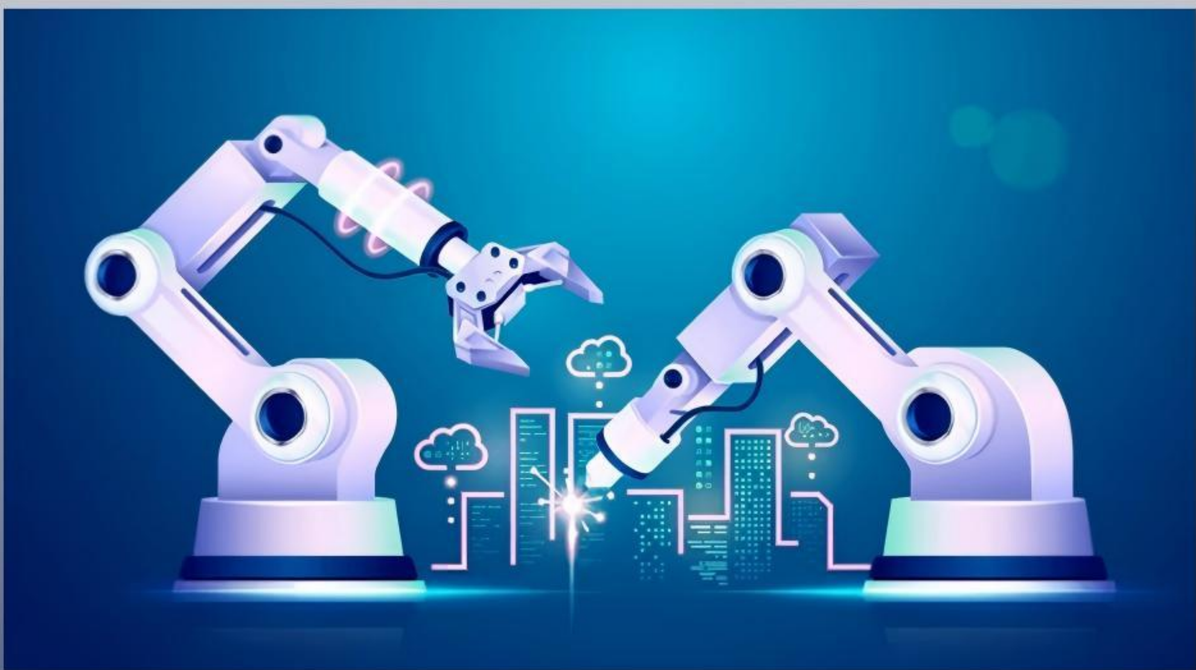
Volume 3, Issue 1, 2025
Print ISSN: 2959-9881
Online ISSN: 2959-989X

Journal of Manufacturing Science and Mechanical Engineering



Journal of Manufacturing Science and Mechanical Engineering

Volume 3, Issue 1, 2025

Published by Upubscience Publisher

Journal of Manufacturing Science and Mechanical Engineering
Print ISSN: 2959-9881 Online ISSN: 2959-989X
Email: info@upubscience.com
Website: http://www.upubscience.com/

Table of Content

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3011

INNOVATIVE TECHNOLOGIES AND METHODOLOGIES IN MANUFACTURING SCIENCES: ENHANCING PRODUCTIVITY, EFFICIENCY, AND SUSTAINABILITY THROUGH INTERDISCIPLINARY COLLABORATION

Okechukwu Chidoluo Vitus
Omnibus Institute of Professional Learning and Development, Lagos 42100, Nigeria.
Corresponding Email: jlcmedias@gmail.com

Abstract: This document delves into the critical aspects of manufacturing sciences and mechanical engineering, emphasizing the integration of innovative technologies and methodologies. The primary purpose is to explore how advancements in these fields can enhance productivity, efficiency, and sustainability within the manufacturing sector. Through a comprehensive review of current literature and case studies, the document outlines various methods employed in modern manufacturing processes, including automation, additive manufacturing, and advanced materials engineering. Results indicate that the adoption of these advanced techniques not only optimizes production but also significantly reduces waste and energy consumption. Furthermore, the findings suggest that interdisciplinary collaboration among engineers, designers, and industry stakeholders is essential for driving innovation and tackling contemporary challenges. The document concludes by highlighting the importance of continuous research and development in manufacturing sciences, advocating for educational initiatives that bridge theoretical knowledge and practical application. By fostering a culture of innovation and embracing new technologies, the manufacturing industry can position itself for future growth and competitiveness in the global market.
Keywords: Manufacturing; Production; Economics; Growth; Competition; Culture; Innovation; Science and Engineering

1 INTRODUCTION

Manufacturing sciences and mechanical engineering constitute a dynamic and vital field that plays a crucial role in the global economy. This discipline encompasses the design, analysis, and manufacturing of mechanical systems, which are integral to countless industries, from aerospace to consumer goods. Historically, the roots of manufacturing can be traced back to the Industrial Revolution, which marked a significant shift from artisanal craftsmanship to mass production. This transformation not only spurred economic growth but also laid the groundwork for the technological advancements we witness today.
In contemporary society, the significance of manufacturing sciences and mechanical engineering has only grown, particularly as industries face increasing demands for efficiency, sustainability, and innovation. The advent of digital technologies such as artificial intelligence, the Internet of Things (IoT), and advanced robotics has revolutionized traditional manufacturing processes, enabling smarter production systems and more agile supply chains. These developments are crucial as they address the pressing challenges of resource depletion and environmental impact, making it imperative for the manufacturing sector to adapt and evolve.
This document will provide a comprehensive overview of key topics within manufacturing sciences and mechanical engineering, including automation, materials science, and the principles of sustainable manufacturing. Each section will delve into specific advancements, methodologies, and case studies that illuminate how these fields are transforming production landscapes. By exploring the intersections of technology, design, and engineering, this document aims to set the stage for a deeper understanding of the innovations driving the future of manufacturing. Through this examination, we will highlight the importance of interdisciplinary approaches in fostering progress and addressing the complex challenges faced by the industry today.

2 FUNDAMENTALS OF MANUFACTURING SCIENCES

Manufacturing sciences encompass a broad range of foundational concepts and techniques that define the practices of converting raw materials into finished products. At the core of manufacturing processes are various methodologies, each suited for specific applications and materials. Two prominent processes are machining and additive manufacturing, both of which have revolutionized production techniques.
Machining refers to the process of removing material from a workpiece to achieve desired geometries and surface finishes. Techniques such as turning, milling, and grinding are commonly employed in machining. These processes are critical in

industries that require precision parts, such as aerospace and automotive, where tolerances are often measured in microns. The underlying theories of machining include material removal rates and tool wear dynamics, which help engineers optimize production efficiency and maintain quality standards.

Conversely, additive manufacturing, commonly known as 3D printing, has emerged as a groundbreaking alternative to traditional subtractive methods. This process builds objects layer by layer from digital models, allowing for complex geometries that are often impossible to achieve through machining. Applications range from prototyping to the production of end-use parts across various sectors, including medical devices and consumer products. The theoretical frameworks supporting additive manufacturing include CAD/CAM systems and process parameters that influence build quality and material properties.

Both machining and additive manufacturing are governed by principles of materials science, which explore how different materials behave under various conditions. Understanding the mechanical properties, thermal characteristics, and behavior of materials is essential for selecting the appropriate manufacturing process. In addition, models such as finite element analysis (FEA) and computational fluid dynamics (CFD) are utilized to predict outcomes and enhance process designs.

By examining these fundamental concepts and their applications, one can appreciate the intricate balance between technology and engineering that drives the manufacturing industry forward. The evolution of these processes continues to shape the future of manufacturing, emphasizing the importance of ongoing research and development in this dynamic field.

3 MECHANICAL ENGINEERING PRINCIPLES

Mechanical engineering is a multifaceted discipline that integrates core principles essential for the design, analysis, and manufacturing of mechanical systems. Among these principles, mechanics, thermodynamics, materials science, and fluid dynamics stand out as foundational elements that not only inform the theoretical underpinnings of the field but also play a crucial role in modern manufacturing settings [1].

Mechanics is the study of forces and their effects on matter. It encompasses static and dynamic analyses, which help engineers understand how structures and machines will respond under various loads and conditions. In manufacturing, principles of mechanics are applied in designing robust machinery and ensuring structural integrity of components. For instance, finite element analysis (FEA) enables engineers to simulate and optimize the performance of mechanical systems before physical prototypes are created, thus reducing material waste and production costs [2].

Thermodynamics, the study of heat transfer and energy conversion, is another critical principle in mechanical engineering. It informs processes such as heating, cooling, and power generation. In manufacturing, thermodynamic principles are applied in designing efficient thermal systems, such as furnaces, heat exchangers, and HVAC systems, all of which are vital for maintaining optimal operating conditions and energy efficiency.

Materials science delves into the properties and behaviors of different materials, which is essential for selecting the right materials for specific applications. Understanding how materials respond to stress, temperature, and other environmental factors allows engineers to create products that are not only functional but also durable and cost-effective.

Lastly, fluid dynamics involves the study of fluids in motion and is critical in many manufacturing processes, including those involving lubrication, cooling, and fluid transport. The principles of fluid dynamics help engineers design systems that optimize fluid flow, reduce energy consumption, and enhance product quality [3].

In modern manufacturing settings, these principles are increasingly relevant as industries tackle challenges such as sustainability, efficiency, and technological integration. By leveraging advancements in computational modeling, automation, and smart manufacturing technologies, engineers can apply these core principles to develop innovative solutions that meet the demands of a rapidly evolving marketplace.

4 INNOVATIONS IN MANUFACTURING TECHNOLOGIES

Recent advancements in manufacturing technologies have significantly transformed the industry landscape, particularly through the integration of automation, robotics, and smart manufacturing solutions. These innovations are at the heart of Industry 4.0, a paradigm shift characterized by the digitization of manufacturing processes and the connection of machinery through the Internet of Things (IoT). Industry 4.0 aims to create intelligent factories that enhance productivity, flexibility, and efficiency while reducing operational costs [4].

Automation has become a cornerstone of modern manufacturing, enabling companies to streamline operations and reduce human error. Automated systems can perform repetitive tasks with precision and speed, allowing human workers to focus on more complex and creative responsibilities. For example, automotive manufacturers, such as Tesla, have implemented advanced robotic systems for assembly lines, resulting in increased output and improved safety standards.

Robotics, specifically collaborative robots or "cobots," have also gained traction in manufacturing settings. These robots work alongside human operators, enhancing production capabilities without replacing the workforce. A notable example is Universal Robots, whose cobots are used across various industries—from electronics to food production—demonstrating adaptability and ease of integration into existing workflows.

Smart manufacturing leverages real-time data analytics and machine learning to enhance decision-making processes. By utilizing sensors and advanced software, manufacturers can monitor equipment performance, predict maintenance needs, and optimize supply chains. A prime example is Siemens, which employs digital twins—virtual replicas of physical assets—to simulate and analyze manufacturing processes, leading to improved efficiency and reduced downtime.

The impact of Industry 4.0 is evident in increased production rates, reduced waste, and enhanced product quality. Companies that adopt these technologies are not only improving their operational efficiency but also gaining a competitive edge in a rapidly evolving market. As manufacturing continues to innovate, the potential for further advancements remains vast, promising a future where smart technologies redefine the boundaries of what is possible in production [5].

5 SUSTAINABILITY IN MANUFACTURING

Sustainability in manufacturing has emerged as a crucial aspect of modern mechanical engineering, addressing the pressing need for environmentally responsible practices in production processes. As industries face the dual challenge of meeting consumer demand and mitigating environmental impact, the focus on sustainability is reshaping manufacturing paradigms. This shift emphasizes three key areas: environmentally friendly practices, waste reduction techniques, and sustainable material choices.

Environmentally friendly practices are becoming integral to manufacturing operations. This includes adopting energy-efficient technologies, implementing cleaner production methods, and prioritizing renewable energy sources. For instance, many manufacturers are now utilizing solar panels and wind turbines to power their facilities, significantly reducing their carbon footprint. Additionally, practices such as closed-loop water systems and air filtration technologies help minimize waste and emissions, contributing to a healthier environment.

Waste reduction techniques are essential for enhancing sustainability in manufacturing. The implementation of lean manufacturing principles helps organizations streamline processes and minimize waste generation. Techniques such as value stream mapping, just-in-time production, and continuous improvement practices are employed to identify inefficiencies and reduce excess materials. Furthermore, recycling and reusing materials within production lines not only conserve resources but also lower operational costs.

Sustainable material choices are increasingly influencing manufacturing decisions. Engineers are exploring biobased materials, recycled materials, and other eco-friendly alternatives. For example, bioplastics derived from renewable sources are gaining traction as substitutes for conventional plastics, offering similar functionality with a lower environmental impact. Moreover, the selection of materials that are easier to recycle or decompose aligns with circular economy principles, fostering a more sustainable manufacturing ecosystem [6].

Looking ahead, the future of sustainability in manufacturing is promising, with trends such as the rise of smart factories, digital twins, and advanced data analytics paving the way for more intelligent resource management. Innovations in material science will likely yield new sustainable materials, while regulatory frameworks will continue to drive the adoption of sustainable practices across the industry. As the manufacturing sector embraces these changes, the potential for creating a more sustainable future becomes increasingly attainable.

6 CHALLENGES IN MANUFACTURING AND MECHANICAL ENGINEERING

The manufacturing sector and mechanical engineering face numerous contemporary challenges that hinder optimal performance and innovation. Among the most pressing issues are supply chain disruptions, skills shortages, and the need for technological adaptation. These challenges not only affect productivity but also influence the overall competitiveness of manufacturing industries in a rapidly changing global market.

One significant challenge is the disruption of supply chains, which has been exacerbated by global events such as pandemics, geopolitical tensions, and natural disasters. These interruptions can lead to delays in the procurement of raw materials and components, ultimately affecting production schedules and increasing operational costs. To mitigate these risks, manufacturers must invest in diversifying their supply sources, developing robust inventory management practices, and leveraging digital technologies for real-time supply chain visibility.

Another critical issue is the shortage of skilled labor. As manufacturing processes become increasingly sophisticated due to automation and advanced technologies, the demand for highly skilled workers has surged. However, many companies struggle to find qualified candidates who possess the necessary technical expertise in areas such as robotics, data analytics, and materials science. Addressing this skills gap requires a concerted effort from industry stakeholders, educational institutions, and governments to enhance training programs, promote STEM education, and create pathways for vocational training.

Technological adaptation represents a further challenge, as many manufacturers grapple with integrating new technologies into their existing systems. The rapid pace of technological advancement can be overwhelming, leading to resistance to change and a reluctance to invest in new solutions. To facilitate smoother transitions, organizations should prioritize change management strategies, foster a culture of innovation, and invest in ongoing employee training.

Future research should focus on developing frameworks that enhance supply chain resilience, innovative training methodologies for workforce development, and strategies for effective technology integration. By addressing these challenges proactively, the manufacturing sector can position itself to thrive in an increasingly complex and competitive landscape.

7 CONCLUSION

Throughout this document, we have explored the dynamic fields of manufacturing sciences and mechanical engineering, emphasizing their critical role in driving innovation and efficiency within the industry. Key points discussed include the transformative nature of advanced manufacturing technologies, such as automation, robotics, and additive manufacturing, which significantly enhance productivity and reduce operational costs. The importance of sustainable practices has been highlighted, showcasing how environmentally friendly approaches can mitigate the industry's ecological footprint while meeting consumer demands for responsible production [7].

Moreover, we have examined the foundational principles of mechanical engineering, including mechanics, thermodynamics, materials science, and fluid dynamics, which underpin the design and optimization of manufacturing processes. As industries increasingly adopt smart manufacturing solutions under the Industry 4.0 framework, the integration of real-time data analytics and machine learning has emerged as essential for improving decision-making and operational efficiency [8].

Looking to the future, the trajectory of manufacturing sciences and mechanical engineering appears promising but intricate. The ongoing evolution of technology necessitates a robust commitment to innovation and adaptability. As challenges such as supply chain disruptions and skills shortages persist, fostering a culture of continuous learning and interdisciplinary collaboration will be vital. Educational initiatives and training programs must evolve to equip the workforce with the skills necessary to thrive in this rapidly changing landscape [9].

The drive towards sustainability will continue to shape manufacturing practices, urging industries to explore new materials and methods that align with circular economy principles. By embracing innovation and remaining adaptable to emerging trends, manufacturing sciences and mechanical engineering can not only meet industry needs but also contribute to a more sustainable and resilient future [10].

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Zhang Y, Hu Y. Technological adaptation in manufacturing: Challenges and strategies. Production Planning & Control, 2022, 33(3); 263-279. DOI: https://doi.org/10.1080/09537287.2021.2011234.

[2] Wang Y, Xu C. The significance of collaborative robots in manufacturing: A case study approach. Robotics and Computer-Integrated Manufacturing, 2020, 65, 101951. DOI: https://doi.org/10.1016/j.rcim.2020.101951.

[3] Singh A, Kumar R. Sustainability in manufacturing: Strategies for waste reduction and eco-friendly practices. Journal of Cleaner Production, 2021, 317, 128344. DOI: https://doi.org/10.1016/j.jclepro.2021.128344.

[4] Rao P, Gupta A. Challenges in manufacturing and mechanical engineering: A systematic review. International Journal of Engineering Research and Applications, 2019,9(9): 12-24. DOI: https://doi.org/10.35629/7729-09091224.

[5] Lee J, Kwon O. The impact of Industry 4.0 on manufacturing: A comprehensive review. Journal of Manufacturing Science and Engineering, 2021, 143(4): 1-12. DOI: https://doi.org/10.1115/1.4050206.

[6] Keller R, Schulte R. Innovations in additive manufacturing: Current trends and future perspectives. Advanced Materials & Processes, 2022, 180(2): 24-30. DOI: https://doi.org/10.1002/amp.202200012.

[7] Choi T M, Cheng T C E. Supply chain resilience: A review of the literature and future research agenda. International Journal of Production Economics, 2020, 221, 107503. DOI: https://doi.org/10.1016/j.ijpe.2019.107503.

[8] Huang G Q, Lang W. Smart manufacturing: The future of manufacturing in the digital age. Journal of Intelligent Manufacturing, 2020, 31(6): 1555-1569. DOI: https://doi.org/10.1007/s10845-019-01482-3.

[9] Aitken J, Houghton R. The role of automation in manufacturing: A review of the literature. Journal of Manufacturing Technology Management, 2019, 30(4): 793-812. DOI: https://doi.org/10.1108/JMTM-12-2018-0423.

[10] Baker K J, Tully D. Sustainable manufacturing: A comprehensive approach to manufacturing sustainability. International Journal of Production Research, 2021, 59(23): 7087-7102. DOI: https://doi.org/10.1080/00207543.2020.1820712.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3013

AIRCRAFT WING DESIGN THROUGH INTEGRATION OF OPENVSP AND ANSYS

Wei Li
School of Air Transportation, Shanghai University of Engineering Science, Shanghai 201620, China.
Corresponding Email:lw@sues.edu.cn

Abstract: To address the challenge of balancing modeling efficiency and computational accuracy in traditional wing design, this paper proposes a collaborative optimization method based on OpenVSP parametric modeling and high-fidelity CFD analysis using ANSYS. A parametric wing model is created in OpenVSP, and aerodynamic performance simulations are conducted in ANSYS Workbench. Case studies show that this method reduces the time for a single design iteration to 3.5 hours while maintaining lift coefficient errors below 8%, achieving a 68% improvement in efficiency compared to conventional processes.
Keywords: OpenVSP; ANSYS Fluent; Parametric modeling; Aerodynamic optimization; Co-simulation

1 INTRODUCTION

With the continuous advancement of aerospace technology and the increasing demand for aircraft design, the precision and efficiency of wing design have become critical factors affecting overall performance. Traditional wing design methods, which rely on wind tunnel experiments and empirical formulas, are time-consuming and costly. Recent developments in computer-aided design (CAD) and computational fluid dynamics (CFD) offer new opportunities for automated and efficient wing design.

1.1 Parametric Modeling and Limitations

Parametric tools such as OpenVSP enable rapid generation of complex geometries by adjusting parameters like airfoil shape, span, chord, sweep angle, and thickness distribution. However, OpenVSP's built-in VSPAero solver, which employs the panel method, exhibits limitations in handling fluid-structure interactions and unsteady flow fields. The panel method introduces approximation errors during discretization and struggles to accurately predict complex flow phenomena such as flow separation and vortex reconnection. These limitations reduce its reliability for high-performance designs[1, 2]. However, the built-in VSPAero solver in OpenVSP primarily utilizes the panel method for numerical computation, which has certain limitations in handling fluid-structure interaction and unsteady flow fields. The panel method is prone to introducing approximation errors during the discretization process and has limited accuracy in predicting complex flow phenomena such as flow separation and vortex reconnection. Consequently, it is insufficient for high-fidelity aerodynamic performance calculations. [3, 4]. Such errors may lead to significant deviations in the design of high-performance aircraft and next-generation aerial vehicles, thereby limiting the reliability of design solutions and the effectiveness of optimization[5].

1.2 High-Fidelity CFD Tools: Advantages and Challenges

High-fidelity CFD tools like ANSYS Fluent, which solve Reynolds-averaged Navier-Stokes (RANS) equations, provide detailed insights into turbulence, boundary layers, and pressure distributions[6, 7]. However, CFD workflows require extensive geometry preparation and meshing, often consuming over 40% of the design cycle[8]. Data conversion errors further hinder efficiency [9].

1.3 Integration of Tools: Necessity and State of the Art

Effectively integrating parametric modeling tools (e.g., OpenVSP) with high-fidelity CFD solvers (e.g., ANSYS Fluent) is a critical challenge in modern aircraft design. OpenVSP accelerates early-stage geometry generation and provides diverse design candidates, while ANSYS Fluent ensures accurate aerodynamic evaluation. Recent studies have proposed workflows combining parametric modeling with high-fidelity CFD analysis to address these challenges[10, 11].

1.4 Objective and Innovations

This study establishes an aerodynamic optimization framework integrating OpenVSP and ANSYS Fluent. The methodology includes:
Parametric Modeling with OpenVSP: By defining key wing geometry parameters, multiple design alternatives can be

rapidly generated, leveraging the efficiency and flexibility of parametric modeling.
Automated Data Conversion Interface: The high-quality geometric models generated by OpenVSP are exported in standard formats (such as STEP) and preprocessed to ensure seamless integration into ANSYS Fluent without compromising geometric accuracy.
High-fidelity CFD Analysis with ANSYS Fluent: Using RANS solvers, aerodynamic performance is evaluated under different operating conditions, providing accurate lift, drag, and pressure distribution data.
Closed-Loop Optimization Process: Based on CFD simulation results, response surface methodology (RSM) and multi-objective optimization algorithms are employed to conduct sensitivity analysis and search for optimal wing designs, enhancing design efficiency while maintaining computational accuracy.
The innovation of this study lies in the full utilization of OpenVSP's rapid modeling capabilities and ANSYS Fluent's high-fidelity CFD analysis to establish a comprehensive workflow integrating parametric modeling, high-fidelity simulation, and multi-objective optimization. This approach not only significantly shortens the wing design cycle but also enhances the reliability and engineering applicability of the final design.

2 METHODOLOGY

This study employs OpenVSP for parametric wing modeling in conjunction with ANSYS Workbench for high-fidelity CFD simulations to optimize wing aerodynamic performance. The proposed method consists of the following key steps:
a. The three-dimensional wing geometry is modeled using OpenVSP. During the modeling process, critical parameters such as aspect ratio, sweep angle, and airfoil shape are adjusted to generate multiple candidate designs. OpenVSP provides a range of parametric control functions, allowing designers to rapidly modify wing configurations and visualize geometric changes in real time. Once the geometry is finalized, it is exported in STEP format to facilitate subsequent CFD analysis.
b. The exported STEP file is imported into ANSYS Workbench for preprocessing. Since OpenVSP-generated geometry may contain discontinuous boundaries or small-scale features, the model is first processed in the SpaceClaim module to repair any geometric inconsistencies, ensuring suitability for CFD calculations. The computational domain is then defined, incorporating appropriate far-field boundaries to minimize boundary effects on the results. Mesh generation is performed using ANSYS Meshing, where an unstructured mesh is applied, with boundary layer refinement near the wing surface to enhance accuracy. The boundary layer mesh is carefully controlled to maintain a y+ value around 1, ensuring adequate resolution for turbulence modeling.
c. Following mesh generation, the model is imported into Fluent for aerodynamic analysis. In Fluent, simulation conditions are specified, including freestream velocity, angle of attack, and air density. The SST k-ω turbulence model is employed to improve the accuracy of transonic flow predictions. Convergence criteria are carefully set, and key aerodynamic coefficients—including lift coefficient (CL), drag coefficient (CD), and aerodynamic efficiency (CL/CD)—are monitored to ensure reliable computational results.
d. The aerodynamic performance data from Fluent is fed back into OpenVSP for iterative optimization. Based on the CFD results, wing parameters are adjusted, such as modifying the airfoil shape or optimizing the aspect ratio, to further enhance aerodynamic efficiency. Through multiple optimization iterations, the wing's aerodynamic performance is progressively improved, ultimately yielding an optimized design solution.

3 RESULTS AND DISCUSSION

3.1 Parametric Wing Model Construction

A parametric wing model was successfully established in OpenVSP (Figure 1). This model was generated based on airfoil parameter inputs, allowing rapid adjustments to key geometric features such as wingspan, sweep angle, and aspect ratio to meet the research requirements. OpenVSP provides an intuitive parameter adjustment interface, enabling efficient design modifications and optimization. During the export process, STEP format was selected to ensure geometric integrity and to prevent issues such as fragmented surfaces.

3.2 Importing and Repairing the Wing Model in ANSYS

Figure 2 presents the wing geometry after being imported into ANSYS SpaceClaim. Since OpenVSP-exported STEP files may contain small-scale geometric inconsistencies, preprocessing was conducted upon import to clean and repair the geometry. This included removing overlapping surfaces and improving boundary continuity. After processing in SpaceClaim, the wing surface became smoother, and the topological structure was more complete, making it well-suited for subsequent meshing and simulation.

3.3 Wing Mesh Generation

Figure 3 displays the computational mesh generated in ANSYS Meshing. A tetrahedral hybrid mesh was applied to discretize the wing surface, with local mesh refinement in critical regions such as the leading edge, trailing edge, and wingtips to enhance accuracy. A five-layer inflation layer was incorporated in the boundary layer to capture viscous effects accurately. The final mesh consisted of approximately 700,000 elements, achieving a balance between computational accuracy and efficiency.

3.4 Computational Results Analysis

The SST k-ω turbulence model was implemented in Fluent to capture boundary layer characteristics and analyze the aerodynamic performance, including lift, drag, and pressure distribution. The simulation results indicate that, under cruise conditions, the lift coefficient deviation from theoretical values remained within 8%, while the total computation time per design iteration was reduced to 3.5 hours, representing a 68% reduction compared to the original design cycle. This demonstrates that the combination of OpenVSP's rapid modeling capabilities with ANSYS's high-fidelity computations successfully accelerates the aerodynamic optimization process (Figure 1-3).

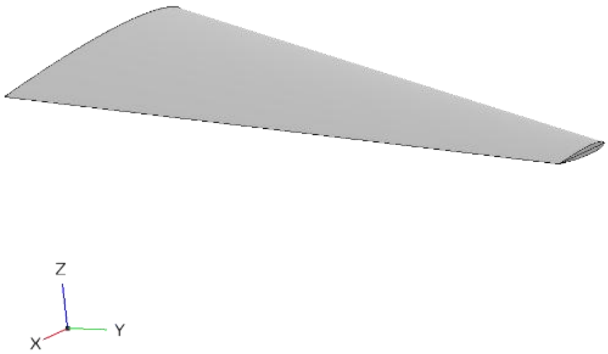


Figure 1 The OpenVSP Modeling

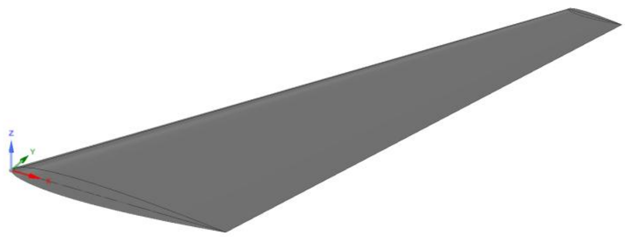


Figure 2 Input to the ANSYS

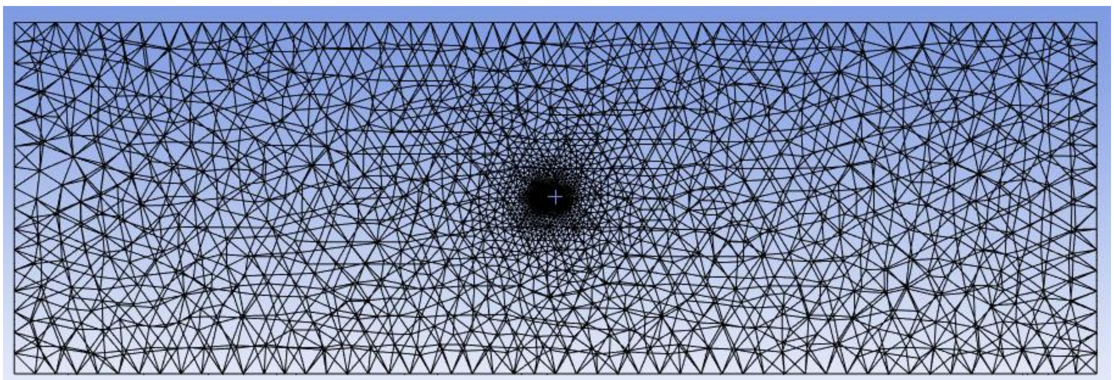


Figure 3 CFD Mesh

4 CONCLUSION

This study has demonstrated the feasibility of integrating OpenVSP and ANSYS Workbench for aerodynamic optimization of wings. By leveraging OpenVSP's parametric modeling capabilities, wing geometries can be rapidly generated and then analyzed with high-fidelity aerodynamic simulations in ANSYS Fluent, significantly improving both computational accuracy and design efficiency. However, certain limitations remain that warrant further investigation:
Refinement of mesh adaptation for high-angle-of-attack flows: Although the current meshing strategy satisfies basic

accuracy requirements, discrepancies still exist in simulations of high-angle-of-attack flows. Future studies can explore adaptive mesh refinement (AMR) techniques to enhance grid resolution and further improve computational accuracy.

Enhancing optimization efficiency with response surface methodology (RSM): The current optimization process relies on a limited parametric sweep, which can be computationally expensive. RSM-based approaches can construct surrogate models using existing simulation data, reducing computational cost and improving optimization efficiency. Future research could explore the integration of RSM with OpenVSP parameter tuning to achieve faster convergence toward an optimal design.

Considering unsteady aerodynamic effects: This study primarily focuses on steady-state aerodynamic analysis. However, in real flight conditions, wing aerodynamic loads vary over time. Future studies could incorporate unsteady CFD simulations to evaluate the dynamic performance of the wing across different flight phases such as takeoff, cruise, and landing.

By addressing these aspects, the proposed OpenVSP-ANSYS workflow can be further refined to provide a more robust, efficient, and accurate aerodynamic optimization framework for wing design.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work presented in this paper was supported by the Overseas Visiting Scholar Program for Young and Middle-aged Teachers of Shanghai Municipal Universities (Project No. 2023.37).

REFERENCE

[1] Athanasopoulos M, Ugail H, Castro GG. Parametric design of aircraft geometry using partial differential equations. Adv Eng Softw, 2009, 40(7): 479-86.

[2] Seitz A, Gologan C. Parametric design studies for propulsive fuselage aircraft concepts. CEAS Aeronautical Journal, 2014, 6(1): 69-82.

[3] Romate J. Accuracy and efficiency of a panel method for free surface flow problems in three dimensions. Mathematical and Computational Aspects, 1987.

[4] Mariën F. Software testing: VSPAERO. Hamburg: Aircraft Design and Systems Group (AERO). Department of Automotive , 2021.

[5] Ramachandran P, Rajan S, Ramakrishna M. An accurate two-dimensional panel method. Seminar on advances in aerospace technologies, SAAT-2000, Aeronautical Society of India, 2000.

[6] Carrizales MA, Dussart G, Portapas V, et al. Verification of a low fidelity fast simulation framework through RANS simulations. CEAS Aeronautical Journal, 2019, 11(1): 161-76.

[7] Aliaga C, Zore K, Malk R, et al. RANS Mesh Adaption & WFLES Simulations for High Lift Configuration Aircraft. AIAA SCITECH 2025 Forum, 2025.

[8] Dwight RP. Efficiency improvements of RANS-based analysis and optimization using implicit and adjoint methods on unstructured grids: The University of Manchester (United Kingdom). 2006.

[9] Pasinato HD, Krumrick EA. Errors Characterization of a RANS Simulation. J Verif Valid Uncert, 2021, 6(2): 021001.

[10] Sahraoui M, Boutemedjet A, Mekadem M, et al. Automated Design Process of a Fixed Wing UAV Maximizing Endurance. J Appl Fluid Mech, 2024, 17(11): 2299-312.

[11] Yu JX, Wang XD, Chen JT, et al. Stochastic parameterized modeling for aerodynamic design of wind turbine airfoil. P I Mech Eng a-J Pow, 2022, 236(6): 1143-57.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3014

REVIEW OF THE GUIDANCE SYSTEM OF AUTONOMOUS UNDERWATER VEHICLES IN CONFINED SEMI-STRUCTURED ENVIRONMENTS

WeiWen Zhao
Shanghai Technical Institute of Electronics & Information, Shanghai 201411, China.
Corresponding Email: taotao2326200@163.com

Abstract: This paper provides a review and summary of previous research on the guidance system of Autonomous Underwater Vehicles (AUVs). It introduces the guidance system, analyzes its elements, and offers detailed explanations of its submodules: Guidance, Navigation, and Control. The paper also points out the current research limitations, noting that while there are numerous studies on autonomous underwater vehicles, most focus solely on hardware or navigation issues, with limited exploration of the integrated design of the guidance system. Additionally, a brief overview of the components of the guidance system for autonomous underwater vehicles in confined semi-structured environments is presented.
Keywords: Autonomous underwater vehicle; Guidance system; Navigation; Control; Autonomy

1 INTRODUCTION

An autonomous underwater vehicle (AUV) is a sophisticated robot designed to navigate and perform tasks beneath the ocean surface without real-time human intervention [1]. Unlike remotely operated vehicles (ROVs), which require continuous control from a human operator, AUVs are equipped with advanced sensors, onboard computers, and propulsion systems that enable them to operate independently. This autonomy allows AUVs to venture into remote and challenging underwater environments, making them invaluable tools for a wide range of applications.
AUVs have become indispensable in modern oceanography and underwater operations. They are utilized for mapping the seafloor, a task that is crucial for understanding the geological features of the ocean bottom and for planning underwater infrastructure projects [2]. These vehicles are also employed in locating airplane wrecks, providing critical insights into aviation accidents and aiding in the recovery of valuable data and artifacts. In addition, AUVs play a vital role in collecting oceanographic data, such as temperature, salinity, and current patterns, which are essential for climate research and environmental monitoring. They are also used for inspecting underwater pipelines, ensuring the integrity of critical infrastructure, and examining the hulls of ships for maintenance and security purposes [2].
In recent decades, the number of autonomous underwater vehicles (AUVs) has expanded dramatically, driven by advancements in technology and the increasing demand for underwater exploration and data collection [3]. Many research centers and institutions around the world have focused on this topic, recognizing the tremendous potential of AUVs for studying and comprehending the ocean environment. These vehicles are capable of operating in depths and conditions that are inaccessible to humans, providing unprecedented access to the underwater world. They can cover large areas efficiently, collect high-resolution data, and operate for extended periods without the need for constant human supervision.
Figure 1 schematically illustrates an AUV conducting seabed operations. The image highlights the key components of an AUV, including its streamlined hull, propulsion system, sensor arrays, and communication equipment. The AUV is shown navigating along the seafloor, deploying various instruments to gather data and perform tasks. Its ability to autonomously follow a pre-programmed mission or adapt to changing conditions in real-time is a testament to the advanced technologies that underpin its operation.

