, analysis, and conclusions were produced by the authors, who take full responsibility for the integrity, accuracy, and originality of the work.

Data Availability

The data used in this study are publicly available within the MURA dataset. Any processed results or model outputs can be shared by the corresponding author upon reasonable academic request.

Authors Biosketch

Mohadeseh Mahtabi is a Master's student in Computer Engineering – Artificial Intelligence at the University of Mazandaran, Iran. Her research interests include deep learning, ensemble learning, and medical image analysis, with a focus on automated bone fracture detection.

Sekineh Asadi Amiri is an Associate Professor in the Department of Computer Engineering at the University of Mazandaran, Iran. She received her Ph.D. in Computer Engineering from Shahrood University of Technology in 2016. Her research interests include image processing, deep learning, and intelligent computer-aided diagnostic systems, with a particular focus on medical image analysis and deep neural network–based models.

Samira Mavaddati is an Associate Professor in the Department of Electrical Engineering at the University of Mazandaran, Iran. She received her Ph.D. from Amirkabir University of Technology (Tehran Polytechnic), Iran, in 2016. Her research interests include image processing, signal processing, deep learning, pattern recognition, and intelligent data analysis, with applications in engineering and biomedical systems.

REFERENCES

1. Lu S, Wang S, Wang G. A novel deep learning approach for image segmentation. Multimedia Tools and Applications. 2022;81:44487-503. https://doi.org/10.1007/s11042-022-13287-z
2. Saggi SS, Kuah LZD, Toh LCA, Shah MTM, Wong MK, Abd Razak HRB. A novel approach to structural health monitoring using deep learning. Journal of Civil Engineering. 2023;50(2):115-32.
3. Ioppolo F, Rompe JD, Furia JP, Cacchio A. Clinical application of shock wave therapy (SWT) in musculoskeletal disorders (MSDs). European Journal of Physical and Rehabilitation Medicine. 2014;50(2):217-30.
4. Karthik K, Kamath SS. MSDNet: A deep neural ensemble model for abnormality detection and classification of plain radiographs. Journal of Ambient Intelligence and Humanized Computing. 2023;14(12):16099-113. https://doi.org/10.1007/s12652-022-03835-8
5. Karthik K, Kamath SS. A deep neural network model for content-based medical image retrieval with multi-view classification. The Visual Computer. 2021;37(7):1837-50. https://doi.org/10.1007/s00371-020-01941-2
6. Ying J, Dutta J, Guo N, Hu C, Zhou D, Sitek A, Li Q. Classification of exacerbation frequency in the COPDGene cohort using deep learning with deep belief networks. IEEE Journal of Biomedical and Health Informatics. 2016;24(6):1805-13. https://doi.org/10.1109/JBHI.2016.2642944
7. Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J. Long radiology workdays reduce detection and accommodation accuracy. Journal of the American College of Radiology. 2010;7(9):698-704. https://doi.org/10.1016/j.jacr.2010.03.004
8. Mukesh BR, Harish T, Mayya V, Kamath S. Deep learning based detection of diabetic retinopathy from inexpensive fundus imaging techniques. In: 2021 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); 2021. 1-6. https://doi.org/10.1109/CONECCT52877.2021.9622703
9. García-Floriano A, Ferreira-Santiago Á, Camacho-Nieto O, Yáñez-Márquez C. A machine learning approach to medical image classification: Detecting age-related macular degeneration in fundus images. Computers and Electrical Engineering. 2019;75:218-29. https://doi.org/10.1016/j.compeleceng.2017.11.008
10. Nedumkunnel IM, George LE, Sowmya KS, Rosh NA, Mayya V. Explainable deep neural models for COVID-19 prediction from chest X-rays with region of interest visualization. In: 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC); 2021. 96-101. https://doi.org/10.1109/ICSCCC51823.2021.9478152

11. Mayya V, Karthik K, Sowmya KS, Karadka K, Jeganathan J. Coviddx: AI-based clinical decision support system for learning COVID-19 disease representations from multimodal patient data. In: International Conference on Health Informatics (HEALTHINF); 2021. p. 659-66.
12. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL, Langlotz C. MURA: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv. 2017. https://doi.org/10.48550/arXiv.1712.06957
13. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems. 2012;25:1097-105.
14. Ebsim R, Naqvi J, Cootes TF. Automatic detection of wrist fractures from posteroanterior and lateral radiographs: A deep learning-based approach. In: International Workshop on Computational Methods and Clinical Applications in Musculoskeletal Imaging; 2018. p. 114-25. https://doi.org/10.1007/978-3-030-11166-3_10
15. Chada G. Machine learning models for abnormality detection in musculoskeletal radiographs. Reports. 2019;2(4):26. https://doi.org/10.3390/reports2040026
16. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. In: Intelligent Computing – Proceedings of the Computing Conference; 2019. p. 971-81. https://doi.org/10.1007/978-3-030-22871-2_69
17. Yadav DP, Rathor S. Bone fracture detection and classification using deep learning approach. In: International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC); 2020. 282-85. https://doi.org/10.1109/PARC49193.2020.236611
18. Meena T, Roy S. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. Diagnostics. 2022;12(10):2420. https://doi.org/10.3390/diagnostics12102420
19. Spahr A, Bozorgtabar B, Thiran JP. Self-taught semi-supervised anomaly detection on upper limb X-rays. In: IEEE International Symposium on Biomedical Imaging; 2021. 1632-36. https://doi.org/10.1109/ISBI48211.2021.9433771
20. Dahal S, Thapa R, Panth S. A hybrid deep learning model for musculoskeletal abnormality detection. In: Proceedings of the 15th IOE Graduate Conference; 2024. 1-6.
21. Hardalaç F, Uysal F, Peker O, Çiçeklidağ M, Tolunay T, Tokgöz N, Kutbay U, Demirciler B, Mert F. Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors. 2022;22(3):1285. https://doi.org/10.3390/s22031285
22. Beyaz S, Yayli SB, Kılıc E, Doktur U. The ensemble artificial intelligence method: Detection of hip fractures in AP pelvis plain radiographs by majority voting using a multi-center dataset. Digital Health. 2023;9:1-14. https://doi.org/10.1177/20552076231216549
23. Ghosh M, Hassan S, Debnath P. Ensemble based neural network for the classification of MURA dataset. Journal of Nature, Science and Technology. 2021;4:1-5.
24. Lu X, Chang EY, Du J, Yan A, McAuley J, Gentili A, Hsu CN. Robust multi-view fracture detection in the presence of other abnormalities using HAMIL-Net. Military Medicine. 2023;188(Supplement 6):590-7. https://doi.org/10.1093/milmed/usad252
25. Stanford ML Group. MURA dataset. Available from: https://stanfordmlgroup.github.io/competitions/mura/
26. Zhang P, Yang L, Li D. EfficientNet-B4-Ranger: A novel method for greenhouse cucumber disease recognition under natural complex environment. Computers and Electronics in Agriculture. 2020;176:105652. https://doi.org/10.1016/j.compag.2020.105652
27. Albelwi SA. Deep architecture based on DenseNet-121 model for weather image recognition. International Journal of Advanced Computer Science and Applications. 2022;13(10):559.
28. Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in Neural Information Processing Systems. 2018;31.
29. Agnihotri D, Verma K, Tripathi P, Singh BK. Soft voting technique to improve the performance of global filter based feature selection in text corpus. Applied Intelligence. 2019;49(4):1597-619. https://doi.org/10.1007/s10489-018-1349-1
30. Mim MA, Majadi N, Mazumder P. A soft voting ensemble learning approach for credit card fraud detection. Heliyon. 2024;10:e25466. https://doi.org/10.1016/j.heliyon.2024.e25466
31. Asadi Amiri S, Mohammadpoory Z, Nasrolahzadeh M. A novel content-based image retrieval system using fusing color and texture features. Journal of AI and Data Mining. 2022;10(4):559-68. https://doi.org/10.22044/jadm.2022.12042.2353
32. McHugh ML. Interrater reliability: The kappa statistic. Biochemia Medica. 2012;22(3):276-82.
33. Mavaddati S, Razavi M. An optimized YOLO-ViT hybrid model for enhanced precision in rice classification and quality assessment. International Journal of Engineering, Transactions A: Basics. 2025;38(10):2435-50. https://doi.org/10.5829/ije.2025.38.10a.19
34. Mohammadi M, Talebpour A, Hosseinsabet A. Presenting effective methods in classification of echocardiographic views using deep learning. International Journal of Engineering, Transactions B: Applications. 2024;37(11):2150-61. https://doi.org/10.5829/ije.2024.37.11b.02
35. Farsi H, Noursoleimani S, Mohamadzadeh S, Barati A. Multimodal biomedical image segmentation by using multi-path U-Net. International Journal of Engineering. 2025;38(1):179-93. https://doi.org/10.5829/ije.2025.38.01a.17

Persian Abstract

چکیده

پیشرفت‌های اخیر در شبکه‌های عصبی عمیق موجب بهبود چشمگیری در تحلیل تصاویر پزشکی شده است، با این حال **تشخیص شکستگی استخوان در تصاویر رادیوگرافی** همچنان یک چالش اساسی محسوب می‌شود؛ زیرا ساختار پیچیده استخوان، الگوهای ظریف شکستگی و محدودیت داده‌های برچسب‌خورده، عملکرد مدل‌ها را با دشواری روبه‌رو می‌کند. در این پژوهش، یک چارچوب پیشرفته یادگیری عمیق تلفیقی ارائه می‌شود که دو معماری بهینه‌شده EfficientNet-B4 و DenseNet-121 را از طریق راهبرد ترکیب نرم (Soft-Voting Fusion) برای تشخیص خودکار و دقیق شکستگی‌ها ادغام می‌کند. در این ساختار هیبریدی، از وزن‌دهی تطبیقی و لایه‌های متراکم بهینه‌شده استفاده شده است تا تمایز ویژگی‌ها تقویت شود و توان شبکه در تشخیص جزئیات ظریف شکستگی‌ها افزایش یابد. علاوه بر این، با به‌کارگیری یادگیری انتقالی و تنظیم دقیق (Fine-tuning)، عدم‌توازن داده‌ها کاهش یافته و تعمیم‌پذیری مدل در نواحی مختلف آناتومیکی بهبود یافته است. نتایج حاصل از آزمایش‌های جامع بر روی مجموعه‌داده MURA که شامل تصاویر رادیوگرافی از هفت ناحیه مختلف بدن است، نشان می‌دهد که مدل پیشنهادی با دستیابی به ۸۳/۵۲٪ دقت و ۹۰/۷۶٪ حساسیت عملکردی برتر نسبت به مدل‌های پایه ارائه می‌دهد. پایداری نتایج در تنظیمات مختلف آموزشی، قابلیت اطمینان مدل را برای کاربردهای بالینی تأیید می‌کند. در مجموع، این پژوهش یک **سامانه نوآورانه مبتنی بر یادگیری عمیق تلفیقی** معرفی می‌کند که با بهره‌گیری از ترکیب تطبیقی ویژگی‌ها و نقاط قوت معماری‌های مکمل، دقت بالایی در تشخیص شکستگی‌ها فراهم می‌آورد و گامی مؤثر در جهت توسعه **سامانه‌های هوشمند پشتیبان تصمیم‌گیری رادیولوژی** به‌شمار می‌رود.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2602-2616

International Journal of Engineering

Journal Homepage: www.ije.ir

Optimal Design of Excitation Inputs for Identifying the Dynamics of Fixed-wing Aircraft Considering Active Control in MIMO Systems

A. Mohseni[a], M. Mortazavi*[b], N. Sayyaf[a], S. A. Askari[a]

[a] *Department of Mechanical Engineering, University of Isfahan, Isfahan, Iran*
[b] *Department of Electrical Engineering, University of Isfahan, Isfahan, Iran*

PAPER INFO

Paper history:
Received 16 September 2025
Received in revised form 25 October 2025
Accepted 08 January 2026

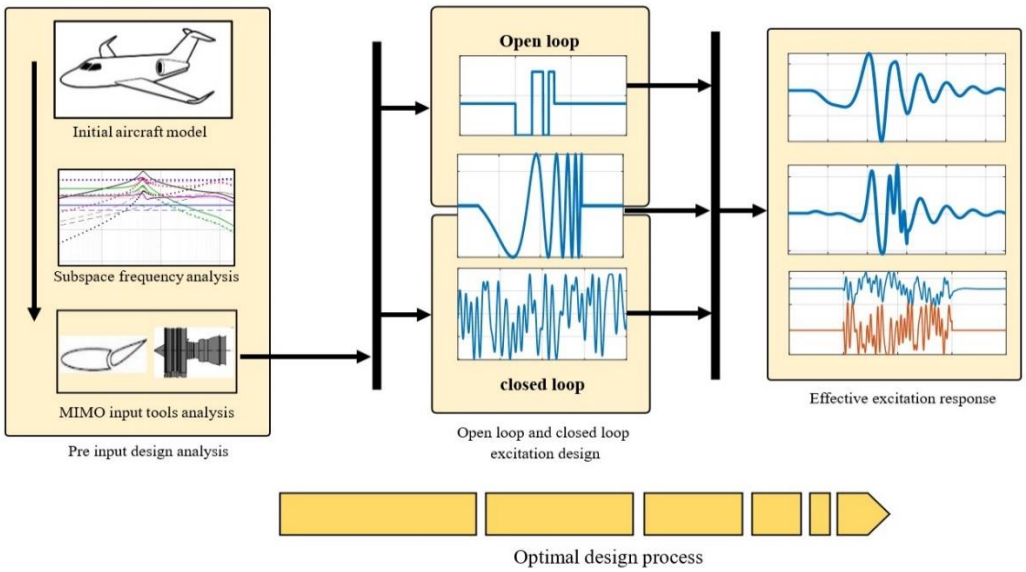
Keywords:
Aircraft
System Identification
Excitation Input Design
Parameter Estimation
Multi-input Multi-output Systems
Closed Loop

ABSTRACT

This study presents a systematic and optimal framework for designing excitation inputs for fixed-wing aircraft system identification, with a particular focus on multi-input multi-output (MIMO) systems under active flight control. Excitation signals such as pulse, doublet, and multi-step are commonly selected due to their simplicity and widespread use in practical applications. However, their selection is often empirical and may not provide optimal excitation for accurate and robust parameter estimation. The proposed methodology systematically determines the most effective input signals, designs their frequency content and amplitude characteristics, and evaluates their performance under both open- and closed-loop control conditions. This approach facilitates accurate parameter estimation across multiple identification techniques and is generalizable to other aircraft as well as potentially to other complex dynamic systems. The effectiveness of the proposed method is demonstrated through high-fidelity six-degree-of-freedom (6-DOF) simulations in MATLAB/Simulink, showing its capability to efficiently and accurately excite the relevant aircraft modes for precise model identification.

doi: *10.5829/ije.2026.39.10a.20*

Graphical Abstract



*Corresponding Author Email: ma.mortazavi@eng.ui.ac.ir (M. Mortazavi)

Please cite this article as: Mohseni A, Mortazavi M, Sayyaf N, Askari SA. Optimal Design of Excitation Inputs for Identifying the Dynamics of Fixed-wing Aircraft Considering Active Control in MIMO Systems. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2602-16.

1. INTRODUCTION

System identification is a fundamental discipline that has consistently played a pivotal role in the design and development of aerospace vehicles, with enduring applications from historical projects to contemporary advanced aircraft. This methodology is employed across a diverse spectrum of areas, including understanding and optimizing aircraft behavior, designing and refining control systems, fault detection and diagnosis, developing high-fidelity simulators, establishing handling qualities criteria, and planning efficient aerial missions (1-5). A core tenet of system identification is that flight data must be rich in information about the dynamic characteristics to be identified; as a fundamental rule, "if the characteristics are not in the data, do not expect to be able to identify them in the model" (6). Consequently, the design of appropriate excitation inputs to stimulate the aircraft's dynamics during flight tests, — thereby ensuring the collection of informative data— represents one of the most critical stages in the identification process. This design must be executed systematically and aligned precisely with the objectives of the system identification endeavor.

The most prevalent excitation inputs in aircraft system identification include pulse, doublet, multi-step, sinusoidal, and multi-sinusoidal signals, each serving specific purposes (7-10). For instance, the doublet and multi-step inputs are widely utilized due to their simplicity and effectiveness in exciting a desired frequency range, making them particularly suitable for identifying short-period and Dutch roll modes (11-14). Conversely, the pulse input is more appropriate for exciting and characterizing the phugoid mode (15, 16). Another commonly used input is the frequency sweep, which covers a broad range of frequencies and is especially valuable when prior knowledge of the system dynamics is limited (17, 18). Among more recent developments, the Orthogonal Optimized Multisine (OOM) method (7) has been introduced as a practical, versatile, and effective approach for flight test applications. It has been employed both theoretically and experimentally for the excitation of aircraft dynamics and the estimation of dynamic parameters (19, 20). Given their predominant application in exciting aircraft system dynamics, these input types form the primary focus of the present investigation.

In Multi-Input Multi-Output (MIMO) systems, a crucial consideration is determining the impact of different control effectors on the identification process. For example, in an aircraft featuring a V-tail configuration, longitudinal dynamics can be excited using the elevator, rudder, or engine thrust. A review of the existing literature reveals that excitation inputs are often designed and applied based on traditional or empirical methods, frequently leading to suboptimal performance and increased time and cost for flight testing and identification (21-23). While these references address input design for identifying aircraft dynamics, the critical questions of how to apply these inputs across multiple channels and their subsequent effects are often not thoroughly examined. Furthermore, besides the application strategy, key input characteristics—such as frequency content, amplitude, duration, and signal shape—must be meticulously designed with respect to the specific aircraft dynamics. This aspect of tailored input design has also been largely overlooked in favor of more conventional approaches .

Another critical factor in aircraft system identification is the state of the flight control system, which can be either active or inactive, corresponding to closed-loop or open-loop identification, respectively. In open-loop identification, the control system is disengaged, allowing excitation inputs to be applied directly to the aircraft, stimulating its dynamics according to the input's frequency content. However, for most modern aircraft, mission requirements and safety constraints often preclude open-loop operation due to their reliance on active stability augmentation and flight control systems. Consequently, the active controller significantly influences the system's response to excitations, making open-loop identification challenging and often impractical (4, 24, 25). While necessary, closed-loop testing introduces specific challenges; primarily, the feedback control action can dampen the excitation signals and introduce correlations between the control input and process noise, complicating the identification process (4). Under these conditions, excitation inputs must be designed to optimally overcome these limitations and effectively excite the system dynamics, a challenge that is addressed in this work.

System identification and precise dynamic modeling, in addition to the mathematical model and flight motion equations, have novel applications in other parts of an aircraft as well. Recent advancements in UAV-based sensing and control systems highlight the growing importance of precise dynamic modeling and identification for complex aerial platforms. For example, reference (26) developed a UAV-based structural health monitoring system combining vision-based localization and deep learning models to detect structural defects in GNSS-denied environments. Similarly, Jiang et al. (27) integrated intelligent sensors with UAVs to enhance operational safety and spatial awareness. These studies underscore the increasing reliance on UAVs as intelligent dynamic systems that require accurate system identification and effective input design, motivating the development of the systematic framework proposed in this work.

Despite the variety of existing methods for designing excitation inputs, a persistent challenge in aircraft system

identification is their optimal design, particularly when multiple control effectors are involved and the control system is active. Flight testing is inherently risky, costly, and time-consuming. Therefore, it is imperative to design excitation inputs purposefully and efficiently to maximize information yield and minimize the number of required test flights. Most existing studies focus on specific excitation types or single-input cases and do not provide a unified procedure for evaluating and comparing their effectiveness under different flight conditions. To further clarify the importance of this topic, Table 1 lists some studies that have employed excitation inputs, along with the identified design limitations.

The main contribution of this paper is to present a systematic and optimal framework for designing and evaluating excitation inputs for fixed-wing aircraft. The method accounts for both open- and closed-loop conditions in MIMO systems. In this context, the proposed process not only determines suitable excitation frequency ranges but also evaluates which control effectors (e.g., elevator or throttle) are more effective for exciting specific aircraft modes and extracting aerodynamic derivatives. The novelty of this work lies in addressing all key challenges simultaneously — including the closed-loop nature of the system, the MIMO interactions between control effectors, and the difficulty of properly exciting relevant modes while maintaining safe flight conditions. The proposed framework provides an integrated process that can be applied to various aircraft configurations and potentially extended to other dynamic systems.

The paper is structured as follows: the first section details the derivation of the initial aircraft model. The second section presents the design of the excitation frequency range based on this model and provides a method for selecting the most effective input in MIMO systems. The third section develops various excitation inputs according to the specified frequency range and analyzes their performance through simulation. The fourth section addresses the specific design of excitation inputs for closed-loop system identification. Finally, the fifth section designs an optimal excitation input for the target aircraft and validates its effectiveness through comprehensive six-degree-of-freedom (6-DOF) simulations in the MATLAB/Simulink environment.

2. DERIVATION OF INITIAL MODEL

The design process for excitation inputs begins with the development of a mathematical model representing the system at its intended operational condition. Following this approach, the longitudinal state-space model of the fixed-wing aircraft is formulated according to Equation 1. (28).

$$\begin{cases} \dot{x} = Ax + Bu_{inp} \\ y = Cx + Du_{inp} \end{cases}, \tag{1}$$

Where matrices A and B are defined as follows:

$$A = \begin{bmatrix} X_u + X_{Tu} + \frac{X_{\dot{\alpha}}(Z_u + Z_{Tu})}{U_1 - Z_{\dot{\alpha}}} & X_\alpha + \frac{X_{\dot{\alpha}} Z_\alpha}{U_1 - Z_{\dot{\alpha}}} & X_q + \frac{X_{\dot{\alpha}}(U_1 + Z_q)}{U_1 - Z_{\dot{\alpha}}} & -g \\ \frac{Z_u + Z_{Tu}}{U_1 - Z_{\dot{\alpha}}} & \frac{Z_\alpha}{U_1 - Z_{\dot{\alpha}}} & \frac{U_1 + Z_q}{U_1 - Z_{\dot{\alpha}}} & 0 \\ M_u + M_{Tu} + \frac{M_{\dot{\alpha}}(Z_u + Z_{Tu})}{U_1 - Z_{\dot{\alpha}}} & M_\alpha + M_{T\alpha} + \frac{M_{\dot{\alpha}} Z_\alpha}{U_1 - Z_{\dot{\alpha}}} & M_q + M_{\dot{\alpha}} + \frac{U_1 + Z_q}{U_1 - Z_{\dot{\alpha}}} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{2}$$

$$B = \begin{bmatrix} X_{\delta e} + \frac{X_{\dot{\alpha}} Z_{\delta e}}{U_1 - Z_{\dot{\alpha}}} & X_T + \frac{X_{\dot{\alpha}} Z_T}{U_1 - Z_{\dot{\alpha}}} \\ \frac{Z_{\delta e}}{U_1 - Z_{\dot{\alpha}}} & \frac{Z_T}{U_1 - Z_{\dot{\alpha}}} \\ M_{\delta e} + \frac{M_{\dot{\alpha}} Z_{\delta e}}{U_1 - Z_{\dot{\alpha}}} & M_T + \frac{M_{\dot{\alpha}} Z_T}{U_1 - Z_{\dot{\alpha}}} \\ 0 & 0 \end{bmatrix} \tag{3}$$

TABLE 1. Comparison of issues in the design of excitation inputs in selected references

Refs.	Input type	Closed loop excitation	Optimal Control Effector in MIMO	Frequency range estimation	Optimal process
(1)	MS	*	*	*	*
(4)	FS	✓	*	*	*
(7)	OOM	*	*	✓	*
(8)	OOM 3211	*	*	✓	*
(9, 11)	D 3211	*	*	*	*
(12, 14)	MS OOM	✓	*	✓	*
(15)	OOM MS	✓	*	✓	*
(19)	OOM	✓	*	✓	*

D: Doublet, FS: Frequency sweep, MS: Multi sign

Also, the states vector $x = [u \quad \alpha \quad q \quad \theta]$ denotes longitudinal speed, angle of attack, pitch rate and flight pitch angle respectively. The control inputs $u_{inp} = [\delta e \ \ \delta_{th}]$ represent elevator and throttle. For simplifying, the matrix A and B in Equations 2 and 3 are expressed as Equations 4 and 5, respectively. This simplification is accepted for aircraft dynamics and is documented in reputable aerospace references (29, 30), as the effect of the omitted terms is negligible compared to the retained terms. Since the objective is to determine the effects of aerodynamic terms on the forces and moments applied to the aircraft and their respective frequency ranges, this simplification provides an appropriate approximation of each term independently.

$$A_{simplified} = \begin{bmatrix} X_u & X_\alpha & X_q & -g \\ \frac{Z_u}{U_1} & \frac{Z_\alpha}{U_1} & 1 & 0 \\ M_u & M_\alpha & M_q & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (4)$$

$$B_{simplified} = \begin{bmatrix} X_{\delta e} & X_T \\ Z_{\delta e} & Z_T \\ M_{\delta e} & M_T \\ 0 & 0 \end{bmatrix} \quad (5)$$

The coefficients in the aforementioned equations can be estimated through computational fluid dynamics (CFD) analysis (31, 32), wind tunnel testing (33, 34), or empirical methods such as DATCOM and vortex lattice methods like AVL (35-38). Although these techniques provide initial estimates for the system parameters, the resulting values often lack the precision required for high-fidelity modeling. The primary objective of system identification is to refine these parameter estimates accurately by leveraging flight test data. To this end, excitation inputs must be strategically designed and applied to ensure the system response comprehensively captures the relevant dynamic modes. This is achieved by utilizing the initial model to calculate the magnitude of these dynamic terms across the frequency range of interest. In the present study, longitudinal data from a scaled model of a Bonanza fixed-wing aircraft is employed. The key specifications of the aircraft under trim conditions are summarized in Table 2.

