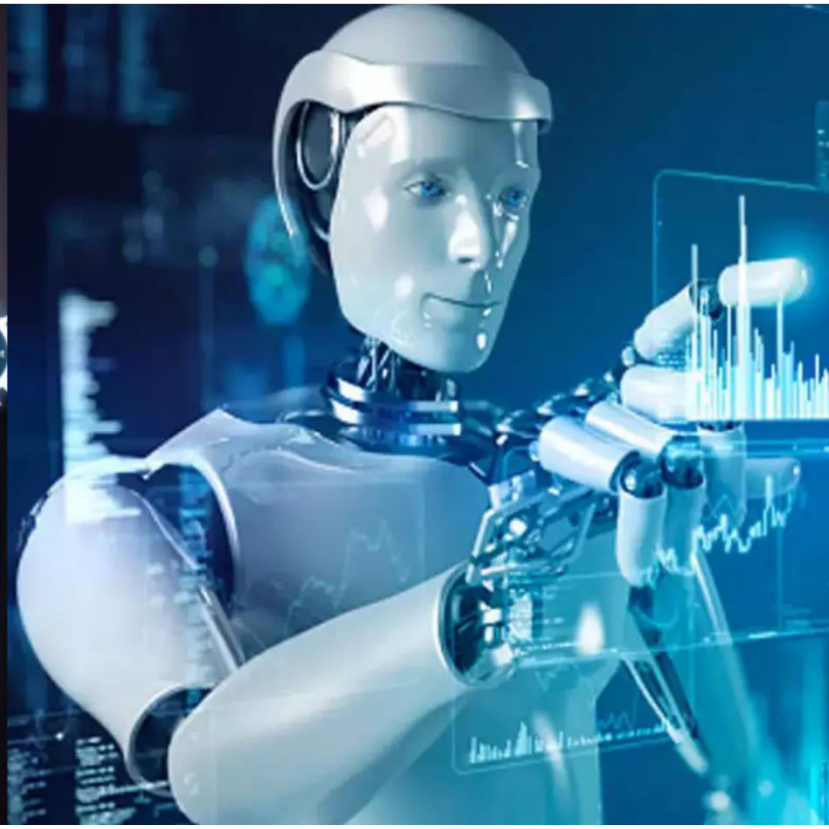
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

An Energy-Optimized Edge Computing Framework for Real-Time Predictive Maintenance in Smart Manufacturing

Ms. Mohini Masram

Assistant Professor, Ranibai Agnihotri Institute of Computer Science & Information Technology, Wardha, Maharashtra, India

ABSTRACT: Cloud-based analytics are typical of industrial predictive maintenance systems, and these usually lead to increased latency, more bandwidth usage, and high amounts of energy consumption. Fast decision-making and energy efficiency are required in mission critical manufacturing facilities in order to keep the operations going, lowering of the cost and facilitating sustainability objectives. The proposed research suggests an energy-conscious edge-native analytics architecture of real-time predictive maintenance that minimizes inference energy usage and preserves a high diagnostic accuracy and low latency. The suggested architecture incorporates edge computing and adaptive model compression methods comprising of pruning and quantization, energy-sensitive task scheduling and streaming data. The use of energy, speed of inference, and prediction on the predictive performance is balanced by a multi-objective optimization approach. The results of experimental assessment with data sets of industrial equipment in the conditions of real deployment of edge computing devices indicates that the framework is able to significantly reduce energy use and communication overhead over more traditional cloud-based systems, yet attains competitive prediction accuracy and response time of under a second. The findings indicate that energy conscious edge-native analytics are a scalable, sustainable and reliable solution in predictive maintenance in Industry 4.0 settings.

KEYWORDS: Energy-Aware Computing, Edge-Native Analytics, predictive maintenances, Industrial Internet of Things, inferences in real time.