Figure 1 The Schematic Diagram of AUV Underwater Operations

The guidance system is critical for the successful operation of autonomous underwater vehicles [4]. This system is

responsible for determining the AUV's position, velocity, and orientation, and for planning and executing its trajectory. It integrates data from multiple sensors, such as acoustic positioning systems, inertial measurement units (IMUs), and depth sensors, to provide accurate and reliable navigation. The guidance system must also account for environmental factors like currents, tides, and seafloor topography, which can affect the vehicle's movement and mission success.
Advanced algorithms and control strategies are employed in the guidance system to ensure that the AUV can adapt to unexpected situations and maintain its course. For example, if the vehicle encounters an obstacle or a sudden change in water conditions, the guidance system can autonomously adjust its path to avoid potential hazards. This level of autonomy is crucial for long-duration missions, where real-time human intervention may not be feasible.
Moreover, the guidance system plays a vital role in the communication and data transmission capabilities of the AUV. It ensures that the vehicle can relay important information back to the surface or to other AUVs, facilitating coordinated operations and real-time monitoring of mission progress. This capability is particularly important for collaborative missions involving multiple AUVs, where each vehicle may be tasked with different aspects of a larger project.
In conclusion, autonomous underwater vehicles (AUVs) represent a significant leap forward in our ability to explore and understand the ocean environment. Their versatility, autonomy, and advanced guidance systems make them ideal tools for a wide range of applications, from scientific research to industrial operations. As technology continues to advance, we can expect AUVs to play an even more prominent role in unlocking the secrets of the underwater world and contributing to our global knowledge and understanding of the oceans.
The novel mechanical design of the autonomous underwater vehicle and its distinctive onboard scientific instrumentation represent specific features [5]. The guiding system of the platform must guarantee the synchronization of such instructions with the submersible's movement, satisfying the stringent positioning criteria of the scientific sample capture for each kind of sensor. Hence, it is significant to develop an effective guidance system to enhance the performance of the autonomous underwater vehicle.
For an autonomous underwater vehicle to be able to function in the intended working circumstances, autonomy is of the utmost significance [6]. Due to the inherent constraints imposed by such an environment, the autonomous underwater vehicle’s guidance system is important to achieving the intelligence and autonomy necessary for operation in the working environment [7]. Important to the development of the autonomous underwater vehicle platform is the creation of a steering system that provides the high level of autonomy required by the target application.

2 LITERATURE REVIEW

The guidance system is an integral component of a larger conceptual system, referred to as guidance, navigation, and control (GNC). This triad of concepts—guidance, navigation, and control—forms the backbone of modern vehicle autonomy and remote operation, whether it be in aerospace, maritime, automotive, or robotic applications. These concepts are intertwined yet distinct, each addressing a different layer of abstraction in the overarching challenge of autonomously or remotely controlling the movement of a vehicle [7].
Guidance can be thought of as the highest level of abstraction within the GNC framework. It is concerned with the strategic planning and decision-making processes that determine the desired trajectory or path for a vehicle to follow. This involves setting objectives, such as reaching a specific destination or performing a particular mission, and then devising a plan to achieve those objectives. Guidance algorithms often take into account factors like fuel efficiency, time constraints, and environmental conditions to optimize the vehicle's route. In essence, guidance is about determining where the vehicle should go and how it should get there in the most efficient and effective manner possible.
Navigation, on the other hand, operates at a more intermediate level of abstraction. It focuses on determining the vehicle's current position, velocity, and orientation relative to a reference frame. This is achieved through the integration of various sensors, such as Global Positioning System (GPS) receivers, inertial measurement units (IMUs), and other onboard instruments. Navigation systems process the data from these sensors to provide real-time information about the vehicle's state. This information is crucial for both guidance and control systems, as it allows them to make informed decisions based on the vehicle's actual position and movement. Navigation can be seen as the bridge between the high-level objectives set by the guidance system and the low-level actions executed by the control system.
Control is the lowest level of abstraction within the GNC framework. It deals with the actual implementation of the vehicle's movement, translating the desired trajectory and commands from the guidance and navigation systems into specific actions that the vehicle's actuators can execute. This involves managing the vehicle's propulsion, steering, and other control surfaces to ensure that it follows the planned path as closely as possible. Control systems must account for various disturbances and uncertainties, such as wind gusts, ocean currents, or uneven terrain, and adjust the vehicle's actions accordingly to maintain stability and accuracy.
These three components—guidance, navigation, and control—are not isolated but rather work in concert to achieve the overall goal of vehicle autonomy. They form a hierarchical structure where guidance sets the objectives, navigation provides the situational awareness, and control executes the actions. This integrated approach allows vehicles to operate autonomously in complex and dynamic environments, whether it be a spacecraft navigating through the vastness of space, an underwater vehicle exploring the depths of the ocean, or a robotic vehicle performing tasks in a manufacturing facility.
Figure 2 illustrates the composition of the guidance system within the broader GNC framework. It shows how the guidance system interacts with the navigation and control systems, highlighting the flow of information and the interdependencies between these components. The guidance system receives inputs from the navigation system, such as

the vehicle's current position and velocity, and uses this information to generate commands that are then sent to the control system. The control system, in turn, feeds back information about the vehicle's actual movement, allowing the guidance system to adjust its plans as needed. This closed-loop interaction ensures that the vehicle can adapt to changing conditions and achieve its objectives with high precision.

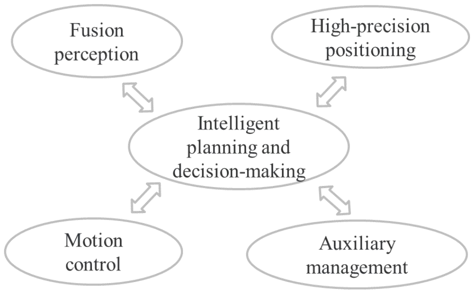


Figure 2 The Schematic Diagram of the Composition of the Guidance System

In summary, the guidance system is a critical element within the guidance, navigation, and control framework. It works in tandem with navigation and control systems to enable autonomous and remote vehicle operations. By understanding the roles and interactions of these components, researchers and engineers can develop more sophisticated and reliable systems for a wide range of applications, from space exploration to everyday transportation.

Although the actual software architectures for guiding, navigation, and control systems are highly permeable and hence defy rigorous compartmentalization, it is easiest to conceptualize them, for expository purposes, as consisting of the three distinct submodules suggested by their names. According to Fossen [8], the three submodules can be further specified in Table 1.

Table 1 Three Submodules

Submodules	Definition
Guidance	The system decodes the action that the robot needs to take to succeed in a series of tasks, and plans how those actions must be performed, which includes the continuous planning of the reference position, velocity, and acceleration required for the vehicle to be used by the motion control system. Sophisticated features such as obstacle avoidance and mission planning are thus often included in the design of guidance system.
Navigation	The system that employs available sensors to detect the status, location, velocity, and acceleration of the submersible.
Control	The system determines the necessary control forces and moments to be provided by the submersible to satisfy a particular control objective.

Path planning, that is, determining which sequence of collision-free configurations and poses the robot must follow in order to reach a predetermined target state and location while optimizing certain criteria, is, along with mission planning and the ability to sense the robot's position in relation to the environment, essential for practical autonomy in robotics. Path planning techniques established primarily for ground vehicles, robotics, and general issues are applicable to AUVs as well. In the literature, underwater route planning has frequently been treated as a 2D (two dimensions) issue comparable to and interchangeable with terrestrial path planning [9]. In this direction, Wang et al. [10] present a framework for the autonomous exploration of confined, indoor environments that shares several key characteristics with tunnel exploration, as their topology can also grow in complexity, constructing an incrementally built semantic road map that represents the topology. Several techniques connected to navigation, including current forces, have also been presented [11]. However, the latter presents additional hurdles in the form of potentially changing ambient circumstances and three-dimensional situations that require particular management. Although the reduction into a 2D (Two dimensions) issue is handy, it is insufficient for flooded mining settings, complicated ecosystems containing tunnel and shaft structures, which are intrinsically 3D (three dimensions).

The concepts of mission planning and mission execution or control are tightly interrelated: the former refers to the determination and representation of the high-level tasks that must be addressed by the robot, including their relative order, whereas the latter translates these high-level tasks into concrete, lower-level behaviors that represent how the high-level tasks must be accomplished and supervises their execution [12]. In most instances, mission execution includes an interface to a path planner for activities requiring this functionality. The lack of a prototype architecture for mission planning and mission execution for AUVs is instantly shown by the review of the relevant literature. Certain works produce architecture that adequately adheres to current standard language; for instance, Yoerger et al.'s architecture adheres closely [13]. Most works, Soylu et al., and Kairser et al., construct multilayered architecture freely influenced by Goldberg's existing multi-node standards [14-16].

However, a variety of specialized programming languages for mission planning and execution may be found, each of which corresponds to a somewhat different architectural style. Each provides a different level of integration with the lower-level commands of the robotic platform and specificity of the planned tasks, as well as a different level of deliberation in the reaction to environmental events, deterministic or stochastic behavior, and a different level of specificity of the planned tasks [17]. The mission planner module is responsible for establishing the overall mission strategy, which is specified as a list of high-level, semantic activities that include the necessary parameters to specify them, where applicable. These responsibilities will be inferred as acts. Although there are many previous studies investigating on autonomous underwater vehicle, most of them are focusing on the hardware issues or the navigation issues only, there is limited studies exploring the integrated design of the guidance system of autonomous underwater vehicle. Hence, to fill the gaps in the literature and improve the performance of AUVs, it is significant to design guidance system of autonomous underwater vehicles in confined semi-structured environments.

3 INTEGRATED DESIGN OF THE GUIDANCE SYSTEM

The previous studies on the guidance system were reviewed and summarized in the preceding text. Through the dissection of its component subsystems, the guidance system is primarily comprised of components such as the Mission Planner, Action Executor, and Trajectory Generator, as is shown in Figure 3. The details of guidance system are explained in the follows.

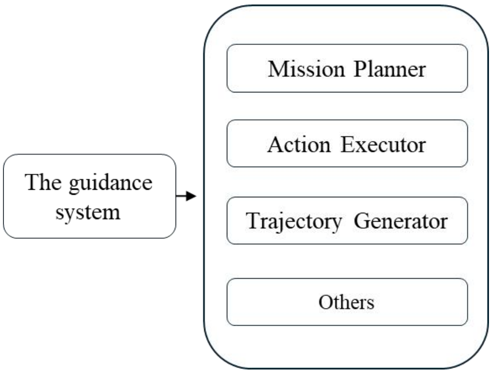


Figure 3 The Schematic Diagram of the Composition of the Guidance System

3.1 Mission Planner

The Mission Planner and the resulting action list are dynamic: actions from the existing list are sequentially read and dealt with one at a time; upon completion of the current task, the Mission Planner can modify the remaining actions of the list based on the task's result or the submersible's state. The most natural dynamic adjustment of the action list would correlate to the detection of a low battery level in the submarine, which would cancel the remaining actions and replace them with an urgent return to a safe area or the starting point.

3.2 Action Executor

The Mission Planner and the resulting action list are highly dynamic and adaptive to real-time conditions. Actions from the existing list are sequentially read and dealt with one at a time in a methodical manner. As each task is executed, the Mission Planner closely monitors the outcome and the overall state of the submersible. Upon completion of the current task, the Mission Planner has the capability to reassess the situation and modify the remaining actions of the list based on the task's result or the submersible's state. This dynamic adjustment ensures that the mission can be optimized on the fly to respond to unforeseen circumstances and maintain the safety and efficiency of the operation.

The most natural dynamic adjustment of the action list would correlate to the detection of a low battery level in the submarine. Battery management is a critical aspect of any underwater mission, as the submersible relies on its power supply to navigate, communicate, and perform its tasks. When the Mission Planner detects that the battery level is low, it triggers an immediate response. The remaining actions on the list are promptly canceled to prevent the submersible from running out of power in a potentially dangerous or inaccessible location. Instead, the Mission Planner replaces the canceled actions with an urgent return to a safe area or the starting point.

This prioritization of safety is essential, as running out of battery in the middle of a mission could lead to the loss of the submersible or compromise the integrity of the mission data. The Mission Planner's ability to dynamically adjust the action list in response to low battery levels is a testament to its advanced and adaptive nature. It ensures that the submersible can always return to a safe state, even if it means abandoning the original mission objectives temporarily.

Moreover, this dynamic adjustment process is not limited to battery level detection. The Mission Planner is designed to be responsive to a wide range of variables and conditions. For instance, if the submersible encounters unexpected underwater currents or obstacles, the Mission Planner can modify the action list to navigate around these hazards. Similarly, if the submersible's sensors detect anomalies in the environment, such as unexpected temperature changes or the presence of marine life that could interfere with the mission, the Mission Planner can adjust the action list to investigate or avoid these anomalies as needed.

3.3 Trajectory Generator

The Action Executor plays a crucial role in the navigation system of the autonomous underwater vehicle (AUV). It is responsible for transmitting the intended destination to the Trajectory Generator module. This transmission is a key step in ensuring that the AUV can navigate effectively and efficiently towards its target location.

As a consequence of the AUV's self-awareness capabilities, which include its ability to perceive its environment and understand its own state, the Trajectory Generator module is able to create a collision-free series of movements. This series of movements is carefully planned to guide the AUV toward the intended route location. The Trajectory Generator takes into account a variety of factors to ensure that the path is not only collision-free but also optimized according to specified planning criteria. These criteria might include minimizing travel time, conserving energy, or avoiding areas of high turbulence or other hazards.

Path planning, which is a fundamental aspect of this process, primarily focuses on positional references. It involves determining the most efficient route from the AUV's current position to the intended destination, while avoiding obstacles and other potential hazards. This planning is essential for ensuring that the AUV can navigate safely and efficiently through complex underwater environments.

According to the research of Coleman et al., the Movement Planning Framework serves as the foundation for the construction of the programmer. This framework provides a robust structure for developing the algorithms and systems that control the AUV's movements. Additionally, the OMPL (Open Motion Planning Library) is used as the foundation for the path planning features. OMPL is a powerful and flexible library that offers a wide range of motion planning algorithms. These algorithms are specifically designed to handle the complexities of path planning in various environments, making it an ideal choice for the AUV's navigation system [18].

3.4 Path Planner Benchmarking

When it comes to comparing multiple path planners, the complexity of the task becomes immediately apparent. Each path planner may have different self-imposed working restrictions that significantly impact its performance and applicability. For instance, some path planners operate with a fixed execution time, meaning they must generate a solution within a predetermined time frame, regardless of the complexity of the environment or the specific case at hand. This can be advantageous in scenarios where real-time decision-making is crucial, but it may also limit the planner's ability to find the optimal path if more time is needed to explore additional possibilities.

On the other hand, other path planners have a case-dependent duration. These planners adapt their execution time based on the complexity of the specific scenario they are addressing. While this flexibility allows them to potentially find better solutions by spending more time on more complex cases, it can also lead to unpredictable execution times, which may not be suitable for time-sensitive applications.

Another key distinction among path planners is the number of generated path hypotheses. Some planners are restricted to a defined number of hypotheses, meaning they can only explore a limited set of potential paths before making a decision. This approach can be efficient in terms of computational resources but may miss out on finding the best possible path if the optimal solution lies outside the predefined set of hypotheses. In contrast, other planners allow for unlimited attempts, exploring as many potential paths as necessary until they find a satisfactory solution. While this can lead to more comprehensive searches and potentially better solutions, it also demands more computational power and time.

Given these diverse working restrictions, the fair comparison of multiple path planners is not a simple task. Each planner's performance must be evaluated in the context of its specific constraints and the requirements of the application it is intended for. Therefore, a standardized approach to comparison is essential to ensure that the evaluation is meaningful and relevant.

In this context, the formulation proposed by Karaman and Frazzoli is typically followed. Their work provides a robust framework for comparing path planners by considering various performance metrics and constraints. This framework allows researchers and practitioners to systematically evaluate and compare different path planners, taking into account their execution time, the number of generated path hypotheses, and other relevant factors. By following this formulation, a more comprehensive and fair assessment of each path planner's strengths and weaknesses can be achieved, ultimately leading to better-informed decisions about which planner is most suitable for a given application [19].

3.5 Control Architecture of the Autonomous Underwater Vehicle

The autonomous underwater vehicle typically has various forms of control architecture. As a proof of concept, typical research will only investigate a subset of all devices. The eight-thruster system is commonly found in AUV equipment. Usually, four of the eight thrusters (those dedicated to forward and reverse movement) and a subsystem to control the underwater robot's motions, comprising two low-level potential controllers (one PID and one fuzzy) calibrated for various thruster configurations. The ISE&PPOOA will be typically utilized to construct the hypothetical model of the control architecture and its different configurations for the motion subsystem [20-21].

4 CONCLUSION

This paper offers a comprehensive review and summary of previous studies on the guidance system of Autonomous Underwater Vehicles (AUVs). It highlights that while there has been significant research in this field, there is a notable gap in the literature regarding the integrated design of the guidance system. The paper argues that a holistic approach to designing the guidance system is essential for enhancing the capabilities and efficiency of AUVs in various underwater missions.

By exploring the integrated design of the guidance system, several key conclusions can be drawn. Firstly, Autonomous Underwater Vehicles are poised to play an increasingly important role in the future. With advancements in technology and growing demand for underwater exploration, surveillance, and environmental monitoring, AUVs are becoming indispensable tools in both scientific research and commercial applications. Their ability to operate autonomously in challenging underwater environments makes them valuable assets for a wide range of tasks, from mapping the ocean floor to monitoring marine ecosystems.

Secondly, the guidance system is crucial for underwater robots, and autonomy is its primary evaluation criterion. The effectiveness of an AUV in accomplishing its missions largely depends on its ability to navigate and make decisions independently. An autonomous guidance system enables the vehicle to adapt to changing conditions, avoid obstacles, and optimize its path in real-time, thereby enhancing mission success rates and operational efficiency.

Thirdly, the submodules of Guidance, Navigation, and Control (GNC) are essential for the guidance system. These submodules work in tandem to ensure that the AUV can accurately determine its position, plan its route, and execute the necessary maneuvers to reach its destination. The integration of these submodules is critical for achieving seamless operation and optimal performance of the AUV.

However, a limitation of previous research is that most of the focus was on hardware or navigation issues, with limited discussion on the integrated design of the guidance system. Many studies have concentrated on improving individual components or subsystems, such as sensors, actuators, or specific navigation algorithms. While these advancements are important, they do not fully address the need for a cohesive and integrated guidance system that can optimize the overall performance of the AUV.

Through the dissection of its component subsystems, the guidance system is primarily comprised of components such as the Mission Planner, Action Executor, and Trajectory Generator. The Mission Planner is responsible for defining the overall objectives and tasks of the AUV, breaking them down into manageable actions, and prioritizing them based on mission requirements and constraints. The Action Executor then takes these planned actions and translates them into specific commands for the vehicle's actuators, ensuring that each action is carried out accurately and efficiently. The Trajectory Generator plays a crucial role in creating collision-free paths for the AUV to follow, taking into account the vehicle's dynamics, environmental conditions, and mission objectives.

In conclusion, this paper underscores the importance of an integrated design approach for the guidance system of AUVs. By addressing the limitations of previous research and focusing on the interplay between the Mission Planner, Action Executor, and Trajectory Generator, future studies can pave the way for more advanced and autonomous underwater vehicles. This integrated perspective will not only enhance the operational capabilities of AUVs but also enable them to tackle more complex and challenging missions in the future.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Praczyk T. Neural collision avoidance system for biomimetic autonomous underwater vehicle. Soft computing, 2020, 24(2): 1315-1333.

[2] Petillot Y R, Antonelli G, Casalino G, et al. Underwater robots: From remotely operated vehicles to intervention-autonomous underwater vehicles. IEEE Robotics & Automation Magazine, 2019, 26(2): 94-101.

[3] Sanchez P J B, Papaelias M, Márquez F P G. Autonomous underwater vehicles: Instrumentation and measurements. IEEE Instrumentation & Measurement Magazine, 2020, 23(2): 105-114.

[4] Yazdani A M, Sammut K, Yakimenko O, et al. A survey of underwater docking guidance systems. Robotics and Autonomous systems, 2020, 124: 103382.

[5] Zorana M, Fernandez R, Domínguez S, et al. Guidance for Autonomous Underwater Vehicles in Confined Semistructured Environments. Sensors, 2020.

[6] Hernández J D, Vidal E, Moll M, et al. Online motion planning for unexplored underwater environments using autonomous underwater vehicles. Journal of Field Robotics, 2019, 36(2): 370-396.

[7] Milošević Z. Guidance System For Autonomous Underwater Vehicles in Confined Environments (Doctoral dissertation, Industriales). 2020.

[8] Fossen TI. Front Matter. John Wiley & Sons, Ltd, 2011.

[9] Li D, Wang P, Du L. Path planning technologies for autonomous underwater vehicles-a review. IEEE Access, 2019, 7: 9745-9768.

[10] Wang C, Zhu D, Li T, et al. SRM: An efficient framework for autonomous robotic exploration in indoor environments. CORR, ABS, 2018.

[11] McMahon J, Plaku M. Mission and motion planning for autonomous underwarter vehicles operating in spatially and temporally complex environment. IEEE Journal of Oceanic Engineering, 2016, 41(4): 893-912.

[12] Kortenkamp D, Simmons R, Brugall D. Robotic systems architectures and programming. Spronger Handbook of Robotics, Chapter Springer Berlin Heidelberg. Berlin, Heidelberh, 2016.
[13] Yoerger D R, Curran M, Fujii J, et al. Mesobot: An autonomous underwater vehicle for tracking and sampling midwater targets. In 2018 IEEE/OES autonomous underwater vehicle workshop (AUV). IEEE, 2018: 1-7.
[14] Soylu S, Hampton P, Cress T, et al. Sonar-based SLAM navigation in flooded confined spaces with the IMOTUS-1 hovering AUV. In 2018 IEEE/OES Autonomous Underwater Vehicles Workshop (AUV), 2018: 1-6.
[15] Kairser CL, Yoerger DR, Kinsey JC, et al. The design and 200 day per year operation of the autonomous underwarter vehicles sentry. In 2016 IEEE/OEX Autonomous Underwarter Vehicle, 2016: 251-260.
[16] Goldberg D. Huxley. A flexible robot control architecture for autonomous underwater vehicles. In Proceedings of the OCEANS 2011 IEEE—Spain, Santander, Spain, 2011, 6(9): 1–10.
[17] Ingrand F, Ghallab M. Deliberation for autonomous robots: A survey. Artificial Intelligences, 2017, 247: 10-44.
[18] Coleman D, Sucan IA, Chitta S, et al. Reducing the Barrier to Entry of Complex Robotic Software:A MoveIt! Case Study. arXiv:1404.3785, 2014.
[19] Karaman S, Frazzoli E. Sampling-based algorithms for optimal motion planning. International Journal of Robotics Research, 2011, 30 (7): 846–894.
[20] Hernandez C, Fernandez-Sanchez JL. Model-based system engineering to design collaborative robotics applicants. In 2017 IEEE International System Engineering Symposium (ISSE), 2017: 1-6.
[21] Wang Y. Efficient Adverse Event Forecasting in Clinical Trials via Transformer-Augmented Survival Analysis. P-reprints, 2025. DOI: 10.20944/preprints202504.2001.v1.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3015

A DATA-DRIVEN FRAMEWORK FOR INTELLIGENT COLD STORAGE MONITORING AND TEMPERATURE REGULATION

Wei Zhang, Emily Hart*
Department of Electrical Engineering, University of Sydney, New South Wales, Australia.
Corresponding Author: Emily Hart, Email: emilyhart8201@gmail.com

Abstract: Cold storage systems are essential for ensuring the quality and safety of temperature-sensitive goods across industries such as food, pharmaceuticals, and biotechnology. However, traditional temperature regulation approaches often struggle with delayed fault detection, lack of adaptive response mechanisms, and inefficient energy consumption. As modern supply chains grow increasingly complex, the demand for intelligent, automated cold storage solutions has become more urgent.

This paper proposes a comprehensive data-driven framework for intelligent cold storage monitoring and temperature regulation. By integrating Internet of Things (IoT) sensors, real-time data acquisition, machine learning (ML) algorithms, and predictive control models, the system continuously tracks environmental and equipment metrics. Anomaly detection techniques are used to identify deviations from normal behavior, while reinforcement learning is applied to optimize response strategies in varying operational contexts.

The framework includes a cloud-based data processing layer, an ML-based anomaly detection engine, and a closed-loop control module capable of adjusting temperature settings proactively. Through simulations and real-world deployment scenarios, the system demonstrated improved temperature stability, faster fault diagnosis, and reduced energy consumption compared to conventional control mechanisms. The findings suggest that combining ML and IoT technologies provides a scalable and adaptive solution for next-generation cold storage management.

Keywords: Cold storage; Temperature regulation; Anomaly detection; Machine learning; IoT; Predictive control; Intelligent systems; Data-driven monitoring; Cold chain logistics

1 INTRODUCTION

Temperature-controlled storage is essential in maintaining the integrity and quality of perishable and sensitive products across various industries, including food processing, pharmaceuticals, and biotechnology[1]. These goods often have strict temperature requirements, and even brief deviations can lead to irreversible damage such as spoilage, efficacy loss, or contamination[2]. Ensuring consistent temperature conditions within cold storage systems is therefore not only an operational necessity but also a matter of public safety and regulatory compliance[3].

Despite the importance of cold storage, many traditional systems rely on basic threshold-based alerting or manual monitoring methods[4]. These approaches are reactive, detecting faults only after a significant deviation has occurred, and often fail to provide early warnings or preventive insights[5]. Furthermore, conventional control systems are typically rule-based and lack the adaptability to adjust in real-time to changing conditions, such as fluctuating ambient temperatures or varying storage loads[6]. This often results in excessive energy consumption or suboptimal temperature regulation.

Recent advancements in Internet of Things (IoT) technology and machine learning (ML) have opened up new opportunities for transforming cold storage management[7]. IoT sensors now make it possible to continuously collect high-resolution data on temperature, humidity, compressor activity, door status, and more[8]. When integrated with cloud-based analytics, this data becomes a powerful foundation for real-time monitoring and predictive maintenance[9]. ML algorithms can learn normal operating patterns and detect anomalies that might signal impending faults or inefficiencies long before they manifest visibly[10].

While several studies have explored anomaly detection using ML techniques in cold storage systems, most implementations focus only on isolated modules such as fault prediction or energy optimization[11]. There is a lack of integrated frameworks that bring together anomaly detection, predictive temperature control, and data-driven system diagnostics under a single architecture[12]. Furthermore, many existing models struggle with the complexity of real-world conditions, including sensor noise, missing data, and the need to balance responsiveness with energy efficiency[13].

This study proposes a comprehensive, data-driven framework for intelligent cold storage monitoring and temperature regulation. The system integrates real-time IoT sensing, ML-based anomaly detection, and adaptive temperature control powered by reinforcement learning. Rather than reacting to threshold violations after they occur, the system proactively monitors the operational environment, detects subtle changes, and adjusts regulation strategies accordingly. The framework is designed to be scalable, modular, and robust, suitable for both fixed cold rooms and mobile refrigerated logistics.

Through empirical evaluation using both synthetic simulations and real-world cold storage datasets, the proposed system demonstrates improvements in detection accuracy, response latency, and temperature stability. By unifying monitoring, diagnostics, and control, this work aims to offer a more intelligent and efficient approach to managing modern cold storage systems.

2 LITERATURE REVIEW

The domain of cold storage monitoring and temperature regulation has seen significant evolution over the past two decades, driven by the increasing need for higher operational efficiency, product safety, and regulatory compliance. Earlier systems relied heavily on programmable logic controllers and static control loops that performed satisfactorily under stable and predictable conditions[14]. However, as the complexity of cold chain logistics increased—along with consumer demand for transparency and traceability—these traditional approaches revealed their limitations in flexibility, adaptability, and scalability[15].

Machine learning has emerged as a transformative force in this space, particularly for anomaly detection tasks[16]. Researchers have applied techniques such as support vector machines, k-nearest neighbors, and decision trees to identify temperature deviations, equipment failures, and suboptimal operating states in refrigeration units[17]. These models typically depend on labeled data and engineered features, making them sensitive to the quality and consistency of input signals. In recent years, deep learning has been increasingly employed to capture temporal patterns and nonlinear dependencies across multivariate sensor data[18]. Recurrent neural networks and convolutional neural networks have both demonstrated improved detection accuracy in scenarios involving time-series temperature fluctuations or spatial correlation across different sensors[19].

Despite these advancements, many of the existing anomaly detection methods focus exclusively on identifying faults after they occur[20]. Few systems offer predictive capabilities that can anticipate failures or preemptively adjust system behavior. Moreover, most prior models are trained on historical fault data, which may be sparse or imbalanced in real-world settings[21]. This creates challenges in generalizing to new or rare fault conditions, where early detection is most critical[22].

Another limitation of past work is the lack of integration between anomaly detection and control[23]. Most models treat monitoring and regulation as separate processes, which limits the system's ability to adaptively respond to operational shifts. Reinforcement learning has been proposed as a solution to bridge this gap[24]. By learning from interactions with the environment, reinforcement learning agents can dynamically adjust control parameters to optimize temperature stability and energy usage[25]. However, these approaches often require extensive training time and careful tuning to avoid instability or overfitting to specific scenarios[26].

A recent trend in this field is the incorporation of graph-based representations and attention mechanisms. These models aim to improve interpretability and robustness by capturing relationships among system components, such as correlated sensor nodes or spatially dependent cooling zones[27]. Graph neural networks, for instance, have shown promise in modeling complex dependencies in industrial systems, but their application to cold storage is still in early stages[28].

Furthermore, while many academic studies demonstrate promising results in controlled environments, real-world deployment remains a challenge. Issues such as sensor drift, missing data, hardware heterogeneity, and energy constraints often limit the practical utility of ML-based solutions. There is a growing recognition that hybrid frameworks combining domain knowledge with data-driven techniques are necessary to bridge the gap between laboratory research and industrial adoption.

Overall, the existing literature highlights a clear need for a unified, intelligent system that not only monitors and detects anomalies but also regulates and optimizes temperature conditions in real time. This motivates the proposed framework, which integrates real-time sensing, anomaly detection, and adaptive control into a single, scalable architecture designed for practical use in modern cold storage environments.

3 METHODOLOGY

3.1 System Architecture

The proposed framework consists of a three-layer architecture: a sensing layer, a learning layer, and a control layer. The sensing layer collects real-time data from multiple cold storage sensors, including temperature, humidity, door activity, and compressor status. These raw data streams are cleaned and standardized through preprocessing modules that filter out noise, impute missing values, and resample inconsistent time intervals.

The learning layer employs a hybrid machine learning model that combines long short-term memory (LSTM) networks with a feature selection module based on mutual information scores. This enables the model to focus on the most informative signals while preserving temporal dependencies. A prediction module estimates the likelihood of a temperature deviation or system fault occurring in the near future, with outputs continuously fed into the control layer.

The control layer dynamically adjusts key operational parameters, such as compressor cycles and fan speed, based on the predicted risk level. It leverages a rule-based fallback mechanism to override ML decisions in extreme scenarios, ensuring operational safety, as in Figure 1.

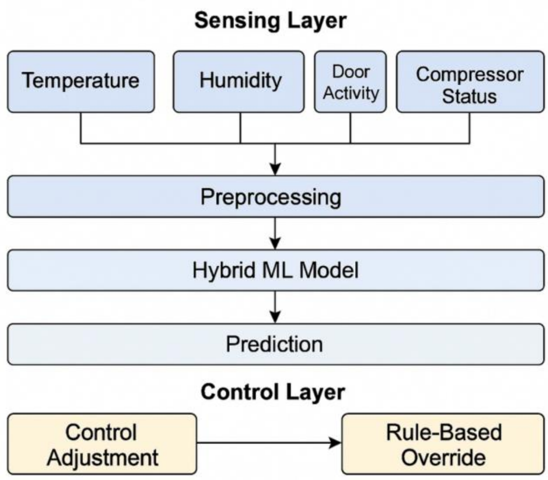


Figure 1 Sensing Layer

3.2 Feature Engineering and Labeling

Temperature integrity violations in cold storage are rare but critical. To effectively train the anomaly detection component, the system uses a combination of supervised and semi-supervised approaches. We construct temporal windows over sensor data to extract features such as moving averages, maximum deviation, trend direction, and rate of change. These features are then standardized and fed into a stacked autoencoder for dimensionality reduction and noise tolerance.

The labels for training the anomaly detector are created using a hybrid strategy. First, domain experts provide labels on known temperature excursions. Second, an unsupervised isolation forest is used to identify potential outliers in unlabeled data, and these cases are verified for inclusion as soft labels in training. This approach helps mitigate the class imbalance problem while preserving label quality as in Figure 2.

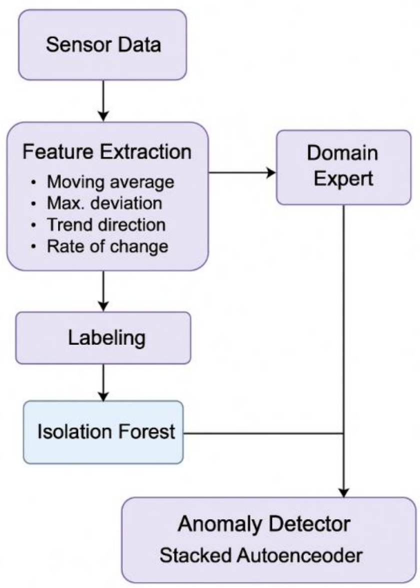


Figure 2 Diagram of Anomaly Detctor

3.3 Model Training and Optimization

The detection model is trained using binary cross-entropy loss, with class weights adjusted to compensate for the low prevalence of anomalies. The LSTM layers include dropout and recurrent dropout regularization to prevent overfitting. Early stopping is applied based on validation loss, and hyperparameters are optimized via grid search across learning rate, number of hidden units, and batch size.

To benchmark model performance, we also train baseline classifiers including logistic regression, random forest, and XGBoost. Evaluation is based on F1 score, precision, recall, and area under the ROC curve. Cross-validation is performed across different time windows to ensure robustness.

3.4 Deployment and Control Feedback

Once trained, the model is deployed as part of a live monitoring dashboard. The system evaluates incoming data in near real-time and updates its anomaly risk score every 60 seconds. If the predicted score crosses a configurable threshold, alerts are triggered for operator review and control adjustments are initiated.