TABLE 2. Bonanza scale model aircraft reference data

Parameter	Value	Unit
Mass	8.5	Kg
Wing span	2.9	M
Wing area	0.55	M^2
Mean aerodynamic chord	0.19	M
Trim altitude (MSL)	2000	M
Trim speed (U1)	25	m/s

Hence, Equations 6 to 8 provide the values of the state-space matrices A and B and the aircraft dynamic mode poles, where sp and ph denote short period and phugoid modes respectively. In this study, these values were obtained using the DATCOM method, which provides an initial approximation of the model. This approximation is suitable for initial estimating the desired frequency range for exciting the aircraft modes.

$$A = \begin{bmatrix} -0.1107 & 4.1015 & 0 & -9.8 \\ -0.0312 & -4.5948 & 0.9695 & 0 \\ 0.0360 & -74.927 & -5.370 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \quad (6)$$

$$B = \begin{bmatrix} -0.2747 & 0 \\ -0.1052 & 0 \\ -28.9236 & -0.094 \\ 0 & 0 \end{bmatrix} \quad (7)$$

$$\begin{cases} s_{1.2} = -4.99 \pm 8.51 \, , \zeta_{sp} = 0.5 \ , \omega_{n_{sp}} = 9.87 \frac{rad}{s} \\ s_{3.4} = -0.047 \pm 5.0 \, , \zeta_{ph} = 0.09 \, , \omega_{n_{ph}} = 0.5 \frac{rad}{s} \end{cases} \quad (8)$$

3. CALCULATION OF DESIRED FREQUENCY RANGE

The design of excitation inputs for system identification and parameter estimation of fixed-wing aircraft follows a systematic three-stage methodology. The initial stage involves determining the frequency range essential for accurate parameter estimation. In the second stage, the most effective control surfaces (inputs) for exciting the aircraft's dynamics are selected. The final stage entails synthesizing specific excitation signals that adequately cover the designated frequency bandwidth. The target frequency range is determined by analyzing the contribution of each aerodynamic parameter to the force and moment equations. This analysis is performed using Bode diagrams, which pinpoint the frequency bands where the influence of each parameter is most observable. This ensures that the designed excitation inputs effectively target these critical frequencies. Furthermore, comparative analysis of Bode diagrams across different control inputs facilitates the selection of the optimal input channel for estimating each specific parameter. The frequency response of each state-space parameter is calculated using the transfer function approach outlined in Equation 9. For example, for the parameter M_α, the magnitude of the term $\left|(M_\alpha \alpha) \ / \ u_{inp}\right|$ is computed accordingly (39).

$$\begin{cases} y = \alpha = Cx + Du \\ \alpha = [0\ 1\ 0\ 0] * [u\ \alpha\ q\ \theta]' + [0\ 0] * [\delta e\ T] \\ M_\alpha \alpha = [0\ M_\alpha\ 0\ 0] * x + [0\ 0] * u_{inp} \\ C = [0\ A(3.3)\ 0\ 0] = [0\ M_\alpha\ 0\ 0] \\ D = [0\ 0] \end{cases} \quad (9)$$

So, the $\left|M_\alpha \alpha / u_{inp}\right|$ can be calculated using space state:

$$\left|M_\alpha \alpha / u_{inp}\right| = |C(SI - A)^{-1}B + D| \quad (10)$$

Which A and B matrices has been presented in Equations 2 to 7. It is worth noting that Equation 10 can be calculated using (ss2tf(A, B, C, D)) in MATLAB. Also, the term A12=$M_\alpha + M_{T\alpha} + \frac{M_{\dot{\alpha}} Z_\alpha}{U_1 - Z_{\dot{\alpha}}}$ is simplified as M_α.

Similar to Equations 9 and 10, the magnitude of the terms corresponding to other state-space parameters is calculated in the same manner as Table 3.

The frequency responses of the parameter magnitudes for both control inputs—the elevator and engine thrust—are presented in the Bode diagram shown in Figure 1. In this figure, input 1(In(1)) corresponds to the elevator input, and input 2 (In(2)) corresponds to the engine throttle. It should be noted that the frequency range determined in Figure 1, corresponds to the elements of the state-space matrix A. For example, the element A_{33} comprises the sum of $M_q + M_\alpha$. These terms must be separated and distinguished post-identification using mathematical methods.

TABLE 3. State-Space Representation for Individual Parameter Effects

Row	Parameter	Simplified	C Matrix	D Matrix
1	$\left\|\frac{\dot{u}}{u_{inp}}\right\|$	$\left\|\frac{\dot{u}}{u_{inp}}\right\|$	$[A_{11}\ A_{12}\ A_{13}\ A_{14}]$	$[B_{11}\ B_{12}]$
2	$\left\|\frac{A_{11}u}{u_{inp}}\right\|$	$\left\|\frac{X_u u}{u_{inp}}\right\|$	$[A_{11}\ 0\ 0\ 0]$	$[0\ 0]$
3	$\left\|\frac{A_{12}\alpha}{u_{inp}}\right\|$	$\left\|\frac{X_\alpha \alpha}{u_{inp}}\right\|$	$[0\ A_{12}\ 0\ 0]$	$[0\ 0]$
4	$\left\|\frac{A_{13}q}{u_{inp}}\right\|$	$\left\|\frac{X_q q}{u_{inp}}\right\|$	$[0\ 0\ A_{13}\ 0]$	$[0\ 0]$
5	$\left\|\frac{B_{11}u_{inp}}{u_{inp}}\right\|$	$\|X_{\delta e}\|$	$[0\ 0\ 0\ 0]$	$[B_{11}\ 0]$
6	$\left\|\frac{\dot{\alpha}}{u_{inp}}\right\|$	$\left\|\frac{\dot{\alpha}}{u_{inp}}\right\|$	$[A_{21}\ A_{22}\ A_{23}\ A_{24}]$	$[B_{21}\ B_{22}]$
7	$\left\|\frac{A_{21}u}{u_{inp}}\right\|$	$\left\|\frac{Z_u u}{u_{inp}}\right\|$	$[A_{21}\ 0\ 0\ 0]$	$[0\ 0]$
8	$\left\|\frac{A_{22}\alpha}{u_{inp}}\right\|$	$\left\|\frac{Z_\alpha \alpha}{u_{inp}}\right\|$	$[0\ A_{22}\ 0\ 0]$	$[0\ 0]$
9	$\left\|\frac{A_{23}q}{u_{inp}}\right\|$	$\left\|\frac{Z_q q}{u_{inp}}\right\|$	$[0\ 0\ A_{23}\ 0]$	$[0\ 0]$
10	$\left\|\frac{B_{21}u_{inp}}{u_{inp}}\right\|$	$\|Z_{\delta e}\|$	$[0\ 0\ 0\ 0]$	$[B_{21}\ 0]$
11	$\left\|\frac{\dot{q}}{u_{inp}}\right\|$	$\left\|\frac{\dot{q}}{u_{inp}}\right\|$	$[A_{31}\ A_{32}\ A_{33}\ A_{34}]$	$[B_{31}\ B_{32}]$
12	$\left\|\frac{A_{31}u}{u_{inp}}\right\|$	$\left\|\frac{M_u u}{u_{inp}}\right\|$	$[A_{31}\ 0\ 0\ 0]$	$[0\ 0]$
13	$\left\|\frac{A_{32}\alpha}{u_{inp}}\right\|$	$\left\|\frac{M_\alpha \alpha}{u_{inp}}\right\|$	$[0\ A_{32}\ 0\ 0]$	$[0\ 0]$
14	$\left\|\frac{A_{33}q}{u_{inp}}\right\|$	$\left\|\frac{M_q q}{u_{inp}}\right\|$	$[0\ 0\ A_{33}\ 0]$	$[0\ 0]$
15	$\left\|\frac{B_{31}u_{inp}}{u_{inp}}\right\|$	$\|X_{\delta e}\|$	$[0\ 0\ 0\ 0]$	$[B_{31}\ 0]$
16	$\left\|\frac{B_{32}u_{inp}}{u_{inp}}\right\|$	$\|M_T\|$	$[0\ 0\ 0\ 0]$	$[0\ B_{32}]$

When parameters exhibit overlapping magnitudes in a Bode magnitude plot, their differentiation becomes challenging. In such cases, phase plots provide a powerful complementary tool for discrimination, as they often reveal unique phase behaviors—such as distinct phase shifts or steep phase slopes near natural frequencies—that are characteristic of individual parameters. By synergistically combining both magnitude and phase information, a unique dynamic signature for each parameter can be established, greatly aiding the identification process. However, challenges persist when parameters exhibit similar phase characteristics or when phase measurements are corrupted by noise. Under these conditions, the application of advanced techniques, such as modal analysis or time-domain parameter estimation, becomes necessary. Employing high-resolution frequency data alongside sophisticated computational tools can further enhance the accuracy and reliability of parameter differentiation.

As evidenced by the Bode diagram in Figure 1, the optimal frequency range for parameter estimation can be effectively determined. A comparison of the parameter magnitudes in response to elevator versus throttle inputs reveals that the elevator generates significantly larger magnitudes across key parameters. This indicates the superior effectiveness of the elevator for exciting the system dynamics and achieving accurate parameter identification.

Furthermore, it is observed that certain parameters, such as specific control derivatives, exhibit inherently small magnitudes, rendering them more challenging to identify.

The frequency spectrum can be partitioned into two distinct regions corresponding to the aircraft's short-period and phugoid (long-period) dynamic modes (Equation 8). Consequently, the elevator has been selected as the primary control input for excitation, and the target frequency range for input design is determined to be 0.1 to 10 rad/s (base on Figure 1), as summarized in Table 4.

4. EXCITATION INPUT

The primary function of an excitation input is to actively perturb the aircraft's dynamic modes. To ensure effective and safe excitation, the input must be designed to maintain the aircraft within a linear range around its trim conditions. This requirement implies that the input amplitude must remain within predefined limits, and the input signal must begin and end at the trim value to avoid introducing a net change in the aircraft's state.

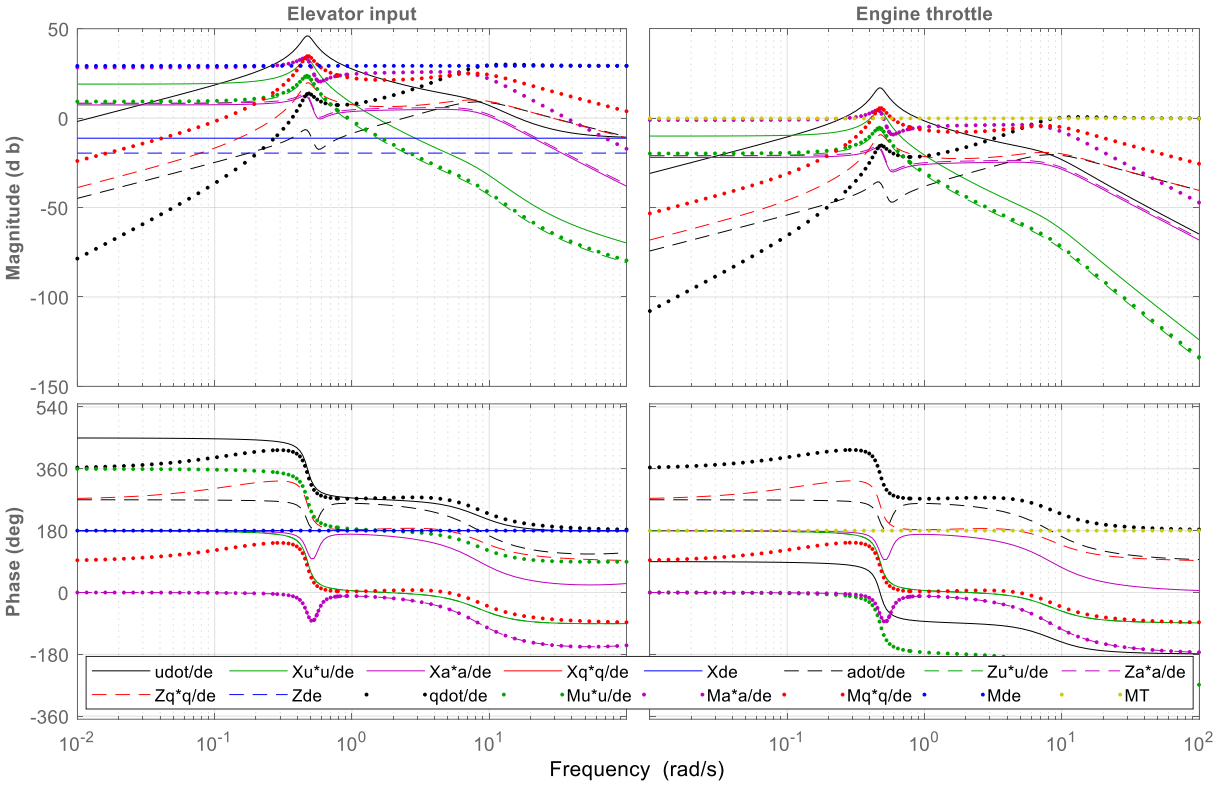


Figure 1. Frequency Response (Magnitude & Phase) of Individual Parameters for Different Control Inputs

TABLE 4. Determining frequency range for aircraft dynamic excitation

Parameter	Simplified	Frequency range (rad/s)
A_{11}	X_u	0.1 to 0.7
A_{12}	X_α	0.1 to 10
A_{13}	X_q	---
A_{21}	Z_u	0.1 to 0.5
A_{22}	Z_α	0.1 to 10
A_{23}	Z_q	0.2 to 10
A_{31}	M_u	0.1 to 0.7
A_{32}	M_α	0.2 to 10
A_{33}	M_q	0.2 to 10
B_{31}	$M_{\delta e}$	All frequency Range
$A_{11}.A_{21}.A_{31}$	$X_u.Z_u.M_u$	Phugoid: 0.1 to 0.7
$A_{12}.A_{22}.A_{23}.$ $A_{32}.A_{33}$	$X_\alpha.Z_\alpha.Z_q.$ $M_\alpha.M_q$	Short period: 8 to 10

Consequently, a conventional step input is generally unsuitable for flight test applications due to its persistent offset. Alternative signals, synthesized from combinations of pulse functions—such as pulse, doublet, and multi-step inputs—are employed instead. It is critical that the selected input possesses sufficient energy spectral density across the targeted frequency band to adequately excite the relevant system dynamics. For inputs derived from pulse combinations, the total energy of the signal can be quantified using Equation 11 (39).

$$E(w) = \frac{2\Delta t^2(1-\cos\Omega)}{\Omega^2}\left[\sum_{i=1}^{N} V_i^2 + 2\sum_{j=1}^{N-1}\cos j\Omega \sum_{i=1}^{N-j} V_i V_{i+j}\right]. \quad \Omega > 0 \tag{11}$$

where $\Omega = \omega\Delta t$ is the normalized frequency, $T = N\Delta t$ is the total duration of the input consisting of N impulses each of duration Δt, and V_i is the amplitude for the current impulse. Consequently, for each input, the energy versus frequency graph is plotted to determine which input provides suitable energy within the desired frequency range. Based on an established empirical criterion, the frequency band over which the signal's energy exceeds 50% of its maximum value can be defined as the effectively excited frequency range (39).

The amplitude and the time interval Δt for an input can be determined based on the load factor applied to the aircraft and the deviation from trim conditions. The load factor should not exceed the allowable limit specified in the aircraft flight manual. Additionally, the applied excitation should not drive the aircraft's flight conditions significantly away from the pre-excitation state. In this

study, these considerations have been evaluated through simulation, and accordingly, the amplitude and time interval have been appropriately limited.

4. 1. Pulse Input A pulse input is one of the simplest and effective excitation signals used in aircraft system identification. By applying a short-duration control surface deflection, the aircraft dynamics are excited over a frequency range, allowing the transient response to reveal information about the underlying aerodynamic and stability derivatives. This method is particularly attractive due to its ease of implementation during flight tests and its ability to provide data for parameter estimation with minimal pilot workload. The pulse input is illustrated in Figure 2. To design this input, two parameters, Δt and amplitude A, must be determined. In Figure 2, the energy of this input is shown for different A and Δt values.

The amplitude of the excitation input must be constrained to a range that prevents significant deviation from trim conditions and ensures the resulting load factor remains within acceptable structural and safety limits. Simulation results for a pulse input with a duration of $\Delta t = 1.2$ s and an amplitude of $A = 0.2$ radians are presented in Figure 3. It is evident that employing a larger input amplitude induces a higher load factor and causes substantial deviation from the nominal trim condition. The Power Spectral Density (PSD) analysis in Figure 3 indicates that the pulse input is particularly suitable for exciting low-frequency dynamics, such as the phugoid mode. Furthermore, the total energy of the pulse input, calculated from the aircraft's pitch rate response, is 19.11 $(\text{rad/s})^2$. Each figure consists of six subplots, arranged in 3 rows and 2 columns. The first row, first column shows the aircraft velocity in meters per second; the first row, second column shows the pitch rate in degrees per second; the second row, first column depicts the load factor applied to the aircraft; the second row, second column illustrates the aircraft pitch angle in degrees; the third row, first column shows the signal energy or power spectral density across different frequencies; and the third row, second column represents the input applied to the aircraft. The inputs are applied starting from the 10th second, with the aircraft trimmed during the initial 10 seconds. The data prior to this period are not shown for clarity.

4. 2. Doublet Input A doublet input is constructed by summing two consecutive pulse signals. This input waveform, along with its spectral energy distribution for a specific amplitude and time step, is illustrated in Figure 4. The horizontal axis represents the normalized frequency, ωΔt, which renders the plot dimensionless and independent of the specific time step value.

In Figure 4, the analysis focuses on the solid curves (Amplitude = 0.2 rad). Applying the previously mentioned empirical criterion, a suitable effective frequency range is identified from approximately 1.5 to 3.5 rad/s. This relationship can be expressed as:

$$\omega\Delta t = 1.5 \rightarrow \Delta t = \frac{1.5}{\omega}$$
$$\omega\Delta t = 3.5 \rightarrow \Delta t = \frac{3.5}{\omega} \quad (12)$$

Analysis of the initial model indicates a short-period natural frequency of approximately 10 rad/s and a phugoid frequency near 0.5 rad/s. Consequently, the required Δt values for effective mode excitation range from 0.15 s to 0.35 s for the short-period mode, and from 3 s to 7 s for the phugoid mode. Δt values smaller than 0.5 s produce insufficient energy magnitude in the spectral plot and fail to excite the dynamics adequately.

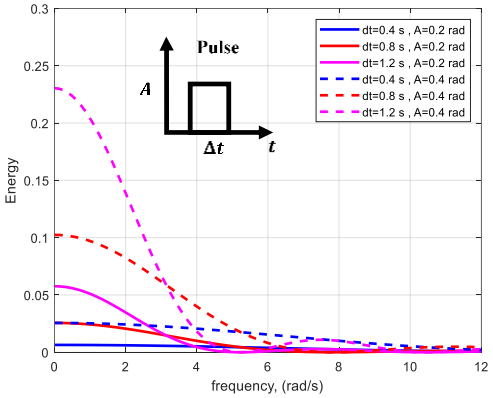


Figure 2. Pulse input and its energy across the frequency range for various Δt and A

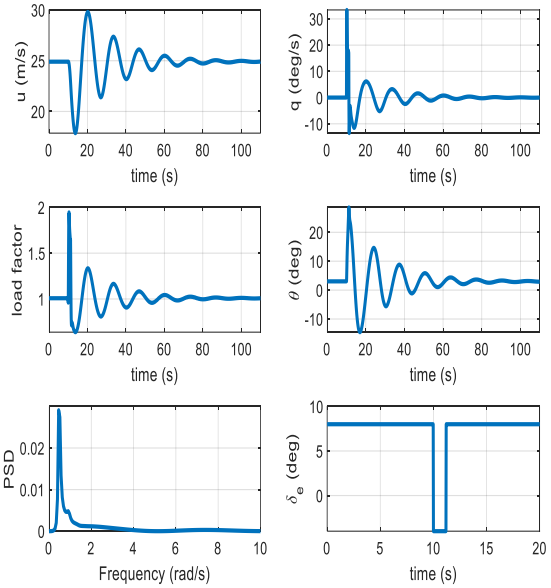


Figure 3. Pulse input (Δt=1.2 sec. and A=0.2 rad) and the aircraft response

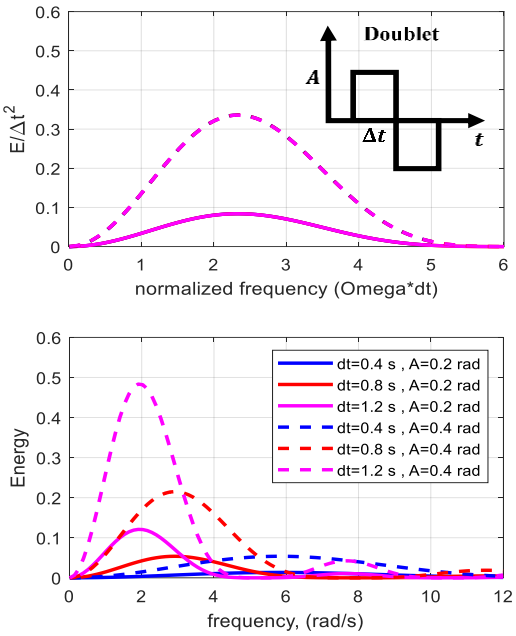


Figure 4. Doublet input and its energy across the frequency range for various Δt and A

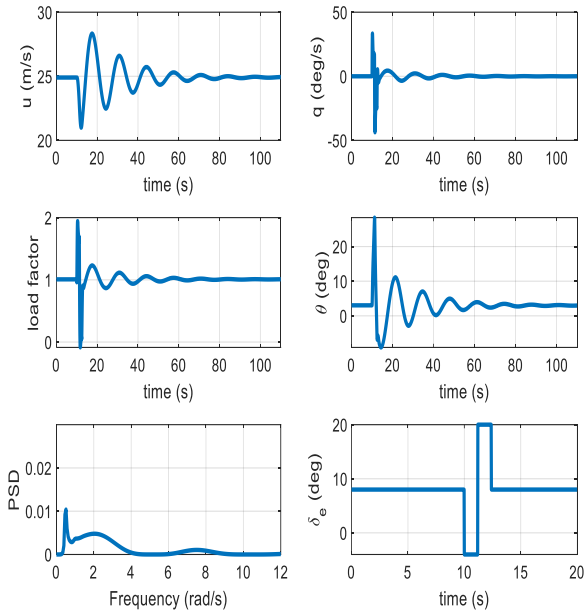


Figure 5. Doublet input (Δt=1.2 sec. and A=0.2 rad) and the aircraft response

Excessively large Δt values, conversely, cause significant deviations from trim conditions. Simulation results in Figure 5 suggest an optimal compromise at $\Delta t \approx 1.2$ s, which effectively excites frequencies in the 1 to 3 rad/s range. This makes the doublet input suitable for this specific bandwidth, though it lacks the capability to excite the short-period mode for this particular aircraft. According to Table 4, however, it remains applicable for estimating certain phugoid mode parameters. The total energy of the doublet input in Figure 5, calculated from the aircraft pitch rate response, is 27.17 $(rad/s)^2$.

4. 3. Multi Step The multi-step input is synthesized from a sequence of multiple pulse signals. The most prevalent variant is the 3211 input, which integrates both positive and negative pulses. In contrast to simple pulse and doublet inputs, this waveform is designed to excite a broader frequency spectrum. Figure 6 presents this input signal alongside its corresponding energy distribution for various amplitudes and time steps (Δt).

Consistent with the methodology applied to the doublet input in the previous section, the effective frequency range for the 3211 input is defined by the bandwidth where its normalized energy exceeds 50% of its maximum value. Analysis indicates that this corresponds to a normalized frequency (ωΔt) range of approximately 0.3 to 2.7. Consequently, by identifying the target frequency (ω) for excitation, the appropriate time step (Δt) can be systematically determined.

Numerous combinations of amplitude and Δt for the 3211 input were investigated through simulation.

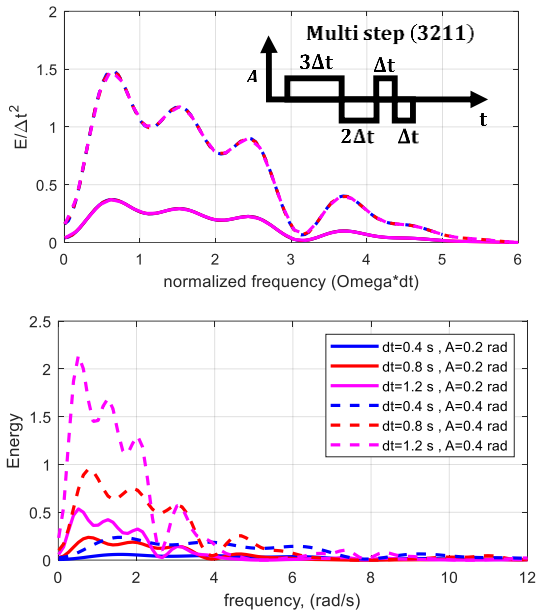


Figure 6. Multi step input and its energy across the frequency range for various Δt and A

Considering amplitude constraints inherent to flight testing, a configuration with an amplitude of 8 degrees and $\Delta t = 1\ s$ was selected for implementation. The application results of this specific input to the aircraft are depicted in Figure 7. The results demonstrate that while this input excites a wider frequency range compared to simpler inputs, the energy amplitude across this spectrum

remains relatively low. For the specific aircraft configuration in this study, the 3211 input proved ineffective in exciting the short-period mode, primarily exciting frequencies below 5 rad/s. The total energy of the multi-step input in Figure 7, calculated from the aircraft's pitch rate response, is 47.22 $(rad/s)^2$.

4. 4. Frequency Sweep The frequency-sweep technique is extensively employed to excite an aircraft's dynamic response across a broad frequency spectrum. This method utilizes a continuous input signal, typically a sinusoid with a linearly or exponentially varying frequency over time, facilitating the comprehensive identification and analysis of key dynamic characteristics, including natural frequencies and damping ratios. Frequency-sweep excitation is particularly advantageous when preliminary information regarding the aircraft's dynamics is limited or unavailable. The mathematical formulation for the frequency-sweep input is given by (6):

$$\begin{cases} u_{sweep} = Asin(\theta(t)) \\ \theta(t) \equiv \int_0^T \omega(t)dt \\ \omega = \omega_{min} + K(\omega_{max} - \omega_{min}) \\ K = C_2\left[\exp\left(\frac{C_1 t}{T}\right) - 1\right] \end{cases} \quad (13)$$

where ω_{min}, ω_{max} and T denote the minimum frequency, maximum frequency and total excitation duration, respectively. The values $C_1 = 4.0$ and $C_2 = 0.0187$ have been established as suitable for a wide range of applications (6). The time-history of the frequency-sweep input for a duration of $T = 50\ s$ and a frequency range from 0.3 to 12 rad/s is illustrated in Figure 8.

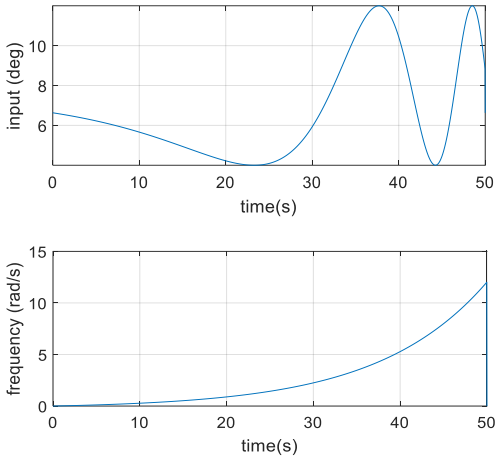


Figure 8. Frequency sweep input and its frequency progress

The system response to the applied frequency-sweep input (Figure 8) is presented in Figure 9. A primary limitation of this method is the constraint on input amplitude, which must be restricted to prevent significant deviation from trim conditions. For the specific aircraft configuration in this study, the resulting excitation amplitude across the frequency spectrum was found to be insufficient for robust identification of all modes. Consequently, while the input excites a wide frequency band, the energy distribution is not adequate for high-quality parameter estimation, particularly for modes requiring stronger excitation. The total energy of the frequency-sweep input in Figure 9, computed from the aircraft's pitch rate response, is 31.30 $(rad/s)^2$.