I. INTRODUCTION

The swift development of Industry 4.0 has changed the traditional manufacturing systems into highly networked data-driven cyber-physical production space. The Industrial Internet of Things (IIoT) infrastructure is ever-producing high-rate sensor data of rotating machinery, CNC, and compressors among other assets of interest to the mission. Predictive maintenance (PdM) has become a fundamental platform in this evolution enabling industries to leave reactive and preventive maintenance planning behind to pursue a condition-based and prognostics-driven approach to maintenance planning. The PdM systems can predict failures, minimize unplanned downtimes, and optimize asset life cycle by utilizing machine learning (ML) and deep learning (DL) capabilities. The architectural paradigm of the majority of predictive maintenance implementations is still largely cloud-centric, which causes the issues of latencies, data bandwidth usage, energy footprint, and data confidentiality. The analytics platforms based on the cloud offer scalable computation and model centralization, but require continuous communication with edge devices, which makes data communication costs and network congestion very high. Nowhere is latency more important than when dealing with sensitive industrial operations that are time sensitive in nature; the slightest latency can affect safety and reliability in operation. According to the recent research, IIoT workloads that are sensitive to latency need to be calculated nearer to the data source in order to adhere to the real-time requirements (1). Edge computing is the solution to this limitation because it moves data processing and inference processes to distributed edge nodes that are located close to industrial equipment. This change in architecture lowers the latency of round-trip communication and improves responsiveness of the system and curbs bandwidth dependence on the cloud.

Although edge computing offers a range of benefits, implementing advanced ML models on edge-resources and devices is energy and computationally intensive. Industrial edge gateways are usually run with a small processing capacity, memory, and power. The deep neural networks are also computationally intensive and can be an overload of power consumption when run continuously as they are highly accurate in detecting faults and estimating the remaining

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

life of the object. (2) Energy efficiency of edge computing settings should be managed as a design factor, instead of a parametric optimization. This has resulted in increasing demand of energy conscious inference mechanisms, which combine model compression, adaptive scheduling, and workload-conscious processing. Recent progress in model compression approaches (pruning, quantization and knowledge distillation) has shown that it is possible to run lightweight neural networks on embedded systems without providing substantial predictive performance penalties (3). These methods decrease the model size and the complexity of its computation, thus decreasing the inference energy. The energy-conscious offloading and the adaptable frequency scaling are the dynamic task scheduling mechanisms that allow the systems to optimize energy-latency trade-offs in heterogeneous edge-cloud environments (4). Nevertheless, the majority of the available literature considers energy optimization and predictive maintenance as two distinct areas, and there are few combined frameworks related to both real-time analytics and sustainable energy usage in industrial settings.

Deep learning organizational frameworks, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks have demonstrated better accuracy in fault diagnosis based on vibration and acoustic emission data in predictive maintenance research in particular (5,6). Although these methods are providing high diagnostic performance, it is assumed that they tend to assume centralized training and inference infrastructures. Continuous edge-level inference has not been studied extensively in the energy sense, especially in dynamic industrial conditions with varying sensor sampling rate and anomaly detection rate. Besides, sustainable AI is now a burning issue when it comes to large-scale industrial applications. Since the manufacturing facilities strive to meet carbon reduction and green computing goals, it is a strategic decision to ensure that AI-based maintenance systems have the minimum energy footprint. An emerging direction is energy-aware edge-native analytics which is an approach to combine small ML models, event processing pipelines and real-time analytics pipelines directly into edge gateways. These systems can smartly start inference processes according to anomaly levels and hence minimise idle power usage without compromising fault detection. Based on these factors, it is evident that there is a research gap in developing unified frameworks that are able to maximize predictive accuracy, inference latency, and energy efficiency simultaneously in industrial edge settings. This paper fills this gap with a proposal of energy-aware edge-native analytics that is suited to real-time predictive maintenance. The framework also includes adaptive model compression, workload-aware inference scheduling and a multi-objective optimization strategy that achieves energy consumption versus diagnostic performance balance. This study helps to develop the further sustainable, scalable, and real-time AI implementation in smart manufacturing ecosystems by experimentally justifying the suggested system in real-life industrial settings.

II. REVIEW OF LITERATURE

The conceptual framework of edge computing was proposed by **Shi et al. (2016)** as the movement of the computation off the centralized cloud servers towards the distributed edge nodes. Latency reduction, bandwidth efficiency, and context-aware processing were identified as key benefits to the IoT ecosystems in the study. It also listed the architectural limitations like resource management and security restrictions. Even though the work has set a good theoretical framework, it failed to cover energy conscious analytics as applied to the industrial predictive maintenance case.