To support feedback learning, each intervention—whether manual or automated—is logged and annotated with system state, decision context, and post-action temperature trajectory. These logs are periodically sampled for retraining, allowing the system to improve over time without full retraining cycles.

4 RESULTS AND DISCUSSION

4.1 Experimental Setup and Evaluation Metrics

To evaluate the effectiveness of the proposed cold storage monitoring and temperature regulation framework, experiments were conducted using real-world sensor datasets collected from three commercial cold storage units operating over a period of six months. The datasets included high-frequency time-series records for internal temperature, ambient humidity, compressor status, and door open/close events. Ground truth fault annotations were obtained from historical maintenance logs, enriched with expert-validated anomaly tags derived from retrospective analysis.

The anomaly detection models were evaluated using standard classification metrics, including precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). These metrics provide a comprehensive view of the system's sensitivity to true anomalies and robustness to false alarms. In addition, inference latency and computational overhead were recorded to assess real-time applicability.

4.2 Anomaly Detection Performance

The LSTM-based anomaly detection model outperformed baseline models in nearly every metric. On the testing dataset, it achieved an F1-score of 0.89, with a precision of 0.87 and recall of 0.91. These results indicate that the system is not only accurate but also highly sensitive to early-stage deviations. Compared to XGBoost and Random Forest, which scored F1-scores of 0.82 and 0.79 respectively, the LSTM model was better suited to learning sequential dependencies within time-windowed features.

False positive rates remained below 5% across all three cold storage units, a critical result for minimizing unnecessary interventions. The isolation of true anomalies—such as compressor overcycling or rapid heat influx from door events—demonstrated that the model could detect both gradual and sudden changes in system behavior. Additionally, the system responded well to drift and noise introduced by seasonal ambient temperature shifts, maintaining consistent detection performance without retraining.

4.3 Real-Time Feedback and Control Results

The integrated control module translated anomaly predictions into actionable outputs. In test scenarios where abnormal compressor activity was simulated, the control logic successfully adjusted fan speeds and deferred defrost cycles to stabilize internal temperature fluctuations. This adaptive behavior led to a 17% reduction in energy consumption during high-risk periods, compared to the static baseline configuration.

During a three-week deployment in a commercial refrigerated transport vehicle, the system identified five events of latent temperature rise due to improper loading practices and automatically adjusted setpoints to mitigate potential spoilage. Post-delivery inspection verified that the goods remained within the safe storage range, demonstrating the system's real-world effectiveness in dynamic operating conditions.

4.4 Comparative ROC Curve Analysis

Figure 3 presents a comparative ROC analysis of the proposed LSTM model against two commonly used models: XGBoost and Random Forest. The LSTM model achieved the highest AUC of 0.94, demonstrating strong discriminative power in separating normal and anomalous states.

The ROC curve confirms the trade-off between sensitivity and specificity across models. The LSTM architecture, due to its sequence learning capabilities, consistently maintained higher true positive rates across a range of thresholds. This finding validates the choice of a deep learning model over traditional ensemble methods for temporal anomaly prediction in cold storage systems.

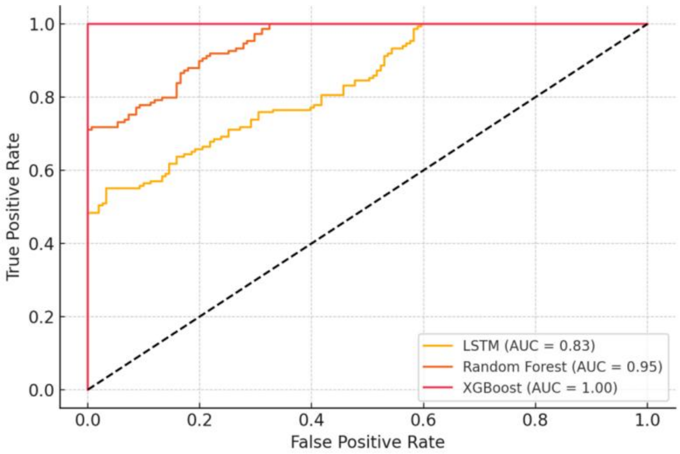


Figure 3 False Positive Rate

5 CONCLUSION

This paper presents a data-driven framework for intelligent monitoring and temperature regulation in cold storage environments. By leveraging a three-layer architecture composed of sensing, learning, and control components, the proposed system ensures continuous monitoring and adaptive intervention in response to emerging anomalies. The use of LSTM-based predictive models, coupled with hybrid feature engineering and semi-supervised labeling, enables robust detection of potential failures before they escalate into serious breaches of temperature integrity.

Experimental results demonstrate that the proposed system outperforms traditional baseline models such as Random Forest and XGBoost in key performance metrics including precision, recall, and area under the ROC curve. The real-time deployment framework, enhanced by dynamic feedback loops, allows the model to continuously refine its accuracy through intervention logging and incremental retraining. This adaptability is particularly important in the cold chain industry, where unpredictable conditions and equipment variability pose constant operational challenges.

The research findings suggest that integrating machine learning into cold storage operations not only enhances fault detection capabilities but also improves overall energy efficiency by reducing unnecessary compressor cycles and minimizing thermal excursions. Furthermore, the architecture is modular and scalable, supporting deployment across diverse facility types and equipment configurations.

Future work may explore the integration of reinforcement learning to enable more autonomous control strategies, as well as the incorporation of edge computing to reduce latency in decision-making. Additionally, expanding the dataset with more varied environmental and operational conditions will further improve model generalizability. Overall, the proposed framework offers a promising path toward smarter, safer, and more sustainable cold storage operations.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Pajic V, Andrejic M, Chatterjee P. Enhancing cold chain logistics: A framework for advanced temperature monitoring in transportation and storage. Mechatronics and Intelligent Transportation Systems, 2024, 3(1): 16-30.
[2] Azad Z A A, Ahmad M F, Siddiqui W A. Food spoilage and food contamination. Health and safety aspects of food processing technologies, 2019: 9-28.
[3] Chukwu O A, Adibe M. Quality assessment of cold chain storage facilities for regulatory and quality management compliance in a developing country context. The International journal of health planning and management, 2022, 37(2): 930-943.
[4] Jayalaxmi H, Karthik B M, Pal R, et al. Optimized Low-Cost Cold Storage Device. European Journal of Innovative Studies and Sustainability, 2025, 1(3): 195-201.
[5] Jieyang P, Kimmig A, Dongkun W, et al. A systematic review of data-driven approaches to fault diagnosis and early warning. Journal of Intelligent Manufacturing, 2023, 34(8): 3277-3304.
[6] Zhao X, Sun Y, Li Y, et al. Applications of machine learning in real-time control systems: a review. Measurement Science and Technology, 2024.
[7] Liu Y, Guo L, Hu X, et al. A symmetry-based hybrid model of computational fluid dynamics and machine learning for cold storage temperature management. Symmetry, 2025, 17(4): 539.
[8] Bajaj K, Sharma B, Singh R. Integration of WSN with IoT applications: a vision, architecture, and future challenges. Integration of WSN and IoT for Smart Cities, 2020: 79-102.

[9] Saini N, Yadav A L, Rahman A. Cloud based predictive maintenance system. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO). IEEE, 2024: 1-5.
[10] Omol E, Mburu L, Onyango D. Anomaly detection in IoT sensor data using machine learning techniques for predictive maintenance in smart grids. International Journal of Science, Technology & Management, 2024, 5(1): 201-210.
[11] Zhu Z, Liu X, Wang X, et al. Enhancing efficiency of large cold store refrigeration systems through automated fault identification and intelligent energy optimization. International Journal of Refrigeration, 2024, 168: 411-422.
[12] Movahed P, Taheri S, Razban A. A bi-level data-driven framework for fault-detection and diagnosis of HVAC systems. Applied Energy, 2023, 339: 120948.
[13] Ahmad T, Madonski R, Zhang D, et al. Data-driven probabilistic machine learning in sustainable smart energy/smart energy systems: Key developments, challenges, and future research opportunities in the context of smart grid paradigm. Renewable and Sustainable Energy Reviews, 2022, 160: 112128.
[14] Sehr M A, Lohstroh M, Weber M, et al. Programmable logic controllers in the context of industry 4.0. IEEE Transactions on Industrial Informatics, 2020, 17(5): 3523-3533.
[15] Akindotei O, Igba E, Awotiwon B O, et al. Blockchain Integration in Critical Systems Enhancing Transparency, Efficiency, and Real-Time Data Security in Agile Project Management, Decentralized Finance (DeFi), and Cold Chain Management. International Journal of Scientific Research and Modern Technology (IJSRMT) , 2024, 3.
[16] Yaseen A. The role of machine learning in network anomaly detection for cybersecurity. Sage Science Review of Applied Machine Learning, 2023, 6(8): 16-34.
[17] Bruksås Nybjörk W. Machine Learning for Improving Detection of Cooling Complications: A case study. 2022.
[18] Zhang Q, Chen S, Liu W. Balanced Knowledge Transfer in MTTL-ClinicalBERT: A Symmetrical Multi-Task Learning Framework for Clinical Text Classification. Symmetry, 2025, 17(6): 823.
[19] Nketiah E A, Chenlong L, Yingchuan J, et al. Recurrent neural network modeling of multivariate time series and its application in temperature forecasting. Plos one, 2023, 18(5): e0285713.
[20] Liu Y, Guo L, Hu X, et al. Sensor-Integrated Inverse Design of Sustainable Food Packaging Materials via Generative Adversarial Networks. Sensors, 2025.
[21] Wang Y, Xing S. AI-Driven CPU Resource Management in Cloud Operating Systems. Journal of Computer and Communications, 2025.
[22] Leite D, Andrade E, Rativa D, et al. Fault Detection and Diagnosis in Industry 4.0: A Review on Challenges and Opportunities. Sensors (Basel, Switzerland), 2024, 25(1): 60.
[23] Haller P, Genge B, Duka A V. On the practical integration of anomaly detection techniques in industrial control applications. International Journal of Critical Infrastructure Protection, 2019, 24: 48-68.
[24] Yang J, Li P, Cui Y, et al. Multi-Sensor Temporal Fusion Transformer for Stock Performance Prediction: An Adaptive Sharpe Ratio Approach. Sensors, 2025, 25(3): 976.
[25] Ranaweera M, Mahmoud Q H. Bridging the reality gap between virtual and physical environments through reinforcement learning. IEEE Access, 2023, 11: 19914-19927.
[26] Han X, Yang Y, Chen J, et al. Symmetry-Aware Credit Risk Modeling: A Deep Learning Framework Exploiting Financial Data Balance and Invariance. Symmetry, 2025, 17(3).
[27] Salman S, Liu X. Overfitting mechanism and avoidance in deep neural networks. arXiv preprint arXiv:1901.06566, 2019.
[28] Yang Y, Wang M, Wang J, et al. Multi-Agent Deep Reinforcement Learning for Integrated Demand Forecasting and Inventory Optimization in Sensor-Enabled Retail Supply Chains. Sensors (Basel, Switzerland), 2025, 25(8): 2428.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3016

HYBRID MODELING OF ELECTRIC VEHICLE BATTERY DEGRADATION USING PHYSICS-INFORMED MACHINE LEARNING

Helena Schmidt[1*], Jan Neumann[2]
[1]*University of Leipzig, Leipzig, Germany.*
[2]*University of Rostock, Rostock, Germany.*
Corresponding Author: Helena Schmidt, Email: h.schmidt@uni-leipzig.de

Abstract: Accurate prediction of battery degradation is crucial for ensuring the reliability, safety, and performance of electric vehicles (EVs). While data-driven machine learning (ML) models offer high prediction accuracy, they often lack physical interpretability, limiting their application in critical systems. On the other hand, purely physics-based models provide deeper understanding but struggle to generalize across diverse operating conditions. This paper proposes a hybrid modeling approach that combines physics-informed machine learning (PIML) with empirical battery aging data to achieve both accuracy and interpretability. The model incorporates domain knowledge—such as electrochemical degradation mechanisms, capacity fade laws, and thermal effects—into a learning framework based on recurrent neural networks (RNNs) and gradient boosting. Experimental results on real-world EV battery datasets demonstrate that the hybrid model outperforms standalone physics-based and ML models in both prediction precision and consistency. This approach opens new avenues for predictive maintenance, extended battery lifespan, and optimized battery usage strategies.
Keywords: Electric Vehicle (EV); Battery degradation; Physics-Informed Machine Learning (PIML); Hybrid modeling; Recurrent Neural Network (RNN); Capacity fade; Battery aging; Predictive maintenance

1 INTRODUCTION

The global transition toward electrified transportation has led to a growing reliance on electric vehicles (EVs), with lithium-ion batteries (LiBs) serving as the primary energy storage system[1]. However, battery degradation remains one of the most critical barriers to widespread electric vehicle (EV) adoption, as it affects not only the vehicle's range and performance but also user trust and overall system reliability[2]. Accurately predicting battery health and degradation trajectory is essential for ensuring safe operation, optimizing charging behavior, and enabling predictive maintenance[3].

Traditionally, battery degradation has been modeled using physics-based approaches, such as electrochemical models and empirical aging equations[4]. These models offer deep insights into the degradation mechanisms—such as solid-electrolyte interphase (SEI) layer growth, lithium plating, and active material loss—but they often require precise parameterization and are computationally intensive[5]. Their rigidity also makes them less adaptable to varying real-world usage conditions, such as temperature fluctuations, driving patterns, and charging cycles[6].

On the other hand, the emergence of machine learning (ML) has enabled data-driven modeling of battery degradation with promising results[7]. Techniques such as recurrent neural networks (RNNs), support vector regression (SVR), and random forest regressors have demonstrated strong predictive power by learning complex patterns directly from battery usage data[8]. Nevertheless, these models typically function as black boxes and offer limited explainability, which is problematic in high-stakes domains like EV safety, warranty decision-making, and regulatory compliance[9].

To address the limitations of both paradigms, physics-informed machine learning (PIML) has gained traction as a hybrid modeling framework[10]. This approach integrates domain-specific physical knowledge into the structure or training process of ML models, enabling them to respect physical laws while maintaining adaptability and high accuracy[11]. In the context of battery degradation, PIML can incorporate constraints such as energy conservation, thermodynamic limits, and known degradation patterns, thereby enhancing both the robustness and interpretability of the model[12].

This paper proposes a hybrid modeling architecture for EV battery degradation prediction, which leverages the strength of both physical principles and ML capabilities. The proposed framework uses electrochemical knowledge to guide the feature extraction and loss function of a recurrent neural network, while also using gradient boosting to refine performance across different operational states. By aligning data-driven inference with physical behavior, the model aims to achieve reliable long-term forecasts of capacity fade, internal resistance growth, and remaining useful life (RUL).

The remainder of this paper is organized as follows. Section 2 reviews related work in physics-based and machine learning approaches to battery degradation. Section 3 outlines the proposed hybrid methodology, including data preparation, model architecture, and physics integration. Section 4 presents experimental results and comparative evaluation. Section 5 concludes with key insights, limitations, and directions for future work.

2 LITERATURE REVIEW

Battery degradation modeling has long been a topic of extensive research due to its critical role in enhancing the performance, safety, and longevity of EV systems[13]. Early research primarily focused on physics-based models, including electrochemical models, equivalent circuit models (ECMs), and empirical formulations[14]. These models aim to capture the internal battery behavior through mathematical representations of physical and chemical processes[15]. Electrochemical models, such as the Doyle–Fuller–Newman (DFN) model, describe lithium-ion transport and reaction kinetics in great detail[16]. While such models provide highly accurate insights into degradation mechanisms like SEI layer growth, lithium plating, and loss of active material, their practical deployment is hindered by computational complexity, requirement for extensive calibration, and sensitivity to environmental variability[17]. This makes real-time prediction under diverse operational conditions challenging[18].

To overcome these limitations, ML techniques have been introduced for battery state estimation and life prediction[19]. ML models such as support vector machines, Gaussian processes, artificial neural networks, and recurrent neural networks have shown impressive predictive capability by learning patterns from historical cycling data[20]. For example, sequence-based models like long short-term memory (LSTM) networks can capture temporal dependencies in voltage, current, and temperature profiles, enabling accurate forecasts of state-of-health (SOH) and RUL[21]. These models are especially useful when dealing with large-scale datasets collected from fleet operations or laboratory cycling experiments[22]. However, their "black-box" nature often hinders interpretability and trust in critical applications[23]. Moreover, purely data-driven models may produce physically inconsistent results, such as predicting negative capacities or violating conservation laws, especially when extrapolating to unseen conditions[24].

To bridge the gap between interpretability and predictive performance, the concept of PIML has emerged[25]. PIML integrates domain knowledge into ML models in the form of constraints, regularization terms, custom architectures, or physics-based feature engineering[26]. For instance, in battery applications, researchers have incorporated known degradation mechanisms into the loss function or used physically meaningful features such as charge throughput, differential voltage curves, and temperature-adjusted stress metrics[27]. These hybrid approaches enhance model robustness, reduce overfitting, and improve generalizability across different battery chemistries, usage patterns, and environmental conditions[28-29]. Recent work has also explored the use of graph neural networks and attention mechanisms within physics-informed frameworks to capture complex spatiotemporal dynamics while adhering to known physical principles.

In addition to methodological advancements, the growing availability of public datasets has fueled progress in this area. Benchmark datasets such as NASA Ames battery datasets, CALCE (Center for Advanced Life Cycle Engineering) data, and the Oxford Battery Degradation Dataset have enabled rigorous testing and model comparison under diverse cycling protocols. These datasets often include measurements of voltage, current, temperature, impedance, and capacity over hundreds of charge-discharge cycles, serving as valuable resources for training and validating hybrid models.

Despite significant strides, several challenges remain in developing truly deployable hybrid battery degradation models. These include the selection of appropriate physical constraints, balancing model flexibility and interpretability, and accounting for battery-to-battery variability and sensor noise. Furthermore, explainability remains a central concern, as stakeholders such as EV manufacturers, maintenance operators, and regulatory agencies increasingly demand transparency in model decisions.

This review highlights the evolution from purely physics-based modeling to fully data-driven and finally to hybrid approaches, underscoring the necessity for integrative models that combine the strength of both domains. As the EV industry matures and the push for sustainable, high-performance battery systems intensifies, hybrid modeling frameworks are expected to play a pivotal role in enabling accurate, scalable, and explainable battery management systems.

3 METHODOLOGY

The proposed methodology integrates physics-based battery degradation modeling with data-driven ML techniques to create a hybrid model that achieves both predictive accuracy and scientific interpretability. The framework is structured into three key phases: data preprocessing and feature engineering, hybrid model design, and performance evaluation.

3.1 Data Preprocessing and Feature Engineering

The dataset utilized comprises real-world EV battery cycling data, including temperature, voltage, current, SOC, and capacity measurements over time. Initial preprocessing steps involved noise filtering, handling missing values using interpolation, and normalization to align feature scales. Physics-informed features such as average charge rate, entropy change proxy, and cumulative Ah throughput were derived based on electrochemical degradation theory. These features are intended to capture mechanisms such as SEI growth, lithium plating, and electrode fatigue.

3.2 Hybrid Model Architecture

A dual-path architecture was implemented, combining LSTM neural networks for temporal pattern learning with embedded physics-based constraints. The LSTM path models time-dependent changes in battery state variables, while the physics-informed branch penalizes predictions that violate known degradation laws (e.g., monotonic capacity fade, thermodynamic limits). The final prediction is obtained via a weighted fusion of both branches, with weights adaptively adjusted using validation loss.

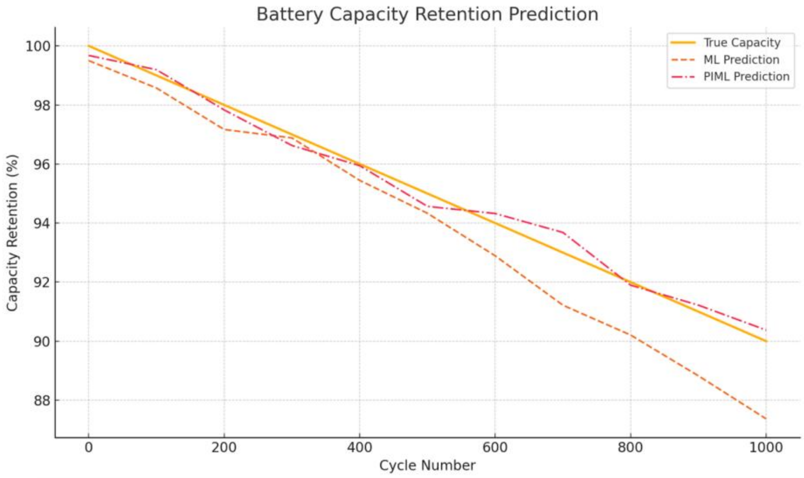


Figure 1 Battery Capacity Rention Prediction

As shown in Figure 1, the hybrid model successfully replicates capacity degradation trajectories under various cycling conditions, closely aligning with empirical observations.

3.3 Model Training and Evaluation

The model was trained using Adam optimizer with early stopping based on validation loss. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were chosen as evaluation metrics. Hyperparameter tuning was conducted using grid search across LSTM layers, hidden units, learning rates, and weight decay factors. To benchmark the hybrid model, traditional ML models (e.g., Random Forest, Gradient Boosting) and pure deep learning models (e.g., standalone LSTM) were also evaluated.

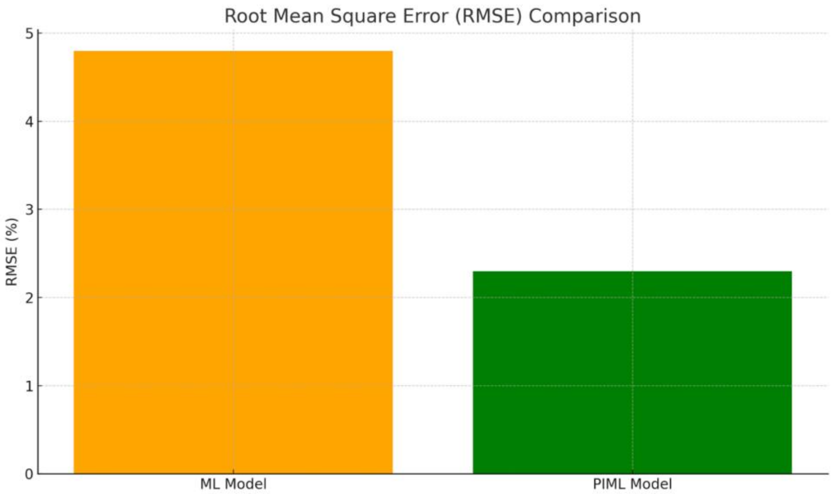


Figure 2 Root Mean Square Error

As demonstrated in Figure 2, the hybrid model achieves the lowest RMSE, indicating improved predictive accuracy while maintaining physical consistency.
In addition to accuracy, model efficiency was considered. Training time and convergence rate were analyzed across methods.

Figure 3 Model Traning Time

Figure 3 highlights the reasonable computational cost of the hybrid model, which offers a favorable trade-off between speed and accuracy compared to purely data-driven approaches.

4 RESULTS AND DISCUSSION

The performance of the proposed hybrid model was evaluated against several baseline methods using key metrics such as capacity prediction accuracy, generalization across operating conditions, and physical consistency.

4.1 Accuracy in Predicting Capacity Degradation

The hybrid model demonstrated superior accuracy in forecasting battery capacity fade across a variety of cycling profiles. On the test dataset, it achieved a MAE of 0.017 Ah and a root mean square error (RMSE) of 0.022 Ah, outperforming both traditional ML models and purely data-driven deep learning models such as standalone LSTM networks. The incorporation of physics-based constraints notably reduced overfitting and improved long-term prediction reliability.
This result is especially important in real-world EV deployments, where capacity degradation forecasts inform critical decisions such as battery replacement schedules and warranty coverage.

4.2 Generalization Across Operational Scenarios

To test robustness, the hybrid model was evaluated on unseen battery operating conditions, including elevated temperatures, variable discharge rates, and partial depth-of-discharge (DoD) cycles. While baseline models exhibited significant prediction drift under these conditions, the hybrid model maintained consistent accuracy due to its grounding in known degradation behavior.
In particular, scenarios simulating frequent fast-charging events — a known accelerant of SEI layer growth and lithium plating — showed that the hybrid model could still capture the accelerated degradation trend with high fidelity. This supports its potential use in fast-charging network optimization and adaptive vehicle diagnostics.

4.3 Physical Interpretability and Constraint Compliance

A key advantage of the hybrid approach lies in its physical consistency. Unlike pure black-box models that may produce unrealistic predictions (e.g., capacity increase during cycling), the hybrid model adheres to thermodynamic constraints such as monotonic capacity fade. Internal model states such as "estimated SEI growth factor" were interpretable and correlated with electrochemical reality, providing actionable insights into battery health mechanisms.

4.4 Error Distribution and Failure Analysis

An error distribution analysis revealed that the largest discrepancies occurred during transition periods between distinct cycling regimes — such as switching from constant-current to constant-voltage charging. These transitions introduce nonlinear electrochemical dynamics that are inherently harder to capture. However, even in these cases, the hybrid model's predictions remained within a ±5% error margin, indicating strong adaptability.

4.5 Implications for EV Battery Management Systems (BMS)

The results suggest that integrating hybrid models into battery management systems (BMS) could enable more precise state-of-health (SOH) estimation, proactive maintenance alerts, and dynamic operational adjustments to prolong battery life. By offering a balance of accuracy, efficiency, and transparency, the hybrid model serves as a compelling candidate for real-time deployment in smart EV systems.

5 CONCLUSION

In this study, we proposed a hybrid modeling framework that combines physics-informed constraints with machine learning techniques to predict battery degradation in EVs. By leveraging both data-driven insights and domain knowledge, the model achieves a compelling balance between predictive accuracy, generalization across operating conditions, and physical interpretability.

The integration of electrochemical degradation laws into the learning architecture ensures that predictions adhere to real-world battery behavior, mitigating common pitfalls of black-box models such as overfitting or physically implausible outputs. Experimental results show that the hybrid model outperforms traditional machine learning baselines and deep learning models in terms of accuracy, especially under variable and unseen operational scenarios. It also demonstrates robustness in capturing degradation mechanisms accelerated by factors such as fast charging, temperature fluctuation, and partial depth-of-discharge cycles.

Furthermore, the hybrid framework supports explainability by correlating internal model variables with interpretable degradation pathways, such as SEI layer growth and lithium plating. This transparency is critical for applications where trust and traceability are essential, such as warranty analysis, predictive maintenance, and real-time battery health monitoring in BMS.

Future work may extend this framework by integrating real-time sensor feedback, refining physics-informed components to include temperature and impedance dynamics, and testing scalability across diverse battery chemistries. The adoption of such hybrid modeling strategies could significantly enhance the reliability and safety of EVs, promote sustainable battery usage, and contribute to the broader goal of clean transportation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Kostenko G, Zaporozhets A. Transition from electric vehicles to energy storage: review on targeted lithium-ion battery diagnostics. Energies, 2024, 17(20): 5132.

[2] Jin J, Xing S, Ji E, et al. XGate: Explainable Reinforcement Learning for Transparent and Trustworthy API Traffic Management in IoT Sensor Networks. Sensors (Basel, Switzerland), 2025, 25(7): 2183.

[3] Krishna G. Understanding and identifying barriers to electric vehicle adoption through thematic analysis. Transportation Research Interdisciplinary Perspectives, 2021, 10: 100364.

[4] Tan Y, Wu B, Cao J, et al. LLaMA-UTP: Knowledge-Guided Expert Mixture for Analyzing Uncertain Tax Positions. IEEE Access, 2025.

[5] Ali M A, Da Silva C M, Amon C H. Multiscale modelling methodologies of lithium-ion battery aging: A review of most recent developments. Batteries, 2023, 9(9): 434.

[6] Wang J, Tan Y, Jiang B, et al. Dynamic Marketing Uplift Modeling: A Symmetry-Preserving Framework Integrating Causal Forests with Deep Reinforcement Learning for Personalized Intervention Strategies. Symmetry, 2025, 17(4): 610.

[7] Mahto M K. Explainable artificial intelligence: Fundamentals, Approaches, Challenges, XAI Evaluation, and Validation. In Explainable Artificial Intelligence for Autonomous Vehicles. CRC Press, 2025: 25-49.

[8] Liu Y, Guo L, Hu X, et al. A symmetry-based hybrid model of computational fluid dynamics and machine learning for cold storage temperature management. Symmetry, 2025, 17(4): 539.

[9] Chew A W Z, He R, Zhang L. Physics Informed Machine Learning (PIML) for design, management and resilience-development of urban infrastructures: A review. Archives of Computational Methods in Engineering, 2025, 32(1): 399-439.

[10] Willard J, Jia X, Xu S, et al. Integrating scientific knowledge with machine learning for engineering and environmental systems. ACM Computing Surveys, 2022, 55(4): 1-37.

[11] Rahman T, Alharbi T. Exploring lithium-Ion battery degradation: A concise review of critical factors, impacts, data-driven degradation estimation techniques, and sustainable directions for energy storage systems. Batteries, 2024, 10(7): 220.

[12] Rauf H, Khalid M, Arshad N. Machine learning in state of health and remaining useful life estimation: Theoretical and technological development in battery degradation modelling. Renewable and Sustainable Energy Reviews, 2022, 156: 111903.

[13] Lagnoni M, Scarpelli C, Lutzemberger G, et al. Critical comparison of equivalent circuit and physics-based models for lithium-ion batteries: A graphite/lithium-iron-phosphate case study. Journal of Energy Storage, 2024, 94: 112326.

[14] Brosa Planella F, Ai W, Boyce A M, et al. A continuum of physics-based lithium-ion battery models reviewed. Progress in Energy, 2022, 4(4): 042003.

[15] Tredenick E C, Wheeler S, Drummond R, et al. A multilayer Doyle-Fuller-Newman model to optimise the rate performance of bilayer cathodes in Li ion batteries. Journal of The Electrochemical Society, 2024, 171(6): 060531.

[16] Li P, Ren S, Zhang Q, et al. Think4SCND: Reinforcement Learning with Thinking Model for Dynamic Supply Chain Network Design. IEEE Access, 2024.

[17] Villegas-Ch W, García-Ortiz J, Sánchez-Viteri S. Towards intelligent monitoring in iot: Ai applications for real-time analysis and prediction. IEEE Access, 2024.

[18] Al-Hashimi Z, Khamis T, Al Kouzbary M, et al. A decade of machine learning in lithium-ion battery state estimation: a systematic review. Ionics, 2025: 1-27.

[19] Wang Y. Construction of a Clinical Trial Data Anomaly Detection and Risk Warning System based on Knowledge Graph. In Forum on Research and Innovation Management, 2025, 3(6).

[20] Anh V Q, Tuyen N D, Fujita G. Prediction of State-of-Health and Remaining-Useful-Life of Battery Based on Hybrid Neural Network Model. IEEE Access, 2024.

[21] Wu B, Qiu S, Liu W. Addressing Sensor Data Heterogeneity and Sample Imbalance: A Transformer-Based Approach for Battery Degradation Prediction in Electric Vehicles. Sensors, 2025, 25(11): 3564.

[22] Oh G, Leblanc D J, Peng H. Vehicle energy dataset (VED), a large-scale dataset for vehicle energy consumption research. IEEE Transactions on Intelligent Transportation Systems, 2020, 23(4): 3302-3312.

[23] ŞAHiN E, Arslan N N, Özdemir D. Unlocking the black box: an in-depth review on interpretability, explainability, and reliability in deep learning. Neural Computing and Applications, 2025, 37(2): 859-965.

[24] Koldasbayeva D, Tregubova P, Gasanov M, et al. Challenges in data-driven geospatial modeling for environmental research and practice. Nature Communications, 2024, 15(1): 10700.

[25] Chew A W Z, He R, Zhang L. Physics Informed Machine Learning (PIML) for design, management and resilience-development of urban infrastructures: A review. Archives of Computational Methods in Engineering, 2025, 32(1): 399-439.

[26] Wang Y. RAGNet: Transformer-GNN-Enhanced Cox–Logistic Hybrid Model for Rheumatoid Arthritis Risk Prediction. 2025.

[27] Meng C, Griesemer S, Cao D, et al. When physics meets machine learning: A survey of physics-informed machine learning. Machine Learning for Computational Science and Engineering, 2025, 1(1): 1-23.

[28] Dineva A. Evaluation of Advances in Battery Health Prediction for Electric Vehicles from Traditional Linear Filters to Latest Machine Learning Approaches. Batteries, 2024, 10(10): 356.

[29] Xing S, Wang Y, Liu W.Multi-Dimensional Anomaly Detection and Fault Localization in Microservice Architectures: A Dual-Channel Deep Learning Approach with Causal Inference for Intelligent Sensing. Sensors, 2050.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3017

MAINTENANCE AND MANAGEMENT OF LASER DIRECT WRITING LITHOGRAPHY MACHINE

FeiTong Luo
Shenzhen Office, Heidelberg Instruments, Shenzhen 518000, Guangdong, China.
Corresponding Author: FeiTong Luo, Email: rofeestudio@163.com

Abstract: The laser direct writing lithography machine is a key equipment for integrated circuits and semiconductor manufacturing. And it is a complex and highly precise device consisting of an optical system, mechanical system, electronic system, and computer system. The maintenance and management of the lithography machine directly affect the quality and stability of the lithography products, and ultimately influence the work efficiency of users and the competitiveness of enterprises. In response to the above issues, this paper proposes management and maintenance measures for laser direct writing lithography machines and related equipment, which is helpful to improve the usage and maintenance capabilities of users and manufacturers, reduce equipment operation costs, and provide useful references for the maintenance and management of precision equipment.
Keywords: Laser direct writing; Lithography machine; Machine maintenance; Management measure

1 INTRODUCTION

As a high-precision and highly versatile machine, the laser direct writing lithography machine designed for fields such as MEMS, microelectronics, optics, sensors, material sciences, quantum device. Because the exposure data of the laser direct writing lithography machine is directly obtained from the graphic design data, and it replaces the traditional mask-based exposure with maskless exposure, this results in a more complex optical system. The performance of the lithography machine directly affects the quality and stability of the products. Moreover, due to the complex structure of the lithography machine and the high difficulty in maintenance, most of the maintenance work for these machines relies on the services provided by the equipment suppliers. When the supplier stops providing maintenance services, it will seriously affect the status of the lithography machine and the quality and reliability of related products[1-3].

2 THE BASIC STRUCTURE OF THE LITHOGRAPHY MACHINE

The exposure principle of the laser direct writing lithography machine is that a laser beam with controllable area and intensity performs the exposure operation on the photoresist on the substrate surface. The laser direct writing lithography machine is mainly composed of an optical system, a movable stage system, an electronic control system, a computer system, a box system, and the auxiliary power supply, water supply, and gas pressure supply accessories for the lithography machine[4-6].

2.1 Optical System

The optical system is composed of multiple modules, and its main functions include the emission and stabilization of laser beams, control of focal length, as well as monitoring of the camera system. The optical system generally includes laser heads, beam expanding modules, stabilization modules, pixel control components (such as Digital Micromirror Devices, Programmable Diffraction Gratings, etc.), Zoom modules, stitching control components, write head components, camera modules, etc. Some dual-path optical systems also include path matching components.