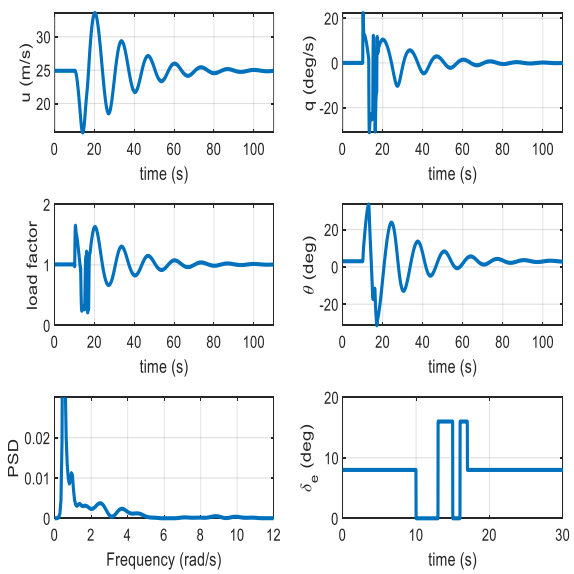


Figure 7. 3211 input (Δt=1 sec. and A=0.14 rad) and the aircraft response

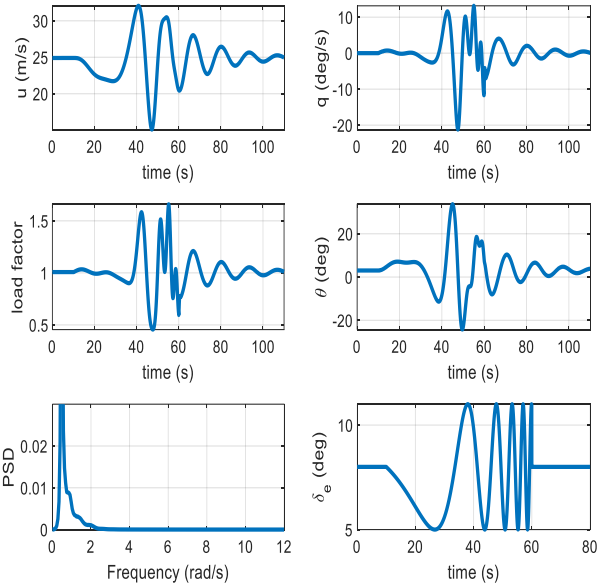


Figure 9. Frequency sweep input (T=50 sec and A=0.0524 rad) and the aircraft response

4. 5. Orthogonal Optimized Multisine Morelli (7) introduced an Orthogonal Optimized Multisine (OOM) input is. The authors claim that this input can be employed without significant operational limitations during flight and effectively excites the desired frequency range. This signal can be superimposed as a disturbance on the control surfaces while maintaining the aircraft near its trim conditions. Further comprehensive details regarding its formulation are provided by Morelli (7). The design equation for this input is given by Equation 14:

$$u = \sum_{k\in 1.2.....M} A\sqrt{P_k}\sin\left(\frac{2\pi kt}{T}+\phi_k\right) \quad (14)$$

where M is the total number of available harmonic frequencies, T is the excitation time length, A is the multisine amplitude, P_k is the power fraction for the kth sinusoidal component, and t is the time vector. The frequency $\omega_M = 2\pi M/T$ defines the upper limit of the excitation frequency band, and $\omega_1 = 2\pi/T$ defines the lower limit. In general, the superposition of harmonic components can result in large amplitude excursions, which is undesirable as it may displace the aircraft from its reference flight condition. To mitigate this, the phase angles ϕ_k in Equation 14 are optimized to minimize the Relative Peak Factor (RPF), defined as follows (7).

$$RPF(u) = \frac{[\max(u)-\min(u)]}{2\sqrt{2}\,rms(u)} \quad (15)$$

The optimal RPF value is unity. Based on Figure 1 and Table 4, a frequency range of [0.3, 10] rad/s with $T = 20\,s$ was utilized. The phase angles ϕ_k for RPF minimization were computed using a genetic algorithm. Specific details of the designed excitation input are summarized in Table 5.

The designed input signal and its frequency spectrum are shown in Figure 10. The aircraft's response to this input is presented in Figure 11. The excitation signal was combined with the elevator control input to form the total command. The total energy of the OOM input in Figure 11, calculated from the aircraft's pitch rate response, is 21.89 (rad/s)^2.

It is observed that this input successfully excites the target frequency range. Unlike conventional inputs, the OOM input enables more optimal excitation across various frequencies. Furthermore, the excitation level can be enhanced by increasing the input amplitude without necessarily compromising trim conditions.

TABLE 5. Elevator excitation input specification

A (deg)	P_k	k	$\phi_k(rad)$	$\omega(r/s)$	RPF
6	0.1	1	-3.626	0.3142	1.2
		4	3.202	1.2566	
		8	6.033	2.5133	
		11	2.911	3.4558	
		15	4.724	4.7124	
		18	0.4512	5.6549	
		22	-0.1469	6.9115	
		25	1.735	7.8540	
		29	-0.2262	9.1106	
		32	-2.035	10.053	

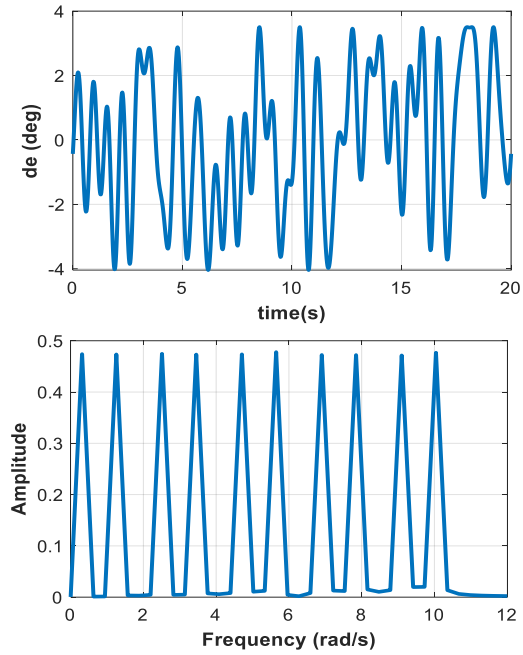


Figure 10. Designed input signal and its specific frequency

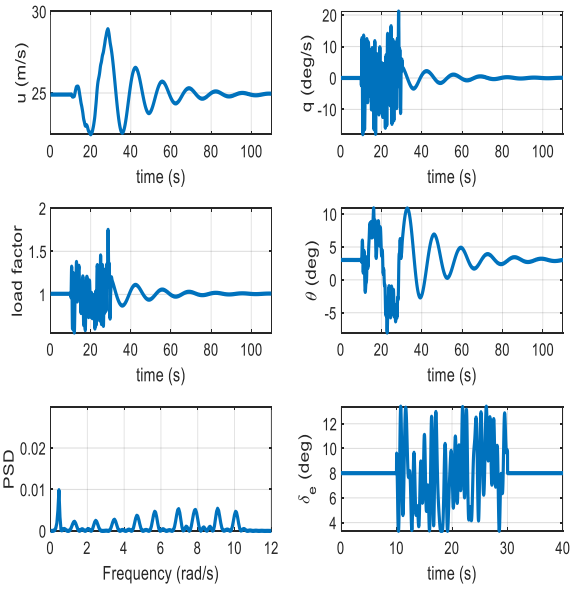


Figure 11. OOM input (T=20 sec and A=0.07 rad) and the aircraft response

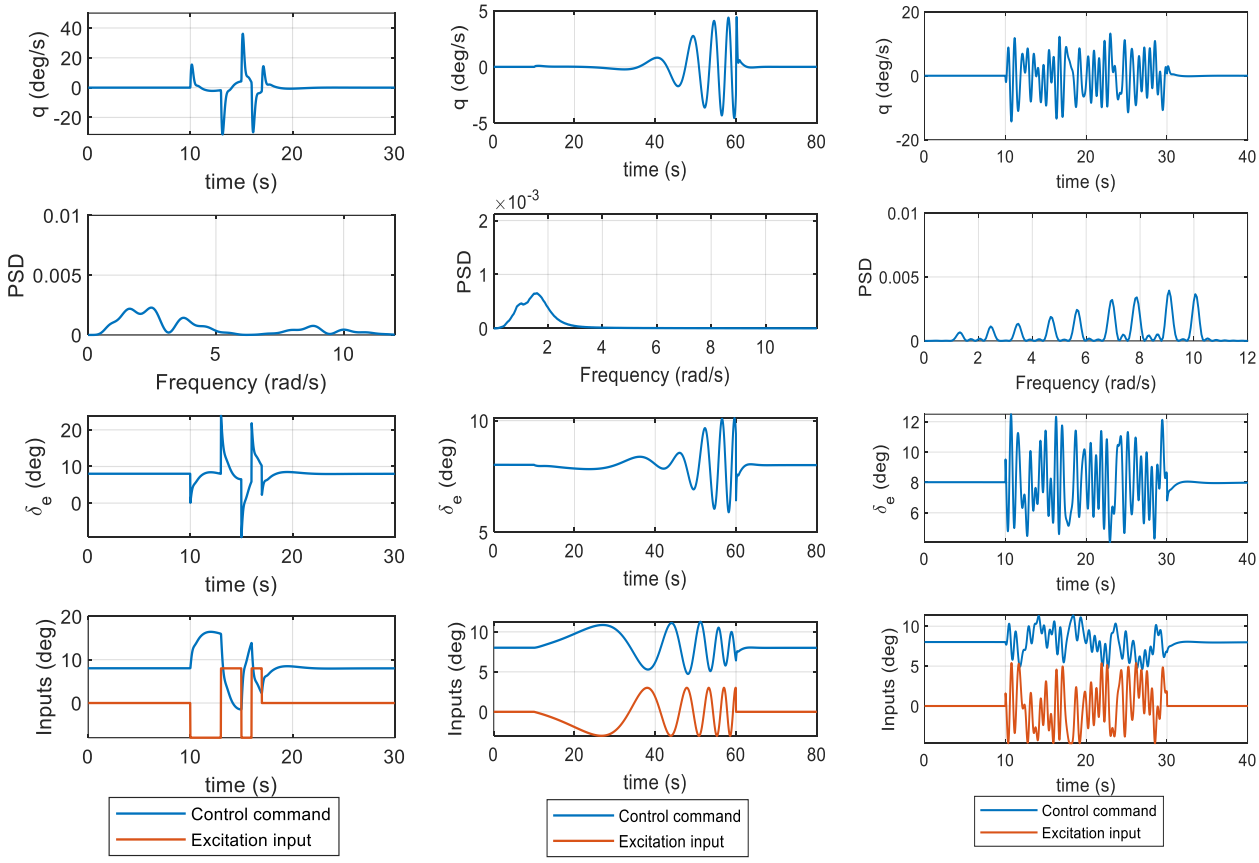
As observed in this section, under open-loop excitation, all input signals are capable of stimulating the aircraft dynamics; however, the OOM, 3211, and frequency sweep inputs can excite a wider frequency range. Among these, the OOM input provides a more optimal response, as it effectively excites higher frequencies and the short-period mode.

5. CLOSED LOOP SYSTEM EXCITATION

The investigations presented thus far have focused on exciting the aircraft in open-loop configuration, without considering the presence of stability augmentation or flight control systems. However, for many modern aircraft, mission profiles and operational requirements necessitate the continuous engagement of control systems during flight. Under these conditions, the active controller significantly alters the effective excitation inputs commanded to the aircraft. In Figures 12a to 12c, the inputs designed in the previous sections are applied to the aircraft with an active control system. The results demonstrate the controller's effect on the inputs. Pulse and doublet inputs have little impact, as they are easily damped and ineffective at exciting the closed-loop system, especially at higher frequencies. Since the multi-step input encompasses these signals, only the multi-step, sinusoidal, and OOM inputs were analyzed.

One established method to mitigate the controller's influence is to inject the excitation signal directly to the control surfaces, bypassing the closed-loop controller; this technique is referred to as Surface Servo Excitation (SSE). An alternative approach involves reducing the controller gains. For the purposes of this study, the SSE method has been employed.

Analysis of the results indicates that the control system's attenuation effect diminishes as the input frequency increases. Among the tested inputs, the Orthogonal Optimized Multisine (OOM) input exhibits the greatest robustness to controller interference. Furthermore, transformation of the controller-affected OOM input into the Fourier domain confirms that higher-frequency components remain effectively excited. This demonstrates the OOM input's superior performance in closed-loop identification scenarios.



(a) Closed loop aircraft response to 3211 excitations

(b) Closed loop aircraft response to frequency sweep excitation

(c) Closed loop aircraft response to OOM excitation

Figure 12. Closed loop aircraft response to various excitation inputs

A comparative summary of the excitation inputs investigated in this article is presented in Table 6. To facilitate a quantitative comparison, a performance indices system ranging from 0, 5 and 10 is utilized, where a score of 10 indicates full satisfaction of a criterion, and a score of 0 denotes a failure to meet it. A score of 5 means the criterion is acceptable but needs some improvements or additional review.

This scoring is based on the simulation results, the capability of the input signals to excite the aircraft dynamics, and the targeted excitation range shown Figure 1. For example, as illustrated in Figure 4, the doublet input is capable of exciting the frequency up to 3.5 rad/s, which is suitable for stimulating the phugoid mode of the considered aircraft (Figure 1). This behavior is also demonstrated in the simulation results presented in Figure 5; therefore, a score of 10 has been assigned to this case.

As the table shows, the OOM input scores higher overall than the other inputs, demonstrating its superiority. It should be noted that the OOM input cannot be applied manually by a pilot and must instead be computationally pre-programmed and injected into the system automatically.

It is important to note that the results presented in this study are specific to the aircraft configuration investigated. In this paper, an optimal process for excitation input design has been developed and implemented for the aircraft under study. Although the obtained numerical results are not directly applicable to other aircraft, the proposed input design process itself has a general and adaptable structure that can be utilized for different airframes. To apply this approach to another aircraft, the steps described in Sections 2 and 3 should first be performed for the target aircraft to derive its corresponding mathematical model and to determine the desired frequency range to be excited. Then, the same optimization-based procedure introduced in this paper can be followed to design the excitation inputs. Naturally, the numerical outcomes will differ depending on the aircraft's characteristics; however, the optimal procedure for input design remains identical.

TABLE 6. Comparison of the types of excitation inputs for considered aircraft

Criteria	Pulse	Doublet	Multi step	Frequency sweep	OOM
Phugoid excitation capability	10	10	10	10	10
Short period excitation capability	0	0	5	5	10
Open loop excitation	10	10	10	10	10
Closed loop excitation	0	0	5	10*	10
Appling continuously during flight**	0	0	0	0	10
Ease of the use	10	10	10	5	5

* Just in high frequencies
** base on reference (7)

6. CONCLUSION

This study proposed a systematic and optimized framework for designing excitation inputs for longitudinal system identification of a fixed-wing MIMO aircraft under both open-loop and closed-loop configurations. A scaled Bonanza aircraft was used as the test case, and several excitation signals—pulse, doublet, multi-step, frequency sweep, and Orthogonal Optimized Multisine (OOM)—were evaluated through high-fidelity MATLAB/Simulink simulations to identify the most effective input strategy.

Results show that multi-step and OOM inputs provide the best excitation performance, yielding the most accurate identification of longitudinal dynamics while minimizing flight test effort. Frequency-domain analysis confirmed that the elevator input is more effective than the throttle for exciting longitudinal modes; thus, all final tests were performed using elevator actuation.

Based on the simulation results, experimental observations, and supporting literature, the following Practical recommendations are proposed:

- The findings presented here are specific to the aircraft configuration studied. The design process should be repeated tailored to each unique aircraft system.
- Excitation inputs should ideally be designed to initially induce a pitch-up motion.
- The exact time-domain form of the input is not critical. Minor asymmetries during manual application may even be beneficial in certain cases.
- All inputs must be applied around trimmed flight conditions. It is recommended to allow a 10-second trim period before and after each input application.
- The OOM input can be superimposed as a small-amplitude disturbance during closed-loop operation for continuous excitation, provided its amplitude remains within safe limits.
- Based on the comprehensive analysis, the use of OOM and multi-step inputs is strongly recommended for the subject aircraft. In particular, the OOM input proves highly suitable for closed-loop identification

7. FUTURE WORK

In this paper, the process of designing aircraft excitation signals for MIMO, nonlinear, opened-loop and closed-loop systems was discussed, along with the main criteria that should be considered. In future work, we intend to develop an intelligent and automated framework that can perform this process autonomously based on the defined objective. The proposed intelligent system will evaluate performance indices and automatically design the optimal excitation input according to the target identification goal.

Acknowledgements

The author appreciates Department of Mechanical Engineering, University of Isfahan for providing necessary facilities to conduct present research.

Funding

This work was supported by Aviation Research and Education Center (AREC), Department of Aerospace Engineering, University of Isfahan, Isfahan, Iran.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The author declares no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this manuscript, the author used ChatGPT and Grok exclusively for minor language editing to improve readability. After using this tool, the author carefully reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Authors Biosketches

Mahdi Mortazavi is an Associate Professor in the Department of Aerospace Engineering, Faculty of Engineering, University of Isfahan. He received his Ph.D. in Aerospace Engineering from the Moscow Aviation Institute (MAI), Russia. His research interests include flight mechanics, control, guidance, and navigation.

Negin Sayyaf is an Assistant Professor in the Department of Electrical Engineering, Faculty of Engineering, University of Isfahan. She received her Ph.D. in control engineering from Sharif University of Technology. Her research interests include reinforcement learning, deep learning and artificial intelligence.

Seyed Abdorreza Askari is an Adjunct Professor at the Faculty of Engineering, University of Isfahan. He received his degree in Aerospace Engineering from Amirkabir University of Technology. His research interests include Flight mechanics and control.

Abdollah Mohseni is a Ph.D. candidate in Mechanical Engineering in the Department of Mechanical Engineering, University of Isfahan. His research interests include artificial intelligence, Flight mechanics and control, and system identification.

REFERENCES

1. Ayyad A, Chehadeh M, Awad MI, Zweiri Y. Real-time system identification using deep learning for linear processes with application to unmanned aerial vehicles. IEEE Access. 2020;8:122539-53. https://doi.org/10.1109/ACCESS.2020.3006277.
2. Belge E, Kaba H, Parlak A, Altan A, Hacıoğlu R. Estimation of small unmanned aerial vehicle lateral dynamic model with system identification approaches. Balkan Journal of Electrical and Computer Engineering. 2020;8(2):121-6. https://doi.org/10.17694/bajece.654499.
3. Hosseini B, Steinert A, Hofmann R, Fang X, Steffensen R, Holzapfel F, et al. Advancements in the theory and practice of flight vehicle system identification. Journal of Aircraft. 2023;60(5):1419-36. https://doi.org/10.2514/1.C037269.
4. Matt JJ, Chao H, Shawon MH, Svoboda BC, Hagerott SG. System Identification for Small Flying-Wing Unmanned Aircraft Using Open-Loop and Closed-Loop Flight Data. Journal of Aircraft. 2025:1-17. https://doi.org/10.2514/1.C038147.
5. Morelli EA, Klein V. Aircraft System Identification: Theory and Practice: Sunflyte Enterprises; 2016. https://doi.org/10.1017/S0001924000087194.
6. Tischler MB, Remple RK. Aircraft and rotorcraft system identification: American Institute of Aeronautics and Astronautics Reston; 2012. https://doi.org/10.2514/4.868207.
7. Morelli EA. Practical aspects of multiple-input design for aircraft system identification flight tests. AIAA Aviation 2021 Forum. https://doi.org/10.2514/6.2021-2795.
8. Lichota P, Ohme P, Sibilski K. Simultaneous multi-step excitations for aircraft system identification. AIAA Atmospheric Flight Mechanics Conference; 2016. https://doi.org/10.2514/6.2016-3704.
9. Gim H, Lee B, Huh J, Kim S, Suk J. Longitudinal system identification of an avian-type UAV considering characteristics of actuator. International Journal of Aeronautical and Space Sciences. 2018;19:1017-26. https://doi.org/10.1007/s42405-018-0084-5.

10. Aslandoğan OH. Development of a Rotorcraft Time Domain System Identification Software: Middle East Technical University (Turkey); 2023. https://hdl.handle.net/11511/101957.
11. Han Y, Kim Y. Optimal input design for online parameter estimation for aircraft with multiple control surfaces. Engineering Optimization. 2011;43(5):559-80. https://doi.org/10.1080/0305215X.2010.502937.
12. Gresham JL, Simmons BM, Fahmi J-MW, Woolsey CA. Remote uncorrelated pilot inputs for nonlinear aerodynamic model identification from flight data. AIAA Aviation 2021 Forum. https://doi.org/10.2514/6.2021-2792.
13. Fu J, Huang J, Song L, Yang D. Experimental study of aircraft achieving dutch roll mode stability without weathercock stability. International Journal of Aerospace Engineering. 2020;2020(1):8971275. https://doi.org/10.1155/2020/8971275.
14. Gresham JL, Simmons BM, Fahmi J-MW, Hopwood JW, Woolsey CA. Remote Uncorrelated Pilot Input Excitation Assessment for Unmanned Aircraft Aerodynamic Modeling. Journal of Aircraft. 2023;60(3):955-67. https://doi.org/10.2514/1.C036942.
15. Roeser MS, Fezans N. Method for designing multi-input system identification signals using a compact time-frequency representation. CEAS Aeronautical Journal. 2021;12(2):291-306. https://doi.org/10.1007/s13272-021-00499-6.
16. Kapeel EH, Safwat E, Kamel AM, Khalil MK, Elhalwagy YZ, Hendy H. Physical modeling, simulation and validation of small fixed-wing UAV. Unmanned Systems. 2023;11(04):327-50. https://doi.org/10.1142/S2301385023500152.
17. Tischler MB. System identification methods for aircraft flight control development and validation. Advances in aircraft flight control.2018:35-69. https://doi.org/10.21949/1403895.
18. Grauer JA, Boucher MJ. Aircraft system identification from multisine inputs and frequency responses. Journal of Guidance, Control, and Dynamics. 2020;43(12):2391-8. https://doi.org/10.2514/1.G005131.
19. Morelli EA. Optimal input design for aircraft stability and control flight testing. Journal of Optimization Theory and Applications. 2021;191(2):415-39. https://doi.org/10.1007/s10957-021-01912-0.
20. Hosseini S, Hofmann R, Holzapfel F, Rauleder J. Optimal Input Design for Aircraft Parameter Estimation-Flight Test Results. AIAA SCITECH 2025 Forum; 2025. https://doi.org/10.2514/6.2025-2173.
21. Lichota P. Multi-Axis Inputs for Identification of a Reconfigurable Fixed-Wing UAV. Aerospace. 2020;7(8):113. https://doi.org/10.3390/aerospace7080113.
22. Millidere M. Optimal input design and system identification for an agile aircraft: Middle East Technical University (Turkey); 2021. https://hdl.handle.net/11511/93115.
23. Mazhar MF, Wasim M, Abbas M, Riaz J, Swati RF. Aircraft system identification using multi-stage PRBS optimal inputs and maximum likelihood estimator. Aerospace. 2025;12(2):74. https://doi.org/10.3390/aerospace12020074.
24. Simsek O, Bayrak AI, Karatoprak S, Dogan A. Comparison of Open Loop and Closed Loop System Identification of a UAV. 2018 Atmospheric Flight Mechanics Conference; 2018. https://doi.org/10.2514/6.2018-3476.
25. Wang Z, Li A, Mi Y, Lu H, Wang C. Aircraft Closed-Loop Dynamic System Identification in the Entire Flight Envelope Range Based on Deep Learning. Journal of Aerospace Engineering. 2024;37(6):04024079. https://doi.org/10.1061/JAEEEZ.ASENG-5657.
26. Chu, T. S. C., Sorilla, J., & Chua, A. Y. UAV-Based Structural Health Monitoring Using a Two-Stage CNN Model with Lighthouse Localization in GNSS-Denied Environments. HighTech and Innovation Journal. 2025; 6(2), 398-410. https://doi.org/10.28991/HIJ-2025-06-02-03.
27. Jiang, Y., Pang, X., Zhang, Z., Jing, H., Wei, L., Su, J., & Zhang, P. Integrating Intelligent Sensors for Safe UAV Distribution: Design and Evaluation of Ranging System. HighTech and Innovation Journal. 2024; 5(3), 603-13. https://doi.org/10.28991/HIJ-2024-05-03-04.
28. Schmidt DK. Modern flight dynamics: McGraw-Hill New York; 2012.
29. Stevens BL, Lewis FL, Johnson EN. Aircraft control and simulation: dynamics, controls design, and autonomous systems: John Wiley & Sons; 2015. https://doi.org/10.1002/9781119174882.
30. Roskam J. Airplane flight dynamics and automatic flight controls: DARcorporation; 1998.
31. Bottigliero R, Rossano V, De Stefano G. Transonic Dynamic Stability Derivative Estimation Using Computational Fluid Dynamics: Insights from a Common Research Model. Aerospace. 2025;12(4):304. https://doi.org/10.3390/aerospace12040304.
32. Nguyen NT, Xiong J. CFD-Based Frequency Domain Method for Dynamic Stability Derivative Estimation with Application to Transonic Truss-Braced Wing. AIAA AVIATION 2022 Forum; 2022. https://doi.org/10.2514/6.2022-3596.
33. Shang T, Chen B, Yanling W, Ting Y, Hailiang L, Lixin W. Identification of aircraft longitudinal aerodynamic parameters using an online corrective test for wind tunnel virtual flight. Chinese Journal of Aeronautics. 2024;37(9):261-75. https://doi.org/10.1016/j.cja.2024.05.031.
34. Tai S, Wang L, Wang Y, Lu S, Bu C, Yue T. Identification of lateral-directional aerodynamic parameters for aircraft based on a wind tunnel virtual flight test. Aerospace. 2023;10(4):350. https://doi.org/10.3390/aerospace10040350.
35. Hassan AA-B. Calculating Aerodynamic & Stability Characteristics for Small fixed-wing UAVs based on Digital DATCOM. The International Undergraduate Research Conference; 2022: The Military Technical College. https://doi.org/10.21608/iugrc.2022.303029.
36. Ahmad M, Hussain ZL, Shah SIA, Shams TA. Estimation of stability parameters for wide body aircraft using computational techniques. Applied Sciences. 2021;11(5):2087. https://doi.org/10.3390/app11052087.
37. Granja DM. Hamiltonian Dynamics applied to aerobatic behavior of a high-performance airplane: Universidade do Porto (Portugal); 2024.
38. Rezazadeh Movahhed S, Hamed MA. Aerodynamic coefficients computation for fixed wing aircrafts using datcom software with emphasize on rudderless flying-wing uavs. Journal of Aerospace Science and Technology. 2024;17(1):60-71. https://doi.org/10.22034/jast.2023.398894.1152.
39. Jategaonkar RV. Flight vehicle system identification: a time domain methodology: American Institute of Aeronautics and Astronautics; 2006. https://doi.org/10.1017/S0001924000011015.