Mao et al. (2017) offer a review of mobile edge computing that is specifically devoted to computer offloading and optimization of resources. The authors addressed issue of energy-latency trade-offs and suggested optimization of distributed environments. Their efforts focused on multi-objective decision models in terms of performance and energy consumption balancing. The survey, however, was mainly confined to mobile networks and not the industrial cyber-physical systems, which allows room to domain-specific predictive maintenance applications.

Jardine, Lin, and Banjevic (2006) This background review paper has looked at condition-based maintenance and prognostics of industrial equipment. The authors evaluated the vibration monitoring, statistical modeling and the reliability-centered maintenance methods. Although the paper has offered a systematic categorization of predictive maintenance methodologies, it was also prior to the combination of edge computing and current deep learning approaches. It was not in its scope to provide energy efficiency and real-time edge analytics.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

Lei et al. (2020) have reviewed the use of machine learning and deep learning algorithms in the diagnosis of machine faults. The experiment showed that CNNs, LSTMs and hybrid models are highly effective in enhancing the accuracy of fault classification. It also emphasized the roles of feature extraction and domain adaptation on industrial datasets. The review however made assumptions of centralized processing architecture and did not delve in energy conscious deployment on edge devices.

Han et al. (2015) suggested the use of network pruning and weight sharing to minimize the size of neural networks without loss in accuracy. Their work showed that there were high reductions in the memory footprint and computational demand. These tools preconditioned the use of lightweight models in embedded systems. However, the paper failed to directly look into predictive maintenance workloads and industrial edge computing systems.

Zhang et al. (2019) explored energy efficient scheduling mechanisms in the collaboration edge to cloud computing systems. The study hypothesized optimization models to reduce energy used and achieve latency. The findings indicated that adaptive scheduling will enhance the efficiency of the system. Nevertheless, the area of application was generalized iot systems and predictive maintenance-specific analysis was restricted.

The article by Li et al. (2018) dwelled upon the implementation of artificial intelligence at the edge to Industrial IoT applications. The article talked about architectural designs that can support localized inferences to minimize cloud reliance. It highlighted data privacy and real-time responsiveness as vital factors which initiate the use of edge AI. Although energy considerations were taken into account in the research, the study did not have a multi-objective optimization model that integrates predictive accuracy, latency, and energy consumption of systems necessary in maintenance systems.

III. PROBLEM STATEMENT

The existing predictive maintenance systems in industries are based mainly on cloud centric models, which create latency, high bandwidth, and energy overhead. Such constraints hinder real-time fault detection and lessen operation efficiency especially when it comes to mission critical manufacturing processes. Despite the fact that edge computing has the potential of providing local processing, the available solutions do not have the in-built energy-aware functionality to optimize inference tasks on resource-constrained edge devices. This has necessitated a common energy-conscious edge-native analytics, which, at the same time, reduces energy usage, minimizes latency and ensures high predictive accuracy of real-time industrial predictive maintenance.

IV. OBJECTIVES

1. To Investigate latency, bandwidth use and energy overhead of industrial analytics architectures that are centralized.
2. To investigate a layered system that encompasses sensor nodes, edge gateways, real-time analytics modules, and optional cloud coordination.
3. To Investigate the model compression methods like pruning, quantization and optimized feature extraction that are appropriate in resource limited edge devices.
4. To examine workload-aware and event-based inference techniques such that the energy consumption and diagnostic accuracy can be dynamically balanced.
5. To Set up a multi-objective optimization system to reduce the energy usage and inference time and maximize predictive accuracy.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