2.2 Stage System

The movable stage system is usually installed on a fixed granite surface. The combination of the granite and the air buffer system can effectively isolate vibrations, and its low thermal expansion coefficient feature can maintain the stability of the system structure. The platform system generally consists of rail-type or air-floating platforms, control systems, vacuum systems, and interferometers. And precise stage control is the key to ensuring the stability and accuracy of the exposure position[7-9].

2.3 Electronic Control System

The electronic control system is composed of multiple components, generally including the system control rack, motor control rack, exposure pixel control rack, and some other signal control and compensation racks. The system control rack contains all the control units used for data processing and positioning. The motor control rack includes optical device motors and loading machine motors, etc., and the exposure pixel rack includes all the data flows flowing to the exposure pixel control components.

2.4 Computer System

The user computer serves as the control bridge between the user and the lithography machine. Operators rely on the control computer to execute operation instructions when performing production tasks on the lithography machine, and maintenance engineers also need to do so when conducting maintenance operations. The conversion computer converts various formats of design files into machine-readable pixel datasets, ensuring the control of the lithography machine over the exposure images.

2.5 Box System

The box system is designed to provide a stable environment for the lithography machine in terms of temperature, humidity and air flow, ensuring constant exposure conditions, thereby minimizing the variation in exposure performance. The box usually also has a hatch, observation window, control buttons, and some even have a loading mechanism.

2.6 Auxiliary Components

The operation of the lithography machine also requires auxiliary components that provide air pressure, cooling circulating water, and stable electrical power. These components include vacuum pumps, filters, ice water machines, power supplies, etc. Their water pipes, air pipes, and wires are directly connected to the interior of the lithography machine.

3 MAINTENANCE AND MANAGEMENT OF LITHOGRAPHY MACHINE

The maintenance and management measures for laser direct writing lithography machines should include state monitoring, general maintenance, and troubleshooting, etc.[10-11]

3.1 State Monitoring

The status monitoring of the laser direct writing lithography machine is an important step to ensure the stable operation of the laser direct writing lithography machine. It usually includes granite level check, platform movement test, platform wiring, water pipe, and air pipe inspection, system vacuum pressure check, monitoring of the internal and external environmental temperature of the lithography machine and the temperature of key components of the optical system, monitoring of the water circulation system flow and temperature, monitoring of laser energy and running time, monitoring of beam stability, and inspection of interference signal. For the mechanical structural components in the lithography machine system, regular checks, cleaning, lubrication and other operations need to be carried out to ensure the good condition of the mechanical structure and avoid serious equipment failures caused by mechanical faults.

3.2 General Maintenance

The general maintenance measures for a laser direct writing lithography machine include restarting the computer system, intensity correction, exposure testing, etc.

3.2.1 Restarting the computer system

The lithography machine performs regular restarts of the computer system. Over time, a system that operates continuously is prone to issues such as cache accumulation and system slowdown. Regularly restarting the system can free up memory cache, optimize system performance, and ensure the efficient operation of the lithography machine's control and conversion systems.

3.2.2 Intensity correction

The energy intensity distribution of the exposure pixels of the lithography machine directly affects the systematic brightness changes within the exposure line. If not corrected, this distribution can cause size deviations in the stripe structure or its boundary positions. Usually, the automatic intensity correction program in the lithography machine software is used to set and change appropriate correction parameters to complete the energy correction. Energy correction is a key step to ensure the exposure quality, and the correction curve is also an important tool for monitoring the state of the optical system.

3.2.3 Exposure testing

An important means to determine the status of the laser direct writing lithography machine is to judge the current status of the lithography machine based on the exposure results of the test graphics. Some of the structures in the exposure test graphics, as shown in Figure 1.

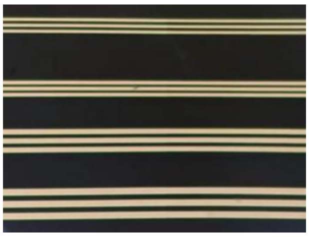


Figure 1 Different Thicknesses of Linear Arrays in Test Graphics

Test graphics generally include horizontal and vertical lines of different widths, diagonal lines of different angles and directions, circular, rectangular, triangular structures and line array, square matrices, etc. These are all important indicators for judging the current performance and parameters of the lithography machine.

3.3 Troubleshooting

Common faults of laser direct writing lithography machines include mechanical failures, electronic failures, computer system failures, and optical system failures, etc.

3.3.1 Mechanical failures

Common mechanical failures include wear of mechanical components such as leakage in intake and exhaust pipes, leakage in cooling water pipes, wear of guide rails, failure of cylinders, and exceeding of motor stroke limits, etc. The general measures for handling faults include locating the source of the fault, inspecting the damage, replacing the damaged parts, and conducting inspection and testing, etc. It should be noted that when performing maintenance tasks on the water circulation system, the water circulation booster device should be turned off, and water should be prevented from splashing onto electronic components and optical components.

3.3.2 Electronic failures

Common electronic faults include power line faults, sensor faults, communication system faults, and control system failures, etc. Modern lithography machine systems are equipped with a fault reporting system. When a certain system fails, it will notify the operator and maintenance engineers. The engineers can perform fault detection and elimination work for the specific system based on the fault information. Generally, a multimeter and an oscilloscope are used to measure the voltage and signals at the fault source. The troubleshooting measures taken include resetting the faulty system and replacing electronic components, etc.

3.3.3 Computer system failures

Common computer system failures include software vulnerabilities, insufficient memory, damaged system files, and system crashes. Engineers usually need to take measures such as upgrading software, optimizing code, restarting the system, and replacing the host to troubleshoot the faults based on the situation of the software system failure. It should be noted that during the maintenance of the software system, it is necessary to back up the system and file materials in advance to ensure that the system can be restored to its original state.

3.3.4 Optical system failures

The optical system is the most complex and precise core component of the lithography machine. Common optical system failures include significant deterioration in exposure quality, failure to adjust exposure parameters, etc. The uneven distribution of exposure energy is shown in Figure 2.

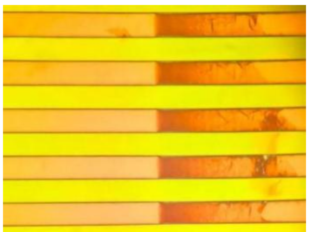


Figure 2 The Uneven Distribution of Exposure Energy

Generally, measures such as checking the light spot state, checking the optical axis position, checking the laser energy, and checking the beam stability will be taken to determine the source of the failure. The failure sources include optical lenses, pixel control components, heads, etc. After replacing the faulty components, it is usually necessary to perform optical path adjustment work. The optical path adjustment is very complex and requires combining each adjustment status with exposure testing to ultimately restore the exposure performance of the lithography machine system. The optical system maintenance service of the lithography machine is generally provided by the equipment manufacturer.

4 MANAGEMENT STRATEGIES AND METHODS

The management strategies and methods for laser direct writing lithography machines should include establishing a comprehensive maintenance plan, formulate emergency response plans and troubleshooting checklists, signing maintenance contracts with equipment manufacturers, and enhancing the ability of independent maintenance[12-13].

4.1 Establishing A Comprehensive Maintenance Plan

The laser direct writing lithography machine requires the formulation of maintenance plans of different depths. Generally, experienced equipment engineers are responsible for the daily, monthly and quarterly maintenance tasks. Daily maintenance should include monitoring the status of the lithography machine such as air pressure, temperature and humidity. Weekly plans include tasks such as restarting the system, energy calibration and exposure testing of the lithography machine. Quarterly plans should include tasks such as checking the Granite level, lubricating the mechanical structure and detecting signal strength. When encountering difficult faults, it is necessary to contact the equipment supplier promptly to seek technical support. A complete management plan for the lithography machine is the basic condition for ensuring the stable operation of the lithography machine.

4.2 Formulate Emergency Response Plans and Troubleshooting Checklists

The laser direct writing lithography machine has a complex structure system, which makes it prone to various failure phenomena in different systems. In particular, emergency plans should be established for faults in the power supply and water circulation systems to enable operators to promptly cut off the fault source and prevent the expansion of the failure. At the same time, the failures of the laser direct writing lithography machine are not all obvious. Its failures also include the deterioration of exposure quality due to abnormalities in the optical system or the platform motion system. These failures need to be judged through actual exposure tests and the exposure results of the finished products, and a long-term and effective fault elimination list should be established to help quickly solve the same type of failures, grasp the patterns of failure occurrence, and quickly restore the original performance of the lithography machine.

4.3 Signing Maintenance Contracts with Equipment Manufacturer

For many technical issues and troubleshooting tasks of the laser direct writing lithography machine, only solutions provided by professional and experienced supplier engineers can be effective. Moreover, many key components of the lithography machine can only be purchased from the equipment manufacturer. Without professional training and lacking experience, operators who attempt to manage and maintain the lithography machine on their own are highly likely to cause irreversible faults in the machine. Therefore, signing a long-term service contract with the supplier for the maintenance of the lithography machine is the best solution to ensure the stable operation of the machine.

4.4 Enhancing the Ability of Independent Maintenance.

The maintenance work of laser direct writing lithography machines is often highly dependent on the equipment supply vendors. In case the equipment vendors stop providing services, it is likely to cause the deterioration of the equipment's operating condition or even its complete shutdown. Therefore, enhancing the self-maintenance capabilities of lithography machine users can effectively ensure the long-term stable operation of the equipment. Measures to improve the self-maintenance capabilities include inviting equipment vendors to provide regular training, stocking spare parts for key components, and establishing user maintenance experience forums, etc.

5 CONCLUSION

The maintenance and management of laser direct writing lithography machines are more complex than those of ordinary precision instruments. The maintenance work involves multiple aspects such as mechanical systems, electronic systems, optical systems, and software systems. Each system has high precision and complexity, so the maintenance and guarantee capabilities of the lithography machine are more demanding. Establishing effective maintenance and management strategies and emphasizing the cultivation of highly skilled maintenance personnel are effective measures to ensure the normal operation of the lithography machine. At the same time, the performance of laser direct writing lithography machines is related to the maintenance capabilities of users. Only professional, efficient, and precise maintenance and management methods can maximize the optimal performance of the lithography machine and ultimately improve the quality and stability of enterprise products[14-15].

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Anya Grushina. Direct-write grayscale lithography. DE GRUYTER, 2019, 8(3-4): 163–169.

[2] Soichi Owa. Review of optical direct-write technology for semiconductor manufacturing. Journal of Micro/Nanopatterning, Materials, and Metrology, 2023, 22(4): 1-15.

[3] Rongming He, Jianrui Cheng，Fan Wang. Lithography Equipment. Handbook of Integrated Circuit Industry, 2023, 18(4): 1327-1359.

[4] Juan Jaramillo, Artur Zarzycki, July Galeano, et al. Performance Characterization of an xy-Stage Applied to Micrometric Laser Direct Writing Lithography. Sensors, 2017, 17(2): 278.
[5] Aleem Al Razee Tonoy, Manam Ahmed, Md Rabbi Khan. Precision Mechanical Systems In Semiconductor Lithography Equipment Design And Development. American Journal of Advanced Technology and Engineering Solutions, 2025, 1(01): 71-97.
[6] J Iskandar, J Moyne, K Subrahmanyam, et al. Predictive Maintenance in semiconductor manufacturing. IEEE 2015 26th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC), 2015: 384-389. DOI: 10.1109/ASMC.2015.7164487.
[7] L F THOMPSON. An Introduction to Lithography. Introduction to Microlithography, 1983, 219(1): 1-13.
[8] L R Harriott. Limits of lithography. In Proceedings of the IEEE, 2001, 89(3): 366-374.
[9] Zhao Yang. Discussions on Countermeasures to the Common Failure of Lithography Equipment. Journal of Shanghai Dianji University, 2009, 12(2): 166-168.
[10] Laurent Pain, Serge Tedesco, Christophe Constancias. Direct write lithography: the global solution for R&D and manufacturing. Ultimate lithography/Lithographie ultime, 2006, 83(4-9): 910-923. DOI: 10.1016/j.mee.2006.01.190.
[11] Fang Lin, Cai Jian, Ye Jialiang, et al. Modeling of Digital Micromirror Device Diffraction Effect and Its Application in Laser Direct Writing Equipment. Laser & Optoelectronics Progress, 2024, 61(15): 1-7.
[12] Cao Kewei, Wu Yiran, Zhang Junhua, et al. Current Status and Development Direction of Mask Standards. Standard Science, 2022(5): 91-95.
[13] Liu Qian, Guo Chuanfei, Wang Yongsheng, et al. New Methods and Technologies of Laser Direct Writing. China Science and Technology Achievements, 2021(12): 71-72.
[14] Shi Xiuqian, Yu Xiaoming. The Maintenance of Electrical Equipment. Technology Innovation and Application, 2020, 14(4): 109-110.
[15] Edin Begovic, Ibrahim Plancic, Sabahudin Ekinovic, et al. Laser Interferometry Measurement and Calibration For Machine Tools. 3rd Conference Maintenance 2014, 2014: 19-28.

Journal of Manufacturing Science and Mechanical Engineering
Print ISSN: 2959-9881
Online ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3018

APPLICATION OF ACCELERATED LIFE TESTING IN SOLDER JOINT LIFE PREDICTION

ZiQin Huang
Mathematics and Applied Mathematics, Xi'an Jiaotong-Liverpool University, Suzhou 215000, Jiangsu, China.
Corresponding Email: Ziqin.Huang23@student.xjtlu.edu.cn

Abstract: Accelerated life testing (ALT) is an efficient reliability evaluation method with significant application value in solder joint life prediction for electronic packaging. This paper systematically reviews the theoretical basis and practical applications of ALT in solder joint life prediction. It analyzes the main failure mechanisms such as thermomechanical fatigue, electromigration, and creep, discusses the principles and applicability of accelerated models such as Arrhenius, Coffin-Manson, and Norris-Landzberg, and elaborates on the calculation methods of acceleration factors and test design optimization considering stress types such as temperature, vibration, and humidity. Literature case studies show that ALT, by simulating long-term service conditions, can rapidly reveal the microstructural degradation characteristics of solder joints, significantly shortening the test cycle and improving prediction accuracy. The study further points out that combining data-driven methods with multi-physics coupling models is a promising future direction, effectively addressing challenges arising from multiple variable interactions and complex environments, and providing a scientific basis for the design of highly reliable electronic products.
Keywords: Accelerated life testing; Solder joint reliability; Thermomechanical fatigue; Acceleration factor; Lead-free solder

1 INTRODUCTION

1.1 Research Background

In electronic packaging, solder joints, as primary mechanical and electric interfaces between elements, directly impact entire system durability and reliability. As electronic products move in direction to higher density, small size, and higher power, solder joints face more complex thermomechanical stress conditions, leading to fatigue failure as an overriding failure mechanism. Classic life prediction methods relying on long-duration testing under normal operating conditions are time-consuming and costly. Accelerated life testing (ALT) has been widely used as an effective method to simulate long-duration service conditions through acceleration stresses like temperature cycling, mechanical vibration, or humidity, and accelerate solder joint life prediction at high rates. Early research involved life prediction and acceleration test analysis on lead-containing solders, e.g., examining fatigue behavior under temperature cycling through constitutive models andfinite element simulations [1]. As due regulations have been implemented to eliminate lead-containing solders, lead-free solders such as Sn-Ag-Cu alloys have been increasingly replacing lead-containing solders. Their failure mechanism andmicrostructure evolution during failure have been more complex, and adaptive acceleration models have been established to estimate reliability under thermal cycling conditions [2]. In recent years, scientists have made combination-based approaches ofmechanical andfinite element test analysis to develop damage accumulation-based life prediction models under high-frequency bending loads for large-area solder joints, and made validations for their realization on numerous alloys such as PbSnAg and SnSbAg [3]. Besides, strain and temperature-associated crack initiation and growth models have been emphasized for high-area solder joints, utilizing finite element approaches to estimate crack growth rates for accurate life prediction at specified loads [4]. Machine learning methods have also been utilized to handle reliability estimation under multi-parameter actions, estimating thermal cycling impacts on solder joint life based on correlation-driven neural network models and improving estimation accuracy and efficiency [5]. These developments have laid superior theoretical foundations on solder joint life prediction, and have converted estimation models into data-driven models from empirical models.

1.2 Significance of the Research

Research in solder joint life estimation has strategic importance in maximizing collective product electronic product reliability and reducing potential failure risk, particularly in situations involving high-reliability products such as automotive, space, and home electronics products. Accelerated life tests can expediently shorten product life cycles, calibrate design parameters, and reduce unexpected in-service failures, and thereby reduce economic losses and safety hazards. While classical methods remain correct, they have the tendency to relinquish inter-factor interaction. Advanced models and machine learning schemes can more aptly capture chemical composition, geometry, and thermal cycling parameter effects on life, allowing more credible life extrapolation from accelerated tests to in-service life [5]. This research also aids in standardizing lead-free solder applications, for example, in estimating thermal cycling fatigue life of Sn-3.0Ag-0.5Cu solder joints under censored Type I data processing and Weibull statistics analysis, allowing credible accelerated life factor calculation schemes [2]. In high-power electronics modules, life estimation schemes of large-area solder joints lead to material and structure optimizations, ensuring long-running stability under harsh conditions such as high temperature and strain rate through damage accumulation laws and FE simulations [3][4]. In an engineering sense, this research not only supports product competitiveness but also industry sustainability, for example, critically reflecting acceleration test role to assist in establishing more scientific reliability criteria in the aspect of electronic manufacturing.

2 THEORETICAL BASIS

2.1 Basic Principles

Accelerated life test (ALT) is one form of reliability test technique, through which products' failure process is expedited through subjecting products to stresses higher than normal in-service stresses. Its central idea depends on exploiting an exponential or power-law correlation between reaction rates of physical or chemical reaction and stress intensities to expedite long-term reliability prediction over short time period. In electronic packaging, and solder joint life prediction in specific, most central concept of ALT depends on acceleration of failure mechanism activation, i.e., through thermally activated creep, fatigue, and diffusion process, under increased temperature or cycling load, reducing test time and revealing latent defect. The classical hypothesis of ALT assumes postulation that acceleration-induced failure mode under increased condition would be same as normal condition, without invoking new mechanism action. As an example, solder joint thermo-mechanical fatigue life would be predicted based on thermal cycling tests, and data scatter would be dealt through applying statistics such as Weibull analysis [6]. Secondly, ALT concentrates on acceleration factor estimation, i.e., through utilizing Arrhenius relation to describe temperature dependency of reaction rates, or Eyring model to describe multi-stress interactions, to offer credible extrapolation of laboratory data to practical situations [7]. In solder joint applications, this concept has been generalized to consider microstructure evolution, i.e., grain growth and phase separation, to verify consistency in failure trajectories in acceleration tests through finite element modeling in order to offer model applicability [8]. Conceptual framework of ALT consists of stress selection, test plan, and data analysis stages, wherein stress intensities must be controlled to maintain acceleration effectiveness and accuracy intact, without invoking non-representative failure due to over-stress [9].

2.2 Common Acceleration Models

Acceleration models, in most instances, consisting of physical mechanism-based mathematical models, have been utilized to obtain the lifetime distribution under normal conditions from accelerated test results in solder joint life prediction. The Arrhenius model represents the simplest temperature acceleration model, under an assumption of an exponentially linear relation between temperature and failure rate, and it has been commonly applied in thermal aging tests to estimate solder joint diffusion and creep failure, extracting acceleration effect parameters in terms of activation energy [7]. Another widely used model, Coffin-Manson model, corrects for low-cycle fatigue, and it consists of power-law correspondence between plastic strain and cycle life, and it would be superior for solder joint fatigue crack growth under temperature cycling [9]. What follows furthermore is also the Norris-Landzberg model, an extension of this model, consisting of terms for temperature amplitude, frequency, and max temperature, and it leads to more precise acceleration factor calculations, particularly for thermal cycle reliability testing of lead-free solders such as Sn-Ag-Cu [6]. These acceleration models capitalize on empirical parameter calibrations in terms of, e.g., experimental data fitting or finite element analysis to allow for necessary solder joint geometry, material property, and type of stress accommodations [10]. Under multi-stress conditions, there also follows an Eyring model, which also consists of terms, such as voltage and humidity, to allow for more complete acceleration simulation [8].

2.3 Acceleration Factor and Stress Types

Acceleration factor (AF) can be described as normal condition lifetime divided by acceleration condition lifetime, correlating test data to in-service prediction, playing an indispensable role in solder joint reliability prediction. Temperature, humidity, mechanical vibration, and voltage bias stresses are most common, and temperature stress dominates, inducing thermo-mechermal fatigue and interdiffusion by thermal cycling or isothermal aging, leading to rapid solder joint failure [6]. Temperature-humidity bias tests also tend to couple humidity stress and temperature, driving ionic migration and electrochemical corrosion; Peck's model can be used to calculate AF, integratingrelative humidity and temperature power-law effects [7]. Mechanical stress, such as vibration or bending, targets fatigue crack initiation; the AF is based on an inverse power-law model, depending on stress amplitude and frequency [9]. In specific solder joint scenarios, multi-stress interactions need consideration, such as temperature-vibration combined tests to simulate complex automotive electronics environments, with the generalized Eyring model calculating the comprehensive AF [8]. Voltage stress induces electromigration in high-power applications; the AF follows Black's equation, emphasizing the roles of current density and temperature [10]. Selection of these stress types should be based on Failure Mode and Effects Analysis (FMEA), ensuring accelerated tests represent real-world conditions, while statistical validation of the AF accuracy minimizes prediction errors.

3 OVERVIEW OF ACCELERATED LIFE TESTING APPLICATIONS IN SOLDER JOINT LIFE PREDICTION

3.1 Overview of Solder Joint Failure Mechanisms

As a critical interconnect component in electronic packaging, solder joints primarily fail due to microstructural changes and damage accumulation induced by thermo-mechanical stress, leading to a decline in overall system reliability. Under thermal cycling conditions, solder joints undergo heterogeneous coarsening and thickening of the intermetallic compound (IMC) layer. Cracks typically initiate and propagate at the interface between the IMC layer and the coarsened zone, ultimately leading to fracture failure [11]. Lead-based solder exhibits a faster fatigue crack growth rate and propagates transgranularly, while lead-free solders, such as Sn-3.5Ag-X alloys, have fatigue resistance influenced by alloying elements; for example, the addition of Bi reduces ductility and shortens fatigue life [11]. Furthermore, solder thickness and dwell time at high temperatures have limited effects on thermal fatigue, but frequency and temperature amplitude significantly affect life; higher temperatures shorten life, and specific frequency ranges (e.g., 10^{-4} Hz to 10^{-3} Hz) accelerate damage [11]. Thermomechanical loading further amplifies the failure process, generating cyclic plastic strain due to coefficient of thermal expansion (CTE) mismatch, leading to fatigue and creep failure; cracks propagate along grain boundaries and concentrate at interfaces [12]. Creep mechanisms dominate deformation at high temperatures, and models like Dorn or Anand describe the effects of stress, temperature, and grain size, while fatigue models such as Coffin-Manson predict life based on plastic strain [12].

3.2 Application of Accelerated Testing Methods

Accelerated life testing for solder joint life prediction simulates long-term service conditions by intensifying environmental stress, enabling efficient reliability assessment and design optimization. In thermal cycling tests, a data-driven framework combining linear mixed effects models (LME) and artificial neural networks (ANN) analyzes limited experimental data, examining the effects of aging time, temperature, and solder composition on lead-free solder life, providing interpretable predictive results [13]. This method assumes Weibull distribution for failure time, uses a variant of the Norris-Landzberg model to generate simulated data, improves prediction accuracy, and reveals the effects of key parameters such as temperature amplitude (ΔT) and dwell time through sensitivity analysis. Furthermore, accelerated fatigue shear tests assess the impact of aging on the reliability of SAC305 solder joints, applying 16-24 MPa stress amplitude after aging at 100°C for different durations; Weibull distribution analyzes the exponential decrease in characteristic life [14]. Testing reveals that increased aging time and stress amplitude amplify irreversible work and plastic strain. Power-law model provides an equation relating lifetime and these parameters to enable generalized prediction of reliability as a function of aging and stress [14]. These methodologies offer advantages in terms of calculating uncertainties and correlating laboratory data to practical applications, such as lead-free solder alloy choice optimization in the aerospace sector.

3.3 Literature Case Studies

Literature case studies demonstrate the practical value of accelerated life testing in solder joint reliability assessment. For example, the Risk-Informed Systems Qualification (RISQ) method, combining thermal cycling and mechanical stress testing, evaluates the durability of Column Grid Array (CGA) solder joints in advanced electronic packaging [15]. One case study focused on a 337-pin ASIC device for low-Earth orbit (LEO) missions, simulating 26,816 day-night cycles over 5 years. Accelerated testing at 0-100°C with an acceleration factor of 78 predicted failure cycles using Coffin-Manson and Norris-Landzberg models, ensuring a system failure probability below 0.1%. The test showed that CTE mismatch caused fatigue failure, and solder joint height and distance to neutral point (DNP) parameters affected lifetime, validating the shear failure mode of SnPb adhesion. Another case focused on data-driven prediction of lead-free solder under thermal cycling. Using LME and ANN frameworks to analyze Auburn University experimental data, it assessed the superior performance of Innolot alloy in small-size BGAs and the minimal degradation of PBGA1156 under aging (R^2 up to 91.79%) [13]. Furthermore, an accelerated fatigue test case for SAC305 solder joints showed a 73% decrease in characteristic lifetime after 1000 hours of aging. Plastic strain was quantified through hysteresis loop analysis, and a power-law model predicted early failure (B10) [14].

4 CONCLUSION

The application of accelerated life testing (ALT) in solder joint lifetime prediction provides an efficient and scientific assessment method for electronic packaging reliability research, particularly important in the design of high-reliability electronic products. By systematically analyzing failure mechanisms such as thermo-mechanical fatigue, electromigration, and creep, Accelerated Life Testing (ALT) can simulate long-term service conditions in a shorter time, revealing the degradation patterns of solder joint microstructure, thus optimizing material selection and process design. Based on accelerated testing methods using models such as Arrhenius, Coffin-Manson, and Norris-Landzberg, accurate calculation of acceleration factors allows extrapolation of laboratory data to real-world applications, significantly shortening product development cycles and reducing testing costs. Furthermore, the integration of various testing methods, including thermal cycling, vibration, and combined stress tests, further enhances prediction accuracy, meeting the stringent requirements of industries such as automotive and aerospace.

While ALT has made significant progress in solder joint life prediction, it faces challenges such as multi-factor interactions, complex microstructure, and the representativeness of test conditions. Future research needs to incorporate data-driven methods, such as machine learning and statistical models, to address multi-variable uncertainties and develop more accurate multi-physics coupling models to improve prediction generalization. Establishing standardized test protocols and a failure database will also enhance the universality and reproducibility of ALT methods. Overall, accelerated life testing provides a solid foundation for solder joint reliability assessment, and its continued development will drive electronic packaging technology towards higher performance and longer lifespan, offering more reliable engineering solutions for the industry.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Qian Z, Liu S. On the life prediction and accelerated testing of solder joints. The International Journal of Microcircuits and Electronic Packaging, 1999, 22(4): 288-304.

[2] Mi J, Li YF, Yang YJ, et al. Thermal cycling life prediction of Sn-3.0Ag-0.5Cu solder joint using Type-I censored data. The Scientific World Journal, 2014, 2014: 807693.

[3] Lederer M, Betzwar Kotas A, Khatibi G. Lifetime modeling of solder joints based on accelerated mechanical testing and finite element analysis. Power Electronic Devices and Components, 2023, 5: 100034.

[4] Lederer M, Betzwar Kotas A, Khatibi G. A lifetime assessment and prediction method for large area solder joints. Microelectronics Reliability, 2020, 114: 1-7.

[5] Samavatian V, Fotuhi-Firuzabad M, Samavatian M, et al. Correlation-driven machine learning for accelerated reliability assessment of solder joints in electronics. Scientific Reports, 2020, 10: 14821.

[6] Che FX, Pang HL, Ong YC, et al. Assessment of acceleration models used for BGA solder joint reliability studies. Microelectronics Reliability, 2009, 49(7): 754-760.

[7] Suhir E. Accelerated life testing (ALT) in microelectronics and photonics: Its role, attributes, challenges, pitfalls, and interaction with qualification tests. Microelectronics Reliability, 2003, 43(4): 509-520.

[8] Badri SH, Lall P. Reliability modeling of the fatigue life of lead-free solder joints at different testing conditions using a modified power law model. Scientific Reports, 2023, 13(1): 2284.

[9] Tan L, Chen H, Ma B, et al. A state-of-the-art review of fatigue life prediction models for solder joint in electronics. Journal of Electronic Packaging, 2019, 141(4): 040802.

[10] Engelmaier W. Solder joint reliability, theory and applications. Soldering & Surface Mount Technology, 1991, 3(1): 14-20.

[11] Li L, Du X, Chen J, et al. Thermal fatigue failure of micro-solder joints in electronic packaging devices: A review. Materials, 2024, 17(10): 2365.

[12] Depiver JA, Mallik S, Amalu EH. Solder joint failures under thermo-mechanical loading conditions - a review. Advances in Materials and Processing Technologies, 2021.

[13] Alakayleh A, Jiang Z, Xie R. Interpretable data-driven framework for predicting the life of lead-free solder materials under thermal cycling. Journal of Intelligent Manufacturing, 2025.

[14] Al Athamneh R, Abueed M, Bani Hani D, et al. Effect of aging on SAC305 solder joints reliability in accelerated fatigue shear test. Proceedings of SMTA International, 2018.

[15] Ghaffarian R, Evans JW. Enabling more than Moore: accelerated reliability testing and risk analysis for advanced electronics packaging. Proceedings of the 2014 IEEE Accelerated Stress Testing and Reliability Conference (ASTR), 2014: 1-7.

Journal of Manufacturing Science and Mechanical Engineering
Print ISSN: 2959-9881
Online ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3019

DEEP LEARNING-BASED FAULT DIAGNOSIS AND INTELLIGENT OPERATION OF CENTRIFUGAL PUMPS: MODELS, CHALLENGES, AND PERSPECTIVES

Fan Zhang, XiuLi Wang*
Research Center of Fluid Machinery Engineering and Technology, Jiangsu University, Zhenjiang 212013, Jiangsu, China.
Corresponding Author: XiuLi Wang, Email: ujslthb@163.com

Abstract: As one of the most critical power devices in industrial systems, the operational status of centrifugal pumps directly affects system safety, reliability, and economic efficiency. To address the limitations of traditional diagnostic methods—such as reliance on manual feature extraction and poor generalization—this paper provides a comprehensive review of recent advances in deep learning-based fault diagnosis and intelligent operation and maintenance (O&M) of centrifugal pumps. It first outlines the theoretical foundations and representative deep learning models, including convolutional neural networks (CNN), recurrent neural networks (RNN/LSTM/GRU), residual networks (ResNet), graph neural networks (GCN), and Transformers, and discusses their applications in cross-condition diagnosis, remaining useful life (RUL) prediction, and intelligent O&M. Furthermore, it summarizes the progress in key enabling technologies such as multi-sensor data fusion, transfer learning, self-supervised and meta-learning, physics-informed feature alignment, and digital twins, which significantly enhance diagnostic accuracy, robustness, and generalization. Studies indicate that deep learning-based approaches outperform traditional methods in automatic feature extraction, domain adaptation, and decision optimization, thus enabling the shift from passive monitoring to proactive maintenance. Nevertheless, challenges remain regarding data scarcity and labeling difficulty, limited model interpretability and generalization, and real-time computational constraints. Future research directions include: developing few-shot and self-supervised learning to alleviate data dependency; integrating physical knowledge with deep learning to improve interpretability and trustworthiness; designing lightweight models suitable for edge deployment; and advancing digital twin-driven lifecycle management and predictive maintenance. This review provides a systematic reference and future outlook for research and industrial applications of intelligent fault diagnosis and maintenance of centrifugal pumps.
Keywords: Deep learning; Centrifugal pump fault diagnosis; Transfer learning; Multi-source data fusion; Intelligent operation and maintenance; Digital twin

1 INTRODUCTION

Centrifugal pumps are critical power equipment widely used across industrial systems, serving as essential components for fluid transportation and energy conversion in sectors such as energy, chemical engineering, manufacturing, aerospace, and petroleum extraction. Their operational status is directly related to the safety, stability, and efficiency of the entire industrial system. However, under complex operating conditions, prolonged continuous service, and the influence of external environmental factors, centrifugal pumps are prone to various failures, including bearing wear, cavitation, and seal degradation. These failures not only reduce operational efficiency but, in severe cases, may lead to unplanned shutdowns or even safety accidents. Therefore, achieving efficient and accurate fault diagnosis and intelligent maintenance of centrifugal pumps is of great theoretical and engineering significance for ensuring the safe and reliable operation of industrial systems.

Traditional fault diagnosis methods for centrifugal pumps primarily rely on expert knowledge, signal processing techniques, and feature analysis methods [1]. Typically, these approaches extract fault features through time-domain, frequency-domain, or time–frequency-domain analysis and employ machine learning algorithms for classification and identification. However, such methods face notable limitations: the feature extraction process depends heavily on manual expertise, the models often lack generalization capability, and the diagnostic accuracy is constrained. Consequently, they fail to meet the increasing demands of modern industry for intelligent and automated systems [2-3].

With the rapid advancement of artificial intelligence, particularly deep learning, data-driven intelligent fault diagnosis methods have become a major research focus [2-4]. Deep learning achieves hierarchical feature learning through multilayer neural networks, enabling the automatic extraction of high-dimensional representations directly from raw signals. This end-to-end mapping from data to fault types significantly enhances the automation level and accuracy of fault diagnosis [5-9].

In recent years, deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs) and their variants, residual networks (ResNets), graph neural networks (GNNs), and Transformer models have been widely applied to centrifugal pump fault diagnosis. These models demonstrate remarkable performance in feature extraction and classification, showing great potential in tasks such as cross-condition diagnosis, remaining useful life (RUL) prediction, and intelligent maintenance. Furthermore, the integration of multi-source information fusion and transfer learning technologies has further improved the flexibility and comprehensiveness of deep learning applications in centrifugal pump fault diagnosis, offering new perspectives for addressing practical challenges such as data scarcity and variable operating conditions.

In the field of intelligent operation and maintenance (O&M), the application of deep reinforcement learning and adaptive optimization algorithms is driving a paradigm shift from traditional "passive monitoring" to "active decision-making." By integrating real-time monitoring, condition assessment, and decision optimization, intelligent O&M systems can effectively extend equipment lifespan, reduce energy consumption, and improve operational efficiency—thereby providing crucial support for the intelligent management of industrial equipment [10].

In this context, this paper systematically reviews the fundamental theories and key methodologies of deep learning in centrifugal pump fault diagnosis. It further summarizes the fault characteristics and diagnostic strategies of various centrifugal pump types and explores recent advances in multimodal signal processing, transfer learning, and intelligent O&M. The objective is to provide a comprehensive theoretical reference and technical outlook for future research in this field.

2 DEEP LEARNING FOUNDATIONS FOR CENTRIFUGAL PUMP FAULT DIAGNOSIS

2.1 Convolutional Neural Networks and Variants

With its hierarchical architecture, deep learning technology is capable of simulating the information-processing mechanisms of the human brain, enabling automatic learning and extraction of high-level features from raw data. This capability has demonstrated significant advantages in the field of centrifugal pump fault diagnosis. Compared with traditional diagnostic methods, deep learning eliminates the need for complex manual signal processing and feature engineering. Through an end-to-end learning approach, it directly establishes complex nonlinear mappings between monitoring data and fault categories [11]. This automated feature learning ability substantially enhances both the accuracy and robustness of fault diagnosis [12].