Persian Abstract

چکیده

در این مقاله یک چارچوب سیستماتیک و بهینه برای طراحی ورودی‌های محرک برای شناسایی سیستم هواپیماهای بال ثابت، با تمرکز ویژه بر سیستم‌های چند ورودی چند خروجی (MIMO) تحت کنترل پرواز فعال ارائه شده است. سیگنال‌های محرک مانند پالس، دابلت و چند مرحله‌ای معمولاً در شناسایی سیستم هواپیما به دلیل سادگی و کاربرد گسترده آنها انتخاب می‌شوند، نه اینکه از یک فرآیند طراحی سیستماتیک طراحی شوند. در نتیجه، انتخاب آنها اغلب تجربی است و ممکن است تحریک بهینه برای تخمین دقیق پارامتر را فراهم نکند. روش پیشنهادی به طور سیستماتیک موثرترین سیگنال‌های ورودی را تعیین می‌کند، محتوای فرکانسی آنها را طراحی می‌کند و عملکرد آنها را در هر دو شرایط حلقه باز و حلقه بسته ارزیابی می‌کند. این رویکرد تخمین دقیق پارامتر را در تکنیک‌های شناسایی متعدد تسهیل می‌کند و قابل تعمیم به سایر پیکربندی‌های هواپیما و حتی سایر سیستم‌های دینامیکی است. اثربخشی روش پیشنهادی از طریق شبیه‌سازی‌های شش درجه آزادی با دقت بالا در MATLAB/Simulink نشان داده شده است که توانایی آن را در تحریک کارآمد و دقیق حالت‌های مربوط به هواپیما برای شناسایی دقیق مدل نشان می‌دهد.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2617-2629

International Journal of Engineering

Journal Homepage: www.ije.ir

Demagnetization Study of Permanent Magnets During the Startup of LSPMSM Considering the Effects of Temperature

L. Anh Tuan[a], T. Duy Khanh[b], N. Y. Do*[c]

[a] *School of Electrical and Electronic Engineering, Hanoi University of Industry, Hanoi, Vietnam*
[b] *Faculty of Electrical Engineering, Saodo University, Haiphong, Viet Nam*
[c] *Faculty of Electro-Mechanics, Hanoi University of Mining and Geology, Hanoi, Vietnam*

PAPER INFO

Paper history:
Received 14 August 2025
Received in revised form 31 December 2025
Accepted 07 January 2026

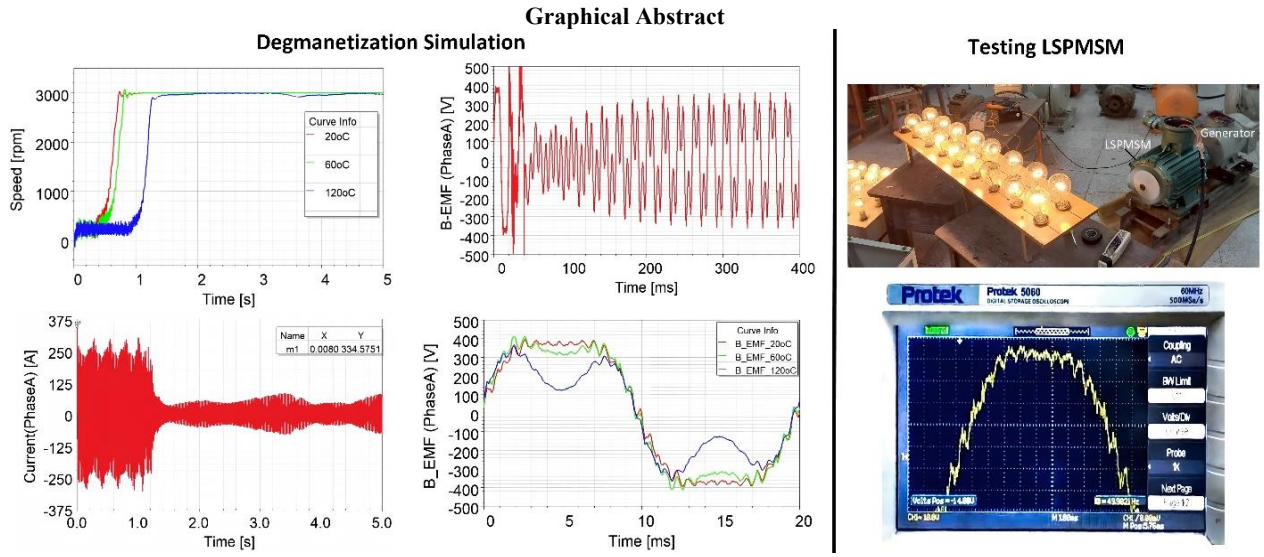
Keywords:
Demagnetization
Permanent Magnet
Temperature
Line-start Permanent Magnet Synchornous Motor
Electric Motor

ABSTRACT

The line-start permanent magnet synchronous motor (LSPMSM) is increasingly being applied due to its high efficiency. However, temperature has a significant effect on the electromagnetic parameters, which in turn affects the operating characteristics of the LSPMSM. Due to the presence of squirrel-cage components in the rotor, LSPMSM can line-start, but the high starting current can lead to localized or complete demagnetization, especially at elevated temperatures. This can result in irreversible demagnetization of the permanent magnets. This paper analyzes the effect of temperature on the electromagnetic parameters of the motor as well as its impact on demagnetization during LSPMSM startup. The theoretical study results indicate the influence of temperature on the electromagnetic parameters of LSPMSM and demonstrate that, under certain temperature conditions, there exists a maximum starting current coefficient to prevent demagnetization. Using FEA simulation for a LSPMSM, 15kW, 2p=2, results show a correlation between theoretical and simulated findings. It is shown that with a high starting current of 337.8A, NdFeB-N38M magnets used in the motor can experience mild demagnetization at 60°C, while visible demagnetization occurs at 120°C. Experimental validation has confirmed the accuracy, as the B-EMF characteristics indicate a close resemblance between experimental and simulated results, with a deviation of less than 7%.

***doi**: 10.5829/ije.2026.39.10a.21*

Graphical Abstract



*Corresponding Author Email: donhuy@humg.edu.vn (N. Y. Do)

Please cite this article as: Anh Tuan L, Duy Khanh T, Do NY. Demagnetization Study of Permanent Magnets During the Startup of LSPMSM Considering the Effects of Temperature. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2617-29.

1. INTRODUCTION

Recently, the line-start permanent magnet synchronous motor (LSPMSM) has been widely investigated and applied as a type of permanent magnet synchronous motor (PMSM). This motor offers several advantages, including high efficiency, self-starting capability upon power supply, reliable operation, and low operating costs (1). Additionally, the trend of applying LSPMSM in manufacturing is driven by the need for motors to meet increasingly stringent efficiency standards for electric motors. In fact, many energy efficiency regulations and energy policies have been enacted, such as the IEC 60034-30-1:2014 standard issued by the International Electrotechnical Commission in June 2014, which specifies the energy efficiency levels for electric motors (2).

However, the long-term operational efficiency of LSPMSM, as well as PMSM, significantly depends on electromagnetic parameters such as d-axis and q-axis synchronous inductances (L_{md}, L_{mq}), leakage inductance (L_{ls}), stator winding resistance (R_s), and remanence of permanent magnet (PM) (B_r) (3). These parameters are strongly influenced by temperature, especially in industrial applications where the motor operates under conditions of significant variations in ambient temperature and load temperature (4). When the operating temperature of the motor is high, it reduces the remanence and coercive force (H_c) of the PM, leading to a decline in parameters such as rotor flux, torque, efficiency, power factor, and so on (5-7).

The demagnetization phenomenon of PMs refers to the reduction of magnetic strength in PMs, which typically occurs when the temperature exceeds the operational limits or when the PMs are subjected to strong opposing magnetic fields (8-10). In LSPMSM, irreversible demagnetization (ID) can occur in two stages. During startup, the large starting current generates a significant opposing magnetic field, combined with the rapid increase in rotor temperature due to Joule losses in the squirrel cage (4, 11). In steady-state operation, high ambient temperatures or prolonged heavy loading can significantly raise the rotor temperature (10). Adly et al. (12) analyzed the impact of partial ID and loss of excitation on the operational capacity of the motor. The results indicate that partial ID can affect the lifespan of the PMs and reduce efficiency. He et al. proposed a method to estimate the reduction in operational efficiency of the motor following the decrease in remanence of the PMs (13). The goal is to reflect the relationship between the complete loss of magnetization in PMSM and the resulting performance fluctuations.

The ID process leads to disadvantages and even failures in the operation of the motor. Therefore, some studies focus on predicting and diagnosing faults to proactively prevent the occurrence of ID or to address issues after ID has occurred. Moosavi et al. (14) synthesized research aimed at developing techniques for detecting demagnetization faults of PM (DFs). These faults diminish the performance, reliability, and efficiency of PMSM drive systems. Ebrahimi et al. proposed a novel method for detecting DFs caused by ID in surface-mounted PMSMs (SMPMSMs) (15). The authors suggested a theoretical analysis based on characteristics to identify DF through torque analysis. Uršič et al. (16) presented a solution for monitoring PMs in PMSM drive systems. Temperature information, along with the machine's current, is incorporated into the demagnetization model of the PMs to determine their demagnetization state. The diagnostic results for potential ID are then used to limit the opposing magnetic field to prevent ID. Pietrzak et al. (17) introduced a fault diagnosis method for DFs based on machine learning (ML) techniques for PMSM drive systems. Two ML-based models were verified and compared during the automatic detection of DFs in operation. Faiz et al. (18) compiled effective technical methods for diagnosing DFs under various operating conditions. The authors analyzed the factors causing faults in PM machines and their impact on efficiency of motor. Finally, the paper presents a general framework for developing DF diagnosis techniques under different operating conditions (static and dynamic). It can be stated that proactively diagnosing the risk of ID during operation is essential. Timely diagnosis allows operators to take appropriate measures in controlling the motor. Additionally, detecting faults due to DFs that has already occurred will enable users to take suitable actions, such as replacing magnets or the motor during operation.

Designing an appropriate motor configuration can be considered the most important step in avoiding ID. Regarding design issues, designers are now applying optimization algorithms to enhance structures that prevent temperature-induced demagnetization while still ensuring operational parameters. Yong-min et al. (19) optimized the shape of a PMSM for electric vehicles by considering the ID characteristics of PMs through a design of experiments (DOE) approach. The paper employs a multi-objective genetic algorithm to minimize the ID rate and maximize the average torque, with constraints on efficiency and torque ripple. Huang et al. (20) analyzed the ID of PMSM used in electric power steering (EPS) systems. The dimensions of the PM were optimized, taking into account magnetic loss performance and production costs. Dandan et al. (21) focused on three main issues: PM material and manufacturing technology, optimizing the electromagnetic design of the motor structure, and examining the operating environment along with necessary preventive measures against ID. Kang et al. (22) proposed an optimal rotor structure that can prevent ID of PMs in spoke-type PMSMs. Thus, research

applying various methods and algorithms in the design of PMs is diverse and serves as a crucial measure to prevent ID during the motor design phase.

In addition to applying optimization algorithms in design, studies also consider proposing motor configurations, primarily rotor configurations, that can prevent ID. Wang et al. (7) proposed a V-shaped combined pole interior PMSM (VCP-IPMSM), in which NdFeB PMs at the outer ends of the V-poles are replaced with ferrite magnets. The authors concluded that compared to traditional IPMSMs, the VCP-IPMSM configuration demonstrates improved magnetic field quality in the air gap and enhanced resistance to ID at high temperatures. Yu et al. (11) analyzed the ID characteristics of NdFeB PMs configured in a 'W' shape for LSPMSMs. The study results indicate that the risk points for ID occur at the corners of the PMs. Kim et al. (23) examined ID of PMs for three rotor types of IPMSMs. The three types of rotors are arranged with PMs in a single layer, V-shape, and double layer, commonly used for PMSMs. Choi et al. (24) investigated the ID of PMs in the rotor of various PM configurations to identify designs with higher resistance to demagnetization. Experimental results confirmed that both localized and global demagnetization can be accurately predicted using finite element analysis (FEA) over time steps. Mahmouditabar et al. (25) studied the dynamic demagnetization process for V-shaped IPMSM, conducting research on two different types of PMs. Thus, along with the application of optimization algorithms in design, the selection of PM configurations is also a key factor in preventing ID.

In addition to diagnostic studies, DF detection, and the selection of appropriate configurations in design to prevent ID, recent research has focused on proposing new methods for studying ID processes. In these studies, the factors of demagnetization and temperature are considered simultaneously. Baranski et al. presented a 2D magnetic-field-electric-circuit model that combines mechanical and thermoelectric phenomena of a LSPMSM (4). This paper applies the proposed model to investigate the demagnetization process of the motor. Zawilak et al. suggested using a magnetic circuit model of electromagnetic phenomena to analyze the demagnetization process, taking into account the effect of temperature on the characteristics of the nonlinear voltage-current and its resistance to ID (10). Jia et al. (26) studied the interaction between electromagnetic fields and thermal fields within the same motor. Additionally, the iron loss curves at different frequencies of electrical steel and the B-H curves at various temperatures of PM were used to establish the electromagnetic model of the PMSM. Palangar et al. compared electromagnetic and thermal characteristics between a LSPMSM with an external magnetic field-LSPMSM and one with an internal magnetic field-LSPMSM (5). In this study, a parameterized thermal network analysis model was proposed to analyze the thermal behavior of the two motors. Thus, it can be stated that research on ID, considering the temperature of PMSMs, has been undertaken. These studies confirm that temperature and ID are two critical factors in the operation of the motor, and these factors need to be examined together to prevent the occurrence of ID in the PMs.

Although many studies have investigated the impact of temperature on PMSMs, research on LSPMSMs, which have specific starting mechanisms and combine squirrel cage with PMs, remains limited. Unlike PMSMs, which are typically accompanied by a inverter that results in a small starting current, the LSPMSM can be directly started from the grid due to its squirrel cage rotor, eliminating the need for an accompanying inverter. Consequently, the starting current is significantly larger, being 5 to 7 times the rated current. This leads to the risk of an ID process in the LSPMSM during startup that is distinctly different from that of a PMSM.

The quantitative relationships between temperature of the starting current, demagnetization levels, and variations in electromagnetic parameters have not been fully clarified. Therefore, to study ID and its mitigation in LSPMSMs from the design phase, this paper evaluates the impact of temperature on electromagnetic parameters and analyzes its effect on ID of PMs in LSPMSM during startup. The research results are based on a theoretical model, simulated using FEA software, as well as experimental evaluations in the laboratory. The findings will provide a scientific basis for designing, operating, and maintaining LSPMSMs under extreme temperature conditions, contributing to enhanced efficiency and lifespan of the motors.

2. MATHEMATICAL MODEL OF LSPMSM CONSIDERING TEMPERATURE

When considering the effects of temperature, the mathematical model must be expanded to reflect the changes in resistance, magnetic flux, and torque due to thermal effects (27). As the temperature increases, the resistance of the stator and rotor also rises. The temperature variation is determined by the following formula:

$$\begin{aligned} R_s(T) &= R_{s0}\left[1+\alpha_s(T-T_0)\right] \\ R_r(T) &= R_{r0}\left[1+\alpha_r(T-T_0)\right] \end{aligned} \quad (1)$$

where R_s, R_r are the stator and rotor resistances at temperature T, R_{s0}, R_{r0} are the stator and rotor resistances at temperature T_0, α is the thermal expansion coefficient.

When the temperature changes, it leads to a change in the magnetic permeability of the PM material, causing the inductances, L_{md} and L_{mq}, to decrease with

temperature. In the case of no magnetic saturation, L_{md} and L_{mq} are determined by the following equations:

$$\begin{aligned} L_{md}(T) &= L_{md}(T_0)\left[1-\alpha_L(T-T_0)\right] \\ L_{mq}(T) &= L_{mq}(T_0)\left[1-\alpha_L(T-T_0)\right] \end{aligned} \quad (2)$$

where α_L is the coefficient of inductance variation with temperature (typically around α_L = 0.0005 to 0.0015/°C depending on the type of magnetic core material). L_{ls} is only slightly affected by temperature, primarily through the physical expansion of the winding. Under the assumption of linearization, the variation of L_{ls} with respect to temperature T is modeled as follows:

$$L_{ls}(T) = L_{ls}(T_0)\left[1+\alpha_{ls}(T-T_0)\right] \quad (3)$$

where α_{ls} is the coefficient of leakage inductance variation with temperature (typically α_{ls}=0.0001 to 0.0005/°C).

The mathematical model of LSPMSM considering the effects of temperature is presented as follows:

$$\begin{cases} u_{ds} = r_1(T)i_{ds} + \dfrac{d\Psi_{ds}(T)}{dt} - \omega_r.\Psi_{qs}(T) \\ u_{qs} = r_1(T)i_{qs} + \dfrac{d\Psi_{qs}(T)}{dt} + \omega_r.\Psi_{ds}(T) \end{cases} \quad (4)$$

$$\begin{cases} u'_{dr} = r'_{dr}(T)i'_{dr}(T) + \dfrac{d\Psi'_{dr}(T)}{dt} = 0 \\ u'_{qr} = r'_{qr}(T)i'_{qr}(T) + \dfrac{d\Psi'_{qr}(T)}{dt} = 0 \end{cases} \quad (5)$$

$$\begin{cases} \Psi_{ds}(T) = (L_{ls}(T) + L_{md}(T))i_{ds} \\ \qquad\qquad + L_{md}(T)i'_{dr} + \Psi'_m(T) \\ \Psi_{qs}(T) = (L_{ls}(T) + L_{mq}(T))i_{qs} \\ \qquad\qquad + L_{mq}(T)i'_{qr} \end{cases} \quad (6)$$

$$\begin{cases} \Psi'_{dr}(T) = L'_{lr}(T)i'_{dr} + L_{md}(T)(i_{ds} + i'_{dr}) + \Psi'_m(T) \\ \Psi'_{qr}(T) = L'_{lr}(T)i'_{qr} + L_{mq}(T)(i_{qs} + i'_{qr}) \end{cases} \quad (7)$$

where ω_r is the rotor angular velocity, i_{ds}, i_{qs} are the stator current of d, q axes, i'_{dr}, i'_{qr} are the rotor equivalent current of d, q axes. $\Psi_m(T)$ *is the* stator flux generated by the PMs, $L_{ls}(T)$ is the leakage inductance of the stator winding, $L_{md}(T)$ and $L_{mq}(T)$ are the magnetizing inductances of d, q axes with respect to temperature.

The mathematical model shows that under the influence of temperature, the electromagnetic parameters of LSPMSM change with varying degrees of significance. Additionally, since LSPMSM has PMs, temperature not only affects the electromagnetic parameters during operation but also has the potential to partially or completely ID of the PMs, leading to a reduction or total loss of magnetic properties. Analyzing the extent of temperature's impact on electromagnetic parameters, as well as the ID capability, is crucial for developing appropriate design and operational strategies.

3. ANALYSIS OF THE IMPACT OF TEMPERATURE ON THE DEMAGNETIZATION IN LSPMSM

3. 1. Demagnetization Mechanism of PMs Considering Operating Temperature

The demagnetization characteristic is the most important property of PMs. The relationship between load line (current) and demagnetization characteristics with respect to temperature for the NdFeB-N38 meterial is described as shown in Figure 1 (28).

The demagnetization characteristic of PMs in the second quadrant is typically represented in two forms: *B(H)* or *J(H)*. On the demagnetization curve, there exists a "knee" point, point a, at a temperature of 20°C. As the temperature increases, the demagnetization characteristic shifts to the right, and the "knee" points also move upward, corresponding to points b, c, d and k at temperatures of 60°C, 80°C, 100°C and 120°C, respectively. The variation of the "knee" point on the demagnetization characteristic of the PM (shown in red) can be expressed by the equation:

$$B_{knee} = \beta H_{knee} + B_k \quad (8)$$

where B_k is the magnetic field strength. β is the slope of the “knee” characteristic with respect to temperature.

When the LSPMSM is operating, the normal working point on the demagnetization characteristic is the point of maximum energy product, which is the intersection of the

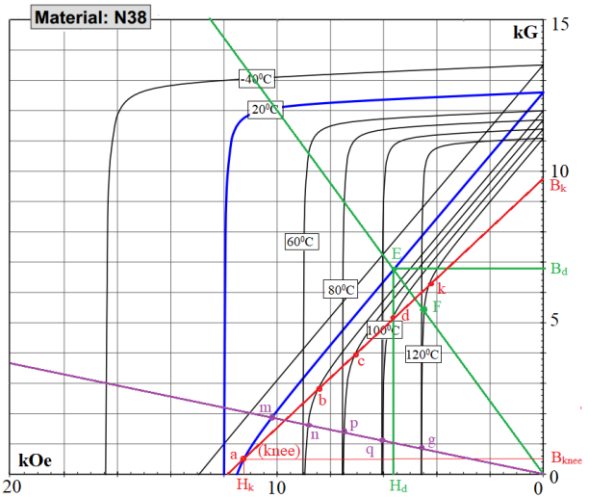


Figure 1. Demagnetization characteristics of NdFeB-N38

load line with the demagnetization curve. The remanence of the PM at the working point is determined by the equation:

$$B_d = B_r + \mu_r \mu_0 H_d \tag{9}$$

where B_d is the magnetic flux density, B_r is the remanence, H_d is the demagnetizing field μ_0 is the vacuum permeability and μ_r is the relative permeability. Demagnetization occurs when the remanence of the PM B_d is less than the remanence at the "knee" point:

$$B_d \le B_{knee} \tag{10}$$

When the motor starts, the starting current increases, causing the load line to shift downward toward the horizontal axis (shown in purple). The remanence of the PM during motor startup (point m at 20°C) is the intersection of the load line and the demagnetization curve. If the temperature of the PM increases, the remanence of the PM during startup changes to points n, p, q and g, corresponding to temperatures of 60°C, 80°C, 100°C, and 120^0C, respectively. The remanence characteristic of the PM varies with the starting current and the temperature of the magnet, as expressed by the equation:

$$B_s \le \gamma H_s \tag{11}$$

where γ is the coefficient related to the magnitude of the starting current, and H_s is the magnetic field strength generated by the stator current during startup. For NdFeB material, if partial ID occurs, then $B_s \le B_{knee}$, meaning that the limit of the magnetic field strength generated by the stator current, considering the effects of temperature, must be maintained to prevent ID:

$$\gamma H_s \le \beta H_{knee} + H_k \tag{12}$$

The starting current multiplier to avoid demagnetization for PMs, considering the effects of temperature:

$$\gamma = \beta \frac{H_{knee}}{H_s} + \frac{H_k}{H_s} \tag{13}$$

Equations 8, 9 and 10 indicate the temperature-dependent demagnetization conditions of the PMs on the demagnetization curve (Figure 1). From Equation 13 and Figure 1, it is evident that demagnetization in PMs is highly dependent on temperature. As the temperature increases, demagnetization can occur even when the LSPMSM is operating at rated load. When the LSPMSM starts, the load line shifts downward, and as the temperature rises, the "knee" point moves upward, making demagnetization more likely to occur. A suitable starting current must be defined for each operating temperature of the LSPMSM to prevent ID in PMs. Therefore, during the design process, the dimensions of the PM must be selected to ensure a correlation between current and temperature to avoid ID.

3. 2. The Effect of Temperature on the Operating Characteristics of the LSPMSM In the design of LSPMSM motors, the motor dimensions are typically selected preliminarily through analytical formulas. The initial calculation of the magnet dimensions is determined using a formula for volume. The volume of the magnet is defined by the Equation 1:

$$V_M = \frac{2.k_{of}.k_{fd}.(1+k_{EC}).P_2}{\pi^2.\xi.2.p.f.B_r.H_C} \tag{14}$$

where P_2 is the output power, f is the input frequency, B_r is the remanence, H_c is the coercive force, k_{ocf} is the overload capacity factor, k_{fd} is the magnetization form factor, k_{EC} is the electromotive force factor, ξ is the magnet utilization factor.

Typically, the thickness of the PM must be ensured according to the following equation to prevent ID, as analyzed above (29):

$$I.\frac{m}{2}.\frac{4}{\pi}.\frac{\sqrt{2}.N.k_w}{2.p.H_C} \le l_M \tag{15}$$

where m is the number of phases, p is the number of poles, I is the peak current value during startup, N is the number of stator windings, and k_w is the winding coefficient.

However, not considering the effects of temperature may be a shortcoming in this equation. This can lead to distorted operating characteristics of the LSPMSM. As the temperature increases, it not only reduces the magnetic flux generated by the PM but can also cause partial ID during operation, especially during the startup phase.

When considering the effects of temperature, the magnetic flux through the stator winding generated by the PM is determined (1):

$$\Psi_m = \frac{4Dl_s k_{w1} N}{\pi p} \sin\left(\frac{\lambda\pi}{2}\right) B_{r0}\left(1-\alpha\Delta T\right) \tag{16}$$

where D is the inner diameter of the stator, l_s is the length of the stator core, and λ is the magnetic pole coefficients.

In the case of elevated temperatures, if the PM experiences partial ID, the magnetic field generated by the magnet continues to decrease. Let ε be the coefficient of partial ID then the magnetic field generated by the PM can be determined by the following equation:

$$B_r \le \varepsilon B_{r0} \tag{17}$$

The magnetic flux linkage of the stator winding generated by the PM, taking into account the temperature and the partial ID of the PM, is determined by:

$$\Psi_m = \frac{4Dl_s k_{w1} N}{\pi p} \sin\left(\frac{\lambda\pi}{2}\right) \varepsilon B_{r0}\left(1-\alpha\Delta T\right) \quad (18)$$

The back-electromotive force (B-EMF) is determined through the magnetic flux Ψ_m as follows (32):

$$E_0 = \frac{2\pi}{\sqrt{2}} f \frac{4Dl_s k_{w1} N}{\pi p} \sin\left(\frac{\lambda\pi}{2}\right) \times \varepsilon B_{r0}\left(1-\alpha\Delta T\right) \quad (19)$$

The electromagnetic torque of the LSPMSM is defined.

$$T_{el} = \frac{3p}{2}\left[\underbrace{\left(L_{md}(T) i'_{dr} i_{qs} - L_{mq}(T) i'_{qr} i_{ds}\right)}_{T_{Ind}} + \underbrace{\Psi'_m(T) i_{qs}}_{T_{Exc}} + \underbrace{(L_{md}(T) - L_{mq}(T)) i_{ds} i_{qs}}_{T_{rel}}\right] \quad (20)$$

where T_{ind}, T_{exc} and T_{rel} are the induction, excitation and reluctance torque component with respect to temperature. The torque equation of the motor:

$$J\frac{d\omega_r}{dt} = T_{rel}(T) - T_L - B\omega_r \quad (21)$$

where T_{rel}, T_L, B, and J are the electromagnetic and load torques, friction coefficient, and inertia with respect to temperature.

In addition to temperature, the partial ID also reduces the B_r of the PMs, leading to a decrease in magnetic flux within the magnetic circuit of the LSPMSM, as described by Equation 16. This, in turn, affects the operating characteristics of the motor, which have been indicated by Equations 19, 20, and 21. From the analysis above, it can be seen that an increase in temperature alters the electromagnetic parameters of the LSPMSM, and there may also be a phenomenon of partial ID of the PM. This affects working parameters such as synchronous speed, B-EMF, and many other parameters of the LSPMSM. Neglecting the influence of temperature in the design calculations of the LSPMSM is a shortcoming that may lead to an impact on the operational characteristics of the LSPMSM.