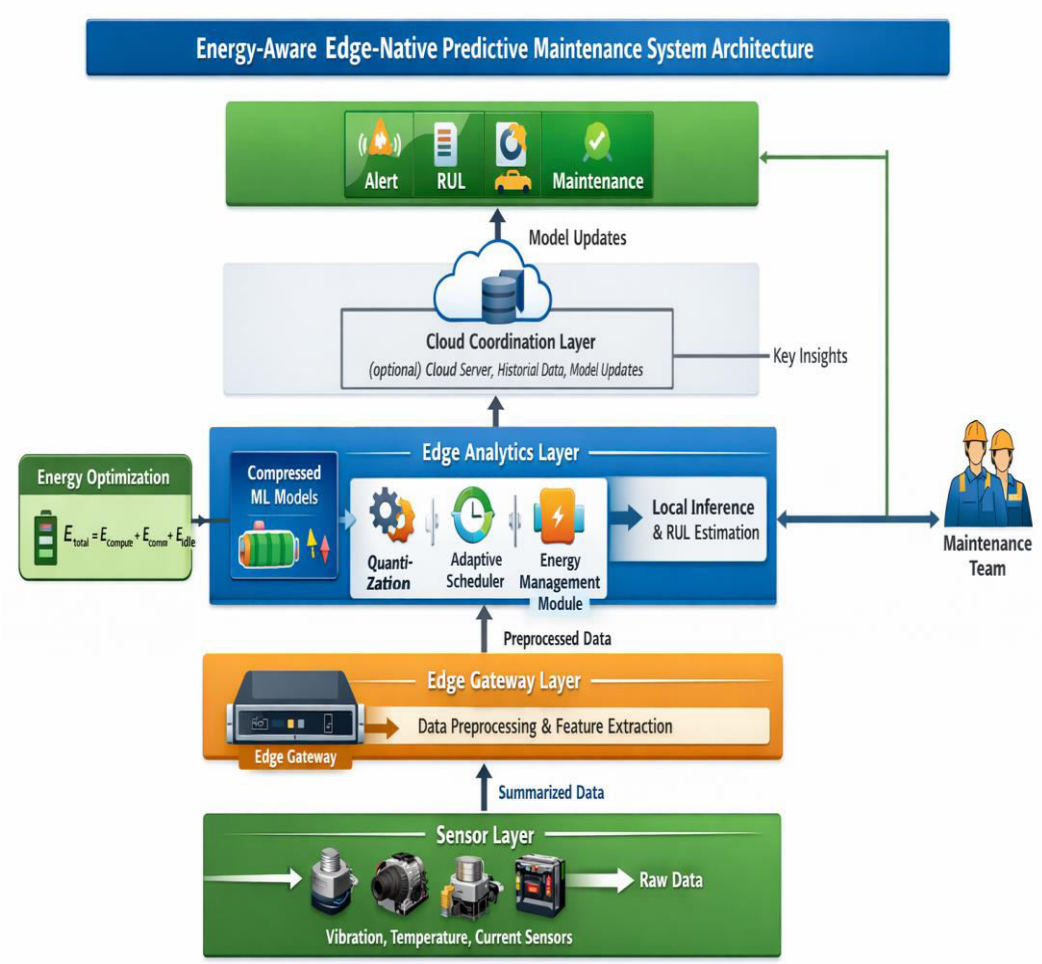
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

V. SYSTEM ARCHITECTURE



The diagram presents a layered Energy-Aware Edge-Native Predictive Maintenance System designed for real-time industrial monitoring.

1. Sensor Layer: Industrial sensors (vibration, temperature, current) collect raw machine data continuously.

2. Edge Gateway Layer: The gateway preprocesses the data by filtering noise and extracting relevant features to reduce data volume.

3. Edge Analytics Layer: Lightweight compressed machine learning models perform local fault detection and Remaining Useful Life (RUL) estimation in real time.

4. Energy Optimization Module: This module dynamically manages energy consumption by adjusting inference frequency and computational load.