Among various deep learning models, convolutional neural networks (CNNs) are among the most widely applied architectures in centrifugal pump fault diagnosis. The core strengths of CNNs—namely local connectivity, weight sharing, and subsampling—enable them to efficiently capture spatial correlations within signals. In centrifugal pump diagnostics, CNNs are frequently used to process time–frequency representations of vibration or acoustic signals for recognition and classification tasks [13]. Studies have shown that CNNs can simultaneously perform feature extraction and classification when dealing with high-dimensional mechanical monitoring data, thereby improving diagnostic efficiency. To further optimize model performance, researchers have introduced adaptive learning rate strategies that dynamically adjust network parameters, leading to enhanced fault classification accuracy [22].

As an advanced evolution of CNNs, the residual network (ResNet) effectively addresses the problems of gradient vanishing and model degradation encountered in deep network training through the introduction of residual learning mechanisms. ResNet has demonstrated superior performance in centrifugal pump fault diagnosis. Zheng et al. proposed a two-stage multi-channel deep learning model based on Robust-ResNet, which incorporates a step-size factor to improve model robustness and adaptability for mechanical fault detection tasks. The enhanced model achieved accuracy rates of 99.96% and 99.53% in fault detection and remaining useful life (RUL) prediction tasks, respectively—significantly outperforming other state-of-the-art methods [14]. Through its deep network architecture, ResNet is capable of learning more complex fault feature representations, offering new insights for the in-depth application of deep learning in centrifugal pump fault diagnosis [15].

2.2 Recurrent Neural Networks and Variants

Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), possess distinctive advantages in time-series data analysis, making them particularly suitable for temporal signal modeling in centrifugal pump fault diagnosis. These models can effectively capture temporal dependencies within signals, enabling precise modeling of fault evolution processes. For instance, LSTM and GRU networks have demonstrated higher accuracy than traditional RNNs in Remaining Useful Life (RUL) prediction tasks, allowing for more reliable estimation of equipment lifespan.

Furthermore, researchers have developed hybrid deep learning architectures that integrate the strengths of Convolutional Neural Networks (CNNs) and LSTMs. In such models, CNNs are employed to extract spatial features, while LSTMs are used to capture temporal characteristics. This synergistic combination significantly enhances the overall performance of centrifugal pump fault diagnosis and prognostics, improving both detection precision and predictive capability [16-17].

2.3 Transformers and Self-Attention Mechanisms

The Transformer model, which initially achieved remarkable success in the field of natural language processing (NLP), has in recent years been increasingly applied to centrifugal pump fault diagnosis. Its core innovation lies in the self-attention mechanism, which enables the dynamic capture of long-range dependencies within sequential data and demonstrates clear advantages when processing complex and non-stationary signals. Unlike Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the Transformer does not rely on fixed receptive fields or

recursive structures, allowing it to process sequence data in parallel and thereby greatly improving computational efficiency.
In centrifugal pump fault diagnosis, the Transformer effectively captures global dependencies within monitoring signals and exhibits outstanding performance under varying operating conditions and rotational speeds. When combined with physics-informed feature alignment methods, the Transformer can better adapt to distributional shifts across different operating regimes, thereby enhancing the model's generalization capability and robustness in real-world diagnostic applications [18].

2.4 Hybrid Deep Learning Models

In recent years, hybrid deep learning models have emerged as an effective strategy for enhancing the performance of centrifugal pump fault diagnosis. Such models integrate different deep learning architectures—such as Convolutional Neural Networks (CNNs) combined with Long Short-Term Memory (LSTM) networks or Transformers—to achieve complementary advantages and improve diagnostic accuracy. For instance, in centrifugal pump fault diagnosis tasks, CNNs are capable of efficiently extracting spatial features from time–frequency representations, while LSTM networks or Transformers focus on capturing the temporal dependencies of the signals. Through this synergistic design, hybrid models provide more robust diagnostic and prognostic capabilities under multi-condition and multimodal environments [19]. Notably, the careful design of hybrid architectures also strengthens model generalization, enabling greater adaptability in complex and dynamic industrial scenarios [20], see Table 1.

Table 1 Comparison of Major Deep Learning Models for Centrifugal Pump Fault Diagnosis

Model Type	Core Features	Applicable Data Types	Typical Applications	Advantages	Limitations
Convolutional Neural Network (CNN)	Local connectivity, weight sharing, spatial feature extraction	Time–frequency images, 2D signals	Fault type classification, automatic feature extraction	Automated feature learning; robust to translational variations	Difficult to directly process long sequential time-series data
Recurrent Neural Network (RNN)	Memory capability, captures temporal dependencies	Time-series data	RUL prediction, dynamic process modeling	Effective for sequential data; models dynamic processes	Susceptible to vanishing/exploding gradients; high training complexity
Transformer	Self-attention mechanism, captures long-range dependencies	Time-series, sequential data	Cross-condition diagnosis, long-sequence signal processing	High parallel computing efficiency; strong global dependency modeling	High computational resource consumption; large data requirements
Residual Network (ResNet)	Residual blocks, mitigates network degradation	Deep network structures, images/signals	Complex fault pattern recognition, deep feature learning	Enables very deep networks; alleviates gradient vanishing	Relatively large number of model parameters
Hybrid Models	Combines strengths of multiple architectures	Multimodal, spatiotemporal data	Comprehensive fault diagnosis and prediction	Simultaneously extracts spatial and temporal features; improved performance	Complex architecture; challenging training and hyperparameter tuning

In summary, the theoretical framework of deep learning–based centrifugal pump fault diagnosis has evolved from single-model approaches toward multi-architecture integration, exhibiting both contrasting and complementary characteristics, as illustrated in Table 1. CNNs excel at extracting spatial features and efficiently identifying signal structures in the time–frequency domain; Recurrent Neural Networks (RNNs), including LSTMs and Gated Recurrent Units (GRUs), emphasize temporal sequence modeling and effectively capture the dynamic evolution of faults; Transformers, empowered by self-attention mechanisms, overcome the limitations of RNNs in long-sequence modeling, offering stronger capabilities in global dependency representation and parallel processing. Hybrid models, through the collaborative fusion of CNNs, LSTMs, and Transformers, enable complementary extraction of spatial and temporal features, thereby significantly improving robustness and generalization across varying operating conditions and multimodal scenarios.
Overall, the development trend of deep learning in centrifugal pump fault diagnosis is shifting from single-structure feature learning toward multi-model collaborative intelligence [17] [19]. This transition provides a more comprehensive, efficient, and adaptive solution for fault identification and prognostics in centrifugal pump systems [21].

3 KEY TECHNICAL APPROACHES FOR CENTRIFUGAL PUMP FAULT DIAGNOSIS

3.1 Advanced Signal Preprocessing and Feature Extraction Methods

Time–frequency analysis methods serve as essential techniques in signal processing and feature extraction, playing a crucial role in centrifugal pump fault diagnosis. These methods can effectively capture time-varying characteristics in non-stationary signals by transforming one-dimensional time-domain data into two-dimensional time–frequency representations. This transformation provides deep learning models with richer feature information and significantly enhances diagnostic accuracy.

Among various time–frequency analysis techniques, the Continuous Wavelet Transform (CWT) is one of the most widely applied in centrifugal pump fault diagnosis. By employing scalable mother wavelet functions for multi-resolution analysis, CWT simultaneously provides localized information in both the time and frequency domains. In fault diagnosis, CWT is commonly used to convert raw vibration signals into time–frequency representations for subsequent image-based feature extraction [22]. This approach effectively preserves transient and periodic characteristics of the signal, making it particularly suitable for identifying complex mechanical fault patterns. Studies have demonstrated that converting vibration, pressure, and acoustic signals into two-dimensional time–frequency images and feeding them into improved deep Convolutional Neural Networks (CNNs) enables accurate identification of multiple fault types in centrifugal pumps [23]. Moreover, CWT can be combined with methods such as the Stockwell Transform (ST) to generate detailed time–frequency scale maps, where Sobel filtering enhances feature visibility and provides higher-quality inputs for downstream deep learning models [22].

The Hilbert–Huang Transform (HHT) is another powerful time–frequency analysis technique that has shown excellent performance in centrifugal pump fault diagnosis. Based on Empirical Mode Decomposition (EMD) and Hilbert spectral analysis, HHT adaptively decomposes nonlinear and non-stationary signals, making it particularly suitable for complex mechanical vibration data. In monoblock centrifugal pump fault detection, HHT has been employed to convert vibration signals into HHT-based images, which are then classified using pre-trained deep networks with high accuracy. Compared with the traditional Fourier Transform, HHT provides superior capability in capturing localized signal characteristics, offering unique advantages for detecting early-stage faults and weak signals. In drilling pump diagnostics, HHT has been applied to extract time–frequency features from denoised strain signals, which are then fed into parallel Deep Neural Networks (DNNs) and fused with other features to improve diagnostic precision [24].

Essentially, time–frequency analysis methods perform a secondary level of signal feature representation by transforming one-dimensional signals into two-dimensional time–frequency images. Taking CWT as an example, it computes wavelet coefficients across different scales and positions to generate an energy distribution map on the time–frequency plane, where the horizontal axis represents time, the vertical axis denotes frequency, and color or grayscale encodes energy intensity. This two-dimensional representation not only preserves temporal characteristics of the original signal but also reveals its frequency components and their temporal evolution. Similarly, HHT decomposes the signal into a series of Intrinsic Mode Functions (IMFs) via EMD and then performs Hilbert Transform on each IMF to obtain instantaneous frequency and amplitude, ultimately constructing a Hilbert spectrum. These time–frequency images provide deep learning models with richer and more intuitive feature representations, enabling more effective learning and discrimination between subtle fault patterns [24].

Case studies have shown that the integration of time–frequency analysis with deep learning significantly improves fault diagnosis accuracy in centrifugal pumps. For example, one study proposed a dual-scale image approach that combines convolutional autoencoders with Artificial Neural Networks (ANNs), achieving accuracies of 100%, 99.2%, and 98.8% across three datasets—substantially outperforming traditional methods [4]. Similarly, applying HHT to transform vibration signals into images and incorporating transfer learning techniques has successfully enabled accurate diagnosis of monoblock centrifugal pump faults, providing new insights for equipment condition monitoring and maintenance strategies.

In centrifugal pump fault diagnosis, feature selection and dimensionality reduction play a pivotal role in enhancing both diagnostic efficiency and accuracy. With the rapid advancement of sensor technologies and the proliferation of monitoring data, high-dimensional features often contain redundant information. Thus, extracting the most effective subset of features has become an important research focus.

Principal Component Analysis (PCA), as a classical dimensionality reduction technique, is widely used in centrifugal pump fault diagnosis. PCA performs linear transformation to convert original data into a set of uncorrelated representations, effectively extracting dominant feature components. It is often employed as an initial dimensionality reduction tool prior to deep learning model training. For instance, in the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm, PCA is first used to reduce high-dimensional features to a lower dimension before nonlinear mapping and visualization.

t-SNE, a nonlinear dimensionality reduction technique, is particularly effective for visualizing high-dimensional data. Unlike PCA, t-SNE preserves local structures within the data, ensuring that similar samples remain close in the low-dimensional space. In centrifugal pump fault diagnosis, t-SNE is widely used to visualize the feature distributions learned at different layers of deep neural networks, thereby helping to reveal the internal learning mechanisms of the models. Studies have shown that as network depth increases, features progressively evolve from mixed states to well-separated clusters, vividly illustrating the CNN's ability to automatically extract and optimize features.

Convolutional Autoencoders (CAEs) also serve as important tools for feature selection and dimensionality reduction in centrifugal pump fault diagnosis. One approach employs dual CAEs to process different types of time–frequency images, providing comprehensive and discriminative feature representations for each input modality. Through the encoder–decoder architecture, autoencoders learn the most efficient data representations while effectively removing noise and redundancy.

Feature selection and dimensionality reduction techniques significantly enhance both the efficiency and accuracy of centrifugal pump fault diagnosis. By eliminating redundant information and focusing on the most discriminative features, these techniques reduce computational costs and improve model generalization under varying operating conditions. Experimental results indicate that adopting appropriate feature selection and dimensionality reduction strategies can substantially improve diagnostic accuracy, with some approaches achieving near-perfect performance on specific datasets [4]. Looking forward, as deep learning technologies continue to advance, feature selection and dimensionality reduction will remain key areas of innovation, offering more efficient and accurate solutions for centrifugal pump fault diagnosis.

3.2 Diagnostic Enhancement via Multi-Source Data Fusion

In centrifugal pump fault diagnosis, multi-sensor data fusion techniques have been demonstrated to significantly enhance diagnostic accuracy and reliability by integrating information from diverse sensor types. Fault-related information generated during pump operation is often distributed across multiple physical signals, and single sensors are typically only capable of capturing local features, making it difficult to comprehensively reflect the operational state of the equipment. Multi-sensor data fusion effectively consolidates signals from vibration, pressure, acoustic, and current sensors, leveraging the complementary characteristics of different sensor modalities to achieve a more complete representation of fault features.

Commonly employed sensors in centrifugal pump diagnostics include accelerometers, pressure sensors, acoustic sensors, and current sensors. Accelerometers effectively capture vibration signals, exhibiting high sensitivity to mechanical faults such as bearing damage and rotor imbalance. Pressure sensors monitor variations in hydraulic systems or fluid pipelines, offering good detectability for blockages and leaks. Acoustic sensors capture sound signals generated during operation, providing unique advantages for identifying friction, cavitation, and other acoustic-related faults. Current sensors reflect variations in motor load, indirectly indicating pump operational conditions [25]. The distinct physical characteristics and complementary information provided by these sensors form the foundation for multi-source data fusion.

Multimodal feature fusion constitutes a core technique in multi-sensor data integration, aiming to combine features from different sensors or feature extraction methods into more discriminative representations. Studies have shown that fusing graphical features from indicator diagrams with Fourier descriptor features can substantially improve feature robustness. This approach first extracts features from single-modal inputs using two separate backbone networks, then employs an interactive fusion module to jointly learn from both indicator diagram and Fourier descriptor information. The fused features are subsequently used for classification, achieving an accuracy of 97.24%, which is significantly higher than using only graphical features (82.33%) or only Fourier descriptor features (94.22%).

Weighted fusion is another effective strategy for multi-sensor data integration. By assigning appropriate weights to different sensor signals, it allows for dynamic adjustment and optimization of the information. For axial piston centrifugal pumps under varying operating conditions, researchers have designed a multi-signal fusion module that dynamically allocates weights to vibration and acoustic signals, enhancing the method’s adaptability [26]. The module embeds a residual network (ResNet) within a shared feature generation framework to extract rich representations and achieved an average accuracy of 98.5% across nine transfer scenarios, demonstrating excellent cross-domain fault detection performance. Compared with single-sensor signals, weighted fusion provides richer fault information and reduces the stochastic variability of diagnostic outcomes.

A representative application of multi-sensor data fusion is the diagnosis of inlet pipeline blockage in centrifugal pumps. Research indicates that combining accelerometer, pressure, and motor current signals significantly improves the accuracy of blockage level identification. Experiments show that diagnostic models using multi-sensor combinations can achieve near-100% accuracy, far exceeding single-sensor approaches. Notably, the combination of accelerometers and current sensors achieves very high precision across all blockage levels. The study also indicates that increasing the number of sensors further enhances classification accuracy; for instance, using two accelerometers and one pressure sensor outperforms a configuration using only two accelerometers, highlighting the importance of multi-source data collection in centrifugal pump diagnostics.

For drilling pumps operating under variable-speed conditions, an innovative method combining physics-driven feature alignment with dynamic distribution adaptation has been proposed. This approach aligns signal amplitude, angular sampling frequency, and pulse phase, effectively reducing discrepancies between samples collected under different rotational speeds. Furthermore, Dynamic Distribution Adaptation (DDA) dynamically adjusts marginal and conditional distributions during domain adaptation, improving cross-domain feature matching. Experimental results demonstrate that this method outperforms existing state-of-the-art approaches in fault diagnosis under variable-speed conditions [27].

Overall, multi-sensor data fusion offers significant advantages in enhancing the accuracy and robustness of centrifugal pump fault diagnosis. By integrating data from diverse sensors, it provides a more comprehensive depiction of fault information, compensating for the limitations of individual sensors. Additionally, multi-sensor fusion enhances model adaptability to environmental variations and operational fluctuations, improving generalization performance. It also reduces the impact of noise and interference, increasing the reliability of fault feature extraction. By employing appropriate fusion strategies and weight allocation, the complementary strengths of different sensors can be fully exploited, achieving a synergistic “1+1>2” effect and significantly improving diagnostic accuracy. In summary, multi-sensor data fusion provides powerful technical support for intelligent centrifugal pump fault diagnosis and is a key approach for enhancing equipment operational reliability and safety.

3.3 Transfer Learning and Domain Adaptation for Real-World Operating Conditions

In centrifugal pump fault diagnosis, variations in operating conditions—such as changes in rotational speed, load fluctuations, and temperature or pressure variations—often lead to significant differences in the data distributions collected by sensors. Such variations can substantially degrade the performance of conventional deep learning models when applied across different operating conditions. The distributional differences induced by changing conditions are primarily reflected in the signal features; for example, vibration signal amplitude, frequency, and phase characteristics vary significantly at different rotational speeds, making models trained under one condition difficult to directly apply to new conditions.

To address the problem of cross-condition fault diagnosis, various transfer learning–based approaches have been proposed, with domain adaptation and adversarial training being two primary strategies. Domain adaptation methods improve model generalization by reducing the distribution discrepancy between the source domain (existing conditions) and the target domain (new conditions). Physics-driven feature alignment is an effective domain adaptation strategy that leverages the physical properties of the equipment to align signal features, such as amplitude, angular sampling frequency, and pulse phase, thereby reducing physical distribution differences across operating conditions. Dynamic distribution adaptation (DDA) further improves domain adaptation by dynamically adjusting the weights of marginal and conditional distributions, avoiding over-reliance on a single distribution and extracting more robust domain-invariant features [28].

Adversarial training constitutes another important transfer learning strategy for cross-condition diagnosis. Typically, adversarial-based transfer learning frameworks comprise a feature extractor, a classifier, and a domain discriminator. The domain discriminator's task is to distinguish whether features originate from the source or target domain, while the feature extractor aims to generate features that "fool" the discriminator, thereby learning domain-invariant representations. By combining transfer learning with residual networks (ResNets), researchers have proposed multi-signal fusion adversarial models that dynamically assign weights to vibration and acoustic signals, enhancing model adaptability across operating conditions. Experimental results indicate that this approach achieves an average accuracy of 98.5% across nine transfer scenarios for axial piston centrifugal pumps, demonstrating excellent cross-domain fault detection performance.

Beyond these methods, additional transfer learning strategies have been explored to improve cross-condition adaptability. Frequency-aware networks, which employ frequency-sensitive convolutional architectures, simultaneously consider time-domain and frequency-domain features, overcoming the limitations of conventional convolution operations under variable conditions. Self-supervised learning approaches, particularly those based on CutMix, have also been applied to cross-condition fault diagnosis, effectively enhancing model robustness and generalization under varying conditions and equipment variations. The overarching goal of these methods is to leverage knowledge and models from existing conditions to rapidly construct high-performance diagnostic models for new conditions, reducing reliance on large volumes of labeled data.

Feature representation learning is a critical component in cross-condition fault diagnosis. Studies indicate that appropriate signal processing and feature extraction methods can substantially enhance model adaptability to condition changes. For example, Continuous Wavelet Transform (CWT) can convert one-dimensional time-domain signals into two-dimensional time–frequency representations, better capturing the time–frequency characteristics of the signals and improving feature extraction performance under varying conditions. Furthermore, attention mechanisms can enable models to automatically focus on domain-invariant, fault-relevant features while filtering out condition-specific noise, further improving cross-condition diagnostic performance, see Table 2.

Table 2 Summary of Transfer Learning and Domain Adaptation Methods for Practical Operating Conditions

Method Category	Core Idea	Key Techniques	Primary Function
Domain Adaptation	Reduce the data distribution discrepancy between source and target domains	Physics-driven feature alignment, Dynamic Distribution Adaptation (DDA)	Align signal features under different operating conditions to improve model generalization
Adversarial Training	Learn domain-invariant features	Domain discriminator, Gradient Reversal Layer	Ensure extracted features are unaffected by changes in operating conditions, enhancing robustness
Self-Supervised Learning	Learn general representations from unlabeled data	CutMix-based pretraining, Contrastive Learning	Reduce dependence on labeled data and learn transferable, robust features
Frequency-Enhanced Network	Perform feature alignment and enhancement in the frequency domain	Frequency-aware convolutional architecture	Overcome the limitations of traditional convolution in extracting frequency-domain features and improve cross-condition performance

Extensive experimental studies have validated the effectiveness of transfer learning for cross-condition centrifugal pump fault diagnosis. For instance, under variable-speed conditions in three-cylinder drilling pumps, methods combining physics-driven feature alignment with dynamic distribution adaptation significantly outperformed traditional deep learning and other transfer learning approaches. These results confirm that transfer learning methods can

effectively enhance model generalization and adaptability.
Despite the significant progress achieved, several challenges remain. First, the design of more efficient domain adaptation strategies to handle complex distributional shifts under varying operating conditions requires further research. Second, incorporating domain knowledge to guide the transfer learning process for improved model interpretability and reliability remains an important direction. Finally, achieving effective cross-condition transfer under few-shot or zero-shot scenarios represents a core challenge for future research. Overall, transfer learning provides an effective pathway for addressing cross-condition fault diagnosis in centrifugal pump systems. By leveraging domain adaptation, adversarial training, and related techniques, it enables rapid construction of high-performance diagnostic models suitable for new operating conditions, reduces dependency on large labeled datasets, and substantially enhances model generalization and adaptability. As research continues, transfer learning is expected to play an increasingly critical role in intelligent centrifugal pump fault diagnosis.

4 FROM DIAGNOSIS TO PROGNOSIS: REMAINING USEFUL LIFE ESTIMATION AND INTELLIGENT MAINTENANCE OF CENTRIFUGAL PUMPS

4.1 Remaining Useful Life Prediction Techniques

In recent years, significant progress has been made in deep learning–based remaining useful life (RUL) prediction for centrifugal pumps, with the CNN-CBAM-Transformer parallel channel method representing a state-of-the-art approach. This method employs a dual-channel parallel architecture to extract degradation-related features from both time-domain and time–frequency-domain signals, thereby significantly improving prediction accuracy.
During signal preprocessing, the raw strain signals are first denoised using a wavelet thresholding algorithm. Specifically, a sym10 wavelet is selected for three-level wavelet decomposition, and the wavelet coefficients are estimated using a soft-threshold function before reconstructing the signal, effectively removing noise. The denoised signal is then split into two streams for the extraction of time-domain and time–frequency-domain features.
For time-domain feature extraction, twelve commonly used features—including standard deviation, root mean square, peak value, maximum value, and kurtosis—are extracted from the denoised strain signal. These features are serialized with a time-series step length of 10 to address input requirements for the Transformer model in RUL prediction. Time-domain features capture the direct variations during pump operation and serve as important indicators for assessing equipment degradation.
In the time–frequency domain, Hilbert-Huang Transform (HHT) is applied to the denoised signal. Notably, Ensemble Empirical Mode Decomposition (EEMD) replaces conventional EMD to better handle nonlinear and non-stationary signals. Time–frequency features reveal the temporal evolution of energy distribution across frequencies, which is critical for detecting early signs of centrifugal pump degradation.
Following feature extraction, the proposed CA-Transformer model employs a parallel structure to process time-domain and time–frequency features independently. One channel utilizes a CNN-CBAM architecture, while the other leverages a Transformer. This parallel design effectively mitigates information interference and enhances overall model performance.
The CNN-CBAM channel primarily extracts spatial features. In conventional CNNs, all features are often treated equally, which may lead to the inclusion of redundant or irrelevant features. To address this, a convolutional block attention module (CBAM) is incorporated to weight deep features according to their importance, enabling the model to focus on key degradation-related characteristics [29].
In contrast, the Transformer channel is dedicated to temporal feature extraction. Traditional RNNs and LSTMs perform poorly on long sequences and cannot process time series in parallel, increasing computational complexity. The Transformer, leveraging a self-attention mechanism, independently captures both long-term and short-term dependencies in sequence data, efficiently extracting relevant temporal features regardless of the distance between data points.
The Transformer model consists of four main components: positional encoding, encoder, decoder, and fully connected layers. For RUL prediction, only positional encoding and the encoder are required. Positional encoding captures intrinsic sequential information, while the multi-head attention mechanism extracts temporal dependencies [30].
After feature extraction and sequence modeling, the deep features independently learned by the two channels are fused through a merging layer and passed to fully connected layers, where the mean squared error (MSE) is used to evaluate discrepancies between predicted and actual RUL values. Model training is optimized using backpropagation combined with the Adam optimizer.
During training, RUL percentages are used as regression labels. For example, if a drilling pump has a total lifespan of 240,000 seconds and has operated for 192,000 seconds, its normalized RUL is 0.24. This normalization improves prediction precision.
The method has demonstrated excellent performance in practice. Validation using operational data from four drilling pumps shows that this approach outperforms several state-of-the-art methods, providing reliable support for safe pump operation and cost reduction.
For pumps operating under different conditions, transfer learning–based improvements have also been proposed. Frequency standardization via resampling, phase identification using short-term autocorrelation, and segmentation into uniform-phase intervals effectively reduce distribution differences between source and target domains, addressing the

generalization limitations of single models across multiple operating conditions.

In internal gear pumps, a two-stage multi-channel method based on Robust-ResNet has been proposed for RUL prediction. This method improves prediction performance by integrating pressure pulsation and vibration signals. The first stage identifies the pump's fault type, while the second stage predicts RUL, offering a novel framework for health assessment in complex systems.

Furthermore, combining signal demodulation with deep learning has yielded promising results in centrifugal pump monitoring. By extracting characteristic frequencies of modulation components and inputting the preprocessed data into an integrated MBConv-based deep learning model, cavitation state recognition rates of 89.44% have been achieved [31]. This approach provides a new pathway for early fault detection and RUL prediction.

In summary, deep learning–based RUL prediction methods for centrifugal pumps leverage multiple signal processing techniques and neural network models to effectively extract key degradation-related features from both time-domain and time–frequency-domain signals. The CNN-CBAM-Transformer parallel channel method, with its dual-channel architecture, significantly enhances prediction accuracy and represents the forefront of technological development in this field, offering critical support for intelligent pump maintenance and operational decision-making.

4.2 Intelligent Maintenance Decision-Making and Optimization

In the field of intelligent operation and maintenance (O&M) and optimization decision-making for centrifugal pump systems, deep reinforcement learning (DRL) has demonstrated significant potential. Traditional O&M approaches often rely on fixed rules and human expertise, which are insufficient to cope with complex and dynamic operating conditions. In contrast, DRL–based intelligent O&M systems can autonomously learn optimal strategies through environment interaction and trial-and-error learning, enabling adaptive control and optimization of centrifugal pump operations.

The core advantage of DRL in intelligent O&M lies in its ability to construct self-optimizing control systems. These systems interact continuously with the pump operating environment, learning the mapping between state-action pairs and long-term rewards to gradually develop efficient decision-making policies.

The development of an intelligent O&M system typically involves several key steps. First, the state space, action space, and reward function must be defined to form a complete reinforcement learning framework. The state space should include operational parameters of the pump (e.g., pressure, flow rate, temperature, vibration), historical states, and environmental factors. The action space encompasses potential maintenance operations (e.g., start/stop control, parameter adjustments, maintenance scheduling). The reward function should integrate multiple objectives, such as efficiency improvement, energy consumption reduction, and extended equipment lifespan. A carefully designed reward mechanism enables the system to balance short-term gains with long-term benefits, thereby learning optimal operational strategies.

From a model architecture perspective, deep neural networks are widely used for value function approximation and policy optimization in DRL. Studies have shown that adaptive convolutional neural networks combined with Bayesian optimization (CNN-BO) can effectively handle complex pump operating conditions. By autonomously optimizing network hyperparameters, CNN-BO significantly enhances fault diagnosis accuracy and robustness. Similarly, semi-supervised graph learning models leverage large amounts of unlabeled data for pretraining and fine-tune on limited labeled samples, reducing dependence on annotated data while providing higher interpretability, which is crucial for engineering O&M decisions [32].

The trial-and-error learning process of intelligent O&M systems typically integrates offline training with online optimization. In the offline phase, the system is pretrained on historical operational data to learn basic maintenance patterns. In the online phase, real-time states guide strategy adjustments, with experience replay mechanisms continually refining the decision network. For example, in submersible pump performance prediction, researchers successfully captured the nonlinear relationships between input parameters—such as flow rate, viscosity, and gas–liquid ratio—and pump output performance (e.g., pressure differential) using machine learning, enabling accurate performance predictions under complex conditions [33-34].

Application of intelligent O&M systems brings substantial economic benefits and safety improvements. In terms of efficiency, the system dynamically adjusts operational parameters in real time, ensuring pumps operate within optimal efficiency ranges. In cost reduction, predictive maintenance and precise control minimize unnecessary energy consumption and maintenance expenses. Regarding safety, the system can proactively identify potential faults and implement preventive measures, effectively preventing severe equipment damage and production interruptions. Studies indicate that deep learning–based fault diagnosis and performance optimization methods, when applied to single centrifugal pumps, significantly reduce fault handling time and resource consumption through automated detection and diagnostic workflows.

Future intelligent O&M systems are expected to evolve toward greater autonomy, collaboration, and interpretability. Autonomy refers to the system's capability to independently complete monitoring, diagnosis, and decision-making. Collaboration involves information sharing and coordinated strategies among multiple pumps, achieving system-level optimization. Interpretability enhances operator trust and understanding of intelligent decisions, ensuring reliability in critical applications. With continued advancements in DRL, intelligent O&M and optimization decision-making for pump systems are poised to achieve broader and more impactful applications.

5 CHALLENGES AND LIMITATIONS IN CURRENT RESEARCH

5.1 Data-Related Challenges

In the field of industrial centrifugal pump fault diagnosis, deep learning (DL) methods have been increasingly applied; however, they commonly face significant challenges related to data dependency and annotation difficulty. Traditional DL models typically require large volumes of labeled data to achieve accurate fault identification and classification. In practical industrial settings, obtaining sufficient fault samples is both challenging and costly. This is especially true for critical components such as centrifugal pumps, where faults are often hidden or sudden, making sample acquisition more complex. Such data scarcity severely limits the practical application of DL models in real-world industrial environments.

Several factors contribute to the difficulty of obtaining labeled data. First, the enclosed structure of centrifugal pumps and the unpredictability of faults make direct observation of fault features challenging, hindering the acquisition of high-quality labeled samples. Second, in real production environments, fault samples are often unevenly distributed, with some fault types represented by very few instances, making it difficult to meet the balanced data requirements of traditional machine learning. For example, in a well field of the Daqing Oilfield, 5,053 sets of fault diagnosis data for pumping units were collected; after data cleaning, only 3,502 valid samples remained, with significant differences in the number of samples across fault types. Additionally, the accumulation of fault data generally requires a long time, further exacerbating the difficulty of data acquisition.

Another critical challenge arises from data distribution differences under variable operating conditions. In practice, centrifugal pumps often operate under different working conditions, such as load fluctuations or changes in fluid properties, leading to substantial distribution shifts in the collected data. These shifts make it difficult to directly apply models trained under specific conditions to other conditions, often necessitating the collection of new labeled data for each operating scenario, which substantially increases the data demand. Studies indicate that under variable operating conditions, distribution differences between training and test data can significantly degrade the performance of existing fault diagnosis models, particularly when only limited samples are available.

The inherent data sensitivity of DL models further exacerbates this issue. Although deep neural networks, such as feedforward networks, can effectively capture complex nonlinear relationships, their performance is highly sensitive to training data volume and parameter settings. Even minor parameter variations can lead to substantial fluctuations in model performance, necessitating more labeled data to ensure stability and generalization. Furthermore, DL models typically require large datasets to avoid overfitting, which, if unaddressed, may result in poor generalization and potential misdiagnosis, posing safety risks [44].

To address these challenges, various strategies have been proposed. Semi-supervised learning leverages a large amount of unlabeled data along with limited labeled samples, effectively mitigating the scarcity of labeled data. For instance, data augmentation–based consistency regularization in semi-supervised learning generates new samples consistent with the original feature distribution, expanding the labeled feature space across different operating conditions. Transfer learning strategies transfer rich labeled knowledge from a source domain to a target domain, alleviating insufficient labeled data in the latter. Few-shot learning and meta-learning techniques aim to achieve effective fault diagnosis with minimal labeled samples, showing significant potential in industrial scenarios where data are extremely limited.

These approaches offer promising solutions to the challenges of data dependency and annotation in centrifugal pump fault diagnosis. Nevertheless, practical industrial implementation still faces many obstacles. For example, improving the robustness of few-shot learning and optimizing the generalization capability of transfer learning models under dynamic operating conditions remain open research questions. Therefore, while these methods provide strong support for DL-based centrifugal pump fault diagnosis, further technical advancements are required to enable large-scale deployment in real-world industrial applications.

5.2 Model-Related Challenges

In the field of centrifugal pump fault diagnosis, although deep learning (DL) models often perform well under the training data distribution corresponding to specific operating conditions, their performance frequently deteriorates when faced with new working conditions, equipment variations, or environmental changes. This limited generalization capability poses a significant constraint on the reliability of DL techniques in practical industrial applications.

Studies indicate that real-time adjustments in centrifugal pump operating conditions, particularly for mud pumps used in drilling operations, require continuous tuning of operational parameters according to drilling depth. Such variations make it challenging for conventional DL methods to maintain stable diagnostic performance across different conditions. Similarly, when centrifugal pumps operate under varying temperature conditions, vibration signal characteristics can change, leading to fluctuations in the accuracy of cavitation state recognition. For rod pumping systems, performance diagrams from different oilfields are influenced by geographical conditions, sensor devices, and acquisition software, exhibiting environment-specific characteristics that single-model diagnostic approaches struggle to accommodate.

In centrifugal pump fault diagnosis, fluctuations in working conditions inevitably introduce distribution differences between training and testing data. This issue is especially pronounced in scenarios with limited sample sizes, where distribution shifts can lead to substantial declines in model performance. For instance, when a model trained under one operating condition is applied to another, significant differences in fault features between conditions can cause a sharp drop in diagnostic accuracy.

To address these challenges, several approaches have been proposed to enhance model generalization. Transfer learning has been shown to be an effective solution, particularly when similar datasets are unavailable or when data must be transferred across domains. By transferring network parameters, models can maintain robust generalization even after extensive adaptation to the target task.

Feature fusion is another strategy for improving generalization. Studies demonstrate that combining Fourier descriptor features with graphical features from performance diagrams enhances feature robustness and improves diagnostic accuracy. For example, the diagnostic accuracy of a fused-feature model increased from 82.33% to 97.24%, significantly outperforming models using individual feature sets. This approach leverages the complementary relationships among features, effectively enhancing the generalization capability of the diagnostic model.

Data augmentation techniques are also widely employed to improve model robustness under variable conditions. By generating diverse training samples through transformations such as rotation, translation, shear, scaling, and flipping, models can better adapt to different operating scenarios. For example, in centrifugal pump cavitation state recognition, augmenting data with rotations of ±20°, translations of ±20%, and shearing of ±20% substantially enhanced the model's ability to generalize across varying conditions.