4. SIMULATING THE EFFECT OF TEMPERATURE ON DEMAGNETIZATION DURING THE STARTUP

The report evaluates the influence of temperature on partial ID and the operating parameters of the LSPMSM, 15 kW, 3000 rpm, which has been modified from a similar IM motor with preliminary design calculations for the dimensions shown in Table 1. The main parameters related to the configuration of the motor considered are the inner diameter, outer diameter, number of slots and the core material of the rotor and stator. Additionally, the dimensions of the PM are determined using the formulas 14 and 15 as mentioned above. The length of the PM is defined to be equal to the length of the stator and rotor iron cores. The supply parameters are determined based on the supply from the corresponding induction motors.

TABLE 1. Parameters of experimental LSPMSM-15kW

Parameters	Symbol	Value	Unit
Stator outer, inner diameter	D_{in}, D_{out}	245, 152	mm
Rotor outer and shaft diameter	D, D_t	151, 52	mm
Number of stator, rotor slots	Z_1, Z_2	36, 28	slots
Air gap length	g	0,5	mm
Power supply voltage	V_n	380/660	V
Power supply frequency	f	50	Hz
Steel material	M800-50A		
Permanent magnet material	NdFeB-N38M		
PM dimensions	W x H x L	35x9x170	mm

The temperature distribution in the motor is not uniform, with certain points exhibiting the highest temperatures. During the startup process, the current of the motor increases; however, the heating time of the motor is short, preventing the thermal state from reaching equilibrium. Therefore, the study scenario in the paper assumes that the points with the highest temperatures are 20°C (equal to ambient temperature), 60°C, and 120°C, corresponding to the temperature values on the demagnetization curves of the PMs for analysis.

The paper employs the FEA to simulate the operational characteristics of the LSPMSM under varying temperature conditions. The structure of the LSPMSM, with design parameters outlined in Table 1, is constructed using FEA software, as illustrated in Figure 2. The simulation scenarios investigated involve operating the motor at temperatures of 20°C, 60°C, and 120°C. This scenario is utilized to simulate and evaluate the working characteristics of the LSPMSM.

4. 1. The Effect of Temperature on Speed Characteristics Speed characteristics are one of the important characteristics of motors. This characteristic is

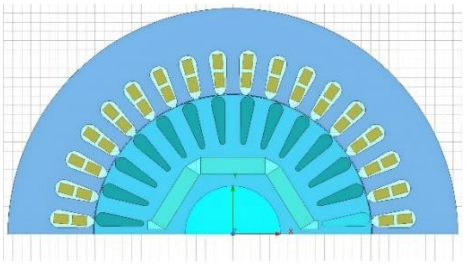


Figure 2. Configuration of LSPMSM 15kW, 2p=2

often used to assess the starting capability and quality of the motor. When the temperature changes, the speed characteristics of the LSPMSM are obtained as shown in Figure 3.

The results in Figure 3 indicate that at a temperature of 20°C, the LSPMSM starts smoothly and easily. As the temperature increases to 60°C, the starting characteristics of the LSPMSM deteriorate compared to those at 20°C due to the temperature effect. At a temperature of 120°C, the high temperature leads to partial ID of the PMs, causing the LSPMSM to lose its synchronization and operate unstably. The analysis of the starting characteristics based on temperature is summarized in Table 2.

The results in Table 2 clearly show the significant impact of temperature on the starting parameters of the LSPMSM. As the temperature increases from 20°C to 60°C, the parameters change noticeably, with the synchronization time increasing by Δt_{st}=0.13s, the oscillation time increasing by Δt_{fl}=0.2s, and the amplitude of oscillation increasing by $\Delta\omega$=50rpm. When the temperature rises further to 120°C, in addition to the impact of temperature, the phenomenon of demagnetization also affects the LSPMSM, causing it to operate unstably at synchronous speed. This instability can lead to significant vibrations, potentially resulting in motor damage.

4. 2. The Effect of Temperature on Current Characteristics Alongside speed characteristics, current characteristics are also an important attribute of motors. This characteristic is often used to assess performance, particularly the transient current. The current characteristics under three temperature conditions studied are presented in Figure 4.

In Figure 4, it can be seen that as the temperature varies at levels of 20°C, 60°C, and 120°C, the current characteristics change and differ significantly. The parameters of the current for each temperature case are presented in Table 3.

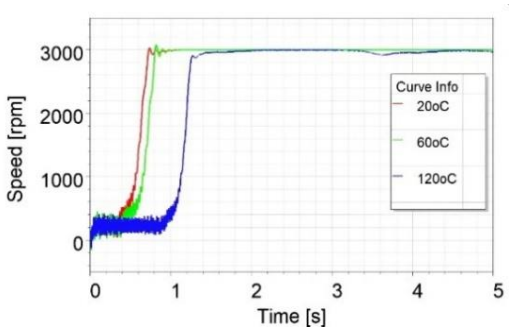


Figure 3. Configuration of LSPMSM 15kW, 2p=2

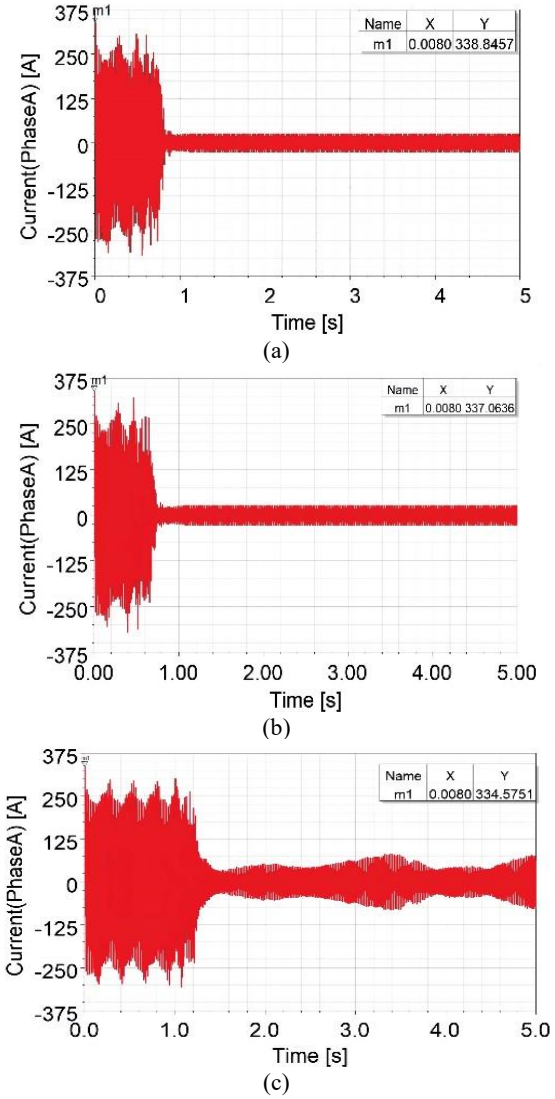


Figure 4. Simulated current characteristics at (a) 20°C, (b) 60°C and (c) 120°C.

TABLE 2. Starting parameters based on temperature

No	$T(^0C)$	Synchronous time t_{st} (s)	Fluctuation time t_{ft} (s)	$\Delta\omega_{max}$ (rpm)
1	20^0C	1.41	0.6	25
2	60^0C	1.54	0.8	75
3	120^0C	Undefined	Undefined	Undefined

TABLE 3. Current based on temperature

No.	$T(^0C)$	Fluctuation time t_{ft} (s)	Maximum starting current (A)
1	20^0C	0.6	338.A
2	60^0C	0.8	337.1
3	120^0C	Undefined	334.6

The results show that the maximum starting current in all cases is approximately 337A, which is due to the fact that the motor operates in an asynchronous state during startup. Additionally, it is observed that at temperatures from 20°C to 60°C, the starting characteristics of the motor do not differ significantly in terms of time and the maximum starting current value. However, when the temperature reaches 120°C, the time required for the motor to start and reach synchronous speed for the first time becomes very large. Once it reaches synchronous speed for the first time, the motor is also unable to operate stably as in the previous two cases. This can be explained by the fact that at a temperature of 120°C, the phenomenon of partial ID has appeared during startup. The flux generated by the PMs undergoes significant changes instead of minor changes as in the case without partial ID, and this process repeats after a few cycles until the operating point of the magnets reaches a new equilibrium state. Therefore, when starting at high temperatures, the time to reach synchronous speed for the first time is much greater. Moreover, due to the partial ID when reaching the new equilibrium state, the coercive force H_c of the PMs is significantly reduced. At this point, even a weak external magnetic field generated by the stator current can considerably alter the operating point of the magnets. This leads to strong fluctuations in the operating current once the motor reaches synchronous speed for the first time. If the LSPMSM continues to operate, leading to a further increase in temperature above 120°C, it may result in complete ID of the PMs, potentially destroying them when the temperature reaches the Curie point.

4. 3. The Effect of Temperature on B-EMF The scenario evaluates the effect of temperature and starting current on B-EMF characteristics by applying a current with an amplitude of '0' to the stator winding of the LSPMSM for the first cycle, then increasing the current amplitude to reach a value equal to the peak current when the motor starts, which is 337.8A, maintained for a second cycle. After 40ms, the phase currents A, B, and C are turned off to observe the recovery process from the remanence of the PMs and the impact of partial ID on the B-EMF characteristics corresponding to operating temperatures of 20°C, 60°C, and 120°C. The current characteristics applied to the stator windings in the simulation are shown in Figure 5.

Under the influence of the current mentioned in Figure 5, the B-EMF characteristics obtained during the simulation at corresponding temperatures of 20°C, 60°C, and 120°C are shown in Figure 6.

Figure 6 shows that the B-EMF characteristics exhibit distinct differences. In the time interval from 0 to 20 ms, when no current is applied to the stator winding, the B-EMF characteristics generated by the PMs at temperatures of 20°C, 60°C, and 120°C have a nearly sinusoidal shape and similar amplitudes. In the interval from 20 to 40 ms, when a current with an amplitude of 337.8A is applied to the stator winding, the amplitude of

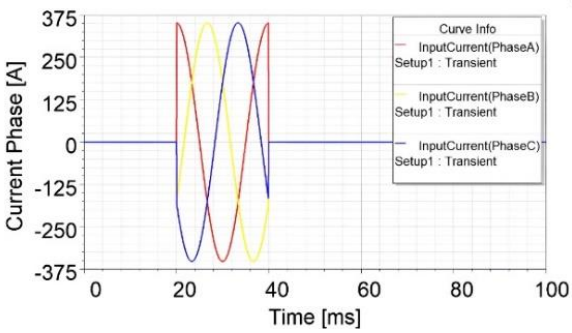


Figure 5. Current scenario in demagnetization simulation

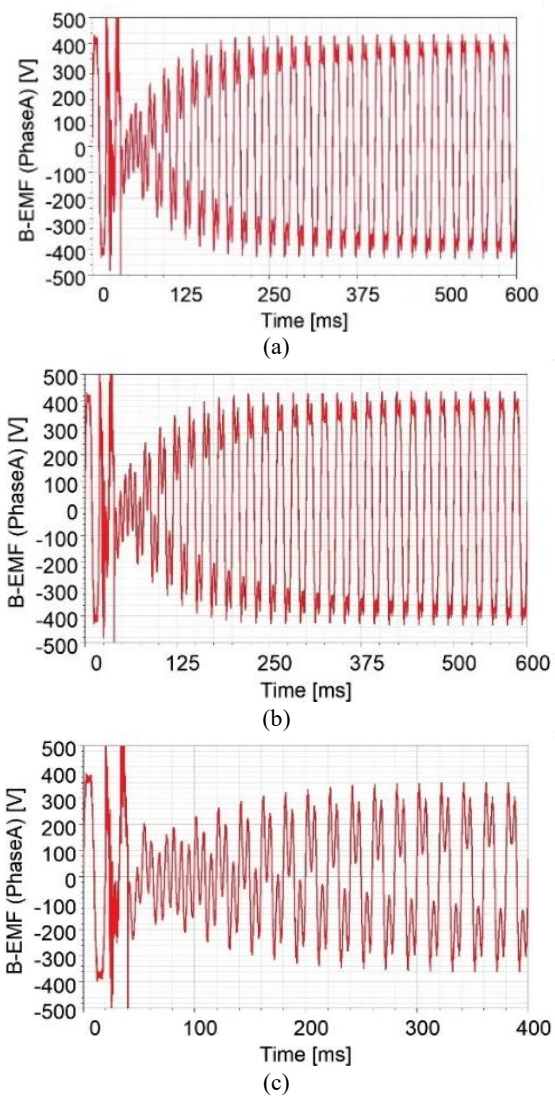


Figure 6. Simulated current characteristics at (a) 20°C, (b) 60°C and (c) 120°C.

the B-EMF decreases under the influence of the current and temperature, with the least reduction at 20°C and the most significant at 120°C. However, after removing the current from the stator at 40 ms, the amplitude of the B-EMF at 20°C is fully restored, the B-EMF amplitude at 60°C decreases slightly due to the thermal effects, and the B-EMF at 120°C shows a pronounced reduction, indicating that under the influence of the high temperature of 120°C and the large starting current of 337.8A, the PMs have experienced partial ID.

The simulation results of the waveforms at the t=0÷20ms (without applying an external magnetic field) and t=380÷400ms are presented. The interval of 380÷400 ms represents the time when the B-EMF waveform generated by the PM magnetic field recovers to a stable value in all cases considered. The waveforms for one cycle at the two time points are shown in Figure 7.

The results shown in Figure 7 (a) indicate that during the t=0÷20ms, when no external magnetic field is applied, the B-EMF waveform generated by the PMs, despite exhibiting many notches due to the influence of the teeth, stator slots, and rotor, remains fundamentally close to a sine wave. During this period, under the influence of temperature, the B-EMF characteristics decrease slitghtly at 120°C; however, the level of reduction is not significant. This can be explained by the fact that as the temperature increases, the remanence of the PMs decreases, leading to a reduction in B-EMF, which is entirely consistent with the theory presented above.

The results shown in Figure 7(b) illustrate the impact of the external magnetic field generated by the stator current during the time t=20÷40 ms on the partial ID of the PMs. Examining the B-EMF waveform at the time t=380÷400ms reveals that when operating at a temperature of 20°C, the B-EMF has fully recovered to its original shape. This indicates that partial ID of the PMs does not occur at this temperature. At 60°C, the B-EMF waveform shows a slight reduction, with the amplitude not returning to its original shape, demonstrating that at 60°C, the PMs have undergone partial ID due to the simulated starting current. At 120°C, the B-EMF waveform is significantly altered, exhibiting a dip in the middle of each half-cycle (from 360V to 120V). This indicates that substantially partial ID of the PMs has occurred under the simulation conditions. Thus, these simulation results validate the theory of partial ID dependent on the external magnetic field generated by the stator current and the operating temperature of the PMs, as analyzed above.

The B-EMF waveform was analyzed using FFT for evaluation. The B-EMF waveforms analyzed using FFT in two cases at the time t=0÷20ms and 380÷400ms are shown in Figure 8.

The $\Delta E\%$ is the measure of deviation between the amplitude value of the fundamental waveform and the

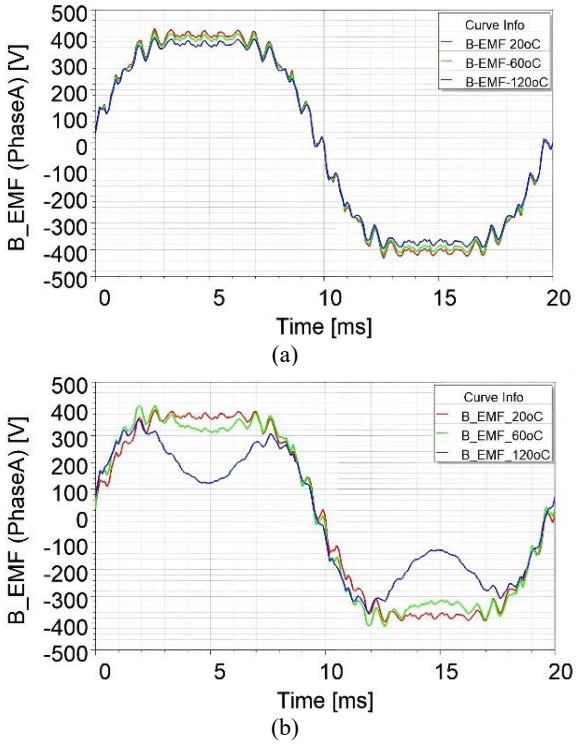


Figure 7. B-EMF waveforms at the times (a) t=0÷20ms, (b) t=380÷400ms

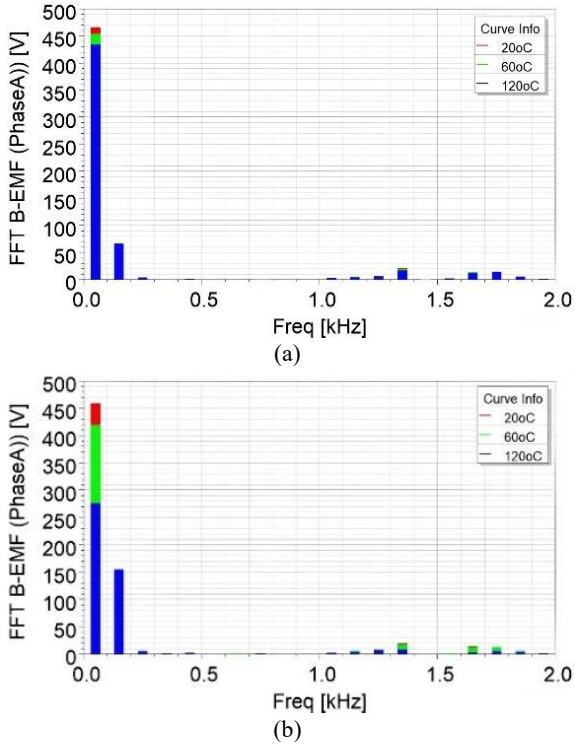


Figure 8. FFT expansion of the B-EMF waveform at the times (a) t=0÷20ms, (b) t=380÷400ms

nominal phase voltage of the power system. *ΔE%* is expressed (1):

$$\Delta E\% = \left| \frac{E_{1NB} - 380}{E_{1NB}} 100\% \right| \quad (22)$$

where E_{1NB} is the amplitude value of the fundamental waveform of B-EMF waveform. The calculation results of *ΔE%* are in Table 4.

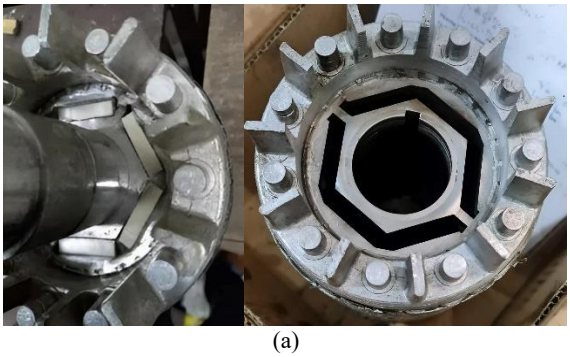
FFT analysis of the B-EMF waveform at the time t=0÷20ms shows that the fundamental waveform decreases gradually, while the third harmonic waveform increases with temperature. Concurrently, the difference between the amplitude value of the fundamental waveform and the nominal phase voltage also increases with temperature. This can be explained by the fact that as the temperature rises, the magnetic field of the PMs decreases, leading to changes in the harmonic components of the B-EMF. During the time t=380÷400ms at a temperature of 20°C, the amplitude of the fundamental waveform remains constant. At 60°C, ΔE% increases and reaches a value of 28.2%. At 120°C, ΔE% increases significantly to 94.7%, which is due to the severely partial ID of PMs at high temperatures.

5. DEMAGNETIZATION EXPERIMENT OF LSPMSM

The author team constructed a LSPMSM 15 kW, 2p=2 for experimentation in the laboratory. The LSPMSM was tested to analyze and evaluate the impact of temperature on motor operation. The rotor structure and the experimental model are shown in Figure 9.

TABLE 4. ΔE% at different operating temperatures

T(°C)	*t*=0÷20ms		*t*=380÷400ms	
	E_{1NB}	*ΔE%*	E_{1NB}	*ΔE%*
20°C	328	15.9	328	15.9
60°C	320	18.6	296	28.2
120°C	306	23.8	195	94.7



(a)

(b)

Figure 9. Rotor and testing model of LSPMSM 15kW (a) rotor configuration, (b) testing model

The PMs are considered the source of magnetic flux (self-excitation) for the LSPMSM. If the PMs undergo demagnetization, the B-EMF of the LSPMSM will be reduced in both amplitude and sinusoidal shape. The testing scenario is conducted in two steps:

Step 1: Perform the motor starting test at a laboratory temperature of 20°C.

Step 2: Start the LSPMSM under high-temperature conditions. Due to the difficulty of heating the motor and considering that the LSPMSM is custom-built and will be used in several subsequent studies, heating the motor to 120°C could damage other components and lead to motor failure. Therefore, to avoid damaging the motor while still being able to investigate partial ID, the author group examined the motor with a temperature increase up to 60°C.

The prototype motor was fabricated and tested to validate the theoretical research and simulations presented. However, due to limitations in experimental conditions, only two cases at temperatures of 20°C and 60°C were conducted for comparison with the simulation results.

The operating temperature of the motor is controlled through the measured winding resistance. The heating process is carried out by operating the motor under rated load for a sufficient duration, with the cooling fan removed. The motor self-heats to reach approximately 60°C before stopping. After that, the motor is started to investigate the B-EMF characteristics. The results of the EMF characteristics when the LSPMSM starts at a temperature of 60°C are shown in Figure 10.

The experimental results shown in Figure 10, when the motor starts at a room temperature of 60°C, indicate that the B-EMF characteristics exhibit a concave peak shape, similar to that shown in Figure 7(b) when simulating the LSPMSM starting at a temperature of 60°C. This demonstrates that the magnetic field of the permanent magnet has been slightly diminished and cannot be fully recovered, unlike when starting the motor at a room temperature of 20°C. Thus, the PM has experienced partial ID in this case. The analysis of the waveform when testing at room temperature of 20°C and at 60°C is provided in Table 5.

(a) (b)

Figure 10. B-EMF waveform at temperatures (a) 20°C (b) 60°C

TABLE 5. Comparison of testing results and simulation

$T(^0C)$	20^0C	60^0C
E_{1NB}(V) - Simulation	328	296
E_{1NB}(V) - Testing	306	279

The results in Table 5 and Figures 10 show that there is a similarity in the B-EMF waveforms between the experimental results and the simulation results at room temperatures of 20°C and 60°C, with an error of less than 7%. The simulation results and experimental results are also fully consistent with the previously analyzed theory. This demonstrates that temperature has a significant impact on the demagnetization phenomenon of permanent magnets. To ensure the magnetic properties of the LSPMSM during operation, it is essential to consider the temperature factor and select an appropriate type of PM with a strong magnetic field and high-temperature tolerance to provide a reserve against partial ID.

6. CONCLUSION

The study has shown that temperature is a key factor affecting the demagnetization phenomenon and operational parameters of a LSPMSM. The increase in temperature during startup or under heavy load not only degrades the magnetic properties of the PMs but also affects the electromagnetic parameters of the LSPMSM, particularly concerning partial ID during motor startup. When partial ID occurs, the motor will operate with parameters that do not meet the required specifications, potentially even performing worse than an induction motor of the same power rating.

The theoretical research results indicate that, under certain temperature conditions, there exists a maximum starting current coefficient to prevent partial ID. Using FEA for a LSPMSM 15 kW, 2-pole the results showed consistency between theory and simulation, demonstrating that with a high starting current of 337.8 A, the NdFeB-N38M magnets used in the motor can experience mild demagnetization at 60°C, with pronounced partial ID occurring at 120°C. When partial ID occurs at 120°C, the LSPMSM loses synchronization and operates in an unstable mode, leading to increased current and a sudden drop in the B-EMF.

Additionally, the author group also constructed a testing LSPMSM of 15 kW, 2p=2 for experimental purposes. The results indicated that at normal operating temperatures, the B-EMF waveforms were consistent across theoretical, simulated, and experimental results. The experimental results confirmed the accuracy of the B-EMF waveforms, showing a similarity in shape between experimental and simulation data with an error of less than 7%.

The research results indicate that when operating at high temperatures, the theoretical formulas for selecting the dimensions of the PMs may no longer be accurate. partial ID can occur even when the LSPMSM is operating under rated load. When the LSPMSM starts, the load current drops, while the temperature increases, leading to a higher likelihood of partial ID. An appropriate starting current must be defined for each working temperature of the LSPMSM to prevent ID of the PMs. Therefore, during the design process, the dimensions of the PMs must ensure a correlation between current and temperature to avoid partial ID.

Thus, high-temperature resistant materials of PMs can be used to avoid partial ID; however, this will result in increased costs. Additionally, the electromagnetic circuit design can be optimized to reduce the starting current, or magnetic barriers can be employed to adjust the external magnetic field applied to the PMs in the rotor's magnetic circuit.

Funding

The authors declare that no financial support was received for this study.

Ethics Approval and Consent to Participate

This article does not involve any studies with human participants or animals performed by any of the authors. Therefore, ethics approval and consent to participate are not applicable.

Competing Interests

The authors declare no financial or organizational conflicts of interest.

Data Availability

The data that support the findings of this study are available upon reasonable request.

Authors Biosketches

Le Anh Tuan is a lecturer in the field of Electrical Engineering, currently working at the SEEE, HaUI,

Vietnam. He received his Ph.D. in Electrical Engineering from HUST, Vietnam, in 2018. His research areas include electric machines, electric machine simulation and electrical switching devices.

Tran Duy Khanh is a Lecturer in the Department of Electrical Engineering at Sao Do University (SDU), Vietnam. He is currently pursuing the Ph.D. degree in Electrical Engineering at Hanoi University of Mining and Geology, with a research focus on line-start permanent magnet synchronous motors (LSPMSMs).

Do Nhu Y received Ph.D. degree in Electrical Engineering in 2013 from Tula State University, Russia. He is the Head of the Department of Electrification and Deputy Director of the Center for Mine Electromechanical Research at the University of Mining and Geology (HUMG). His research interests include electric machines, power quality and renewable energy.