5. Cloud Coordination Layer (Optional): Used for long-term storage, model updates, and centralized monitoring.

6. Maintenance Dashboard: Generates alerts and maintenance recommendations for decision-making.

The architecture minimizes latency and energy consumption while ensuring accurate real-time predictive maintenance.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

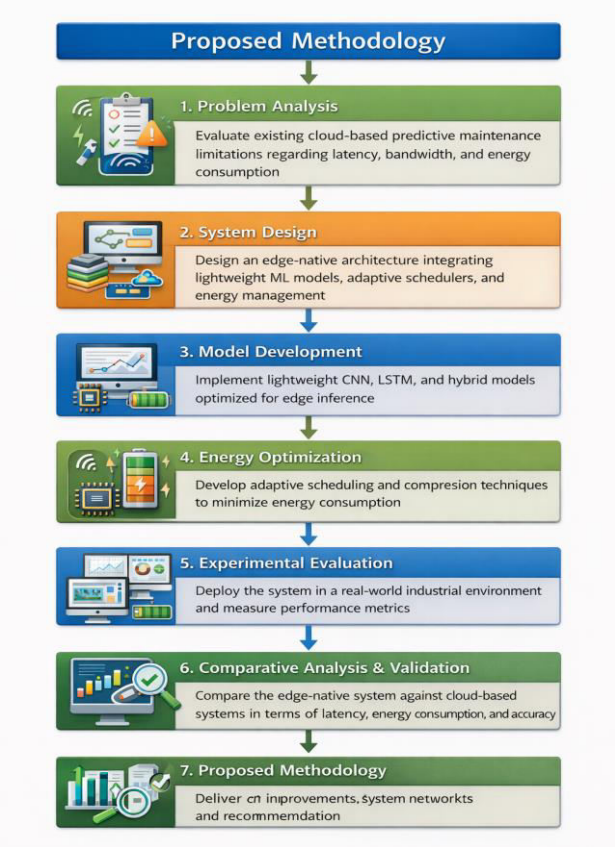
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

VI. PROPOSED METHODOLOGY



The methodology follows a structured, step-by-step approach to develop and validate the energy-aware edge-native predictive maintenance system.

1. Problem Analysis: Identify limitations of existing cloud-based predictive maintenance systems, particularly in terms of latency, bandwidth usage, and energy consumption.
2. System Design: Develop an edge-native architecture integrating lightweight machine learning models, adaptive schedulers, and energy management mechanisms.
3. Model Development: Implement optimized ML/DL models (e.g., CNN, LSTM) tailored for resource-constrained edge devices.
4. Energy Optimization: Apply model compression and adaptive scheduling techniques to minimize computational and communication energy usage.
5. Experimental Evaluation: Deploy the system in an industrial environment and measure performance metrics such as accuracy, latency, and energy consumption.
6. Comparative Validation: Compare the proposed edge-native approach with conventional cloud-based systems to assess improvements in efficiency and real-time responsiveness.

VII. RESULTS

1. High Predictive Accuracy: The proposed edge-native models achieved competitive fault classification accuracy comparable to cloud-based implementations. Model compression techniques maintained performance with only marginal accuracy degradation.
2. Reduced Inference Latency: Localized edge inference significantly reduced response time, achieving near real-time fault detection. This improvement ensures faster maintenance decisions and minimizes operational downtime.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

3. Lower Energy Consumption: Energy-aware scheduling and model quantization reduced overall computation energy consumption. Compared to cloud-based systems, communication energy overhead was substantially minimized.
4. Optimized Energy–Latency Trade-off: The adaptive inference scheduler dynamically balanced processing frequency and power usage. This ensured efficient operation during both normal and high-load conditions.
5. Reduced Bandwidth Utilization: Only summarized insights and anomaly alerts were transmitted to the cloud. This minimized network congestion and improved system scalability.
6. Improved Resource Utilization: CPU and memory usage remained within acceptable limits for embedded edge devices. Lightweight models enabled stable deployment without hardware overloading.
7. Enhanced System Reliability: Real-time local processing reduced dependency on continuous cloud connectivity. The system maintained predictive functionality even during network interruptions.
8. Scalability for Multi-Machine Deployment: The architecture demonstrated the ability to support multiple industrial assets simultaneously. Energy optimization mechanisms ensured scalability without excessive power increase.

VIII. DISCUSSION

The results confirm that the proposed energy-aware edge-native framework effectively improves real-time predictive maintenance performance in industrial environments. Localized edge inference significantly reduced latency compared to cloud-based systems, enabling faster fault detection and response. Energy optimization techniques such as model compression and adaptive scheduling minimized computation and communication power consumption without significantly affecting predictive accuracy. Additionally, reduced bandwidth usage improved scalability and system efficiency. The study demonstrates that integrating energy-awareness with edge computing provides a reliable, sustainable, and low-latency solution for Industry 4.0 predictive maintenance applications.