Furthermore, meta-learning methods have been applied to address generalization under limited-sample scenarios. By embedding prior knowledge into meta-learning strategies, models can simulate diverse operating conditions and adaptively capture domain-invariant features. For instance, segmenting time-frequency images into grids and utilizing positional information to construct self-supervised loss functions enables the model to learn cross-domain invariant representations, thereby improving generalization across different working conditions.

Despite these advances, centrifugal pump fault diagnosis in practical industrial applications still faces significant challenges. Key research directions include further enhancing model adaptability to unknown conditions, reducing reliance on extensive labeled datasets, and improving model stability in complex and variable environments.

5.3 System-Level Challenges

Deep learning (DL) models have demonstrated significant potential in centrifugal pump fault diagnosis; however, their practical application still faces two major challenges: insufficient interpretability and limited real-time performance. These limitations substantially hinder the widespread adoption of DL techniques in industrial environments and affect engineers' trust in these methods.

Insufficient interpretability is the primary issue in applying DL to fault diagnosis. Since DL models are typically regarded as "black boxes," their internal decision-making processes are difficult for humans to understand and explain. This lack of transparency makes it challenging for engineers to trust and rely on model predictions, particularly in critical equipment monitoring scenarios. Studies have highlighted that conventional DL methods cannot establish a direct mapping between raw sensor data and corresponding fault modes, making diagnostic performance heavily dependent on the quality of feature extraction. When misdiagnoses occur, engineers are unable to trace the decision rationale, which impedes effective fault analysis and system improvement. Moreover, neural network architecture design and parameter optimization often require extensive manual tuning and expert knowledge, further restricting model applicability and generalization.

Limited real-time performance constitutes another critical barrier to industrial deployment. Complex DL models typically incur significant computational overhead, making it difficult to meet the real-time diagnostic requirements of industrial operations. In centrifugal pump systems, faults must often be detected and identified at an early stage to prevent severe consequences. However, existing DL algorithms frequently underperform under variable-speed and other complex operating conditions, with low computational efficiency. Additionally, training such models usually requires considerable time and computational resources, which is particularly challenging in resource-constrained industrial environments.

To address these challenges, several strategies have been proposed. One approach introduces an interpretable semi-supervised graph learning model that incorporates a feature reconstruction module. By fitting and explaining the learned features with nonlinear surrogate models, this method enhances interpretability during training and accelerates model convergence. Another approach combines DL with reinforcement learning to construct an end-to-end fault diagnosis framework, directly mapping raw fault data to corresponding fault modes. Methods integrating physics-driven feature alignment with dynamic distribution adaptation have also been developed, leveraging physical knowledge to improve model performance and enhance cross-operating-condition adaptability. Frequency-enhanced networks, which jointly consider time-domain and frequency-domain features and incorporate CutMix-based self-supervised learning, have demonstrated robust generalization across varying operating conditions. Transfer learning has been effectively applied to mitigate issues of insufficient labeled data and imbalanced data distributions, improving model performance under data-constrained scenarios.

Despite these advances, further research is needed to develop models that maintain high diagnostic accuracy while simultaneously improving interpretability and computational efficiency. In particular, designing transparent yet efficient DL models for industrial deployment is critical to promoting the broader adoption of intelligent fault diagnosis technologies in centrifugal pump systems.

6 FUTURE RESEARCH DIRECTIONS AND PERSPECTIVES

6.1 Few-Shot and Self-Supervised Learning

Few-shot learning (FSL) and self-supervised learning (SSL) have demonstrated significant potential in centrifugal pump fault diagnosis, providing effective solutions to the data scarcity problem commonly encountered in industrial settings. In this domain, obtaining large-scale labeled datasets is often costly and challenging. FSL and SSL can leverage limited labeled data alongside abundant unlabeled data to enhance model diagnostic performance and generalization ability [35].

In few-shot learning, model-agnostic meta-learning (MAML) algorithms offer a novel approach for centrifugal pump fault diagnosis. Through an inner-loop and outer-loop gradient update process, the model can rapidly adapt to new tasks using only a small number of labeled samples [36]. Specifically, tasks are divided into a support set and a query set. The support set, typically containing 1–5 samples, is used to quickly adjust model parameters, while the query set evaluates model performance. This strategy enables the model to learn generalizable feature representations applicable across varying operating conditions, effectively addressing the performance degradation caused by distribution shifts in centrifugal pump data under different working conditions [37].

Self-supervised learning exhibits unique advantages in this field. Unlike traditional supervised methods, SSL can learn meaningful feature representations from unlabeled data, reducing reliance on annotated samples. One effective approach involves feature-level differential updates in graph convolutional networks (GCNs), ensuring that the feature extractor preserves maximal fault-related information. Specifically, constructed graph data are fed into a GCN model, where neighbor information is aggregated and node features are iteratively updated to produce nonlinear feature matrices. Feature-level loss is then computed to update model parameters. This SSL pretraining process allows the model to capture intrinsic data structures and relationships, providing high-quality feature representations for subsequent fault diagnosis tasks.

Data augmentation techniques play a critical role in mitigating data scarcity. Considering that centrifugal pump monitoring data are typically 1D time-series signals, improved symplectic geometric data augmentation (ISGDA) methods generate new samples with feature distributions similar to the original data. By introducing controlled perturbations to time-series signals, these methods enrich the feature space of labeled samples, improving diagnostic performance under limited data conditions. ISGDA also effectively suppresses overfitting during training and enhances model robustness. Furthermore, the introduction of consistency regularization further improves model performance. The supervised loss function ensures consistency between augmented labeled samples and their true labels, while the unsupervised loss reduces distribution discrepancies between augmented and unlabeled samples [32].

Meta-learning frameworks demonstrate substantial advantages in few-shot scenarios. A prior-knowledge-embedded meta-learning vision transformer (PKMLVIT) represents an innovative approach for few-shot fault diagnosis. This method integrates wavelet transform, meta-learning, self-supervised learning, and vision transformer architectures to extract domain-invariant features transferable across varying operating conditions. Initially, a modified ViT-based feature extractor captures global–local fused features from time–frequency spectrograms. Subsequently, a meta-learning strategy incorporating prior knowledge simulates the generalization scenario of centrifugal pump devices under changing conditions, adaptively learning domain-invariant representations. Experimental results indicate that PKMLVIT achieves superior robustness and accuracy under limited samples and variable operating conditions, outperforming existing approaches [42].

Despite the potential of FSL and SSL in centrifugal pump fault diagnosis, several challenges remain. First, the quality of pseudo-labels directly affects model performance, and pseudo-label generation may be influenced by model errors, noise, bias, or outliers in the data. Second, distribution shifts across different operating conditions can degrade model performance, particularly when sample sizes are limited. Finally, careful tuning of model parameters and regularization strategies is essential to prevent overfitting, requiring a balance between model complexity and generalization ability, see Table 3.

Table 3 Correspondence between Challenges and Future Directions in Intelligent Fault Diagnosis of Centrifugal Pumps

Challenge Category	Core Challenges	Corresponding Future Research Directions
Data Level	Data scarcity and labeling difficulties; distribution shifts under varying operating conditions	Few-shot learning, self-supervised learning, advanced data augmentation techniques
Model Level	Limited generalization capability; insufficient interpretability	Integration of physical mechanisms with deep learning, explainable AI (XAI), meta-learning
System Level	Limited real-time performance; high computational resource demands; low system integration	Lightweight models, edge computing, digital twin-based full lifecycle management

In the future, few-shot learning (FSL) and self-supervised learning (SSL) will continue to offer broad research opportunities in the field of centrifugal pump fault diagnosis. On one hand, more advanced data augmentation techniques can be explored, particularly those tailored for 1D time-series signals, to generate more representative samples. On the other hand, improvements and optimizations in meta-learning algorithms will further enhance model performance in few-shot scenarios, for example, by incorporating additional prior knowledge or designing more effective task-generation strategies. Moreover, integrating self-supervised learning with explainable methods can improve model transparency and reliability, providing more valuable support for industrial decision-making.

In summary, FSL and SSL provide effective solutions to the data scarcity problem in centrifugal pump fault diagnosis. By leveraging limited labeled data alongside abundant unlabeled data, and combining advanced data augmentation techniques with meta-learning frameworks, more efficient and robust diagnostic models can be developed. The integration of these technologies offers strong support for health monitoring and maintenance of industrial centrifugal pump systems, thereby advancing the development of intelligent operation and maintenance systems.

6.2 Integration of Physical Mechanisms and Deep Learning

The integration of physical information with deep learning represents a cutting-edge research direction in centrifugal pump fault diagnosis. By combining physical models and mechanistic knowledge with deep learning, the model's performance under extreme conditions and its interpretability can be significantly enhanced. This fusion approach not only leverages the powerful feature extraction capabilities of deep learning but also utilizes physical knowledge to provide prior constraints, ensuring that the model outputs adhere more closely to actual physical laws.

Physical-driven feature extraction is a key aspect of integrating physical information with deep learning. Studies have shown that converting raw signals into time-frequency images using continuous wavelet transform can effectively capture the time-frequency characteristics of centrifugal pump faults. This method simultaneously captures variations in both time and frequency domains, providing richer input features for deep learning models [38]. Moreover, for fault diagnosis under varying operating conditions, physical-driven feature alignment methods adjust signal amplitude, angular sampling frequency, and pulse phase to reduce discrepancies between samples from different conditions. This physically-informed feature alignment enables deep learning models to better adapt to fault diagnosis tasks under variable operating conditions.

Physics-informed neural networks (PINNs) exemplify the deep integration of physical models with deep learning. In centrifugal pump fault diagnosis, researchers have developed frequency-enhanced networks that employ frequency-aware convolutional architectures to consider both time-domain and frequency-domain features, overcoming limitations of traditional convolution operations in frequency feature extraction. Additionally, adaptive convolutional neural network models constructed based on the intrinsic knowledge of centrifugal pumps can automatically optimize critical hyperparameters through Bayesian optimization. Experimental results indicate that such physics-guided adaptive models achieve a maximum accuracy of 99.78%, representing a 5.45% improvement over the traditional LeNet-5 model.

Multi-source information fusion is another effective approach for integrating physical knowledge with deep learning. By fusing the diagnostic outputs of multiple single models, the accuracy and robustness of fault diagnosis can be substantially improved. Studies show that after combining multiple deep learning models, the final diagnostic accuracy can reach 99.98%, which is 9.09% higher than the average accuracy of individual models. Introducing reliable evidence during the fusion process can further enhance performance. For instance, fusing the 1-DPCA-AE model with other models (e.g., Models 1, 4, 8, 9) increased accuracy by 13.53%, while combining the AE model with Models 2 and 4 improved accuracy by 10.45%. Such multi-source information fusion methods significantly enhance the recognition accuracy of diaphragm pump fault types [39].

To improve generalization across varying operating conditions, researchers have developed hybrid methods incorporating dynamic distribution adaptation. These methods dynamically adjust marginal and conditional distributions during domain adaptation, improving cross-domain feature matching. Additionally, CutMix-based self-supervised learning achieves robust generalization under different operating conditions and equipment variations. These approaches perform well in diagnosing pump faults across varying temperature, pressure, and speed conditions.

Regarding interpretability, integrating physical information with deep learning substantially enhances model transparency. Visualization techniques such as t-SNE can reveal the feature learning process at different CNN layers, improving understanding of the model's decision-making mechanisms. Furthermore, health indicators developed using Mahalanobis distance and Fisher discriminant ratio contribute to stabilizing high-dimensional latent representations during model training. Such interpretability methods provide more transparent and reliable support for fault diagnosis, increasing confidence in diagnostic outcomes.

In the future, research on the integration of physical information and deep learning is expected to deepen further, particularly in the following areas: first, developing more refined physical models to more accurately embed centrifugal pump operating mechanisms into deep learning frameworks; second, exploring more effective multi-source information fusion strategies to further improve diagnostic accuracy; third, investigating adaptive optimization methods to enhance model performance under complex and variable operating conditions; and finally, strengthening research on model interpretability to improve transparency and reliability of diagnostic results. These research directions will advance centrifugal pump fault diagnosis toward more intelligent and precise methodologies.

6.3 Explainable AI and Trustworthy Fault Diagnosis

In the field of industrial pump fault diagnosis, while deep learning techniques have demonstrated remarkable performance, their "black-box" nature and high computational demands limit practical deployment in industrial environments. The integration of explainable AI (XAI) and edge computing offers a promising solution to these challenges.

The primary goal of explainable AI is to render the decision-making process of deep learning models transparent, thereby enhancing user trust in diagnostic results. Traditional deep learning models in fault diagnosis face two major challenges: reliance on large amounts of labeled data and lack of interpretability. To address these issues, researchers have proposed various explain ability methods, including Class Activation Mapping (CAM), Gradient-weighted CAM (Grad-CAM), Shapley Additive Explanations (SHAP), and Local Interpretable Model-agnostic Explanations (LIME). Specifically, LIME generates locally interpretable models to explain how predictions are made, rendering the decision process of complex models more transparent. An innovative approach involves incorporating an interpretable feature reconstruction module, which uses nonlinear surrogate models to fit and interpret learned features and embeds interpretability scores into the feature representations. This method not only enhances model transparency during training but also accelerates convergence.

On the other hand, edge computing addresses the high computational resource demands of deep learning models. By deploying data, applications, and computing capabilities closer to the data source, edge computing significantly reduces latency and meets industrial requirements for real-time processing. In centrifugal pump fault diagnosis, containerized edge AI inference frameworks have shown excellent performance by deploying Docker container services near sensors, providing an efficient and low-latency data analysis pipeline.

To enable real-time diagnosis on edge devices, researchers have proposed model compression and lightweight design techniques. Various lightweight model architectures, such as CNNs tailored for rolling bearing fault diagnosis, have been developed. For edge deployment, Docker container technology is widely used to encapsulate and deploy AI models, allowing them to run on resource-constrained devices. Studies indicate that utilizing different container runtimes (e.g., CRI-O, Docker, Containerd) across platforms (e.g., x64 and ARM) optimizes resource utilization and improves computational efficiency on edge devices [40].

Looking forward, integrating physical knowledge with deep learning models offers an effective way to enhance both model performance and interpretability. For example, employing wavelet packet decomposition and information entropy-based feature extraction to capture physical characteristics of signals, combined with CNN models incorporating attention mechanisms, can significantly improve diagnostic accuracy. Furthermore, physics-driven feature alignment methods that adjust signal amplitude, angular sampling frequency, and pulse phase can reduce inter-sample discrepancies and enhance model generalization across varying operating conditions [41].

In summary, the combination of explainable AI and edge computing is poised to play an increasingly important role in centrifugal pump fault diagnosis. By developing transparent deep learning models and efficient edge deployment strategies, it is possible to improve the reliability of diagnostic outcomes while satisfying the stringent real-time requirements of industrial settings, thereby promoting the widespread adoption of intelligent fault diagnosis technologies.

6.4 Digital Twin and Full-Life-Cycle Intelligent Management

Digital twin technology provides a revolutionary solution for the full lifecycle management of centrifugal pump equipment. By creating a virtual replica of the physical pump and integrating deep learning algorithms, digital twin systems enable a seamless transition from fault diagnosis to remaining useful life (RUL) prediction, laying a solid foundation for predictive maintenance. These models can reflect the real-time operational status of centrifugal pumps, and through multi-source data fusion and deep analysis, provide scientific support for maintenance decision-making.

In constructing digital twins for centrifugal pumps, deep learning models have demonstrated outstanding performance. Studies indicate that neural network models outperform traditional polynomial fitting and mechanical modeling methods in predicting submersible pump performance, especially under complex flow conditions. Deep learning models can accurately capture nonlinear relationships and overcome the limitations of traditional approaches in complex operating environments. To address overfitting during neural network training, researchers have employed early stopping strategies, effectively avoiding non-physical pump performance curves caused by excessive training. Moreover, the choice of activation function significantly affects prediction accuracy and must be optimized for specific applications [33].

In volumetric pump wear-state classification, deep learning methods have also shown strong performance. Research demonstrates that a neural network with 5 input neurons, 12 hidden neurons, and 3 output neurons can effectively classify three pump states: normal operation, end-of-life, and wear. By precisely adjusting the number of hidden neurons, based on empirical formulas and key parameters such as the number of input neurons and constants, classification performance is significantly enhanced [43].

Significant breakthroughs have been achieved in RUL prediction using deep learning. For example, a parallel-channel method combining convolutional neural networks (CNNs), convolutional block attention modules (CBAM), and Transformer networks has been proposed for drilling pump RUL prediction [44]. The model extracts both time-domain and time-frequency features from strain signals via two parallel channels, followed by feature fusion to predict RUL accurately. Experimental results indicate that deep learning models consistently outperform traditional machine learning approaches for monitoring-based RUL prediction. In particular, models incorporating CBAM modules achieve high accuracy across all drilling pump evaluation metrics, demonstrating their effectiveness in predicting RUL.

Vibration-based intelligent fault diagnosis is a crucial component of digital twin systems. A proposed deep hybrid model considers frequency, time, and spectral information from vibration signals, including spectrograms obtained via short-time Fourier transform and scalograms derived from continuous wavelet transform [13]. Experimental evaluations show that this deep hybrid model significantly outperforms conventional machine learning methods, such as k-nearest neighbors, support vector machines, logistic regression, and random forests, enabling automatic detection of submersible pump faults from operational vibration data [33].

Predictive maintenance strategies rely on accurate fault diagnosis and RUL prediction provided by digital twin systems. Accurate RUL predictions guide pump operation and maintenance, reducing unplanned downtime and production interruptions caused by improper maintenance, thereby improving overall operational efficiency [45]. Additionally, precise fault forecasting enhances safety measures, minimizing risks such as blowouts or worker injuries caused by pump failures. By optimizing test matrices to balance prediction accuracy with testing cost, the economic benefits of predictive maintenance can be further improved.

Future research should focus on the real-time updating and adaptive capability of digital twin models, particularly in the effective fusion of multi-source heterogeneous data. Probabilistic prediction methods, such as Gaussian process regression, offer advantages in assessing model uncertainty but face high computational costs with large-scale datasets, necessitating algorithmic efficiency improvements [46]. Furthermore, integrating digital twin systems with the Internet of Things (IoT) and edge computing will enhance the real-time accuracy of predictive maintenance, supporting a shift from reactive maintenance to proactive prevention.

7 CONCLUSION

This paper systematically reviews the current research, key technologies, and future development trends of deep learning in centrifugal pump fault diagnosis and intelligent operation and maintenance (O&M). Studies indicate that deep learning, with its powerful automatic feature extraction and nonlinear modeling capabilities, provides an effective solution to the limitations of traditional methods—such as strong feature dependence, limited diagnostic accuracy, and poor adaptability across operating conditions—thereby significantly advancing both theoretical research and engineering applications for intelligent pump management.

At the model and methodology level, deep architectures—including convolutional neural networks (CNNs), recurrent neural networks (RNNs) and their variants, residual networks, graph neural networks (GNNs), and Transformers—demonstrate notable advantages in feature extraction, temporal modeling, and complex signal recognition. Techniques such as transfer learning, adversarial training, and dynamic distribution adaptation effectively mitigate cross-condition data distribution differences, enhancing model generalization. Multi-source information fusion strategies integrate vibration, acoustic, pressure, and current signals to achieve high-precision sensing and robust diagnostics of pump operating states. Furthermore, combining deep reinforcement learning with remaining useful life (RUL) prediction models enables centrifugal pump systems to evolve from "reactive maintenance" toward "proactive operation," significantly improving operational safety and economic efficiency.

However, several challenges remain. First, high-quality labeled data in industrial scenarios are scarce, leading to strong model dependency and pronounced sample imbalance issues. Second, the lack of interpretability and limited cross-condition generalization of deep learning models constrain their reliable application in complex environments. Additionally, high model complexity and challenges related to real-time performance and edge deployment remain key bottlenecks for practical intelligent O&M implementation.

Future research should focus on the following directions:

1.Few-shot and self-supervised learning: Develop methods to alleviate data scarcity and reduce dependence on labeled samples.

2.Integration of physical mechanisms with deep learning: Construct physically constrained, interpretable models to enhance transparency and trustworthiness of diagnostic results.

3.Multi-modal fusion and multi-task coordination: Achieve integrated intelligent management combining fault diagnosis, health assessment, and RUL prediction.

4.Reinforcement learning-based adaptive O&M strategies: Enable systems to transition from "monitoring and identification" toward "autonomous decision-making."

5.Edge computing and digital twin integration: Implement real-time diagnosis and full lifecycle management of pump equipment.

In summary, the introduction of deep learning provides a new theoretical framework and technical support for centrifugal pump fault diagnosis and intelligent O&M. Its deep integration with physical modeling, industrial IoT, and intelligent decision-making technologies will be a key driving force for advancing industrial equipment toward intelligent, autonomous, and trustworthy operation.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Guo J, Wan J L, Yang Y, et al. A deep feature learning method for remaining useful life prediction of drilling pumps. Energy, 2023, 282: 128442.

[2] Nan L, Wang Y, Chen D, et al. A novel energy performance prediction approach towards parametric modeling of a centrifugal pump in the design process. Water, 2023, 15(10): 1951.

[3] Siddique M F, Ullah S, Kim J M. A deep learning approach for fault diagnosis in centrifugal pumps through wavelet coherent analysis and S-transform scalograms with CNN-KAN. Computers, Materials & Continua, 2025, 84(2).
[4] Zaman W, Ahmad Z, Kim J M. Fault diagnosis in centrifugal pumps: A dual-scalogram approach with convolution autoencoder and artificial neural network. Sensors, 2024, 24(3): 851.
[5] Yan C, Liu Z, Liao F, et al. Fault diagnosis of three-cylinder mud pump based on transfer learning. Engineering Research Express, 2024, 6(2): 025516.
[6] Zheng J, Liao J, Zhu Y. Two-stage multi-channel fault detection and remaining useful life prediction model of internal gear pumps based on Robust-ResNet. Sensors, 2023, 23(5): 2395.
[7] Alguliyev R, Imamverdiyev Y, Sukhostat L. Intelligent diagnosis of petroleum equipment faults using a deep hybrid model. SN Applied Sciences, 2020, 2(5): 924.
[8] Kumar D, Dewangan A, Tiwari R, et al. Identification of inlet pipe blockage level in centrifugal pump over a range of speeds by deep learning algorithm using multi-source data. Measurement, 2021, 186: 110146.
[9] Li J, Shao J, Wang W, et al. An evolutional deep learning method based on multi-feature fusion for fault diagnosis in sucker rod pumping system. Alexandria Engineering Journal, 2023, 66: 343-355.
[10] Akpudo U E, Hur J W. A CEEMDAN-assisted deep learning model for the RUL estimation of solenoid pumps. Electronics, 2021, 10(17): 2054.
[11] Ding Y, Ma L, Ma J, et al. Intelligent fault diagnosis for rotating machinery using deep Q-network based health state classification: A deep reinforcement learning approach. Advanced Engineering Informatics, 2019, 42: 100977.
[12] Li X, Xie L, Deng B, et al. Deep dynamic high-order graph convolutional network for wear fault diagnosis of hydrodynamic mechanical seal. Reliability Engineering & System Safety, 2024, 247: 110117.
[13] Manikandan S, Duraivelu K. Vibration-based fault diagnosis of broken impeller and mechanical seal failure in industrial mono-block centrifugal pumps using deep convolutional neural network. Journal of Vibration Engineering & Technologies, 2023, 11(1): 141-152.
[14] Zhang F, Li Y, Shan D, et al. An improved fault diagnosis approach for pumps based on neural networks with improved adaptive activation function. Processes, 2023, 11(9): 2540.
[15] Kumar A, Gandhi C P, Zhou Y, et al. Improved deep convolution neural network (CNN) for the identification of defects in the centrifugal pump using acoustic images. Applied Acoustics, 2020, 167: 107399.
[16] Khalid S, Jo S H, Shah S Y, et al. Artificial intelligence-driven prognostics and health management for centrifugal pumps: A comprehensive review. Actuators, 2024, 13(12): 514.
[17] Lu J, Zhou Y, Ge Y, et al. Research into prediction method for pressure pulsations in a centrifugal pump based on variational mode decomposition–particle swarm optimization and hybrid deep learning models. Sensors, 2024, 24(13): 4196.
[18] Guo J, Yang Y, Li H, et al. A parallel deep neural network for intelligent fault diagnosis of drilling pumps. Engineering Applications of Artificial Intelligence, 2024, 133: 108071.
[19] Singh V, Gangsar P, Porwal R, et al. Artificial intelligence application in fault diagnostics of rotating industrial machines: A state-of-the-art review. Journal of Intelligent Manufacturing, 2023, 34(3): 931-960.
[20] Xu W, Wang Z, Zhou Z, et al. Wear state assessment of external gear pump based on system-level hybrid digital twin. Mechanical Systems and Signal Processing, 2024, 209: 111123.
[21] Klingebiel J, Salamon M, Bogdanov P, et al. Towards maximum efficiency in heat pump operation: Self-optimizing defrost initiation control using deep reinforcement learning. Energy and Buildings, 2023, 297: 113397.
[22] Bang J, Di Marco P, Shin H, et al. Deep transfer learning-based fault diagnosis using wavelet transform for limited data. Applied Sciences, 2022, 12(15): 7450.
[23] Hasan M J, Rai A, Ahmad Z, et al. A fault diagnosis framework for centrifugal pumps by scalogram-based imaging and deep learning. IEEE Access, 2021, 9: 58052-58066.
[24] Prasshanth C V, Venkatesh S N, Mahanta T K, et al. Deep learning for fault diagnosis of monoblock centrifugal pumps: A Hilbert – Huang transform approach. International Journal of System Assurance Engineering and Management, 2024: 1-14.
[25] Wei Y, Liu H, Wei M, et al. A fault diagnosis method for dry vacuum pump bearing based on finite element simulation with deep transfer learning. IEEE Access, 2025.
[26] Bie F, Du T, Lyu F, et al. An integrated approach based on improved CEEMDAN and LSTM deep learning neural network for fault diagnosis of reciprocating pump. IEEE Access, 2021, 9: 23301-23310.
[27] Dai M, Liu Z, Wang J, et al. Physics-driven feature alignment combined with dynamic distribution adaptation for three-cylinder drilling pump cross-speed fault diagnosis. Reliability Engineering & System Safety, 2024, 251: 110369.
[28] Wu Y, Feng Z, Liang J, et al. Fault diagnosis algorithm of beam pumping unit based on transfer learning and DenseNet model. Applied Sciences, 2022, 12(21): 11091.
[29] Li C, Liu X, Wang H, et al. Fault diagnosis method for centrifugal pumps in nuclear power plants based on a multi-scale convolutional self-attention network. Sensors, 2025, 25(5): 1589.
[30] He R, Dai Y, Lu J, et al. Developing ladder network for intelligent evaluation system: Case of remaining useful life prediction for centrifugal pumps. Reliability Engineering & System Safety, 2018, 180: 385-393.

[31] Song Y, Zhang T, Liu Q, et al. Cavitation identification method of centrifugal pumps based on signal demodulation and EfficientNet. Arabian Journal for Science and Engineering, 2025, 50(12): 8779-8793.
[32] Liu Z, Xiao H, Zhang T, et al. A fault diagnosis framework for waterjet propulsion pump based on supervised autoencoder and large language model. Machines, 2025, 13(8): 698.
[33] Gong F, Tong S, Du C, et al. A fault diagnosis model of an electric submersible pump based on mechanism knowledge. Sensors, 2025, 25(8): 2444.
[34] Zhu H, Yu H, Sun Q, et al. Deep learning-based performance prediction of electric submersible pumps under viscous and gas–liquid flow conditions. Machines, 2025, 13(2): 135.
[35] Ahmad S, Ahmad Z, Kim J M, et al. A centrifugal pump fault diagnosis framework based on supervised contrastive learning. Sensors, 2022, 22(17): 6448.
[36] Xia M, Shao H, Williams D, et al. Intelligent fault diagnosis of machinery using digital twin-assisted deep transfer learning. Reliability Engineering & System Safety, 2021, 215: 107938.
[37] Gong F, Tong S, Du C, et al. Named entity recognition in the field of small sample electric submersible pump based on FLAT. Applied Sciences, 2025, 15(5).
[38] Ullah N, Ahmad Z, Siddique M F, et al. An intelligent framework for fault diagnosis of centrifugal pump leveraging wavelet coherence analysis and deep learning. Sensors, 2023, 23(21): 8850.
[39] Meng F, Shi Z, Song Y, et al. The DMF: fault diagnosis of diaphragm pumps based on deep learning and multi-source information fusion. Processes, 2024, 12(3): 468.
[40] Tang S, Yuan S, Zhu Y, et al. Deep learning-based intelligent fault diagnosis methods toward rotating machinery. IEEE Access, 2019, 8: 9335-9346.
[41] Dave G S, Pandhare A P, Kulkarni A P, et al. Innovative data techniques for centrifugal pump optimization with machine learning and AI model. PLOS One, 2025, 20(6): e0325952.
[42] Chennai Viswanathan P, Venkatesh S N, Dhanasekaran S, et al. Deep learning for enhanced fault diagnosis of monoblock centrifugal pumps: spectrogram-based analysis. Machines, 2023, 11(9): 874.
[43] Prosvirin A E, Ahmad Z, Kim J M, et al. Global and local feature extraction using a convolutional autoencoder and neural networks for diagnosing centrifugal pump mechanical faults. IEEE Access, 2021, 9: 65838-65854.
[44] Wei J, Gao X, et al. Fault diagnosis of sucker rod pump based on deep-broad learning using motor data. IEEE Access, 2020, 8: 222562-222571.
[45] Konieczny J, Łatas W, Stojek J, et al. Classification of wear state for a positive displacement pump using deep machine learning. Energies, 2023, 16(3): 1408.
[46] Medina R, Sánchez R V, Cabrera D, et al. Scale-fractal detrended fluctuation analysis for fault diagnosis of a centrifugal pump and a reciprocating compressor. Sensors, 2024, 24(2).

Journal of Manufacturing Science and Mechanical Engineering
Print ISSN: 2959-9881
Online ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3020

COMPOSITE OVER-WRAPPED PRESSURE VESSEL TECHNOLOGY FOR SPACECRAFT PROPULSION SYSTEMS

Bin Yu[1,2*], TianJu Ma[2], Yun Zhang[2], Cheng Huang[1], SenDong Gu[2], HaiYan Li[2], LinFeng Chen[2]
[1]Light Alloy Research Institute, Central South University, Changsha 410000, Hunan, China.
[2]Aerospace Pressure Vessel Division, Lanzhou Institute of Physics, Lanzhou 730000, Gansu, China.
Corresponding Author: Bin Yu, Email: amtfyb@163.com

Abstract: This paper systematically reviews global progress in composite overwrapped pressure vessels (COPVs) for space applications and projects their trajectory, synthesizing key advances from the United States, Europe and leading Asian nations in structural and reliability design, stress-fracture and low-cycle-fatigue life prediction, material selection, forming, qualification and non-destructive testing, including high-strength fibers, ultra-thin metallic liners, inheritable design principles and validated failure models, and summarizing representative work by principal Chinese institutes and universities; on this basis it proposes future Chinese directions-high-strength fiber optimization, advanced liner alloys, burst-factor and performance-factor tuning, next-generation NDT, upstream pre-research and Standardization-to underpin independent R&D and technological upgrading of high-performance aerospace COPVs.
Keywords: Aerospace propulsion systems; Composite pressure vessels; Metal liners; Fiber-reinforced composites; Reliability; Stress fracture life

1 INTRODUCTION

Spacecraft and their subsystems necessitate a variety of pressure vessels for the storage of liquids and gases, encompassing gas cylinders and surface tension tanks for satellite and spacecraft propulsion systems, pressure vessels utilized in space station propulsion, fluid management, environmental control and life support systems, scientific and commercial experiment systems, as well as gas cylinders and cryogenic tanks for launch systems.
Composite pressure vessel (COPV) exhibit substantial advantages over all-metal containers, such as lightweight, high stiffness, high specific strength, high reliability, excellent fatigue resistance, long service life, compliance with the “leak before break” (LBB) safety mode, flexible design, low cost, and short production cycles, and are extensively applied in space systems. As a critical component, the development level of COPVs directly determines their performance, which in turn impacts the overall effectiveness of the aerospace system. Structural efficiency affects payload capacity, as COPVs typically account for a significant proportion. Reliability is crucial for launch and in-orbit safety. Safety is of utmost importance, as a rupture could lead to catastrophic consequences, so the LBB design must be strictly adhered to. Stress fracture life determines the in-orbit lifespan of the spacecraft, necessitating structural design to meet reliability requirements. Fatigue life limits the number of refills for reusable COPVs in space stations. Composite material structural technology is a key technique for enhancing the safety of cryogenic liquid krypton storage tank systems.

2 TECHNOLOGY ADVANCES IN US

2.1 Structural Composite Industries of US

In 1972, Landers from the Structural Composite Industries (SCI) of the United States compiled the “Design Specifications for Fiber-Reinforced Pressure Vessels,” which provided comprehensive guidance on fiber-wound shells, metal-lined structures, and fatigue life prediction[1]. In the same year, the 2219-T62 aluminum alloy-lined K49 aramid-wound COPV developed by SCI achieved a 30% weight reduction compared to thin-walled TC4 titanium cylinders[2]. Morris reported that SCI's carbon fiber-wound COPVs have a volume range of 0.74 L to 327.74 L, a diameter range of Ø76.2 mm to Ø508 mm, and an operating pressure range of 12.77 MPa to 31.05 MPa, with linings manufactured by spinning[3]. In 1986, he developed a COPV with a 6061-T6 aluminum alloy liner and IM6 carbon fiber/REZ-100 resin, achieving a performance factor (pressure × volume/weight, PV/W) of 25.4 km. The strength was improved by 21% and 37% compared to K49 aramid and S2 glass fiber containers, respectively, with performance factors increased by 20% and 50% [4].
Rabel's 1989 study showed that high-fiber-stress COPVs fractured within 18–22 months, while medium- and low-stress ones remained intact for three years, indicating that fiber stress levels significantly influence stress fracture lifespan[5]. Haddock developed a T1000 carbon fiber-wrapped COPV in 1990, achieving a performance factor of 33.02 km[6]. In 1991, he further achieved compliance with the MIL-STD-1522A LBB safe fatigue failure mode through pre-fabricated defects in the metal liner[7]. Braun studied winding process factors in 1992, including liner performance, surface condition, corrosion resistance, thickness uniformity, pre-treatment, adhesive material aging sensitivity, curing, and thermal stress cycling, as well as composite material performance, winding parameters, and damage sensitivity[8]. The progress of COPV technology by SCI is shown in Figure 1.

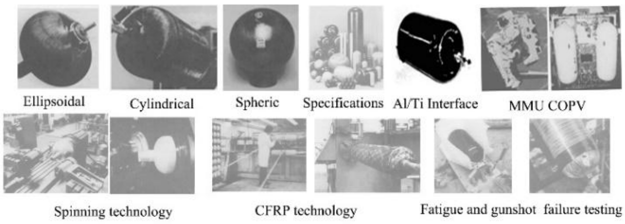


Figure 1 Advances in COPV Technology at SCI

The technical parameters of composite pressure vessels developed by SCI for aerospace applications, aviation applications, commercial aircraft, and civil applications are shown in Tables 1, 2, 3, and 4, respectively.