REFERENCES

1. Le AT, Lan NT, Do NY, Bun HV. Studying the effect of PM thickness on the back-EMF and power factor of LSPMSM. Journal Européen des Systèmes Automatisés. 2025;58(3):493-500. https://doi.org/10.18280/jesa.580307
2. Cepoi RD, Jaşcău FF, Szabó L. Current trends in energy efficient electrical machines. Journal of Electrical and Electronics Engineering. 2017;10(2):13-18.
3. Chaudhari M, Chowdhury A. Improved performance analysis of single-phase line start synchronous reluctance motor derived from induction motor. International Journal of Engineering, Transactions B: Applications. 2022;35(8):1641-50. https://doi.org/10.5829/ije.2022.35.08b.20
4. Baranski M, Szelag W, Jedryczka C. Influence of temperature on partial demagnetization of the permanent magnets during starting process of line start permanent magnet synchronous motor. In: International Symposium on Electrical Machines (SME); 2017. p. 1-6. https://doi.org/10.1109/ISEM.2017.7993535
5. Palangar MF, Soong WL, Mahmoudi A. Outer and inner rotor line-start permanent-magnet synchronous motors: an electromagnetic and thermal comparison study. In: IEEE Energy Conversion Congress and Exposition (ECCE); 2021. p. 1-8. https://doi.org/10.1109/ECCE47101.2021.9595574
6. Quoc VD, Huu HB. Modeling of surface-mounted permanent magnet synchronous motors with inner and outer rotor types. International Journal of Engineering, Transactions C: Aspects. 2024;37(12):2529-37. https://doi.org/10.5829/ije.2024.37.12c.11
7. Wang M, Zhu H, Zhou C, Zheng P, Tong C. Analysis and optimization of a V-shape combined pole interior permanent-magnet synchronous machine with temperature rise and demagnetization considered. IEEE Access. 2021;9:64761-75. https://doi.org/10.1109/ACCESS.2021.3076228
8. Faiza J, Rafieib V. Design of a three-phase outer rotor synchronous motor for hybrid vehicle. International Journal of Engineering, Transactions B: Applications. 2024;37(11):2380-91. https://doi.org/10.5829/ije.2024.37.11b.22
9. Gherabi Z, Toumi D, Benouzza N, Bendiabdellah A. A proposed approach for separation between short circuit fault, magnetic saturation phenomenon and supply unbalance in permanent magnet synchronous motor. International Journal of Engineering, Transactions A: Basics. 2020;33(10):1968-77. https://doi.org/10.5829/ije.2020.33.10a.15
10. Zawilak T, Jędryczka C. Risk of irreversible demagnetisation under transient states of the line start permanent magnet synchronous motor taking into account magnet temperature. Archives of Electrical Engineering. 2023;72(4):1107-19. https://doi.org/10.24425/aee.2023.147429
11. Yu P, Zhu C, Shen Y, Zhang G. Demagnetization analysis of line-start permanent magnet synchronous motors during its starting process. In: International Conference on Electrical, Automation and Mechanical Engineering (EAME); 2018. p. 1-6. https://doi.org/10.2991/eame-18.2018.11
12. Adly AA, Huzayyin A. The impact of demagnetization on the feasibility of permanent magnet synchronous motors in industry applications. Journal of Advanced Research. 2019;17:103-8. https://doi.org/10.1016/j.jare.2019.02.002
13. He H, Zhou N, Sun C. Efficiency decrease estimation of a permanent magnet synchronous machine with demagnetization faults. Energy Procedia. 2017;105:2718-24. https://doi.org/10.1016/j.egypro.2017.03.922
14. Moosavi SS, Djerdir A, Amirat YA, Khaburi DA. Demagnetization fault diagnosis in permanent magnet synchronous motors: a review of the state-of-the-art. Journal of Magnetism and Magnetic Materials. 2015;391:203-12.
15. Ebrahimi BM, Faiz J. Demagnetization fault diagnosis in surface mounted permanent magnet synchronous motors. IEEE Transactions on Magnetics. 2012;49(3):1185-92. https://doi.org/10.1109/TMAG.2012.2217978
16. Uršič L, Nemec M. Permanent magnet synchronous machine demagnetisation prevention and torque estimation control considering rotor temperature. IET Power Electronics. 2019;12(9):2161-9. https://doi.org/10.1049/iet-pel.2018.6162
17. Pietrzak P, Wolkiewicz M. Demagnetization fault diagnosis of permanent magnet synchronous motors based on stator current signal processing and machine learning algorithms. Sensors. 2023;23(4):1757. https://doi.org/10.3390/s23041757
18. Faiz J, Bazrafshan MA, Tabarniarami Z. Demagnetisation fault analysis and diagnosis based on different methods in permanent magnet machines—an overview. IET Electric Power Applications. 2024;18(12):1860-93. https://doi.org/10.1049/elp2.12519
19. You YM, Yoon KY. Multi-objective optimization of permanent magnet synchronous motor for electric vehicle considering demagnetization. Applied Sciences. 2021;11(5):2159. https://doi.org/10.3390/app11052159
20. Huang W, Wang J, Zhao J, Zhou L, Zhang Z. Demagnetization analysis and magnet design of permanent magnet synchronous motor for electric power steering applications. In: China International Youth Conference on Electrical Engineering (CIYCEE); 2020. p. 1-6. https://doi.org/10.1109/CIYCEE49808.2020.9332771
21. Dandan Q, Yanan J, Shuyang Z. The study of permanent magnet demagnetization in permanent magnet synchronous motor. In: International Conference on Advances in Electrical Engineering and Computer Applications (AEECA); 2021. p. 1-6. https://doi.org/10.1109/AEECA52519.2021.9574353
22. Kang DW. Analysis of vibration and performance considering demagnetization phenomenon of the interior permanent magnet motor. IEEE Transactions on Magnetics. 2017;53(11):1-7. https://doi.org/10.1109/TMAG.2017.2708421
23. Kim KC, Kim K, Kim HJ, Lee J. Demagnetization analysis of permanent magnets according to rotor types of interior permanent magnet synchronous motor. IEEE Transactions on Magnetics. 2009;45(6):2799-802. https://doi.org/10.1109/TMAG.2009.2018661
24. Choi G. Analysis and experimental verification of the demagnetization vulnerability in various PM synchronous

machine configurations for an EV application. Energies. 2021;14(17):5447. https://doi.org/10.3390/en14175447

25. Mahmouditabar F, Vahedi A, Ojaghlu P. Investigation of demagnetization effect in an interior V-shaped magnet synchronous motor at dynamic and static conditions. Iranian Journal of Electrical and Electronic Engineering. 2018;14(1):22-27.
26. Jia M, Hu J, Xiao F, Yang Y, Deng C. Modeling and analysis of electromagnetic field and temperature field of permanent-magnet synchronous motor for automobiles. Electronics. 2021;10(17):2173. https://doi.org/10.3390/electronics10172173
27. Bilgin O, Kazan FA. The effect of magnet temperature on speed, current and torque in PMSMs. In: International Conference on Electrical Machines (ICEM); 2016. p. 1-6. https://doi.org/10.1109/ICELMACH.2016.7732809
28. Truong Cong T, Nguyen Vu T, Bui Minh D, Vo Thanh H, Dang Quoc V. Analytical modelling of a six-phase surface mounted permanent magnet synchronous motor. International Journal of Engineering, Transactions A: Basics. 2024;37(7):1274-83.
29. Lee BH, Jung JW, Hong JP. An improved analysis method of irreversible demagnetization for a single-phase line-start permanent magnet motor. IEEE Transactions on Magnetics. 2018;54(11):1-5. https://doi.org/10.1109/TMAG.2018.282850

Persian Abstract

چکیده

موتور سنکرون آهنربای دائم با راه‌اندازی خطی (LSPMSM)به دلیل راندمان بالای خود، به طور فزاینده‌ای مورد استفاده قرار می‌گیرد. با این حال، دما تأثیر قابل توجهی بر پارامترهای الکترومغناطیسی دارد که به نوبه خود بر ویژگی‌های عملیاتی LSPMSM تأثیر می‌گذارد. به دلیل وجود اجزای قفس سنجابی در روتور، LSPMSM می‌تواند با راه‌اندازی خطی کار کند، اما جریان راه‌اندازی بالا می‌تواند منجر به مغناطیس‌زدایی موضعی یا کامل، به ویژه در دماهای بالا، شود. این امر می‌تواند منجر به مغناطیس‌زدایی برگشت‌ناپذیر آهنرباهای دائمی شود. این مقاله به بررسی تأثیر دما بر پارامترهای الکترومغناطیسی موتور و همچنین تأثیر آن بر مغناطیس‌زدایی در هنگام راه‌اندازی LSPMSM می‌پردازد. نتایج مطالعه نظری، تأثیر دما بر پارامترهای الکترومغناطیسی LSPMSM را نشان می‌دهد و نشان می‌دهد که در شرایط دمایی خاص، یک ضریب جریان راه‌اندازی حداکثری برای جلوگیری از مغناطیس‌زدایی وجود دارد. با استفاده از شبیه‌سازی FEA برای یک موتور LSPMSM با توان ۱۵ کیلووات و p=2، نتایج نشان‌دهنده‌ی همبستگی بین یافته‌های نظری و شبیه‌سازی شده است. نشان داده شده است که با جریان راه‌اندازی بالای ۳۳۷.۸ آمپر، آهنرباهای NdFeB-N38M مورد استفاده در موتور می‌توانند در دمای ۶۰ درجه سانتیگراد دچار مغناطیس‌زدایی خفیف شوند، در حالی که مغناطیس‌زدایی قابل مشاهده در دمای ۱۲۰ درجه سانتیگراد رخ می‌دهد. اعتبارسنجی تجربی، دقت را تأیید کرده است، زیرا ویژگی‌های B-EMF نشان‌دهنده‌ی شباهت نزدیک بین نتایج تجربی و شبیه‌سازی شده با انحراف کمتر از ۷٪ است.

IJE TRANSACTIONS A: Basics Vol. 39, No. 10, (October 2026) 2630-2643

International Journal of Engineering

Journal Homepage: www.ije.ir

Optimum design of a Novel Dual-PM Axial-Field Flux Reversal Machine for an In-wheel Direct-Drive Electric Vehicle Application

H. W. F. AL-Ward, S. E. Abdollahi*, S. A. Gholamian

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

P A P E R I N F O

Paper history:
Received 14 August 2025
Received in revised form 11 October 2025
Accepted 20 December 2025

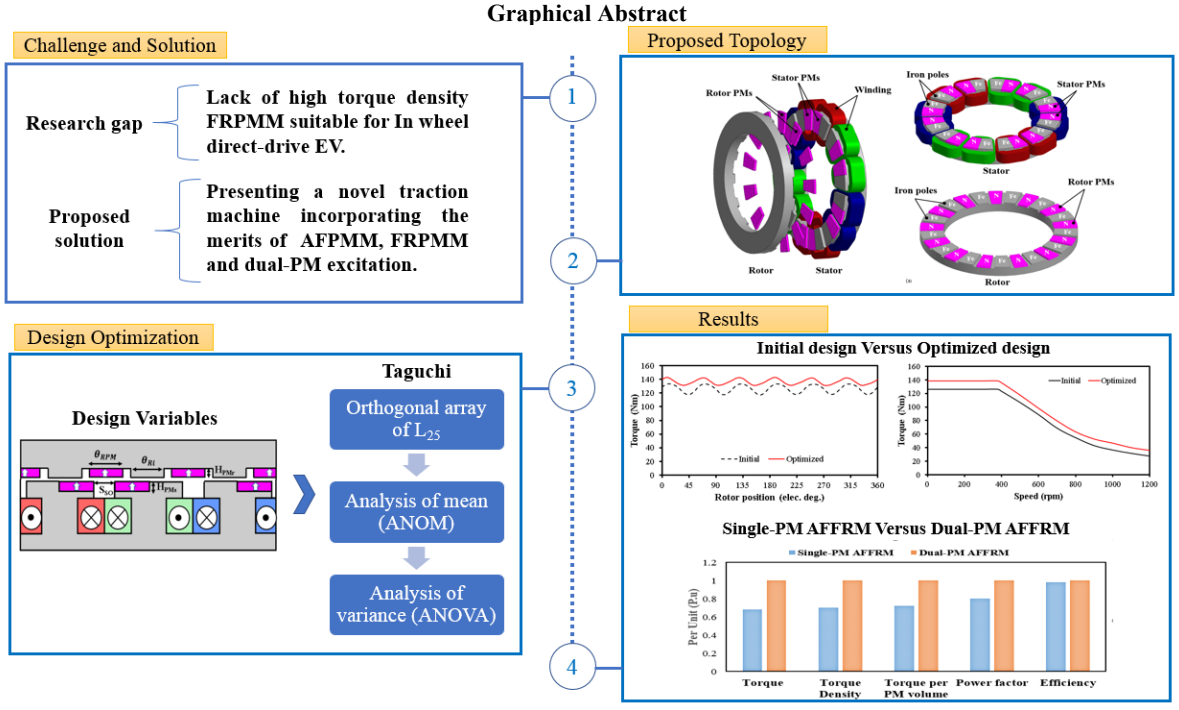
Keywords:
Dual Permanent Magnet
Axial Field Flux Reversal Machine
Torque Density
Taguchi Method
Electric Vehicle

A B S T R A C T

This article presents a novel dual permanent magnet axial field flux reversal machine (dual-PM AFFRM) with a single stator and a single rotor for an In-wheel direct drive electric vehicle (EV) application. The topology and operation principle of the machine were explained, and the impacts of using dual PM configuration on the airgap flux density harmonics were revealed using a simple magnetomotive force (MMF)-permeance model. To obtain the optimum torque characteristic of the machine and reduce the computational effort, the Taguchi method is employed. The electromagnetic performances of the optimized design are assessed via the 3-D finite element method (FEM). The outcomes show that the optimized machine offers high torque capability, low torque ripple, and good flux weakening capability. In addition, the results of a comparative study among eight different traction machines reveal that the dual-PM AFFRM offers the highest torque density with high aspect ratio, and high torque constant, which are crucial for direct-drive In-wheel EV applications.

doi: *10.5829/ije.2026.39.10a.22*

Graphical Abstract



*Corresponding Author Email: e.abdollahi@nit.ac.ir (S. E. Abdollahi)

Please cite this article as: AL-Ward HWF, Abdollahi SE, Gholamian SA. Optimum design of a Novel Dual-PM Axial-Field Flux Reversal Machine for an In-wheel Direct-Drive Electric Vehicle Application. International Journal of Engineering, Transactions A: Basics. 2026;39(10):2630-43.

1. INTRODUCTION

Recently, Flux Reversal Permanent Magnet machine (FRPMM), due to its desired characteristics, is emerging as a promising choice for direct-drive applications (1). The conventional type of FRPMM employs a reluctance rotor, which makes it suitable for applications where mechanical robustness is an area of concern. In addition, with permanent magnets (PMs) mounted on its stator teeth, heat management is easier and less complex. Nevertheless, the root problem of conventional FRPMM is its relatively low torque density caused by a high leakage flux between adjacent PMs (2).

Various types of FRPMMs have been explored and investigated to improve the torque density. A consequent-pole (CP) arrangement of PMs is employed to reduce the leakage flux and PM volume as well as enhance the PM utilization ratio and torque density of a conventional FRPMM (2). The results showed that CP-FRPMM delivers 26% higher average torque compared to the conventional FRPMM. The effects of PM's position on the torque density in CP-FRPMM were investigated by Yang et al. (3) using biased flux modulation theory. Qu et al. (4) introduced a novel type of CP-FRPMM with a larger number of iron poles than PM poles, which can improve the torque capacity of the machine up to 40% in comparison with a conventional CP-FRPMM. In addition, the optimal number of PM pieces on one stator tooth (5), and the adjacent teeth's PM polarities (6) are further studied to increase torque density. Besides, adding a flux-focusing feature to conventional FRPMM is proven to be an effective way to elevate the torque capacity. In this regard, an inset-type FRPMM is developed and studied b Kimy (7). In addition, inserting a quasi-Halbach array PMs (8) and spoke array PMs (9, 10) into the stator slots of FRPMMs is further examined. It is revealed that placing PMs into the stator slots of conventional FRPMMs improves the magnetic loading and torque production capacity. Apart from different arrangements of PMs, other methods have been used to boost the performance of FRPMM. For example. A FRPMM with distributed winding presented by More and Fernandes (11) offers high power density and efficiency. Li et al. (12) proposed a FRPMM with todoidal winding to improve the low pitch factor, which is accompanied by the desired slot-pole combination in the conventional FRPMM with concentrated winding. Another approach to increase the torque density is the utilization of the space occupied by the reluctance rotor in the conventional FRPMM. It is realized by accommodating an auxiliary set of windings inside the rotor slots, turning the conventional type into a novel doubly-fed FRPMM (13-15). Also, the concept of utilizing the inner space of the rotor led to the development of FRPMMs with mechanical offset complementary-stator (16), double-stator with quasi-Halbach array PMs (17), dual-stator with triple-PM excitations (18), partitioned stator hybrid-excited with dual-PM (19), and different topologies of partitioned stator (20-22) in which PMs are moved to the inner stator.

Despite the considerable improvement in the torque performance of FRPMM, the figures in the previous studies are not satisfactory for some demanding applications. For example, in an electric vehicle (EV) with an In-wheel direct-drive system, the traction motor is required to deliver high torque at low speed within the confined space of the wheel. The existing numbers reported for the torque density of the traction motors suitable for small-sized electric vehicles range from 30.2 kNm/m^3 to 37.7 kNm/m^3 (23, 24). Meanwhile, findings by Qu et al. (4) and Wu et al. (22) showed torque density figures for low-speed high-torque FRPMMs to be 17.6 kNm/m^3 and 8.62 kNm/m^3, respectively. Although higher numbers can be obtained in some studies, like Wei et al. (25) reached around 36 kNm/m^3, the complex structure of the stator with a rather long axial length makes it unattractive for In-wheel direct-drive systems, where the structural integrity and short axial length are of utmost importance. Thus, further improvement of torque density is necessary for FRPMM to be a better fit for such applications.

So far, most of the research on FRPMMs has been carried out on radial-flux topologies. It is evidenced by plenty of research investigations that axial flux permanent magnet machines (AFPMMs) offer a higher torque density with a shorter axial length and compact size, making them a viable solution, especially for high-torque low-speed applications (24). Up to now, there is one study conducted on axial field flux reversal permanent magnet machine (AFFRPMM), focusing on torque ripple reduction (26). Therefore, to combine the merits of FRPMM with those of the AFPMMs, this paper presents a new dual-PM axial field flux reversal machine (dual-PM AFFRM) with a single stator and a single rotor for an in-wheel direct drive EV. The structure and operation principle of the proposed machine are illustrated and explained using a simple magnetomotive force (MMF)-permeance model in section 2. The proposed machine requires a 3-D finite element method (FEM) for optimization and performance analysis, which consumes a significant amount of time. To reduce the computational time and obtain the optimum design, the Taguchi method is used in this study, and its implementation is explained in section 3. The electromagnetic performances of the optimized design are presented in section 4. To assess the capabilities of the proposed machine, its performance is compared to that of the AFPMMs and radial-flux FRPMMs designed for low-speed and high-torque applications in section 5. Finally, the main characteristics of the proposed machine are concluded in section 5.

2. TOPOLOGY AND OPERATING PRINCIPLE

Figure 1 shows the exploded model of the proposed dual-PM AFFRM. The machine has a doubly salient structure, and the armature coils are wound around its stator teeth, forming a non-overlapping concentrated winding, which features a short end-winding and low copper loss (27-29). There are two sets of PMs mounted on the stator teeth as well as in the rotor slots. The stator exploits the CP arrangement of PMs with N-Fe/Fe-N order, where "Fes" are ferromagnetic teeth used to reduce the interpolar leakage flux and act as a modulator for the MMF produced by the rotor PMs. Similarly, the rotor is comprised of a CP arrangement of PMs where all the PMs have the same magnetization directions as those of the stator, while ferromagnetic teeth of the rotor operate both as the opposite poles for the rotor PMs and the modulator for the MMF produced by the stator PMS.

The working mechanism of the dual-PM AFFRM is explained by depicting the individual contribution of the stator and the rotor PM excitations to the flux linkage of the stator coil (indicated by the light green) as shown in Figure 2. As can be seen, when the rotor position is θe=0° electrical degree in Figure 2(a), the flux produced by both the PM excitations is short-circuited through the iron teeth of the stator and rotor, and the flux linkage of the coil is zero. As the rotor position changes to θe=90° electrical degree in Figure 2(b), the rotor tooth is aligned with the stator N pole, while the rotor N pole coincides with the stator tooth. As a result, the flux linkage of the coil reaches its positive maximum. It is clear that in this position, both the flux produced by the stator and the rotor PMs strengthen each other, leading to a higher flux linkage in the stator coil. In Figure 2(c), when the rotor position is at θe=180°, the flux of the stator and the rotor PMs is short-circuited and the net flux linkage of the coil drops to zero. In Figure 2(d), the rotor is displaced to θe=270° electrical degree where both the ferromagnetic teeth of the stator and rotor are aligned with each other, creating a path for the flux produced by both excitations.

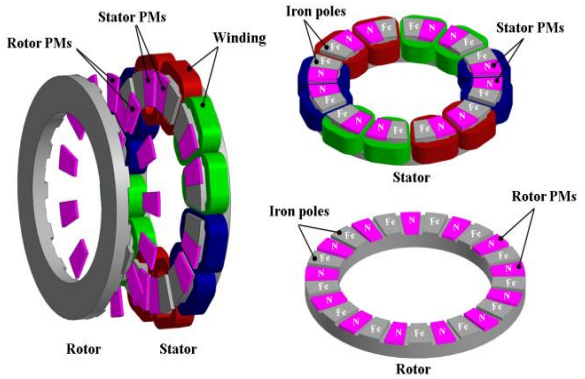


Figure 1. Topology of the proposed dual-PM AFFRM

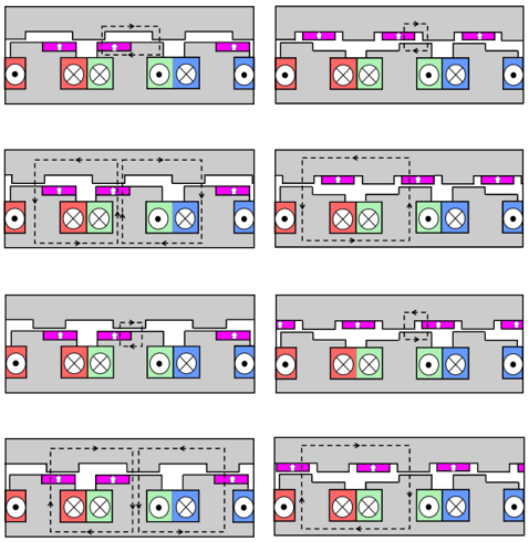


Figure 2. Flux lines of the dual-PM AFFRM. (a) θe=0°. (b) θe=90°. (c) θe=180°. (d) θe=270°

In this position, the flux linkage of the stator coil reaches its negative maximum. As analyzed the flux paths in different rotor positions, the proposed machine has a bipolar flux linkage which makes it suitable for BLAC operation mode.

For the presented dual-PM AFFRM, the magnetic field in the airgap is contributed by two different sources, namely the stator PM-MMF and the rotor PM-MMF. Considering the stator PMs only, the corresponding MMF can be expressed as follows (18):

$$F_{PMs}(\theta) = \sum_{n=1,3,5,\dots}^{\infty} F_{sn} \sin(n\frac{N_s}{2}\theta) \quad (1)$$

where N_s is the number of the stator slots, θ is the mechanical angle in the stator reference frame, and F_{sn} is the magnitude of the nth harmonic order of the MMF excited by the stator PMs (F_{PMs}). In addition, when the rotor PMs are only considered, the rotor PM-MMF can be described as follows (18):

$$F_{PMr}(\theta, t) = \sum_{m=1,3,5,\dots}^{\infty} F_{rm} \sin(mN_r\,(\theta - \omega_r t)) \quad (2)$$

where N_r is the number of rotor slots, ω_r is the angular speed of the rotor, and F_{rm} is the magnitude of the mth harmonic order of the MMF excited by the rotor PMs (F_{PMr}). To calculate the airgap flux density, both the stator and rotor permeances are necessary. The stator permeance function with smoothed rotor iron and slotted stator can be derived as follows (25):

$$\Lambda_s(\theta) = \sum_{i=0,1,2,\dots}^{\infty} \Lambda_{si} \cos(i\frac{N_s}{2}\theta) \quad (3)$$

In addition, the rotor permeance function is obtained with smoothed stator iron while the rotor is slotted (10):

$$\Lambda_r(\theta, t) = \sum_{j=0,1,2,\dots}^{\infty} \Lambda_{rj} \cos(\mathrm{j}\,N_r\,(\theta - \omega_r t)) \quad (4)$$

where Λ_{si} and Λ_{rj} are the magnitudes of the ith and the jth harmonics of the stator and the rotor permeances, respectively. As a result, the airgap flux densities excited by the stator PMs ($B_{ags}(\theta,t)$) and the rotor PMs ($B_{agr}(\theta,t)$) can be written as follows:

$$B_{ags}(\theta,t) = F_{PMs}(\theta)\,\Lambda_r(\theta,t) = \sum_{n=1,3,5,\dots}^{\infty}\sum_{j=0,1,2,\dots}^{\infty}\frac{1}{2}F_{sn}\,\Lambda_{rj}\left[\sin\left((jN_r \pm n\frac{N_s}{2})\theta - jN_r\omega_r t\right)\right] \quad (5)$$

$$B_{agr}(\theta,t) = F_{PMr}(\theta,t)\,\Lambda_s(\theta) = \sum_{m=1,3,5,\dots}^{\infty}\sum_{i=0,1,2,\dots}^{\infty}\frac{1}{2}F_{rm}\,\Lambda_{si}\left[\sin\left((mN_r \pm i\frac{N_s}{2})\theta - mN_r\omega_r t\right)\right] \quad (6)$$

When both the stator PM and rotor PM are taken into account, the no-load airgap flux density can be expressed as follows:

$$B_{ag}(\theta,t) = B_{ags}(\theta,t) + B_{agr}(\theta,t) \quad (7)$$

As can be seen, utilizing the dual-PM excitation in the presented machine gives birth to plenty of airgap flux density harmonics that can play active roles in power conversion. The airgap flux density harmonics can be categorized into four different groups, 1) The harmonics resulted from the interaction between stator PM-MMF components ($n = 1,3,5,\dots$) and the average rotor permeance component ($j = 0$), 2) The harmonics resulted from the interaction between rotor PM-MMF components ($m = 1,3,5,\dots$) and the average stator permeance component ($i = 0$), 3) the modulated harmonics generated by the interaction between stator PM-MMF ($m = 1,3,5,\dots$) and the rotor permeance components ($j = 1,2,3,\dots$), 4) the modulated harmonics generated by the interaction between rotor PM-MMF ($m = 1,3,5,\dots$) and the stator permeance components ($i = 1,2,3,\dots$). More information on the airgap flux density harmonics is included in Table 1.