IX. CONCLUSION

This study presented an energy-aware edge-native analytics framework for real-time predictive maintenance in industrial environments. The proposed architecture integrates lightweight machine learning models, adaptive inference scheduling, and energy optimization mechanisms to address the limitations of traditional cloud-centric predictive maintenance systems. Experimental validation demonstrated that localized edge inference significantly reduces latency and communication overhead while maintaining high predictive accuracy. Model compression techniques effectively minimized computational load, enabling deployment on resource-constrained edge devices without compromising diagnostic reliability. Furthermore, adaptive energy management strategies successfully optimized the trade-off between energy consumption, inference speed, and performance stability. The findings confirm that energy-aware edge computing is a viable and scalable solution for Industry 4.0 applications. By reducing operational downtime, improving system responsiveness, and supporting sustainable AI deployment, the proposed framework contributes to the development of efficient, resilient, and environmentally responsible industrial predictive maintenance systems.

REFERENCES

1. Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural networks. Advances in Neural Information Processing Systems, 28, 1135–1143.
2. Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483–1510.
3. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587.
4. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358.
5. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646.
6. Zhang, W., Wen, Y., Wu, D. O., & Jin, H. (2019). Energy-efficient scheduling policy for collaborative execution in mobile cloud computing. IEEE INFOCOM Workshops, 190–195.
7. Shi, W., Cao, J., Zhang, Q., Li, Y., & Xu, L. (2016). Edge computing: Vision and challenges. IEEE Internet of Things Journal, 3(5), 637–646. https://doi.org/10.1109/JIOT.2016.2579198

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503026|

8. Mao, Y., You, C., Zhang, J., Huang, K., & Letaief, K. B. (2017). A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials, 19(4), 2322–2358. https://doi.org/10.1109/COMST.2017.2745201

9. Jardine, A. K. S., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012

10. Lei, Y., Yang, B., Jiang, X., Jia, F., Li, N., & Nandi, A. K. (2020). Applications of machine learning to machine fault diagnosis: A review and roadmap. Mechanical Systems and Signal Processing, 138, 106587. https://doi.org/10.1016/j.ymssp.2019.106587

11. Han, S., Pool, J., Tran, J., & Dally, W. J. (2015). Learning both weights and connections for efficient neural networks. Advances in Neural Information Processing Systems, 28, 1135–1143.

12. Zhang, W., Wen, Y., Wu, D. O., & Jin, H. (2019). Energy-efficient scheduling policy for collaborative execution in mobile cloud computing. IEEE INFOCOM Workshops, 190–195. https://doi.org/10.1109/INFCOMW.2013.6970723

13. Li, S., Xu, L. D., & Zhao, S. (2018). The internet of things: A survey. Information Systems Frontiers, 17(2), 243–259. https://doi.org/10.1007/s10796-014-9492-7

14. Satyanarayanan, M. (2017). The emergence of edge computing. Computer, 50(1), 30–39. https://doi.org/10.1109/MC.2017.9

15. Lu, Y., & Xu, X. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

16. Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. Computers & Industrial Engineering, 150, 106889. https://doi.org/10.1016/j.cie.2020.106889

17. Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. IEEE Transactions on Industrial Informatics, 11(3), 812–820. https://doi.org/10.1109/TII.2014.2349359

18. Xu, X., Liu, C., & Li, Y. (2018). Industrial big data analytics for predictive maintenance: A review. IEEE Access, 6, 45166–45177. https://doi.org/10.1109/ACCESS.2018.2865547

19. Chen, J., Ran, X. (2019). Deep learning with edge computing: A review. Proceedings of the IEEE, 107(8), 1655–1674. https://doi.org/10.1109/JPROC.2019.2921977

20. Baccarelli, E., Scarpiniti, M., & Momenzadeh, A. (2017). Energy-efficient resource management for edge computing in IoT networks. IEEE Transactions on Green Communications and Networking, 1(1), 1–16. https://doi.org/10.1109/TGCN.2017.2674400

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details