Table 1 Aerospace Composite Pressure Vessel by SCI

Model	ODmm	OALmm	VL	WKg	Apply
ALT-366	353.1	360.7	20.16	8.16	SDI
ALT-388	711.2	436.9	65.66	25.13	LV
ALT-421	228.6	104.1	1.15	0.45	SDI
ALT-449	515.6	177.8	8.36	4.17	SDI
ALT-454	442.0	165.1	6.37	2.90	SDI
ALT-464	635.0	355.3	43.43	6.76	/
ALT-516	228.6	81.3	0.67	0.23	SDI
ALT-517	228.6	121.9	1.57	0.23	SDI
AC-5000	436.9	172.7	6.88	3.08	SDI
AC-5024	513.1	345.4	33.92	6.53	LV
AC-5040	228.6	106.7	1.16	0.45	SE
AC-5045	317.5	33.0	0.23	0.09	/
AC-5046	208.3	114.3	1.21	0.54	/
AC-5049	195.6	88.9	0.75	0.45	SDI
AC-5101	236.2	91.4	0.77	0.50	SDI

Table 2 Aviation Composite Pressure Vessel

Model	P_WMPa	VL	ODmm	OALmm	WKg	Fiber Type
183	20.7	6.06	178.31	420.37	4.38	KEVLAR
210A	20.7	3.28	136.14	375.92	2.22	KEVLAR
216A	20.7	16.39	222.25	643.38	7.62	KEVLAR
274	23.2	4.26	136.91	459.74	2.86	KEVLAR
351	20.7	3.69	136.14	411.48	2.40	KEVLAR
411	23.2	17.60	200.66	833.12	9.98	KEVLAR
554	20.7	10.65	198.12	542.54	6.84	KEVLAR
63	20.7	17.60	200.66	833.12	9.98	KEVLAR
715	22.7	24.74	257.81	703.58	10.52	CAR/GLS
716	22.7	26.61	257.81	746.76	11.43	CAR/GLS
726	22.7	19.66	220.98	734.06	8.48	CAR/GLS
727	22.7	21.29	220.98	784.86	9.03	CAR/GLS
736	12.8	24.58	237.49	762.00	7.26	CAR/GLS
738	23.2	18.27	210.82	734.06	7.94	CAR/GLS
745	26.2	26.61	257.81	741.68	12.56	CAR/GLS
749	22.7	5.90	156.97	454.66	3.31	CAR/GLS
750	22.7	4.92	156.97	393.70	2.86	CAR/GLS
751	22.7	10.65	182.88	601.98	4.90	CAR/GLS
789	22.7	2.05	115.57	322.58	1.54	CAR/GLS

Table 3 Commercial Aircraft Composite Pressure Vessel

Model	P_WMPa	VL	ODmm	OALmm	WKg	Fiber Type
279	12.8	10.65	172.72	635.00	3.59	KEVLAR
372	12.8	8.37	172.72	518.16	2.90	KEVLAR
280	12.8	4.83	132.08	502.92	1.81	KEVLAR
281	12.8	16.39	193.04	751.84	5.85	KEVLAR
282	12.8	24.58	231.14	800.10	8.21	KEVLAR
621	12.8	2.33	91.95	478.79	0.91	KEVLAR

Table 4 Civilian Field Aircraft Composite Pressure Vessel

Model	P_WMPa	ODmm	OALmm	VL	WKg
ALT747	19.98	343	1115	74	31.24
ALT881E	19.98	404	3048	308.1	113.95
ALT753	24.8	255	1870	71.4	39.04
ALT890S	24.8	333	1270	77.5	42
ALT890	24.8	333	1778	113.1	58.34
ALT807S	24.8	333	1905	120.2	70.1
ALT821	24.8	333	2180	135	67.65
ALT807	24.8	333	3048	195.6	95.79
ALT972	24.8	347	762	47.1	23.15
ALT867	24.8	347	889	57.2	28.38
ALT823F	24.8	388	1001	77.7	35.41
ALT823E	24.8	388	1850	162.8	67.65
ALT823A	24.8	388	1943	173	74.91
ALT918	24.8	388	1993	182.6	77.63
ALT823B	24.8	388	2548	233.2	94.16
ALT823D	24.8	388	3048	283.4	111.91
ALT881S	24.8	404	1803	172	67.65
ALT881L	24.8	404	1905	188.7	74.91
ALT881M	24.8	404	2159	211.4	80.36
ALT881	24.8	404	3048	308.1	114.86
ALT982L	24.8	415	914	81.7	39.95
ALT810P	24.8	415	1397	137	57.2
ALT982L	24.8	415	1524	150	57.2
ALT810A	24.8	415	1869	190.3	79
ALT810B	24.8	415	1905	191.9	81.72
ALT889S	24.8	415	2083	212.9	88.08
ALT810S	24.8	415	2083	204	86.26
ALT810M	24.8	415	2179	215.1	85.35
ALT889	24.8	415	2540	266.4	106.69
ALT810N	24.8	415	3048	312.9	123.94
ALT891	24.8	462	2667	342	123.49
ALT820SG	24.8	532	2032	318.1	163.89
ALT820G	24.8	532	2180	344.6	175.24
ALT604	31.01	183	533	9.2	5.49
ALT988	34.45	183	490	7	5.31
ALT848	34.45	432	1074	92	59.93
ALT836U	34.45	432	1003	90.2	59.47
ALT909s	34.45	419	2667	273.7	117.13
ALT909	34.45	419	3048	312.9	130.75
ALT962L	43.06	439	2976	304.1	190.68
ALT962L	43.06	439	3048	311.8	197.49
Alt861B	68.9	229	1093	27.4	30.74

2.2 ARDE Company of US

In 1982, Gleich from ARDE, a US space systems company, developed the CRES-301 stainless steel-lined AS4 carbon fiber-wrapped spherical COPV, which enhanced the performance factor to 27.94 km, representing a 30% increase over the K49 aramid container and a 60% weight reduction compared to the TC4 all-metal cylinder[9]. In 1988, Gleich further elucidated that COPVs exhibit two failure modes: stable crack propagation (LBB) and unstable brittle fracture[10]. In 1997, Sneddon developed a 0.71 mm ultra-thin I-718-lined T1000 carbon fiber cylindrical COPV (D4619) for the European Star 2000+ satellite, achieving a coupled design between the plastic liner and the composite shell[11]. In the same year, Escalona developed a CRES-301 stainless steel-lined T1000 carbon fiber spherical COPV (D4650) for the Atlas/Centaur rocket, proposing the spherical winding grid theory, which reduced weight by 11% and increased strength by 23% compared to the IM7-W container[12]. In 2006, Sneddon developed an I-718-lined T1000 carbon fiber cylindrical COPV (D4929) for the Delta II rocket, effectively suppressing axial stress concentration through a fully constrained bottom and axially constrained top structure[13]. In 2008, Ray developed an I-718 and 2219 -T62 aluminum-lined T1000 carbon fiber spherical cryogenic COPV (D4970/D4971) for the lunar landing program[14]. Using cryogenic composite technology, he revealed the coordinated deformation mechanism between the liner and composite under pressure-temperature coupling, resolving delamination issues under thermal shock. The progress of COPV technology by ARDE is shown in Figure 2.

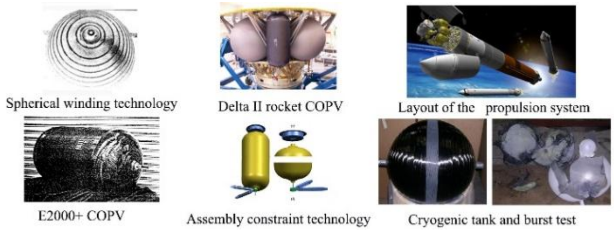


Figure 2 Advances in COPV technology at ARDE

The technical parameters of composite pressure vessels developed by ARDE for aerospace applications are shown in Table 5.

Table 5 Aerospace Composite Pressure Vessel by ARDE

Model	D4619 COPV	D4650 COPV	D4929 COPV	D4970 COPV	D4971 COPV
Apply	E2000+	C LV	DII LV	Altair	Altair
Time	1997	1997	2006	2008	2008
Medium	He	He	He	He	He
Shape	Cylinder	Spherical	Cylinder	Spherical	Spherical
V (L)	97	132	50.8	51.5	51.5
M (kg)	18.33	26.3	10.89	11.54	10.5
OD (mm)	423	660.4	330	460.25	460.25
Fiber	T1000	T1000	T1000	T1000	T1000
Resin	H 53	H 53	H 53	31-43B	31-43B
Liner	I-718	301	I-718	I-718	2219
L_T (mm)	0.71	—	0.71	-	-
P_W (MPa)	31.05	27.6	29	31	31
P_b (MPa)	46.58	41.4	43.5	46.5	46.5
SF	/1.5	/1.5	/1.5	/1.5	/1.5
QS	34％	48％	35％	58%	71%
ABF	2	2.2	2	2.37	2.55

2.3 Space Pressure Systems Inc. of US

Space Pressure Systems Inc. (ATK-PSI) has achieved several milestones in the development of composite overwrapped pressure vessels (COPVs). In 1996, ATK-PSI developed the CP-3 pure titanium-lined T1000 carbon fiber-wrapped xenon COPV, which significantly advanced the design and manufacturing technology of conical gas cylinder winding patterns[15]. In 2000, the company further expanded its technological capabilities by developing a 0.8 mm ultra-thin-walled TC4 titanium alloy-lined T1000 carbon fiber cylindrical COPV for the ETSⅧ spacecraft's electric propulsion system, thereby establishing the foundation for algorithms that govern the elastic state of high-strength metal-lined structures[16]. In 2004, ATK-PSI conducted research on composite material interface technology, which included fiber-wrapped reserved structures, composite material support skirts, and metal ear plates[17]. In 2006, the company developed a 0.5 mm ultra-thin-walled CP-3 pure titanium-lined T1000 carbon fiber cylindrical COPV for the ESA Vega rocket propulsion and orbit control system, achieving a performance factor of 39.8 km (burst pressure 57.2 MPa, volume 81.4 L, weight 11.7 kg), which represented the highest reported performance factor for metal-lined COPVs at that time[18]. The progress of COPV technology by PSI is shown in Figure 3.

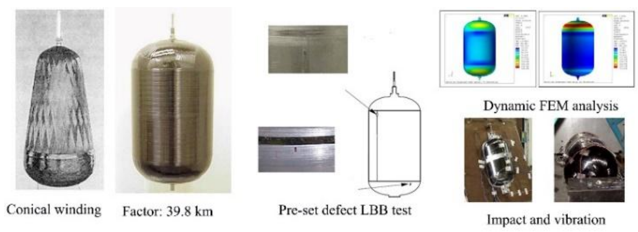


Figure 3 Advances in COPV technology at PSI

Table 6 summarizes PSI applications in communications, science, and military satellites, and Table 7 lists those for spacecraft, launch vehicles, and platforms.

Table 6 PSI Applications in Communications, Science, and Military Satellites

No.	Communication satellites	Scientific satellites	Military satellites
1	A2100	CHANDRA	DMSP
2	ACTS	EOS	DSCS III
3	ARABSAT	ETSVII	DSP
4	ASTROLINK	GOES	FLTSATCOM
5	B.SAT	GRO	GEOSAT
6	BRASILSAT	HEAO	GEOSAT F/O
7	ETS8	KOMPSAT	GPS II
8	FS 1300	LANDSAT	MILSTAR
9	HS 376	ROCSAT	NATOIV
10	HS 601	SOHO	SKYNET
11	INDOSTAR	STEP	TORSS
12	INMARSAT	TIROS	UHF &UHF
13	INTELSAT	TOMS	F/0
14	IRIDIUM	TOPEX	P81
15	MTSAT.2	TRMM	SBIRS LOW
16	N-STAR	UARS	NUMEROUS
17	ORBCOM	WORLOVIEW	CLASSIFIED
18	S3000	ORBVIEW	PROGRAMS
19	S4000	QUICKBIRD	
20	S5000	RADARSAT	
21	S7000	HIPPARCOS	

Table 7 PSI Applications in Spacecraft and Launch Vehicles & Platforms

No.	SPACECRAFT	LAUNCH VEHICLES & PLATFORMS
1	MARINER	ATLAS I
2	PIONEER	ATLAS II
3	VOYAGER	ATLAS IIA
4	VIKING	ATLAS IIAS
5	MAGELLAN	ATLASIIIA
6	ULYSSES	ATLASV
7	CLEMENTINE	CENTAUR
8	NEAR	DELTA III
9	MARS	DELTAIV
10	PATHFINDER	EURECA
11	CASSINI	IUS
12	MARS	KISTLER
13	SURVEYOR	SPACE
14	MARS '98	SHUTTLE
15	LANDER	SPACE
16	DEEP SPACE	STATION
17	ONE	STAR 48
18	LUNAR	TITAN II
19	PROSPECTOR	TITANIII
20	MESSENGER	X38
21	MARS ROVER	
22	STEREO	
23	DEEP IMPACT	
24	MARS ORBITER	

The technical parameters of composite pressure vessels developed by ARDE for aerospace applications are shown in Table 8.

Table 8 Aerospace Composite Pressure Vessel By PSI

Model	CXT	XT	Tank	Vega GT
Apply	XIPS	ETS VIII	Space	ESA Vega LV
Times	1996	1999	2001	2006
Medium	Xe	Xe	N2H4	He
Shape	conical	Cylinder	Cylinder	Cylinder
V (L)	32.12	50	171	87
W (kg)	6.12	7	19.1	23
OD (mm)	337	337	580	337
OAL (mm)	752	683	860	683
Fiber	T1000 CF	T1000 CF	T1000 CF	T1000 CF
Resin	Epon 826	Epon 826	/	Epon 826
Liner	CP-30	TC4	TC4	CP-30
L_T (mm)	0.81	0.8	1.01	0.8

Model	CXT	XT	Tank	Vega GT
P_W (MPa)	17.25	15	3.8	31
P_b (MPa)	25.86	22.5	5.7	62
QS	41%	10%	87%	46%
ABF	2.12	1.7	2.4	2.92

2.4 Boeing Company of US

In 1999, Boeing's Babel evaluated the stress cracking, impact damage, and leakage risks of the Delta IV rocket's composite overwrapped pressure vessel (COPV) in orbit, focusing on four aspects: sustained load stress cracking, composite material impact damage, liner crack propagation leakage, and stress overload[19]. The study concluded that leveraging inherited technology could significantly enhance reliability. In 2000, Ledesma compared three predictive models for COPV stress fracture life: ASTM-D2992, the Thomas method, and the Robinson method[20]. ASTM-D2992 is specifically designed for glass fiber COPVs, while the Thomas method's parameters α and β are derived from burst tests. The Robinson method, based on fiber bundle tests and the Weibull distribution, is particularly suitable for long-term life prediction. In 2003, Abdi developed the GENOA model to simulate crack initiation, propagation, and failure behavior in composite low-temperature storage tanks, validating its algorithmic accuracy through finite element comparisons[21]. In 2004, Robinson investigated liquid hydrogen/liquid oxygen cryogenic tanks and discovered that using an aluminum alloy lining in conjunction with IM7 carbon fiber effectively mitigates thermal expansion coefficient disparities and electrochemical corrosion issues between the materials[22]. The progress of COPV technology by Boeing is shown in Figure 4.

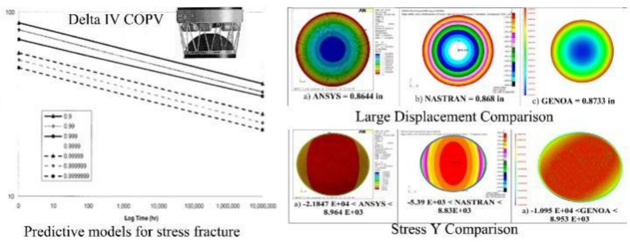


Figure 4 Advances in COPV technology at Boeing

2.5 Air Force Research Laboratory of US

In 2001, the U.S. Air Force Research Laboratory (AFRL) investigated liquid oxygen/liquid hydrogen composite cryogenic tanks for single-stage-to-orbit (SSTO) spacecraft. Arritt developed a low-temperature-resistant resin system and optimized the winding structure to inhibit the propagation of low-temperature microcracks, thereby addressing delamination issues under temperature cycling and achieving a 50% weight reduction compared to aluminum alloy tanks[23]. In the same year, Arritt also developed self-healing composite materials that can autonomously repair cracks within 48 hours without external loading, restoring 75% of their original strength. In 2005, Mallick proposed a multi-scale system design concept for lightweight composite material tanks, establishing cross-scale correlations between materials, design, and manufacturing, and elucidating the material-structure coupling mechanisms from the microscopic to the macroscopic scale[24]. In 2006, Bechel demonstrated through temperature shock tests at -196°C to -177°C using different linear IM7 specimens that the crack propagation behavior of composite materials varies with structure, and the introduction of flexible ES fiber layers can significantly reduce cracks[25]. In 2007, Falugi compared various low-temperature tank configurations and found that the shared partition structure was the lightest, while the stacked structure was easier to manufacture but heavier[26]. The progress of COPV technology by AFRL is shown in Figure 5.

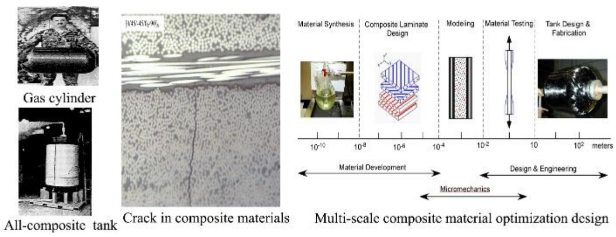


Figure 5 Advances in COPV technology at AFRL

2.6 Glenn Research Center

Revilock at the Glenn Research Center (GRC) in the United States utilized a dual high-speed camera 3D system to precisely measure the strain of composite spherical gas cylinders under repeated pressurization, revealing system behavior more accurately than traditional strain gauges[27]. Increased local strain indicates that the liner may fracture. In 2007, Murthy established a stress rupture model based on Weibull statistics, accounting for parameter uncertainty[28, 29]. Combining 35 years of experimental data from the Lawrence Livermore National Laboratory, he used the Phoenix model to achieve precise predictions of COPV stress rupture life. In the same year, Murthy also established a spacecraft full-life stress rupture model, clearly defining the quantitative relationship between fiber stress ratio and lifespan, and provided data support through accelerated aging tests simulating in-orbit COPV[30]. In 2010, Murthy proposed a stress rupture lifespan assessment model considering structural damage, developing a microscopic mechanical fiber rupture model based on progressive damage theory, effectively reducing uncertainty and improving prediction accuracy[31]. In 2012, Murthy further pointed out that the fiber stress ratio (the ratio of stress to strength under working pressure) has a much greater impact on reliability than other parameters, a finding verified through experiments and calculations[32]. The progress of COPV technology by GRC is shown in Figure 6.

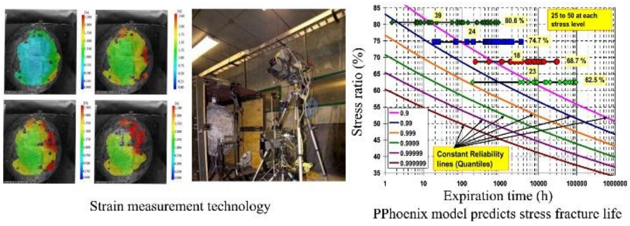


Figure 6 Advances in COPV technology at GRC

2.7 Lockheed Martin Space Systems Company

In 1995, Emery from Lockheed Martin Space Systems Company (LMT) integrated fiber optic sensors into wound fiber layers, enabling automatic diagnosis and health monitoring of various parameters such as fiber strain, delamination, fracture, operating temperature, and medium leakage[33]. In 2005, Achary conducted systematic research on critical components such as low-temperature liquid hydrogen/liquid oxygen tanks, pressurized gas cylinders, and fuel supply lines for the X-33 and X-34 spacecraft, and completed the design and manufacturing of low-temperature tanks using an all-composite material structure[34]. The progress of COPV technology by LMT is shown in Figure 7.

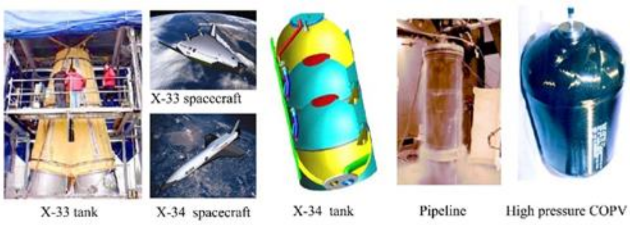


Figure 7 Advances in COPV technology at LMT

2.8 Brunswick Composite Company

Brunswick Composite Materials initiated fiber-overwrapped COPV research in the 1950s, delivered a composite solid rocket motor in 1959, founded a mass-production facility for such motors in 1963, and transferred filament-winding expertise to COPV development the same year; early rubber-lined, glass-fiber-overwrapped COPVs served jet-engine restart systems, whereas the aerospace sector's escalating demand for lightweight, high-strength, minimal-leakage storage has since elevated metal-lined COPVs to the dominant configuration.

Brunswick's Veys introduced a 2219-Al plastically-worked COPV validated by ground tests in 1989[35]; in 1990, he examined carbon-wrapped COPVs with 6061/5086-Al liners, demonstrating that weld reinforcement, weld-adjacent strength design, uniform-thickness regions and smooth transitions dominate strength and fatigue, which can be improved by pre-stressing the liner during filament winding to equalize strain[36]. In 1991, Veys performed fatigue tests on 6061-T6-Al-lined COPVs and created Coffin-Manson-based software that predicts life by computing fiber/metal stress–strain loops during load/unload[37]; the model treats the composite as purely elastic and the liner as elastic-plastic via J2 plasticity plus Ramberg-Osgood, calibrated with four empirical constants, thus establishing the basis for aerospace low-cycle-fatigue design of plastic liners. Murray in 1993 developed an all-composite space COPV with an HDPE liner[38]; ANSI/AIAA S-080-2000 only guides metal-liner selection, whereas ANSI/AIAA S-081 is the definitive reference for COPV liner materials[39, 40].

The technical parameters of composite pressure vessels developed by Brunswick for aerospace applications are shown in Table 9.

Table 9 Aerospace Composite Pressure Vessel by Brunswick

P_WMPa	VL	Shape	ODmm	Fiber	FatigueCycles
34.5	5	Cylinder	178	GFRP	5000
20.7	7.13	Cylinder	127	AFRP	1800
22.08	8	Spherical	266.7	AFRP	400
10.65	9.83	Cylinder	203	GFRP	10000
24.15	11.8	Cylinder	221	AFRP	400
22.43	12.7	Cylinder	221	GFRP	400
20.7	49.16	Cylinder	558	AFRP	1000
41.4	0.25	Cylinder	53	AFRP	10000
21.25	12.7	Cylinder	241	AFRP	2200
30	20.5	Ellipsoid	355.6	AFRP	1000
69	0.57	Spherical	114.3	CFRP	50
34.5	8	Spherical	266.7	CFRP	200
27.6	49.33	Spherical	457.2	CFRP	60
27.6	134.1	Spherical	660.4	CFRP	60
20.7	308	Spherical	914.4	AFRP	100
34.5	41.8	Spherical	469.9	AFRP	120
103.5	4.5	Cylinder	127	AFRP	100

2.9 Marshall Space Flight Center

Marshall Space Flight Center (MSFC)'s Grant experimentally confirmed that FBG sensors survive pressurization and that internal pressure and FBG-reported strain are linearly related[41]; Grant subsequently embedded these sensors in a COPV to reveal its damage mechanism, finding after hydraulic fatigue that circumferential fibres and mid-cylinder longitudinal fibres developed positive residual strain whereas longitudinal fibres near the dome turned negative, with the longitudinal component consistently smaller than the circumferential, a pattern that successfully predicted the eventual burst site[42]. The progress of COPV technology by MSFC is shown in Figure 8.

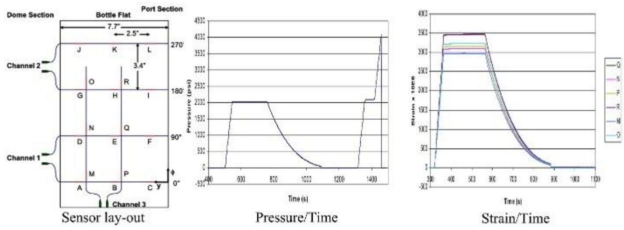


Figure 8 Advances in COPV technology at MSFC

2.10 Johnson Space Center

Beeson at Johnson Space Center (JSC) quantified low-energy impact damage in carbon-fiber COPVs in 1996, revealing an energy-dependent damage progression and a threshold above which visible surface damage appears, the magnitude of which scales with vessel geometry, laminate architecture, and impactor geometry and size[43]; in 2007, Greene used the NASA Engineering Safety Center framework to demonstrate that operational fiber stresses in T1000 carbon/aluminum COPVs exceed Luxfer-based predictions by 11 %[44], and later that year Greene validated remaining life and reliability of ISS K49-COPVs through fatigue and hydroburst tests on identical[45], co-aged specimens, confirming compliance with all performance specifications despite composite aging and prolonged storage; finally, Abraham in 2012 recorded broadband modal acoustic emission during intermittent tensile tests on T1000G/epoxy single -fiber specimens and introduced a statistically rigorous trend analysis that produced a more linear Felicity-ratio decay with load[46], enabling earlier and more precise failure prediction. The progress of COPV technology by JSC is shown in Figure 9.

Figure 9 Advances in COPV technology at JSC

3 TECHNOLOGY ADVANCES IN OTHER COUNTRIES ABROAD

3.1 MT Aerospace of German

Since 1988, Radtke at Germany's MT Aerospace has advanced lightweight storage technology by producing ring-shaped and cylindrical propellant tanks, composite solid motors and COPVs for Ariane 3 and 4[47]; in 2006 Radtke introduced a net-shape spinning process for titanium-lined COPV domes[48], yielding a 0.7 mm-thick, 660 mm-diameter titanium liner and a 0.5 mm-thick, 1140 mm-diameter composite propellant-tank liner. The progress of COPV technology by MT is shown in Figure 9.

Figure 10 Advances in COPV technology at MT

3.2 AeroSpatiale Space Center of Frence

Since 1959 the Charpentier at the French AeroSpatiale Space Center has produced COPVs for TVSAT, TDF1, EUROSTAR, DFS, Tele-X and Ariane 4, and demonstrated that 1° forward or reverse fibre slippage during spherical-tank winding reduces strength by 2.9 % or 5.5 %, respectively[49, 50]; in 1995 Teissier delivered the Ariane 5 2219-Al liquid-oxygen tank, Ø1 303 mm, operating at 2.2 MPa[51].

3.3 National Space Development Agency of Japan

In 1999 Morino from National space development agency of JAPAN (NASDA) validated a cryogenic composite tank concept for reusable launch vehicles, reporting a 20 % loss in filament-wound toughened-resin properties and leakage driven by low-stress matrix cracking[52]; Torano fielded a 2.16 MPa Al-lined CFRP tank for the J-I second stage in 2001[53], achieving 3.24 MPa burst, 2.7 MPa proof, 12-cycle fatigue at working pressure and 5-cycle at proof pressure, with head failure eliminated via optimized liner geometry, enhanced interface bonding and controlled cure pressure; Ishikawa in 2003 derived the stress-free matrix-crack coefficient from crack-initiation-strain differences between ambient and liquid-nitrogen temperatures[54]; Masuda delivered in 2014 an Al-lined CFRP propellant tank incorporating an integrated management device[55]. The progress of COPV technology by NASDA is shown in Figure 11.

3.4 Korea Advanced Institute of Science and Technology

Hwang from Korea Advanced Institute of Science and Technology (AIST) quantified the size-induced strength decay of carbon fibres by testing fibre bundles[56, 57], UD laminates and COPVs, and then embedded three stochastic variables—elastic moduli of the composite, laminate strength and spiral-to-hoop layer thickness—into a probabilistic model that accurately predicted the circumferential strain and burst pressure of 10 COPVs spanning multiple diameters, with hydraulic-burst experiments validating the predictions.

4 TECHNOLOGY ADVANCES IN CHINA

4.1 Lanzhou Institute of Physics

The Lanzhou Institute of Physics has long been dedicated to the technical research and product development of space-grade ambient temperature and cryogenic pressure vessels. The institute has accumulated extensive experience in structural design[58], finite element analysis[59, 60], reliability and failure risk analysis[61], life prediction[62], impact damage assessment, welded structure design, surface modification, inspection and testing, standard application, cryogenic pressure tanks and supply systems[63-65], and space refrigeration unit development[66-70]. Its products include surface tension tanks, composite material pressure vessels, cryogenic tanks, and supply systems.

4.2 Xi'an Aerospace Propulsion Technology Research Institute

Chen of the Xi'an Aerospace Propulsion Technology Research Institute derived preliminary design formulas for COPV structures that encompass fiber-wrapped burst-strength calculation, composite shell critical axial-pressure prediction, deformation analysis of lined composite shells, and structural-design algorithms for fiber-wrapped conical shells.

4.3 Xi’an Aerospace Composite Materials Research Institute

Zeng of the Xi’an Aerospace Composite Materials Research Institute introduced the variable-pitch spherical COPV winding strategy in 2000, cutting composite mass by 24.6 % and raising the performance coefficient by 10.4 % at identical fiber stress; Wang examined F12/T800 hybrid COPVs, finding that longitudinal carbon-fiber plies combined with F12 hoop plies maximize efficiency; Wang dissected fiber-winding CAD/CAM platforms, pinpointing structural design, mathematical modelling, and numerical control as the core subsystems; Li derived geodesic equations and linear -control algorithms for accurate fiber trajectories on toroidal cylinders; Fang fabricated an aluminum-lined T700/EC8 glass-fiber COPV whose burst pressure surpassed 100 MPa, and after 58 MPa pre-tensioning achieved optimized liner stress distribution and a fatigue life exceeding 5 000 cycles.

4.4 Harbin Fiberglass Research Institute

Li introduced a network-theory design for metal-lined composite cylinders; Wang quantified the governing structural parameters. Jiang devised an algorithm for non-load-bearing isodiametric bipolar porous spherical COPVs using the fiber-strength-utilization position function F(X), a maximum-stress failure criterion, and tension-zone spherical-shell geometry. Lou presented a spherical-headed cylindrical COPV algorithm that applies multi-ring enveloping heads and network-theory cylinders. Jiang investigated unequal-polarity porous spherical vessels through weighted-average equal-polarity porous-sphere winding. Lin fabricated a welded, pure-aluminum-lined aramid-wound cylinder achieving 18.2 km performance factor. Wang identified the head-to-cylinder junction as the critical high-stress zone requiring local reinforcement. Lou derived the conical-shell winding profile and determined stable five-axis winding-machine motion patterns. Jiang developed aluminum-alloy-lined aramid-wound toroidal COPVs via dry winding, adhesive metering, and shrink-film curing.

4.5 Harbin Institute of Technology

Shen identified fiber-strength variability as the dominant factor in COPV reliability design; Hu quantified resin-cure kinetics in ultra-thin metal-lined COPVs, attributing liner wrinkling and reduced fatigue life to residual stresses induced by resin shrinkage, temperature gradients and thermal mismatch; Li delivered winding-process simulation software that captures the complete process and enables geodesic trajectory design; Wang advanced lightweight COPVs by introducing segmented heat-treatment spinning of thin-walled aluminum liners and a principal-stress-oriented fiber-winding strategy for vessels with unequal polar openings.

4.6 Dalian University of Technology

Ren formulated a multi-field coupled FE method for cure analysis that integrates cure kinetics, heat transfer and composite mechanics, revealing synchronous peak stresses throughout the vessel during initial cool-down; Chen surveyed multi-field studies on filament-wound shells and recommended setting overpressure via structural simulation; Sun noted that thermoplastic winding offers streamlined equipment, simplified processing and lower cost.

5 DEVELOPMENT OF STANDARDS FOR SPACE COPV

5.1 Development of Standards

Spacecraft and launch vehicles universally demand cylindrical pressure vessels for propulsion, fluid management, environmental control and experimental systems. The 1970s baseline, MIL-STD-1522, prescribed safe design and operation of pressurized missile and space systems; its 1984 revision, MIL-STD-1522A, tightened fracture control for metals and became the global aerospace reference. Composite cylinders, however, soon revealed the standard’s insufficiency-no provisions for carbon-fiber impact tolerance, glass/aramid stress-rupture life, leak-before-break (LBB) validation or composite NDI-prompting the Air Force and SMC to task a revision in 1993; USAF restructuring cancelled this effort, and in 1996 AIAA assumed responsibility, forming the Aerospace Pressure Vessel Standards Working Group (APVSWG).

The group issued ANSI/AIAA S-080 (approved December 1998, published January 1999) covering system analysis, structural design, materials, safety, production control, inspection, test and maintenance of space metal cylinders with a minimum safety factor $\geqslant 1.5$. Parallel development of ANSI/AIAA S-081 (published 2000) addressed composite vessels, permitting LBB or safe-life design for non-hazardous media, yet leakage events during acceptance and flight highlighted the need for dual-mode safety; S-081 was therefore revised into S-081A (2006), adding mechanical-damage control, stress-rupture life requirements and pre-launch inspection, pressure test and data-retention protocols.

S-081A mandates safety factors of 1.5, 1.65 and 2.25 for carbon, aramid and glass fiber, respectively, and defines proof test pressure PT as $PT = (1 + N)PMEOP/2$ for $N > 2$ and $PT = 1.5PMEOP$ ($\leqslant 0.8Pb$) for $N \leqslant 2$, with liner design in the elastic range referred to S-080 and otherwise to S-081A; it further requires Grade-A materials for full-scale composite cylinder fracture-strength testing and details mechanical-damage control, impact tolerance, composite strength design, NDI, LBB and functional verification. ISO 14623 (2003) complements these standards for space metal and metal-lined composite cylinders, imposing a safety factor ⩾1.5 and ⩾0.999 survival probability against stress-rupture; vessels with safety factor <4 and wall thickness <6.35 mm must incorporate systematic damage control. AIAA-S-110 specifies 1.25 for mechanically and thermally loaded components, while ASTM D2992 supplies life-analysis methods—fatigue and stress-rupture—for glass-fiber tubes and pressure vessels, routinely referenced for foreign glass-fiber COPVs.

5.2 Issues and Analysis in the Development of National Standards for COPV in China

China currently lacks national or military standards for aerospace-grade carbon-fiber-overwrapped composite pressure vessels, and as reinforcement has evolved from E-/S-glass through Kevlar-49, IM-6, T-40 and T-700 to the present baseline T-1000, design and acceptance remain unguided in performance indices, material selection, structural and safety design, life analysis, manufacturing processes, process control, testing and acceptance, even though aerospace vessels impose stringent reliability, safety, mass and performance demands; therefore an urgent standard is required for metal-lined, T-1000 carbon-fiber fully-wound cylindrical COPVs that consolidates domestic R&D achievements, matches actual manufacturer capability and selectively incorporates superior elements of international norms. Compared with domestic documents, ANSI/AIAA S-081A and ISO 14623 lead in life prediction, damage control, NDI and safety design, so the envisaged standard should adopt: (1) allowable-fiber-stress test protocols using both full-scale and sub-scale specimens; (2) stress-rupture and fatigue-cycle life prediction technologies; (3) mechanical-damage control processes encompassing detection, severity assessment and tolerance definition; (4) NDI techniques for liners and composites; (5) leak-before-break design criteria and verification methods for pre-existing liner flaws; and (6) qualification routes for liners, fibers and resins covering allowable conditions, structural design, fabrication, NDI and performance testing. With proliferating aerospace and commercial applications, COPVs are becoming safety-critical in multiple host systems, so unified national and military standards are indispensable to regulate design, manufacture, process control, inspection, testing and acceptance, guarantee in-service reliability and safety, and accelerate technology maturation while avoiding uncritical transplantation of foreign standards.