To validate the above analysis, the airgap flux density waveform and its spectra of a dual-PM AFFRM with 12 stator slots and 13 rotor pole pairs are predicted with the aid of 3-D FEM in the average airgap radius and presented in Figure 3. As it is seen from Figure 3(a), the airgap flux density waveform of the proposed machine, which is the sum of the airgap flux densities produced by the stator and rotor PM MMFs, has unequal positive and negative peaks. This has occurred owing to the CP arrangements of PMs in both the stator and rotor. Figure 3(b) illustrates the harmonic spectra of the airgap flux densities produced by the stator PMs, rotor PMs, and dual-PM. As can be noticed, the dominant harmonics order while considering stator PMs are 6th and 18th, which were earlier categorized under type 1. In addition, the 5th, 7th, 19th, and 31st harmonics belong to the type 3 and 4, and as opposed to type 1, are not static, which means they are responsible for the induction of back-EMF in the windings of the proposed machine. On the other hand, in the case of the rotor PMs, the 13th harmonic is the dominant one, which is classified under type 2. Overall, by employing the dual-PM configuration, the new order of airgap flux density harmonics, such as the 13th emerges, which is absent in the conventional FRPMMs with only stator PMs excitation source. Furthermore, the amplitude of the modulated harmonics, such as the 5th, 7th, 19th, and 31st harmonics, increases, thanks to the rotor PM excitation.

TABLE 1. No-load Airgap Flux Density Harmonics

Pole pairs	Mechanical speed	Type
$n\frac{N_s}{2}$	0	(1)
mN_r	ω_r	(2)
$\left\|jN_r \pm n\frac{N_s}{2}\right\|$	$\frac{jN_r\omega_r}{jN_r \pm n\frac{N_s}{2}}$	(3)
$\left\|mN_r \pm i\frac{N_s}{2}\right\|$	$\frac{mN_r\omega_r}{mN_r \pm i\frac{N_s}{2}}$	(4)

3. DESIGN AND OPTIMIZATION OF THE PROPOSED DUAL-PM AFFRM

3. 1. Design Specifications A three-phase 5.5 kW dual-PM axial-field flux-reversal machine (AFFRM) was designed for a four-wheeler city-commuter electric car.

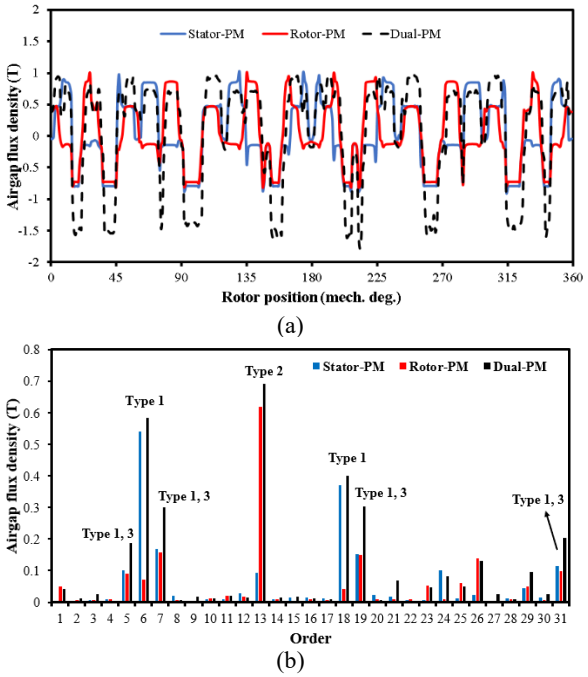


Figure 3. Airgap flux density distribution of the proposed Dual-PM AFFRM. (a) Waveforms. (b) Spectra

The target performance requirements (rated power, torque, and speed) were derived from the drive specifications reported by Kim et al. (30) for a similar class of in-wheel traction systems. For a conventional 225/45R17 tire, the calculated wheel speed at a vehicle speed of 45 km/h corresponds to approximately 380 rpm, which was adopted as the rated speed of the machine. The required torque of 138 Nm was then obtained from the power-speed relationship.

The outer and inner diameters (345.9 mm and 207.5 mm, respectively) and the axial length (78.8 mm) were selected according to the available installation space inside the wheel rim specified by Kim et al. (30). The current density of 5 A/mm² was chosen as a practical limit for liquid-cooled traction machines of this power level to balance copper loss and thermal constraints.

The electromagnetic configuration uses 12 stator slots and 13 rotor poles, a combination shown by Gao et al. (31) to provide a high winding factor (0.933) and nearly sinusoidal back-EMF with low cogging torque. NdFeB magnets with 1.2 T remanence are employed on both stator and rotor sides to achieve high air-gap flux density. The key design parameters are summarized in Table 2. Moreover, the analytical equations for the preliminary design stage are included in the Appendix.

3. 2. Design Optimization Generally, the performance of FRPM machine is highly influenced by multiple design variables (31). Hence, obtaining the desired features requires optimizing the designed model.

TABLE 2. Design Specifications of the Dual-pm AFFRM

Parameters	Value	Unit
Rated power, (P_o)	5.5	Kw
Rated speed, (n_r)	380	rpm
Rated current, ($I_m/\sqrt{2}$)	18	A
Rated voltage, (V_s)	220	V
Outer diameter, (D_o)	345.9	mm
Inner diameter, (D_i)	207.54	mm
Axial length, (l_m)	78.8	mm
Airgap length, (g)	1	mm
Stator yoke height, (h_{ys})	12.6	mm
Stator tooth tip, (h_1)	8	mm
Stator slot height, (h_s)	27	mm
Stator slots, (N_s)	12	-
Rotor pole pairs, (N_r)	13	-
Turns per phase. (n_{ph})	216	Turn
Current density, (J)	5	A/mm^2
Remnenace, (B_r)	1.2	T

Among many methods that have been presented so far, Global optimization based on GA (genetic algorithm) is widely employed to optimize the electromagnetic performances of radial FRPM machines (4, 5, 10). However, the proposed machine has a 3D structure, making its evaluation and optimization costly and a time-consuming task. As a result, to reduce the computational burden significantly and acquire the optimum geometric variable values, the Taguchi approach is chosen in this study and its overall optimization workflow is summarized in Figure 4.

The initial step in the Taguchi method is to set the objectives of the optimization. Since improving the torque characteristics is desirable for In-wheel traction motors, maximizing the average torque is the primary goal of optimization. In addition, to have a smooth operation, minimizing the torque ripple is considered the second goal of optimization in this paper. According to the studies that were conducted on radial FRPM machines, their performance is highly influenced by multiple stator teeth and rotor poles' geometric variables (4, 7, 31). Therefore, in this research, the stator PMs height (H_{PMs}), the rotor PMs height (H_{PMr}), the rotor Iron pole arc θ_{Ri} (elec. deg.), the rotor PM arc θ_{RPM} (elec. deg.) , and the stator slot opening ratio (S_{SO}), which is the ratio of the stator slot opening to the stator slot pitch, are taken as optimization variables (control factors) and shown in Figure 5. Each of the control factors has five levels, which are included in Table 3.

The Taguchi method uses a matrix called orthogonal array in which the impacts of different optimization variables on performance indexes can be evaluated in a limited set of experiments (32). The number of orthogonal array experiments is determined by the number and levels of optimization variables. Considering the optimization variables and their levels, the orthogonal

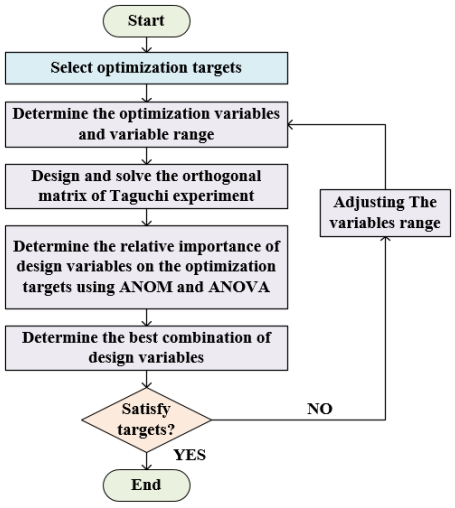


Figure 4. The workflow of implementing Taguchi method

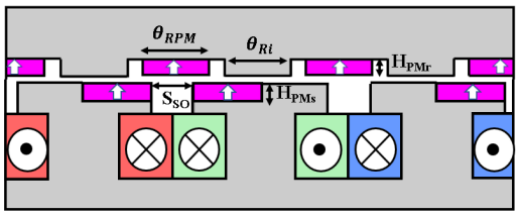


Figure 5. Design variables of the Dual-PM AFRPM machine

TABLE 3. Optimization Variables of the Dual-PM AFFRM

Control factors	Level 1	Level 2	Level 3	Level 4	Level 5
H_{PMs}(mm)	3	3.25	3.5	3.75	4
H_{PMr}(mm)	4	4.25	4.5	4.75	5
θ_{Ri}	126	135	144	153	162
θ_{RPM}	126	135	144	153	162
S_{SO}(%)	0.15	0.2	0.25	0.3	0.35

array of L_{25} is established for this study. The combinations of control factors for each experiment are presented in Table 4. It should be noted that without using the Taguchi method, the number of experiments (3-D simulations) required for the optimization rises to 5^5=3125. This shows that the Taguchi approach is a powerful tool to reduce costs and save time.

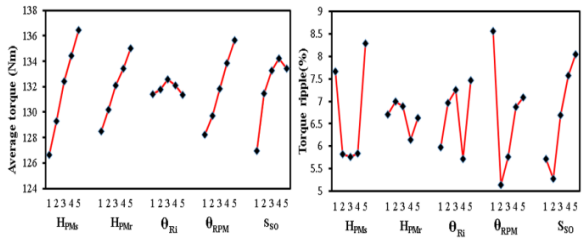
After completing all the experiments with the aid of 3-D FEA and obtaining the results, "Analysis of mean (ANOM)" and "Analysis of variance (ANOVA)" are employed to estimate the impacts of control factors on each optimization goal and determine their relative importance (13). The ANOM begins with calculating the average of each objective function as follows (32):

$$Y_m = \sum_{N=1}^{25} Y_N \quad (8)$$

where Y_m, Y_N, N and are the overall mean of an objective function, the value of the objective function at each experiment, and the number of experiments, respectively. In addition, the mean effect of control factors at different levels on objective functions can be calculated as follows (32):

$$Y_{mxk} = \frac{1}{N_L}(Y_{xk}(A) + Y_{xk}(B) + Y_{xk}(C) + Y_{xk}(D) + Y_{xk}(E) \quad (9)$$

where Y_{mxk} and N_L are the average value of the objective function at the k[th] level of the control factor x and the number of control factor levels, respectively. Besides, A, B, C, D, and E indicate the experiments in which a specific control factor is at the k[th] level. The results of the mean control factors' effects on the average torque and torque ripple are illustrated in Figure 6.



(a) (b)

Figure 6. Mean effects of control factors on (a) average torque (b) torque ripple

TABLE 4. Optimization Variables of the Dual-PM AFFRM

Experiment No	Optimization variables and levels				
	H_{PMs} (mm)	H_{PMr} (mm)	θ_{Ri}	θ_{RPM}	S_{SO} (%)
1	1	1	1	1	1
2	1	2	2	2	2
3	1	3	3	3	3
4	1	4	4	4	4
5	1	5	5	5	5
6	2	1	2	3	4
7	2	2	3	4	5
8	2	3	4	5	1
9	2	4	5	1	2
10	2	5	1	2	3
11	3	1	3	5	2
12	3	2	4	1	3
13	3	3	5	2	4
14	3	4	1	3	5
15	3	5	2	4	1
16	4	1	4	2	5
17	4	2	5	3	1
18	4	3	1	4	2
19	4	4	2	5	3
20	4	5	3	1	4
21	5	1	5	4	3
22	5	2	1	5	4
23	5	3	2	1	5
24	5	4	3	2	1
25	5	5	4	3	2

As can be seen from Figure 5(a), by increasing the levels of control factors H_{PMs}, H_{PMr}, θ_{RPM} and S_{SO} the average torque increases, which is in line with the primary goal of optimization. More specifically, the

combination of the fifth levels of H_{PMs} and H_{PMr}, the third level of θ_{Ri}, the fifth level of θ_{RPM} and the fourth level of S_{SO} maximizes the average torque. However, as shown in Figure 5(b), choosing the third level of H_{PMs}, the fourth level of H_{PMr} and θ_{Ri}, and the second level of θ_{RPM} and S_{SO} is favorable for minimizing the torque ripple. To avoid such a contradiction and obtain the best combination that can simultaneously satisfy both optimization goals, ANOVA is carried out. To reveal the influence share of each control factor on objective functions, the sum of squares (SS) of each control factor for an objective function is calculated using Equation 10, which is presented by Mirnikjoo et al. (32)

$$SS_x = 5\sum_{k=1}^{5}(Y_{mxk} - Y_m)^2 \quad (10)$$

where SSx is the sum of squares of the control factor x The results of the ANOVA analysis are presented in Table 5. As can be deduced, the height of stator PM (H_{PMs}) and rotor PM (H_{PMr}) have larger impacts on the average torque rather than the torque ripple. On the contrary, the proportion influence of the rotor PM arc θ_{RPM}, rotor Iron pole arc θ_{Ri}, and the stator slot opening ratio S_{SO} are larger on the torque ripple. Therefore, based on the calculated control factors effects, the combination with the fifth levels of H_{PMs} and H_{PMr}, the fourth level of θ_{Ri}, the second level of θ_{RPM} and S_{SO} are selected to satisfy both optimization targets simultaneously.

4. EVALUATION OF THE ELECTROMAGNETIC PERFORMANCE OF DUAL-PM AFFRM

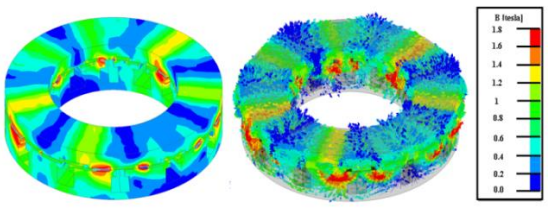
4. 1. No-load Condition The no-load flux density distribution and the vector field of the developed motor are depicted in Figure 7. As can be observed, the flux density is in the allowed range in most parts of the machine. However, due to the interpolar flux leakage and the lower magnetic reluctance of the iron poles, partial saturation occurs in the CP areas of the stator. In addition, the airgap flux density waveforms and the corresponding spectra of main-order harmonics in the initial and optimized designs are illustrated in Figure 8. As it is observed, the amplitudes of the main-order airgap flux density harmonics responsible for producing torque increased remarkably due to the optimization.

TABLE 5. Effects of Different Control Factors on Objective Functions

Optimization variables	Average Torque		Torque ripple	
	SS	Factor effect (%)	SS	Factor effect (%)
H_{PMs}	308.05	38.6	30.8	27.77
H_{PMr}	133.53	16.73	2.29	2.06
θ_{Ri}	4.92	0.62	13.08	11.78
θ_{RPM}	180.75	22.65	34.37	30.96
S_{SO}	170.64	21.4	30.44	27.43

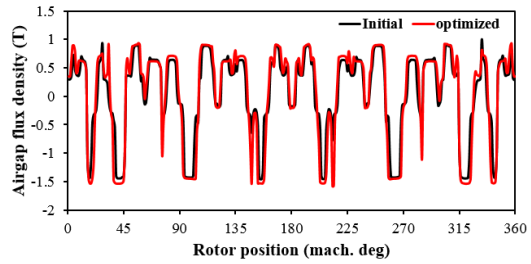
Since the flux reversal machines cannot generate the reluctance torque, the cogging torque is the major contributor to torque ripple (12). The cogging torque profiles of the proposed machine in initial and optimized designs are predicted via 3-D FEM and compared in Figure 9. As a result of using the Taguchi method, the peak-to-peak value of cogging torque is reduced from 13.67 Nm to 2.73 Nm in the optimized design. The phase back-EMF of the dual-PM AFFRM is calculated at the rated speed and presented in Figure 10. As it is clear, the amplitude of the fundamental component of the phase back-EMF increased by 10.67% in the optimized model. In addition, the fifth and seventh harmonics are substantially suppressed, leading to a lower torque ripple in the optimized machine.

4. 2. On-load Condition The output torque waveforms of both the initial and optimized designs of the proposed model are illustrated in Figure 11(a).

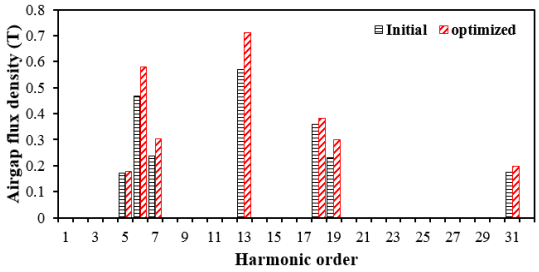


(a) (b)

Figure 7. (a) Flux density distribution of the dual-PM AFFRM (b) Flux vectors of the dual-PM AFFRM



(a)



(b)

Figure 8. Airgap flux density distributions of the initial and optimized design. (a) Waveforms. (b) Spectra

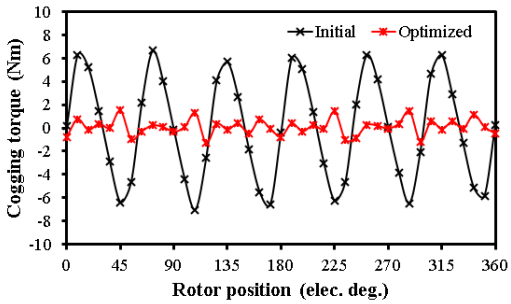


Figure 9. Cogging torque waveforms of the initial and optimized design of the dual-PM AFFRM

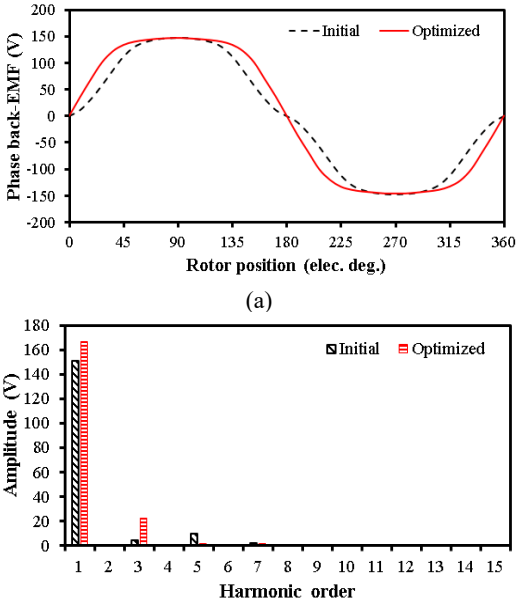


Figure 10. Phase back-EMF in the dual-PM AFFRM (a) Waveform (b) Spectra

Thanks to the Taguchi method, the average torque in the optimized design experienced nearly 10% growth and reached 138.6 Nm. Moreover, the torque ripple decreased from 12.95% to 4.3%. Since the torque ripple is under 5%, it is in the acceptable range for an In-wheel application. Figure 11(b) shows the variations of the average torque with respect to the current angle. It is obvious that, like its radial counterpart, the proposed machine, both in the initial and optimized models, delivers an insignificant reluctance torque that can be neglected. To evaluate the overload capability of the proposed machine, the output torque versus input current curve is depicted in Figure 11(c). As it is seen, the average torque of the proposed machine demonstrates a linear increase over twice the input current. Besides, the optimized model produces a higher torque compared with the initial one over the entire range of input current.

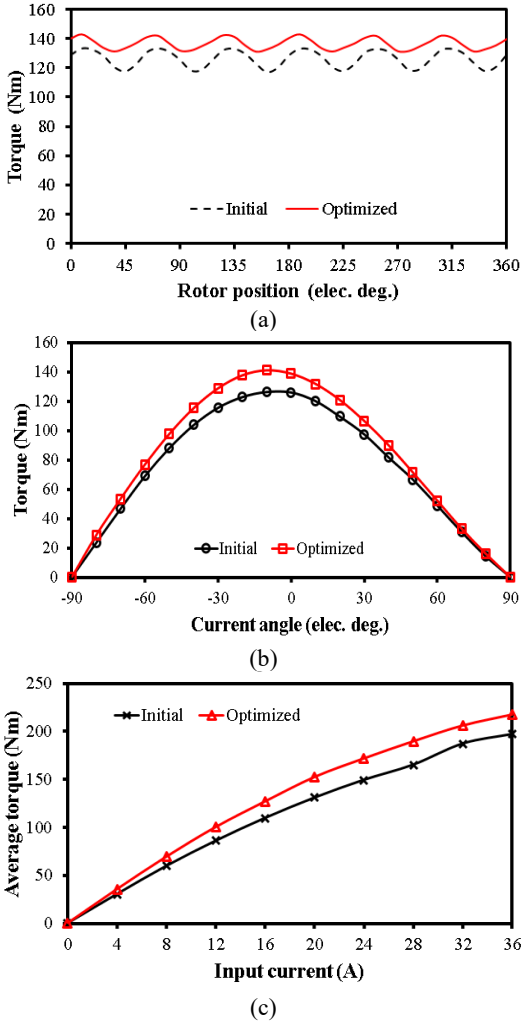


Figure 11. (Torque characteristics of the dual-PM AFFRM. (a) Output torque (b) Torque versus current angle (c) Torque versus input current

4. 3. Flux Weakening Capability According to Qu et al. (10), a K_{fw} factor indicates the field weakening capability of an electric machine and can be expressed as follows:

$$k_{fw} = \frac{L_d I_{rated}}{\Psi_m} \quad (11)$$

Where L_d, Ψ_m, and I_{rated} are the d-axis inductance, PM flux linkage, and the phase rated current. It is preferred k_{fw} to be close to 1 so that the machine has both torque capability in the constant torque region and flux weakening capability in the constant power region (10). For the initial and optimized design, the flux weakening factors are 1.26 and 1.41, respectively. The optimized design shows a figure closer to 1, indicating a better trade-off between the torque and flux weakening capabilities. The values for d and q inductances and PM flux linkages of both models are listed in Table 6.

TABLE 6. Inductances, flux linkages, and flux weakening factors of the initial and optimized designs

	L_d (mH)	L_q (mH)	Ψ_m (mwb)	K_{fw}
Intial	16.08	16.87	321.82	1.26
Optimized	15.95	16.81	289.96	1.41

Figure 12 illustrates the torque-speed and power-speed characteristics of the initial and optimized models. For this purpose, an SVMPWM control strategy with the DC=540 and the maximum allowed current for the windings is employed. As can be observed, both designs can operate over a wide speed range; however, the optimized design exhibits a slightly better performance.

4. 3. Comparison between single-PM AFFRM and dual-PM AFFRM To reveal the impacts of employing dual-PM excitation, the key performance of the proposed machine is compared with a single-PM AFFRM in which PMs are mounted just on the stator, like a conventional radial field FRPMM. It should be mentioned that the torque characteristic of the single-PM AFFRM is optimized using the same method in Section 3. Furthermore, the overall dimensions, input current, number of stator slots, rotor pole pairs, and the number of turns per phase are kept constant for both machines to make a fair analogy. The results of the comparison are tabulated in Table 7. As can be noticed, the average torque of the dual-PM AFFRM is 47% larger compared to its counterpart. As a result, the torque density of the proposed machine grew significantly to 42.92 kNm/m^3 while that of the single-PM AFFRM is 29.19 kNm/m^3. Although the single-PM AFFRM has a lower torque density compared to the dual-PM AFFRM, it ranks higher compared to its radial-flux counterparts developed in the previous studies. In addition, due to the increased number of airgap flux density harmonics in the dual-PM AFFRM, the PM utilization ratio reached 0.54 Nm/cm^3, which is 38.46% higher than in the single-PM AFFRM. Similarly, the power factor of the proposed machine has risen to 0.6 owing to the increase in the airgap flux density harmonics. However, the enriched airgap flux density harmonics led to a higher core loss in the proposed machine. Nevertheless, the dual-PM AFFRM exhibits a higher efficiency compared to its counterpart.

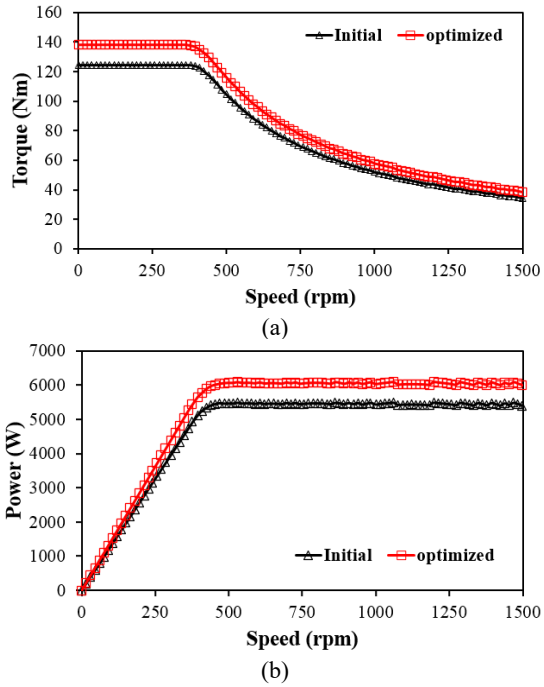


Figure 12. (a) Torque-speed characteristics (b) Power-speed characteristics

5. COMPARATIVE STUDY

Propulsion motors designed for low-speed operation are bulky and heavy. Accordingly, the main concern in In-wheel direct-drive systems is the increased unsprung mass created by the placement of such motors inside the wheel. Therefore, a desirable in-wheel motor must be light and compact, as it simultaneously satisfies other requirements for EV applications. To further investigate the suitability of the proposed motor for in-wheel direct-drive EVs, three different types of high-performance motors designed for low-speed and high-torque applications are chosen from the previous studies for comparison purposes. The candidates are a dual-stator axial-flux spoke-type permanent magnet vernier machine (DS-AFSPMVM) proposed by Zhao et al. (24), indicated by M1 in Table 8, a dual-rotor toroidal winding axial-flux permanent magnet vernier machine (DR-TWAFPMVM) proposed by Zou et al. (23), indicated by M2 in Table 8, and a dual rotor slotless axial-flux permanent magnet

TABLE 7. Comparison of Electromagnetic Performance between Single-PM AFFRM and Dual-PM AFFRM

Items	Single-PM AFFRM	Dual-PM AFFRM
Core loss (W)	56.7	122.6
Copper loss (W)	291.6	
Torque (Nm)	94.3	138.65
Torque ripple (%)	3.57	4.3
Power factor	0.48	0.6
Efficiency (%)	91.5	93
PM volume (cm^3)	244.32	253.355
Torque/PM volume (Nm/cm^3)	0.39	0.54
Torque density (kNm/m^3)	29.19	42.92

synchronous machine (DR-SLAFPMSM) proposed by Si et al. (33), indicated by M3 in Table 8. To draw a fairer analogy, the candidates are chosen to have the operating speed close to that of the proposed motor. In addition, all the motors are air-cooled. As can be seen from Table 8, the torque density of the proposed motor is 13.84%, 31.65%, and 15.22% higher than those of the M1, M2, and M3, respectively. In addition, the torque constant of the Dual-PM AFFRM is 22.22% and 36.52% higher compared to those of the M2 and M3. In terms of efficiency, the proposed motor outperforms both the M1 and M3, while it falls 1.3% below the M2. Another index for comparing the motors is the aspect ratio, which is defined as the ratio of a motor's outer diameter to its axial length. Considering the results, the aspect ratio of the presented motor is remarkably high (4.39). This signifies that the motor has a more pancake shape, which makes it suitable for in-wheel applications. Furthermore, the presented design has a single-sided structure, while other designs have double-sided structures. This indicates that, when it comes to manufacturing complexity and cost, the proposed design is a more attractive candidate in comparison to others.