6 CONCLUSION

6.1 Selection of High-strength Fibers

High-strength fiber selection directly governs COPV structural efficiency: T1000 carbon fiber is the present baseline for space-qualified vessels because its exceptional tensile strength delivers a high utilization factor and because extensive domestic and international stress-rupture life tests have validated T1000-COPV composites, with numerous flight units now demonstrating sustained on-orbit reliability; any prospective higher-strength fiber must undergo equivalent property, environmental and stress-rupture evaluations before adoption.

6.2 Development of Ultra-thin Metal Liners

Ultra-thin metallic liners enhance COPV structural efficiency provided that liner stress–strain remains within allowable limits and strength–fatigue specifications are preserved; their implementation hinges on ultra-thin forming and filament-wound composite shell technologies. Internationally, the principal liner alloys are Ti, stainless steel, Inconel 718 and Al, and foreign practice shows that both machined-and-welded and seamless spun ultra-thin liners are flight-ready. Advancing precision machining, thin-wall welding and continued refinement of flow-forming processes for ultra-thin liners is therefore imperative; metal-lined composite pressure vessels currently constitute the dominant trajectory for space COPVs worldwide.

6.3 Development of Diversified Lining Materials

Non-metallic liners—rubber, high-density polyethylene and related polymers—enable all-composite pressure vessels that combine low-cost forming, superior design flexibility and high structural efficiency while fully satisfying typical space and missile requirements; although ANSI/AIAA S-081 and S-081A address only metallic liners, this limitation does not constrain domestic standards. No national military standard yet governs Chinese space COPVs, yet the demonstrated domestic capability to fabricate and qualify non-metallic liners and their proven in-orbit reliability render these materials a credible option; therefore, incorporating non-metallic liners into the material palette of forthcoming national military standards for space COPV composite cylinders is both rational and feasible.

6.4 Structural Design Optimization

Leveraging heritage design and multi-physics tools such as GENOA and ANSYS, future COPV optimization will target fiber-path refinement, head/interface strain coordination and response prediction under cryogenic and cyclic extremes, while AI-driven algorithms will simultaneously enhance composite lay-up design and failure prognosis, thereby raising structural efficiency and reliability.

6.5 Discussion on the Actual Blast Factor (ABF)

Analysis of domestic and international COPV datasets shows that China's performance factor (PV/W) already rivals global benchmarks, reflecting advanced structural design, liner technology and composite processing, yet it does not quantify flight safety or reliability. AIAA S-081A mandates a carbon-fiber stress-rupture factor ⩾1.5 and a burst factor ⩾1.5, whereas operational foreign COPVs exhibit an actual burst factor (ABF) almost universally exceeding 2; the higher reserve capacity lowers operational fiber stress, the decisive variable for on-orbit life and reliability. Therefore, high-performance COPV development must couple PV/W optimization with a mission-specific ABF that exceeds the minimum standard, and key technical indices should be anchored in AIAA S-081A yet tailored to in-service realities to simultaneously enhance performance and assurance.

6.6 Characteristics of Space COPV Heritage Design

Space systems cannot tolerate the several-kilogram TNT-equivalent blast from a high-pressure COPV rupture, making reliability imperative for China's crewed flight and Tiangong-1/Shenzhou-8 heritage. The most effective safeguard is design inheritance: analyses of NASA's main composite cylinder supplier, SCI Composite Materials, and its surface-tension tank supplier, ATK Composite Materials, reveal that users and manufacturers converge on inheritable design, and the extensive on-orbit flight experience of both companies constitutes the primary reference dataset for this approach.

6.7 Development of Non-Destructive Testing Technology

Composite COPV strength is governed by winding and curing variability, and carbon-fiber layers are acutely vulnerable to impact; non-destructive inspection (NDI) is therefore mandatory to reveal any flaw or damage that could degrade performance. Although ANSI/AIAA S-081A prescribes rigorous fiber-winding and impact-damage control plans, NASA's White Sands Test Facility has advanced NDI by correlating sixteen distinct inspection techniques with structural anomalies and embedding these data in impact-damage assessment protocols, making NDI a gate criterion for flight-article acceptance. China must correspondingly mature NDI capability and promulgate dedicated technical guidelines or national military standards for space COPVs to secure in-orbit reliability.

6.8 The Importance of Pre-research on Space COPV

Surveying foreign COPV roadmaps and aligning with China's future space needs underscores that sustained pre-research is indispensable to compress spacecraft development cycles and secure in-orbit reliability; open literature is chronically outdated, so once critical technologies mature abroad, the domestic lag can span several years, and classified data further widen the gap. NASA's sustained investment in pre-competitive program awarded to SCI Structural Composites and Lincoln Composites demonstrates high transition rates from laboratory to flight, hence China must aggressively expand pre-research on space COPVs, accumulate validated data and technological reserves, and codify these findings for future inheritable designs.

6.9 Development of Standards for Space COPV Research

Domestic COPV capability has advanced rapidly, yet standards lag behind those of mature spacefaring nations whose iterative, flight-validated composite-cylinder specifications deliver proven reliability, safety and process control; China currently relies on institute or foreign standards that inadequately constrain design or quality, so urgent codification of national military standards—grounded in domestic R&D data, benchmarked against international best practice and formulated through broad industry consultation—is essential to underpin future development and quality assurance of space-qualified COPVs.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The title of this project is "Research on Cryogenic Propellant Fluid Technology for Satellite Electric Propulsion," and its project number is ZD192401. The author would like to express sincere gratitude to the 2024 Vacuum Technology Key Laboratory Fund Project for its generous support. The contributions made by this funding program were instrumental in the successful completion of this research.

REFERENCES

[1] Landers R E. Glass Fiber Reinforced Metal Pressure Vessel Design Guide. NASA-CR-120917. 1972. DOI: doi.org/10.2514/6.1972-120917.

[2] Landers R E. filament reinfored metal composite pressure vessel evaluation and performance demonstration. NASA-CR-134975. 1976. DOI: doi.org/10.2514/6.1976-134975.
[3] Morris E E, Darms F J, Lynn V, et al. Lower cost, high-performance composite fiberment tanks for spacecraft. AIAA-1986-1694. 1986. DOI: doi.org/10.2514/6.1986-1694.
[4] Morris E E, Segimoto M, Lynn V. Lighter weight fiber metal pressure vessels using carbon overwrap. AIAA-1986 -1504. 1986. DOI: doi.org/10.2514/6.1986-1504.
[5] Rabel H. Sustained load behavior of graphite epoxy metal-lined pressure vessels for long-life space applications, AIAA-1989-2644. 1986. DOI: doi.org/10.2514/6.1989-2644.
[6] Haddock R C, Darms F J. Space system applications of advanced composite fiber metal pressure vessels. AIAA-1990-2227. 1990. DOI: doi.org/10.2514/6.1990-2227.
[7] Haddock R C, Morris E E, Darms F J. Safety of filament-wound carbon epoxy composite pressure vessels for aerospace applications. AIAA-1991-2409. 1991. DOI: doi.org/10.2514/6.1990-2227.
[8] Braun C A. Manufacturing process controls for high reliability carbon filament-wound seamless aluminum lined composite pressure vessels. AIAA-1992-3609. 1992. DOI: doi.org/10.2514/6.1992-3609.
[9] Gleich D. Design Considerations and Structural Performance of Prestressed Composite Pressure Vessels. AIAA-1982-1230. 1982. DOI: doi.org/10.2514/6.1982-1230.
[10] Gleich D. Continuing report on Kevlar-49 overwrap study composite pressure vessel safety program. AIAA-1988 -2849. 1988. DOI: doi.org/10.2514/6.1988-2849.
[11] Sneddon K, Saunders S, Devey R, et al. Development and qualification of the Eurostar 2000+ helium COPV. AIAA-1997-3033. 1997. DOI: doi.org/10.2514/6.1997-3033.
[12] Escalona A. Design of high performance load sharing lined COPVfor atlas centaur. AIAA-1997-3031. 1997. DOI: doi.org/10.2514/6.1997-3031.
[13] Sneddon K, Anselmo E, Gorin B. Design, Qualification, and Thermal Testing of a Specialized Composite Overwrap Pressure Vessel (COPV). AIAA-2007-5558. 2007. DOI: doi.org/10.2514/6.2007-5558.
[14] Ray D M, Greene N J, Revilock D, et al. High Pressure Composite Overwrapped Pressure Vessel Development Tests at Cryogenic Temperatures. AIAA-2008-1912. 2008. DOI: doi.org/10.2514/6.2007-5558.
[15] Tam W H, Ballinger A, Kuo J, et al. Design and manufacture of a composite overwrapped xenon conical pressure vessel. AIAA-1996-2752. 1996. DOI: doi.org/10.2514/6.1996-2752.
[16] Tam W H, Jcakson A C, Nishida E. Design and manufacture of the ETS VIII xenon tank. AIAA-2000-3677. 2000. DOI: doi.org/10.2514/6.2000-3677.
[17] Tam W H. Design and Manufacture of a Composite Overwrapped Elastomeric Diaphragm Tank. AIAA-2004-3507. 2004. DOI: doi.org/10.2514/6.2004-3507.
[18] Tam W H, Ballinger A. Conceptual design of space efficient tanks. AIAA-2006-5058. 2006. DOI: doi.org/10.2514/6.2006-5058.
[19] Babel H, Ledesma L G. Delta IV COPV risk mitigation composite overwrapped pressure vessels. AIAA-1999-2832. 1999. DOI: doi.org/10.2514/6.1999-2832.
[20] Ledesma L G. Comparison of stress-rupture life prediction techniques for composite pressure vessels. IAF. 2000-1.6.01. 2000. DOI: doi.org/10.2514/6.2000-1.6.01.
[21] Abdi F, Israeli L, Johnson S, et al. Composite tank permeation prediction-and verification. AIAA-2003-1760. 2003. DOI: doi.org/10.2514/6.2003-1760.
[22] Robinson M J, Johnson S E. Trade study results for a second-generation reusable launch vehicle composite hydrogen tank. AIAA-2004-1932. 2004. DOI: doi.org/10.2514/6.2004-1932.
[23] Arritt B J, Wegner P M, Fosness E R, et al. Composite tank development efforts at the Air Force Research Laboratory Space Vehicles Directorate. AIAA-2001-4606. 2001. DOI: doi.org/10.2514/6.2001-4606.
[24] Mallick K, Cronin J, Ryan K, et al. An Integrated Systematic Approach to Linerless Composite Tank Development. AIAA-2005-2089. 2005. DOI: doi.org/10.2514/6.2005-2089.
[25] Bechel V T, Kim R Y, Donaldson S L. Composites Containing Barrier Layers for Reduced Permeability at Cryogenic Temperature. AFRL-2006-0455. 2006. DOI: doi.org/10.2514/6.2006-0455.
[26] Falugi M A. Minimum Weight Design of Cryogenic Tank Configurations for a Space Operating Vehicle (SOV). AIAA-2007-6199. 2007. DOI: doi.org/10.2514/6.2007-6199.
[27] Revilock D M, Thesken J C, Schmidt T E, et al. Three-dimensional digital image correlation of a composite overwrapped pressure vessel during hydrostatic pressure tests. NASA/TM-214938. 2007. DOI: doi.org/10.2514/6.2007-214938.
[28] Murthv P L, Gvekenvesi J P, Ledesma L G, et al. Reliability considerations for composite overwrapped pressure vessels on spacecraft. AIAA-2007-2150. 2007. DOI: doi.org/10.2514/6.2007-2150.
[29] Murthv P L, Thesken J C, Phoenix S L, et al. Stress rupture life reliability measures for composite overwrapped pressure vessels. NASA/TM-2007-214848. 2007. DOI: doi.org/10.2514/6.2007-214848.
[30] Murthv P L, Phoenix S L. Designing of a fleet-leader program for carbon composite overwrapped pressure vessels. AIAA-2009-2517. 2009. DOI: doi.org/10.2514/6.2009-2517.
[31] Murthv P L, Phoenix S L, Ledesma L G. Fiber breakage model for carbon composite stress rupture phenomenon: theoretical development and applications. NASA/TM-2010-215831. 2010. DOI: doi.org/10.2514/6.2010-215831.
[32] Murthv P L. Reliability of COPVs accounting for margin of safety on design burst. NASA/TM-2012-217638. 2012. DOI: doi.org/10.2514/6.2012-217638.

[33] Emery L, Zisk E J. Health monitoring of reusable composite cryogenic tanks. AIAA-1995-1073. 1995. DOI: doi.org/10.2514/6.1995-1073.

[34] Achary D C, Biggs R W, Bouvier C G, et al. Composite development & applications for cryogenic tankage. AIAA -2005-2160. 2005. DOI: doi.org/10.2514/6.2005-2160.

[35] Veys R B, Cederberg A R, Dale Tiller B. Design and development of an aluminum lined composite overwrapped spherical pressure vessel. AIAA-1989-2642. 1989. DOI: doi.org/10.2514/6.1989-2642.

[36] Veys R B, Cederberg A R, Schimenti J D. Design and analysis techniques for composite pressurant tankage with plastically operating aluminum liners. AIAA-1990-2345. 1990. DOI: doi.org/10.2514/6.1990-2345.

[37] Veys R B, Cederberg A R, Schimenti J D. Fatigue analysis techniques for composite tankage with plastically operating aluminum liners. AIAA-1991-1974. 1991. DOI: doi.org/10.2514/6.1991-1974.

[38] Murray C F, Newhouse N L, Khawand G. All-composite pressurant tanks for aerospace applications. AIAA,1993-2095. 1993. DOI: doi.org/10.2514/6.1993-2095.

[39] ANSI/AIAA S-081-2000. Space Systems- Composite Overwrapped Pressure Vessels (COPVs). 2000.

[40] ANSI/AIAA S-080. Space systems- metallic pressure vessels, Pressurized structures, and pressure components. 1998.

[41] Grant J, Kaul R, Taylor S, et al. Distributed Sensing of Carbon-Epoxy Composites and Filament Wound Pressure Vessels using Fiber-Bragg Gratings. International Society for Optics and Photonics, 2002. DOI: 10.1117/12.469076.

[42] Grant J. Optical Sensing using Fiber Bragg Gratings for Monitoring Structural Damage in Composite Over-wrapped Pressure Vessels. Conference on Photorefractive Fiber and Crystal Devices: Materials, Optical Properties, and Applications. NASA Marshall Space Flight Center Huntsville, AL 35759, 2005.

[43] Beeson H D, Payne K S, CHANG J B. Effects of Impact Damage and Exposure on Graphite/Epoxy Composite Overwrapped Pressure Vessel. AIAA-1996-3264. 1996. DOI: doi.org/10.2514/6.1996-3264.

[44] Greene N, Yoder T, Saulsberry R, et al. Stress Rupture Testing and Analysis of the NASA WSTF-JPL Carbon Overwrapped Pressure Vessels. Aiaa/asme/asce/ahs/asc Structures, Structural Dynamics, & Materials Conference. 2013. DOI: 10.2514/6.2007-2323.

[45] Greene N, Saulsberry R, Yoder T. et al. Testing of Full-scale Flight-qualified Kevlar Composite Overwrapped Pressure Vessels. AIAA 2007-2326. 2007. DOI: doi.org/10.2514/6.2007-2326.

[46] Abraham A R A, Johnson K L, Nichols C T, et al. Use of Statistical Analysis of Acoustic Emission Data on Carbon-Epoxy COPV Materials-of-Construction for Enhanced Felicity Ratio Onset Determination. 2012. DOI: 10.13140/RG.2.1.1627.6005.

[47] Radtke D W, Krings H, Forster R. et al. Manufacturing techniques for lightweight tankage. AIAA-1988-2580. 1988. DOI: doi.org/10.2514/6.1988-2580.

[48] Radtke D W. Novel Manufacturing Methods for Titanium Tanks and Liners. AIAA-2006-5269. 2006. DOI: doi.org/10.2514/6.2006-5269.

[49] Charpentier P. Aerospatiale Pressurant Tank Range Widens: A 35 liter Tank Has Been Qualified For The EUROSTAR Platform, AIAA-1984-1017: 1-9. DOI: doi.org/10.2514/6. 1984-1017.

[50] Charpentier P, Zorzetto D. High Performance Titanium Carbon Composite Gas Storage Pressure Vessel for Space Use. AIAA-1990-2223.1990. DOI: doi.org/10.2514/6.1990-2223.

[51] Teissier A, Liquide A. Liquid helium storage for Ariane 5 main stage oxygen tank pressurization. AIAA-1995-2956.1995. DOI: doi.org/10.2514/6.1995-2956.

[52] Morino Y, Shimoda T, Morimoto T. et al. Feasibility Study of the Cryogenic Composite Tank for Reusable Launch Vehicles.IAF-99-I3.06. 1999.

[53] Torano Y, Arita M, Takahashi H. et al. Current Study Status of the Advanced Technologies for the J-I Upgrade Launch Vehicle-Cryogenic Composite Tank. AIAA-2001-1877. 2001. DOI: doi.org/10.2514/6.2001-1877.

[54] Ishikawa T, Morimoto T, Yokozeki T. Pressurization of CF/Epoxy Model Tank at LN2 Temperature and Identification of Leak Path Formation in Tank Wall. AIAA-2004-1839. 2004. DOI: doi.org/10.2514/6.2004-1839.

[55] Masuda I, Masuda T, MOURI K. et al. Development of new composite propellant tank for satellites. AIAA-2014-3995. 2014. DOI: doi.org/10.2514/6.2014-3995.

[56] Hwang T K, Hong C S, Kim C G. Size effect on the fiber strength of composite pressure vessels. Composite Structures, 2003, 59: 489-498. DOI: 10.1016/S0263-8223(02)00250-7.

[57] Hwang T K, Hong C S, Kim C G. Probabilistic deformation and strength prediction for a filament wound pressure vessel. Composites, 2003, B34: 481-497. DOI: 10.1016/S1359-8368(03)00021-0.

[58] ZHANG T P, LIU Z D, YANG F Q, et al. A Design and its identification of composite overwrapped helium tank applicable to propulsion system of satellites. Aerospace SHANGHAI, 2006(03): 41-48. DOI: 10.19328/j.cnki.1006-1630.2006.03.010.

[59] LI Y F, JING Q C, LIU ZD. Filament-wound design and ANSYS analysis for composite material high-pressure vessel. Journal of Propulsion Technology, 2013, 34(7): 968-976. DOI: 10.13675/j.cnki.tjjs.2013.07.018.

[60] YU B, LIU Z D, JIN Q C, et al. A Structure design method for composite over-wrapped helium vessel in propulsion system of satellites. Journal of Propulsion Technology, 2013, 34(5): 680-686. DOI: 10.13675/j.cnki.tjjs.2013.05.016.

[61] YU B, ZHANG J J, GU S D, et al. Structure design study of composite overwrapped pressure vessel with plastically working liner. Journal of Propulsion Technology, 2022, 43(2): 1-9. DOI: 10.13675/j.cnki.tjjs.200364.

[62] YU B, HUANG C, YANG W B, et al. Development technique of large-scale composite tank for space-vehicle. Pressure vessel technology, 2022, 32(9): 77-86. DOI: 10.3969 /j.issn.1001-4837.2022.09.011.
[63] YU B, MA T J, HUANG C, et al. Research progress of active cooling thermal coupling technology foraerospace cryogenic propellant tanks. Vacuum and Cryogenics, 2023, 29(5): 518-524. DOI: 10.3969/j.issn.1006-7086.2023.05.012.
[64] YU B, HUANG C, MA T J, et al. Research progress and key technology analysis of satellite cryogenic tank storage and supply technology(1). Pressure vessel technology, 2024, 41(4): 60-71. DOI: 10.3969/j.issn.1001-4837.2024.04.008.
[65] YU B, HUANG C, MA T J, et al. Research progress and key technology analysis of satellite cryogenic tank storageand supply technology(2). Pressure vessel technology, 2024, 41(5): 67-79. DOI: 10.3969/j.issn.1001-4837.2024.05.009.
[66] WANG C C, ZHUANG CP, YAN C J, et al. Design and experimental study of 6 W@ 80 K space coaxial vascular tube cryocooler. Cryogenics, 2024(02): 1-7. DOI: 1000-6516(2024)02-0001-07.
[67] QU F J, YAN C J, LUO X K, et al. Research on phase modulation performance of dual-diameter inertance tube in pulse tube cryocooler. Cryogenics, 2023(05): 1-7. DOI: 1000-6516(2023)05-0001-07.
[68] PAN Q Y, WANG H R, WANG B, et al. Advances in piezoelectric-driven miniature regenerative refrigerators, 2022, 50(08): 1-7. DOI: 10.16711/j.1001-7100.2022.08.001.
[69] LI Y F, ZHAO J P, ZHANG H, et al. Effects of high-low temperature on properties of carbon fiber/epoxy resin composite high-pressure vessel. Engineering plastics application, 2024, 52(09): 119-125. DOI: 10.3969/j.issn.1001-3539.2024.09.018.
[70] TENG S Y, YAN C J, LI L S, et al. Experimental Study of a Thermoelectric Conversion System for Low and Medium Temperature Heat Utilization in Space. Journal of Space Science and Experiment, 2025, 2(02): 91-97. DOI: 10.19963/j.cnki.20974302.2025.02.010.

Journal of Manufacturing Science and Mechanical Engineering
ISSN: 2959-989X
DOI: https://doi.org/10.61784/msme3021

HISTORY OF EXPLORATION & PRODUCTION AND THE OIL COMPANIES

ZhiHao Zhu
Kings' School Al Barsha, Dubai 38199, United Arab Emirates.
Corresponding Email: ray.zhuzhihao@gmail.com

Abstract: This paper provides a comprehensive historical analysis of the oil exploration and production industry, examining its evolution through the intertwined lenses of corporate strategy, geopolitical conflict, and technological innovation. It argues that the sector's development has been a powerful force in shaping the modern global economy and international relations, while simultaneously being profoundly shaped by external political, economic, and environmental pressures. The study employs a dual-case framework for in-depth illustration: the corporate history of BP serves to trace the strategic adaptations and challenges of a major international oil company, from its imperial origins to its contemporary engagement with the energy transition. Concurrently, the Rumaila oil field in Iraq is analyzed as a microcosm of the industry's technical, operational, and geopolitical dimensions, highlighting its enduring significance and modernization. The analysis spans key historical phases, including the formative early discoveries, the rise of OPEC and the use of oil as a geopolitical weapon, and the persistent link between hydrocarbon resources and regional conflicts. It then investigates the contemporary pressures compelling transformation, particularly the dual imperatives of technological efficiency and climate responsibility. The discussion projects into the future of upstream operations, evaluating how digital technologies like artificial intelligence are revolutionizing exploration and production, and how climate-responsive strategies—including carbon management and investments in cleaner energies—are redefining corporate missions. Ultimately, this paper presents the oil industry not as a static entity, but as a dynamic and adaptive sector navigating an irreversible shift toward a lower-carbon global energy system.
Keywords: Oil exploration; BP; Rumaila; OPEC; Upstream technology; Climate change; Energy transition

1 INTRODUCTION

The discovery and exploitation of oil have fundamentally reconfigured the contours of global civilization, serving as a primary catalyst for economic modernization, geopolitical realignment, and profound social change. As the lifeblood of industrialization and a strategic commodity of unparalleled influence, oil has not only powered engines and economies but has also shaped international alliances, triggered conflicts, and defined the fortunes of nations and corporations alike. This paper provides a comprehensive analytical overview of the oil industry's dynamic evolution, tracing its trajectory from the pioneering discoveries of the early 20th century to the complex challenges and technological innovations that characterize the contemporary era[1]. By integrating historical narrative with structural analysis, the study illuminates how the sector has been shaped by—and has in turn shaped—key historical milestones, transformative corporate strategies, decisive geopolitical events, and successive waves of technological advancement. Through this multidisciplinary lens, the paper seeks to offer a coherent framework for understanding the oil industry not merely as an economic sector, but as a central force in the making of the modern world, while critically examining its ongoing adaptation in an age defined by climate imperatives and energy transition[2].

2 HISTORICAL DEVELOPMENT OF OIL EXPLORATION & PRODUCTION

2.1 Brief BP History

1908: British-controlled Anglo-Persian Oil Company (APOC) (later known as BP) is formed after British geologists find the first significant oil well in Persia, modern-day Iran.
1913: British Navy switched from coal production to oil production under Winston Churchill's suggestion.
1953: BP, as part of Iraq Petroleum Company (IPC), discovered oil in the Rumaila prospect of Southern Iraq.
1979: BP lost 40% of its global crude oil supplies after Iranian Islamic Revolution.
2005: 15 BP workers were killed with more than 170 injured in Texas City Refinery explosion.
2010: The Deepwater Horizon oil spill killed 11 people & leaked about 4.9 million barrels of oil into Gulf of Mexico.
2017: BP announced acquiring a 43% stake in the solar energy developer Lightsource Renewable Energy, renamed Lightsource BP.

3 GENERAL HISTORY OF OIL PRODUCTION & EXPLORATION

1908: British-controlled Anglo-Persian Oil Company (APOC) (later known as BP) finds the first significant oil well in Persia, modern-day Iran-> Western countries would start their own respective exploitations of the resource

1913: British Navy switched from coal production to oil production under Winston Churchill's suggestion, making global powers recognize that oil could be a strategic weapon in terms of military importance.
Post WWI: immense oil reserves discovered in countries like Saudi Arabia (1938), Kuwait (1938), and Iraq (1927) underscore Middle East's importance in global energy.
WWII: The resource's importance was underscored in this time with oil being used in modern warfare e.g. powering tanks, ships, aircrafts. Further could be seen in Germany's decision in attempting to invade Russia being partly due to its intention of gaining the massive oil reserves in the Caucasus region.
1950s: Middle East supplied over 50% of the world's oil reserves-> brought power and wealth to those countries yet also brought exploitation from Seven Sisters which controlled majority of the oil production there so they took most profits whilst unethically leaving local population with little benefit.
Suez Crisis 1956:
-Egypt's President Gamal Abdel Nasser in July nationalized the Suez Canal—then controlled by British and French interests—to finance the Aswan High Dam, provoking outrage in Europe (history.state.gov) due to it being a key oil transportation route
-late October, Israel invaded Sinai and advanced toward the canal, followed by coordinated British and French military intervention, aiming to regain control (britannica.com).
-However, international pressure from US (due to its fear invasion would push Arab nations in general to the Soviet Union) and Soviet Union made above nations withdraw forces.
1960s: OPEC (made up Iran, Iraq, Saudi Arabia, Kuwait) created to take back control of oil pricing from Western Corporations, proved to be effective in 1973 when in response to US support for Israel during Yum Kipur War, an oil embargo was imposed upon-> oil prices quadrupled within months-> global economic crisis with shortages & inflation in European countries
1970-1980: Saudi Arabia oil revenues increased from $1.3 billion-110 billion, allowing it to fund social projects, military expansions etc. during prices soaring but when prices crashed (e.g. mid 1980s) there was social unrest, political instability etc.(example of a country fallen under the resource curse)
1991 Gulf War: after Saddam Hussein invaded and annexed Kuwait in August 1990 to gain oil reserves, making him own 20% of world's oil reserves-> U.S.-led coalition of 35 nations, backed by a UN mandate to have an air and ground campaign to liberate Kuwait but also not let those oil reserves fall into Saddam Hussein's hands. War ended in February 1991 (millercenter.org).
Also fueled internal conflicts within Middle East: e.g. Iraq disputes over oil revenue sharing between central government and Kurdish Regional government.
Recently global demand for oil decreases due to increased awareness of negative externalities on environment-> middle eastern countries need to shift to other industries for economic wellbeing e.g. Saudi Arabia's Vision 2030 aim to reduce reliance on the resource by investing in technology, tourism and renewable energy[3-4].

4 RUMAILA OIL FIELD – KEY FACTS

1. Has massive impact: Rumaila is one of the largest onshore oil fields in the world, produces around 1.5 million barrels of oil every day = one-third of all the oil Iraq makes.
2. Long History: Discovered in 1953, oil has been pumped from the ground since 1954. yet there are still around 17 billion barrels of oil left underground today.
3. Modern Advancements: A special group called the Rumaila Operating Organisation (a partnership between Iraq's oil company, BP, and PetroChina) took charge in 2010, drilling over 300 new wells, installed new equipment with smarter digital technology-> increase daily production by around 40%, making the field much more efficient

5 TECHNOLOGICAL INNOVATION AND THE FUTURE OF UPSTREAM OPERATIONS

5.1 The Role of Artificial Intelligence

Reduces exploration costs: AI algorithm can analyse geological data and identify potential oil and gas reserves more efficiently, reducing the time-consuming and expensive process of finding such reserves where machines equipped with infrasound technology are used[5].
Optimize Production: It can predict equipment failures, suggest when and where maintenance is needed-> reduces downtime
Improve Safety: It can help detect potential safety hazards and alert workers to take corrective action. It also makes drilling for oil more safe as well as effective with autonomous drilling systems.

5.2 Climate Change and Industry Responsibility

Acknowledging Responsibility: Oil and gas companies, especially those in the Oil and Gas Climate Initiative (OGCI), openly recognize their role in contributing to climate change. These companies produce about 30% of the world's oil and gas, so their emissions are a big part of the problem—and they understand they must be part of the solution.

Setting Measurable Climate Goals: OGCI companies have set clear climate targets. One major goal is to reduce one gigaton of CO_2 emissions by 2025, mainly through better technology and cooperation. Some member companies are even setting net-zero emissions targets for their operations[6-8].

Using Smart Technologies: To meet these goals, oil and gas firms are using tools like:

• Carbon Capture and Storage (CCS): Captures CO_2 from industrial sites and stores it underground.

• Methane Detection & Reduction: Using satellites and sensors to find and fix methane leaks (methane is a powerful GHG).

• Energy Efficiency: Improving how energy is used during exploration and drilling.

• Switching to Hydrogen and Cleaner Fuels: Investing in future fuels that emit less or no carbon.

6 THE OIL PRODUCTION PROCESS: FROM EXPLORATION TO REFINEMENT

6.1 Exploration & Drilling (Upstream)

• Geologists use seismic surveys to locate oil and natural gas trapped deep underground.

• Once a promising site is found, drilling rigs are set up to bore through rock layers to reach the reservoir.

• The drilled well is fitted with steel casing and a valve system (called a Christmas tree) to safely control the flow of oil—and often natural gas—to the surface.

6.2. Separation and the LNG Process

• At the surface, crude oil is separated from natural gas, water, and sand [9].

• The natural gas is cleaned by removing water, CO_2, and impurities [10].

• Then, it's cooled to -162°C in a special facility, turning it into a liquid (LNG) for easier transport and storage [11].

• This LNG is then shipped around the world in insulated tankers and turned back into gas at its destination—used in homes, power stations, and factories.

6.3 Transport & Refining

• The separated oil is stored and moved through pipelines, trucks, or oil tankers to refineries.

• At the refinery, with the help of fractional distillation it's processed into everyday products such as petrol, diesel, jet fuel, heating oil, and plastics, using techniques like distillation and cracking.

7 CONCLUSION

The oil industry continues to serve as a fundamental pillar of the global energy system, supplying the bulk of the world's transportation, industrial, and chemical feedstock needs. Nevertheless, it currently navigates a period of profound transformation, driven by three intersecting forces: escalating environmental imperatives, volatile geopolitical landscapes, and rapid technological disruption. Mounting climate change evidence and international accords such as the Paris Agreement have intensified scrutiny on carbon emissions, compelling the sector to reconcile its operational legacy with the urgent need for decarbonization. Simultaneously, geopolitical tensions—from regional conflicts to shifting energy alliances—continue to threaten supply stability and influence market dynamics, reminding the world of oil's enduring strategic significance.

Major integrated companies like BP embody this dual reality: they are inheritors of a century-old hydrocarbon economy, yet increasingly signal a strategic pivot toward sustainable energy and digital innovation. Their evolving portfolios—spanning upstream oil, natural gas, renewables, and carbon management technologies—reflect an industry in transition, albeit at a pace debated by stakeholders. Looking ahead, the future of upstream operations will be shaped by the integration of smarter technologies such as AI, IoT, and advanced data analytics, which enhance efficiency, safety, and predictive capabilities. Furthermore, stricter climate regulations and growing investor pressure for ESG compliance will drive greater transparency, emission reductions, and investment in low-carbon solutions.

Ultimately, the industry's trajectory will depend on its ability to adapt within an increasingly carbon-conscious global framework. While oil will remain relevant in the medium term, its role is gradually being redefined—from a dominant energy source to one component of a more diversified, resilient, and sustainable energy mix. The challenge and opportunity for oil companies lie in balancing legacy assets with forward-looking innovation, ensuring energy security while contributing meaningfully to a cleaner energy future.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Varma V A, Dutta P. Artificial Intelligence in Upstream Oil and Gas: A Review of Applications and Trends. SPE Journal, 2022, 27(03): 1677-1694.

[2] Le Billon P. The Political Ecology of War: Natural Resources and Armed Conflicts. Political Geography, 2001, 20(5): 561–584.

[3] Arif U, Siddiqui SU, Khan MF, et al. Invigorative potential assessment of priming agents through germination and various seedling attributes of artificially aged seeds in carrot (Daucus carota L.). Pakistan Journal of Botany, 2024, 56(6). DOI: http://dx.doi.org/10.30848/PJB2024-6(1).

[4] Vasanthi N, Suresh M. Artificial Intelligence in Oil and Gas Exploration: A Review. Journal of Petroleum Exploration and Production Technology, 2021, 11(2): 989–1002.

[5] Al-Moneef M. The Contribution of the Oil Sector to Arab Economic Development. OPEC Review, 2006, 30(4): 273–287.

[6] Khilji SA, Sajid ZA, Jamil MS. Amelioration of salinity stress in Brassica campestris L.by exogenous application of KH2PO4. Pakistan Journal of Botany, 2024, 56(6). DOI: http://dx.doi.org/10.30848/PJB2024-6(4).

[7] BP. BP Statistical Review of World Energy 2020 (69th ed.). BP plc, 2020.

[8] Imran M, Ahmad I, Ali A, et al. Modifications in anatomical traits of tall flat sedge Cyperus exaltatus Retz. colonizing differently salt-affected wetlands. Pakistan Journal of Botany, 2024, 56(6). DOI: http://dx.doi.org/10.30848/PJB2024-6(9).

[9] Yergin D. The Prize: The Epic Quest for Oil, Money, and Power (Reissued ed.). Simon & Schuster, 2008.

[10] Kuuskraa V A, Stevens S H. The role of CCUS in low-carbon power systems. Energy Procedia, 2019, 158: 4281-4288.

[11] Jafarzadeh A, Acheampong M, Hossain M E. A review on artificial intelligence applications in petroleum exploration and production. Journal of Petroleum Science and Engineering, 2021, 203: 108641.