To have a clearer perspective, some of the best-performed radial-flux FRPMMs are chosen from the previous studies to identify how much the torque density improved in the proposed machine compared to its radial-flux counterparts. The candidates are a dual-PM spoke-type flux reversal machine (dual-PM SFRM) proposed by Wei et al. (25), indicated by M4 in Table 9, an FRPMM with quasi-Halbach array magnets proposed by Qu and Zhu (9), indicated by M5 in Table 9, and a CP-FRPMM proposed by Qu et al. (4), indicated by M6 in Table 9. As before, to make a fair comparison, the candidates are chosen to have a close operating speed to that of the Dual-PM AFFRM. It is seen from Table 8, the torque density of the proposed machine is 16.25%, 16.88%, and 143.86% larger in comparison with those of the M4, M5, and M6, respectively. Also, the proposed model has a considerably higher torque constant than M4 and M6, while it is lower by 15.71% compared to M6. With regard to efficiency, the Dual-PM AFFRM outperforms its other competitors by far. The figures for aspect ratio demonstrate that the Dual-PM AFFRM is more compact than the radial-flux FRPMMs chosen for this study.

Overall, it can be inferred that the Dual-PM AFFRM has a promising performance for In-wheel direct-drive EV applications since it has the largest torque density among the 8 designs.

TABLE 8. Specifications of the proposed, M1, M2 amd M3 motors

Parameters	Proposed	M1	M2	M3
Speed (rpm)	380	300	320	400
Outer diameter (mm)	345.9	334	220	240
Axial length (mm)	78.8	75	91	58
Phase number		3		
Current density (A/mm^2)	5	4.4	5.8	-
Average torque (Nm)	138.65	160.8	106.1	56.44
Torque density (KNm/m^3)	42.92	37.7	32.6	37.25
Torque constant (Nm/A)	7.7	-	6.3	5.64
efficiency	93	86	94.3	84.83
Aspect ratio	4.39	4.45	2.41	4.13

TABLE 9. Specifications of the proposed, M4, M5 amd M6 motors

Parameters	Proposed	M4	M5	M6
Speed (rpm)	380		600	
Outer diameter (mm)	345.9	90	124	90
Axial length (mm)	78.8	25	120	25
Phase number		3		
Current density (A/mm^2)	5	-	-	-
Average torque (Nm)	138.65	5.8	50.1	2.8
Torque density (KNm/m^3)	42.92	36.92	36.72	17.6
Torque constant (Nm/A)	7.7	0.29	9.32	0.53
efficiency	93	85.4	87.2	87.2
Aspect ratio	4.39	3.6	1.03	3.6

5. CONCLUSION

In this study, a dual-permanent-magnet axial-field flux-reversal machine (dual-PM AFFRM) with a single stator and single rotor configuration was presented for in-wheel direct-drive electric vehicle (EV) applications. The operating principle was analyzed using the magnetomotive force (MMF)–permeance model, and the torque characteristics were refined through the Taguchi optimization method to reduce computational effort. After optimization, the average torque increased by approximately 10%, and the torque ripple decreased to 4.3%, which is within a typical range for in-wheel drive systems. Comparative analysis with a single-PM AFFRM showed higher torque density (from 29.19 kNm/m³ to 42.92 kNm/m³) and improved power factor and magnet utilization with the dual-PM structure. Furthermore, benchmarking against several existing axial- and radial-flux machines designed for low-speed, high-torque applications indicated that the proposed configuration provides comparable electromagnetic performance and structural compactness. Overall, the presented design demonstrates the feasibility of the dual-PM AFFRM topology for in-wheel traction applications. Future work will address detailed thermal and mechanical analyses, analytical lifetime estimation, and experimental validation of the proposed concept.

6. APPENDIX

According to the required power for the machine and the application, its overall dimensions are calculated. Based on Huang et al. (34), if the stator resistance and leakage inductance are not considered, the output power of the machine can be defined as follows:

$$P_o = \eta \frac{m}{\mathrm{T}} \int_0^{\mathrm{T}} e(t)\, i(t)\, dt = \eta \frac{m}{2} E_m I_m \quad (12)$$

where η, E_m, I_m, and m are the machine efficiency, the maximum values for EMF and phase current, and the number of phases of the machine, respectively. Assuming the machine is working in the generating mode, the induced EMF in the stator phase winding can be expressed by Equation 13.

$$e(t) = -\frac{d\Psi_m}{dt} = -n_{ph} k_W \frac{d\varphi_m}{d\theta}\frac{d\theta}{dt} = -n_{ph} k_W \frac{d\varphi_m}{d\theta}\omega_r \quad (13)$$

where k_W,Ψ_m, φ_m, and θ are the winding factor, the phase flux linkage created by both PM excitations, the flux linkage of one turn created by both PM excitations, and the rotor angular position, respectively. The phase flux linkage can also be defined as follows:

$$\varphi_m = \varphi_p \cos(N_r \theta) \quad (14)$$

By inserting Equation 14 into Equation 13, the EMF can be rewritten as follows:

$$e(t) = n_{ph} k_W N_r\, \omega_r \varphi_p \sin(N_r\theta) = E_m \sin(N_r\theta) \quad (15)$$

According to Figures 2(b) and 2(d), the maximum flux passing through the stator coil occurs when both PMs on the stator and rotor are aligned with the rotor and stator iron teeth, respectively. In these positions, the area of the airgap (A_g) between the stator and rotor is derived by Equation 16.

$$A_g = \frac{\pi}{8}\left(\frac{w_{PMr}}{N_r} + \frac{w_{PMs}}{N_s}\right)(1 - {k_d}^2)\, {\mathrm{D}_o}^2 \quad (16)$$

In which w_{PMr}, w_{PMs}, and k_d are the ratio of rotor PM width to rotor tooth pitch, the ratio of stator PM width to stator slot pitch, and the ratio of inner diameter to outer diameter of the machine, respectively. Thus, the magnitude of flux per turn can be obtained by Equation 17.

$$\varphi_p = \mathrm{B}_m \frac{\pi}{8}\left(\frac{w_{PMr}}{N_r} + \frac{w_{PMs}}{N_s}\right)(1 - {k_d}^2)\, {\mathrm{D}_o}^2 \quad (17)$$

where B_m is the maximum value of airgap flux density in the no-load condition. By using Equation 17, the E_m can be defined as follows:

$$E_m = n_{ph} k_W N_r \omega_r \mathrm{B}_m \frac{\pi}{8}\left(\frac{w_{PMr}}{N_r} + \frac{w_{PMs}}{N_s}\right)(1 - {k_d}^2)\, {\mathrm{D}_o}^2 \quad (18)$$

The maximum phase current of the machine is related to the stator electrical loading A_s. It can be expressed by the following equation.

$$I_m = \frac{\sqrt{2}\pi A_s}{6\, n_{ph}}\left(\frac{1+k_d}{2}\right)\mathrm{D}_o \quad (19)$$

By inserting Equations 18 and 19 into Equation 12, the output power of the machine can be calculated using Equation 20.

$$P_o = \frac{\sqrt{2}\pi^3}{1920}\eta k_W \mathrm{B}_m A_s n_r \left(w_{PMr} + \frac{w_{PMs}}{N_s}\right)(1 - {k_d}^2)(1 + k_d){\mathrm{D}_o}^3 \quad (20)$$

Also, the output torque of the machine can be obtained by Equation 21.

$$T_o = \frac{\sqrt{2}\pi^2}{64}\eta k_W \mathrm{B}_m A_s \left(w_{PMr} + \frac{w_{PMs}}{N_s}\right)(1 - {k_d}^2)(1 + k_d){\mathrm{D}_o}^3 \quad (21)$$

As it is seen from Figure 13, there are six more geometric design parameters that need to be derived to complete the design process. Hereafter, the equations for the calculation of these design parameters are explained.

If the flux density of the stator yoke (B_{ys}) is fixed, the thickness of the stator yoke can be calculated as below.

$$h_{ys} = \frac{k_d \mathrm{D}_o \mathrm{B}_m}{2\, P_a\, B_{ys}\, K_{stk}} \quad (22)$$

In which P_a and K_{stk} are the stator winding pole pair and the stacking factor of the stator and rotor cores. For the proposed design P_a is calculated via Equation 23.

$$P_a = 3\frac{N_s}{2} - N_r \quad (23)$$

In addition, by assuming the flux density of the stator teeth to be B_{st}, the stator tooth width can be obtained by Equation 24.

$$w_{ts} = \frac{2\, k_d \mathrm{D}_o \mathrm{B}_m}{N_s\, B_{st}\, K_{stk}} \quad (24)$$

To increase the output power/torque and protect the PMs against demagnetization, the stator PMs height ($\mathrm{H_{PMs}}$), rotor PMs height ($\mathrm{H_{PMr}}$) are selected as below:

$$H_{PMs} = H_{PMr} = 4g \sim 6g \quad (25)$$

The optimum ratio for the stator slot opening S_{so} in radial-field FRPMMs is 0.25 (1) and can be calculated by Equation 26. It should be noted that parameters related to the slot area are calculated at the innermost diameter of the stator since the slot area of the stator must be constant in the radial direction.

$$S_{so} = \frac{\pi(k_d \mathrm{D}_o)}{4\, N_s} \quad (26)$$

By using the winding electric loading A_s and the current density J_s, the slot area of the stator can be obtained.

$$A_{slot} = \frac{\pi\, k_d \mathrm{D}_o A_s}{24\, k_f\, J_s} \quad (27)$$

where k_f is the slot filling factor. Also, the stator slot area can be calculated by the stator geometric parameters.

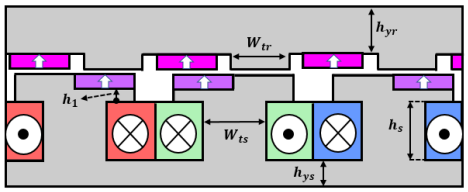


Figure 13. Parametric model of Dual-PM AFFRM

$$A_{slot} = h_s(\frac{\pi\, k_d \mathrm{D}_o}{N_s} - w_t) \quad (28)$$

Equations 27 and 28, the height of the stator slots h_s can be found. Consequently, the axial length of the stator l_s is calculated using Equation 29.

$$l_s = H_{PMs} + h_1 + h_s + h_{ys} \quad (29)$$

Similar to the stator slot calculations, by setting the rotor yoke and teeth flux densities to B_{yr} and B_{rt}, respectively, the rotor yoke height h_{yr} and rotor tooth width w_{tr} are determined as:

$$h_{yr} = \frac{k_d \mathrm{D}_o \mathrm{B}_m}{2\, N_r\, B_{yr}\, K_{stk}} \quad (30)$$

$$w_{tr} = \frac{2\, k_d \mathrm{D}_o \mathrm{B}_m}{N_r\, B_{rt}\, K_{stk}} \quad (31)$$

The rotor length is now calculated by Equation 32.

$$l_r = H_{PMr} + h_{yr} \quad (32)$$

Based on the presented analytical equations, the design of the dual-PM AFFRM can be completed by following the procedure depicted in Figure 14.

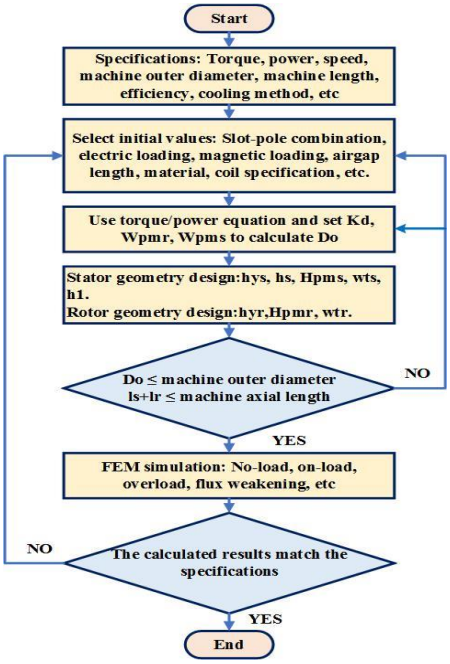


Figure 14. Design flowchart of Dual-PM AFFRM

Funding

The authors declare that no financial support was received for this study.

Declarations Ethics Approval and Consent to Participate

Not applicable. This study did not involve human participants or animals.

Data Availability

All data generated or analyzed during this study are included in this published article.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this manuscript, the authors used ChatGPT to improve the clarity of language and to correct grammar. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

1. Gao Y, Qu R, Li D, Li J. Torque performance analysis of three-phase flux reversal machines. IEEE Transactions on Industry Applications. 2017;53(3):2110-9. https://doi.org/10.1109/TIA.2017.2677356
2. Gao Y, Qu R, Li D, Li J, Zhou G. Consequent-pole flux-reversal permanent-magnet machine for electric vehicle propulsion. IEEE Transactions on Applied Superconductivity. 2016;26(4):1-5. https://doi.org/10.1109/TASC.2016.2514345
3. Yang H, Zhu ZQ, Lin H, Li H, Lyu S. Analysis of consequent-pole flux reversal permanent magnet machine with biased flux modulation theory. IEEE Transactions on Industrial Electronics. 2019;67(3):2107-21. https://doi.org/10.1109/TIE.2019.2902816
4. Qu H, Zhu Z, Li H. Analysis of novel consequent pole flux reversal permanent magnet machines. IEEE Transactions on Industry Applications. 2020;57(1):382-96. https://doi.org/10.1109/TIA.2020.3036328
5. Li HY, Zhu Z. Optimal number of magnet pieces of flux reversal permanent magnet machines. IEEE Transactions on Energy Conversion. 2018;34(2):889-98. https://doi.org/10.1109/TEC.2018.2866765
6. Li H, Zhu Z. Influence of adjacent teeth magnet polarities on the performance of flux reversal permanent magnet machine. IEEE Transactions on Industry Applications. 2018;55(1):354-65. https://doi.org/10.1109/TIA.2018.2867818
7. Kim TH. A study on the design of an inset-permanent-magnet-type flux-reversal machine. IEEE Transactions on Magnetics. 2009;45(6):2859-62. https://doi.org/10.1109/TMAG.2009.2018776
8. Xie K, Li D, Qu R, Yu Z, Gao Y, Pan Y. Analysis of a flux reversal machine with quasi-Halbach magnets in stator slot opening. IEEE Transactions on Industry Applications. 2018;55(2):1250-60. https://doi.org/10.1109/TIA.2018.2873540

9. Qu H, Zhu Z. Analysis of spoke array permanent magnet flux reversal machines. IEEE Transactions on Energy Conversion. 2020;35(3):1688-96. https://doi.org/10.1109/TEC.2020.2986948
10. Qu H, Zhu Z, Shuang B. Influences of PM number and shape of spoke array PM flux reversal machines. IEEE Transactions on Energy Conversion. 2020;36(2):1131-42. https://doi.org/10.1109/TEC.2020.3031030
11. More D, Fernandes B. Power density improvement of three phase flux reversal machine with distributed winding. IET Electric Power Applications. 2010;4(2):109-20. https://doi.org/10.1049/iet-epa.2009.0003
12. Li H, Zhu ZQ, Hua H. Comparative analysis of flux reversal permanent magnet machines with toroidal and concentrated windings. IEEE Transactions on Industrial Electronics. 2019;67(7):5278-90. https://doi.org/10.1109/TIE.2019.2934064
13. Wu L, Zheng Y, Fang Y, Huang X. Novel fault-tolerant doubly fed flux reversal machine with armature windings wound on both stator and rotor teeth. IEEE Transactions on Industrial Electronics. 2020;68(6):4780-9. https://doi.org/10.1109/TIE.2020.2989721
14. Zheng Y, Wu L, Zhu J, Fang Y, Qiu L. Analysis of dual-armature flux reversal permanent magnet machines with Halbach array magnets. IEEE Transactions on Energy Conversion. 2021;36(4):3044-52. https://doi.org/10.1109/TEC.2021.3070039
15. Zheng Y, Wu L, Fang Y, Li T, Zheng W. Comparative analysis of doubly fed flux-reversal permanent magnet machines with different PM arrangements and consequent-pole topologies. IEEE Transactions on Magnetics. 2020;57(2):1-6. https://doi.org/10.1109/TMAG.2020.3019821
16. Yu J, Liu C, Zhao H. Design and optimization procedure of a mechanical-offset complementary-stator flux-reversal permanent-magnet machine. IEEE Transactions on Magnetics. 2019;55(7):1-7. https://doi.org/10.1109/TMAG.2019.2896177
17. Aslani B, Gholamian S, Abdollahi SE. Analysis of a double stator flux reversal permanent magnet machine with Halbach array magnets. In: 14th Power Electronics, Drive Systems, and Technologies Conference (PEDSTC); 2023. p. 1-6. https://doi.org/10.1109/PEDSTC57673.2023.10087099
18. Meng Y, Fang S, Wang H, Pan Z, Qin L. Design and analysis of a new dual-stator consequent-pole flux reversal machine with triple-PM excitation. IEEE Transactions on Magnetics. 2021;57(6):1-4. https://doi.org/10.1109/TMAG.2021.3070855
19. Meng Y, Fang S, Pan Z, Liu W, Qin L. Design and analysis of a new partitioned stator hybrid-excited flux reversal machine with dual-PM. IEEE Transactions on Magnetics. 2021;58(2):1-6. https://doi.org/10.1109/TMAG.2021.3077035
20. Wu Z, Zhu Z. Partitioned stator flux reversal machine with consequent-pole PM stator. IEEE Transactions on Energy Conversion. 2015;30(4):1472-82. https://doi.org/10.1109/TEC.2015.2448796
21. Zhu ZQ, Wu ZZ, Evans DJ, Chu WQ. Novel electrical machines having separate PM excitation stator. IEEE Transactions on Magnetics. 2014;51(4):1-9. https://doi.org/10.1109/TMAG.2014.2358199
22. Wu Z, Zhu Z, Zhan H. Comparative analysis of partitioned stator flux reversal PM machines having fractional-slot nonoverlapping and integer-slot overlapping windings. IEEE Transactions on Energy Conversion. 2016;31(2):776-88. https://doi.org/10.1109/TEC.2016.2525826
23. Zou T, Li D, Qu R, Li J, Jiang D. Analysis of a dual-rotor, toroidal-winding, axial-flux vernier permanent magnet machine. IEEE Transactions on Industry Applications. 2017;53(3):1920-30. https://doi.org/10.1109/TIA.2017.2657493
24. Zhao F, Lipo TA, Kwon BI. A novel dual-stator axial-flux spoke-type permanent magnet vernier machine for direct-drive applications. IEEE Transactions on Magnetics. 2014;50(11):1-4. https://doi.org/10.1109/TMAG.2014.2329861
25. Wei F, Zhu Z, Qu H, Yan L, Qi J. New dual-PM spoke-type flux-reversal machines for direct-drive applications. IEEE Transactions on Industry Applications. 2022;58(5):6190-202. https://doi.org/10.1109/TIA.2022.3190248
26. Zhao J, Wang Q, Han Q. Reducing torque ripple through innovative configuration of permanent magnet based on air gap field modulation theory in a novel axial flux reversal permanent magnet machine. IEEE Access. 2024;12:54901-12. https://doi.org/10.1109/ACCESS.2024.3389699
27. Patel A, Suthar B. Design optimization of axial flux surface mounted permanent magnet brushless DC motor for electrical vehicle based on genetic algorithm. International Journal of Engineering, Transactions A: Basics. 2018;31(7):1050-6. https://doi.org/10.5829/ije.2018.31.07a.07
28. Quoc VD, Huu HB. Modeling of surface-mounted permanent magnet synchronous motors with inner and outer rotor types. International Journal of Engineering, Transactions C: Aspects. 2024;37(12):2529-37. https://doi.org/10.5829/ije.2024.37.12c.11
29. Conga TT, Vua TN, Minha DB, Thanhc HV, Quoc VD. Analytical modelling of a six-phase surface mounted permanent magnet synchronous motor. International Journal of Engineering, Transactions A: Basics. 2024;37(7):1274. https://doi.org/10.5829/ije.2024.37.07a.07
30. Kim JH, Choi W, Sarlioglu B. Closed-form solution for axial flux permanent-magnet machines with a traction application study. IEEE Transactions on Industry Applications. 2015;52(2):1775-84. https://doi.org/10.1109/TIA.2015.2496877
31. Gao Y, Qu R, Li D, Li J, Wu L. Design of three-phase flux-reversal machines with fractional-slot windings. IEEE Transactions on Industry Applications. 2016;52(4):2856-64. https://doi.org/10.1109/TIA.2016.2535108
32. Mirnikjoo SA, Abbaszadeh K, Abdollahi SE. Multiobjective design optimization of a double-sided flux switching permanent magnet generator for counter-rotating wind turbine applications. IEEE Transactions on Industrial Electronics. 2020;68(8):6640-9. https://doi.org/10.1109/TIE.2020.3005106
33. Si J, Zhang T, Hu Y, Gan C, Li Y. An axial-flux dual-rotor slotless permanent magnet motor with novel equidirectional toroidal winding. IEEE Transactions on Energy Conversion. 2021;37(3):1752-63. https://doi.org/10.1109/TEC.2021.3138465
34. Huang S, Luo J, Leonardi F, Lipo TA. A comparison of power density for axial flux machines based on general purpose sizing equations. IEEE Transactions on Energy Conversion. 1999;14(2):185-92. https://doi.org/10.1109/60.766982

Persian Abstract

چکیده

این مقاله یک ماشین جدید شاربرگردان با میدان محوری و آهنربای دائم دوگانه (Dual-PM AFFRM) دارای یک استاتور و یک روتور برای کاربرد در خودروهای الکتریکی درون‌چرخی محرک مستقیم (Direct-drive In-wheel) ارائه می‌دهد. توپولوژی و اصول عملکردی ماشین توضیح داده شده است و تأثیرات استفاده از آهنربای دائم دوگانه بر هارمونیک‌های چگالی شار فاصله هوایی با استفاده از یک مدل ساده نیروی محرکه مغناطیسی –پرمانس آشکار شده است. برای به دست آوردن مشخصه گشتاور بهینه ماشین و کاهش حجم محاسباتی، از روش تاگوچی استفاده شده است. عملکرد الکترومغناطیسی طراحی بهینه شده از طریق روش المان محدود سه‌بعدی ارزیابی شده است. نتایج نشان می‌دهند که ماشین بهینه شده قابلیت گشتاور بالا، ریپل گشتاور کم و قابلیت تضعیف شار خوب را ارائه می‌دهد. علاوه بر این، نتایج یک مطالعه مقایسه‌ای بین هشت ماشین کششی مختلف نشان می‌دهد که ماشین جدید شاربرگردان با میدان محوری و آهنربای دائم دوگانه، بالاترین چگالی گشتاور را با نسبت وجه بالا و ثابت گشتاور بالا ارائه می‌دهد که برای کاربردهای خودروهای الکتریکی درون‌چرخی محرک مستقیم بسیار مهم هستند.

AIMS AND SCOPE

The objective of the International Journal of Engineering is to provide a forum for communication of information among the world's scientific and technological community and Iranian scientists and engineers. This journal intends to be of interest and utility to researchers and practitioners in the academic, industrial and governmental sectors. All original research contributions of significant value focused on basics, applications and aspects areas of engineering discipline are welcome.

This journal is published in three quarterly transactions: Transactions A (Basics) deal with the engineering fundamentals, Transactions B (Applications) are concerned with the application of the engineering knowledge in the daily life of the human being and Transactions C (Aspects) - starting from January 2012 - emphasize on the main engineering aspects whose elaboration can yield knowledge and expertise that can equally serve all branches of engineering discipline.

This journal will publish authoritative papers on theoretical and experimental researches and advanced applications embodying the results of extensive field, plant, laboratory or theoretical investigation or new interpretations of existing problems. It may also feature - when appropriate - research notes, technical notes, state-of-the-art survey type papers, short communications, letters to the editor, meeting schedules and conference announcements. The language of publication is English. Each paper should contain an abstract both in English and in Persian. However, for the authors who are not familiar with Persian, the publisher will prepare the latter. The abstracts should not exceed 250 words.

All manuscripts will be peer-reviewed by qualified reviewers. The material should be presented clearly and concisely:

- *Full papers* must be based on completed original works of significant novelty. The papers are not strictly limited in length. However, lengthy contributions may be delayed due to limited space. It is advised to keep papers limited to 7500 words.
- *Research* notes are considered as short items that include theoretical or experimental results of immediate current interest.
- *Technical notes* are also considered as short items of enough technical acceptability with more rapid publication appeal. The length of a research or technical note is recommended not to exceed 2500 words or 4 journal pages (including figures and tables).

Review papers are only considered from highly qualified well-known authors generally assigned by the editorial board or editor in chief. Short communications and letters to the editor should contain a text of about 1000 words and whatever figures and tables that may be required to support the text. They include discussion of full papers and short items and should contribute to the original article by providing confirmation or additional interpretation. Discussion of papers will be referred to author(s) for reply and will concurrently be published with reply of author(s).

INSTRUCTIONS FOR AUTHORS

Submission of a manuscript represents that it has neither been published nor submitted for publication elsewhere and is result of research carried out by author(s). Presentation in a conference and appearance in a symposium proceeding is not considered prior publication.

Authors are required to include a list describing all the symbols and abbreviations in the paper. Use of the international system of measurement units is mandatory.

- On-line submission of manuscripts results in faster publication process and is recommended. Instructions are given in the IJE web sites: www.ije.ir-www.ijeir.info
- Hardcopy submissions must include MS Word and jpg files.
- Manuscripts should be typewritten on one side of A4 paper, double-spaced, with adequate margins.
- References should be numbered in brackets and appear in sequence through the text. List of references should be given at the end of the paper.
- Figure captions are to be indicated under the illustrations. They should sufficiently explain the figures.
- Illustrations should appear in their appropriate places in the text.
- Tables and diagrams should be submitted in a form suitable for reproduction.
- Photographs should be of high quality saved as jpg files.
- Tables, Illustrations, Figures and Diagrams will be normally printed in single column width (8cm). Exceptionally large ones may be printed across two columns (17cm).

PAGE CHARGES AND REPRINTS

The papers are strictly limited in length, maximum 8 journal pages (including figures and tables). For the additional to 8 journal pages, there will be page charges. It is advised to keep papers limited to 3500 words.

Page Charges for Papers More Than 8 Pages (Including Abstract)	
For International Author ***	**$55 / per page**
For Local Author	**100,000 Toman / per page**

AUTHOR CHECKLIST

- Author(s), bio-data including affiliation(s) and mail and e-mail addresses).
- Manuscript including abstracts, key words, illustrations, tables, figures with figure captions and list of references.
- MS Word files of the paper.

International **J**ournal of **E**ngineering, P.O. Box 13145-1659, Tehran, I.R. Iran
Materials and Energy Research Center, 5 Ahuramazda St., Alvand Ave., Argentina Sq., Tehran, Islamic Republic of Iran
www.merc.ac.ir - www.emerc.ac.ir

THOMSON REUTERS

WEB OF SCIENCE

Scopus®

CAS®
A DIVISION OF THE
AMERICAN CHEMICAL SOCIETY

Google
Scholar