International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Energy-Aware Predictive Maintenance in Industrial Edge Systems Using Pruned LSTM Networks for Sensor-Based Time-Series Data

[1]Ali Mohamed Ali Annas, [1]Mohamed Ali Mohamed Ali Abdulkader, [1]Huda Abdussalam Ali Abdulla, [1]Massoud Ali Abdalhadi, [1]Abdussalam Mohamed Ali Altaher

[1]Department of Information Technology, Higher Institute of Science and Technology, Wadi Al-Ajal, Libya.

Abstract

In modern industrial environments, predictive maintenance has become a vital strategy for ensuring operational reliability, reducing downtime, and optimizing energy utilization. However, existing deep learning (DL) approaches such as CNN, GRU, and hybrid architectures, while accurate, often suffer from high computational complexity and energy consumption, making them unsuitable for real-time edge deployment. To address these limitations, this study proposes an Energy-Aware Predictive Maintenance framework using Pruned LSTM Networks for sensor-based time-series data, designed specifically for industrial edge systems. The model employs structured pruning techniques to reduce redundant parameters and computational overhead while preserving the temporal learning capability of LSTM. The proposed approach was implemented using Python and TensorFlow on the Kaggle Industrial Equipment Monitoring Dataset, which contains multi-sensor readings representing normal and faulty machine states. Experimental results show that the Pruned LSTM model achieves a 98.8% accuracy, marking an increase of approximately 6–7% over conventional models like GRU and CNN, while reducing energy consumption by nearly 40% compared to baseline methods. This improvement demonstrates the model's ability to maintain high precision and reliability under resource constraints. The proposed framework establishes a strong foundation for real-time edge-based predictive analytics, offering both energy efficiency and predictive robustness. In the future, the model will be extended with adaptive transfer learning and federated edge optimization to enable scalable and cross-domain industrial applications, driving the next generation of intelligent and sustainable maintenance systems.

Keywords: *Predictive Maintenance, Energy-Aware Computing, Pruned LSTM Networks, Industrial Edge Systems, Sensor-Based Time-Series Data*

1. Introduction

Industry 4.0 has changed the landscape of industry development because of the introduction of intelligent systems, sensor technologies, and real-time analytics as efficient asset management tools [1]. Predictive Maintenance is one of such innovations that have provided a ground-breaking solution in predicting equipment failures prior to their actual occurrence hence reducing downtimes, enhancing the level of productivity and lowering maintenance expenses [2]. The fast adoption of the Internet of Things (IoT) devices and industrial sensors has produced a huge amount of time-series data, which can help to obtain valuable information regarding the health of the machine [3], [4]. The effective analysis of this data is important in the prediction of faults on time and the planning of maintenance in industrial systems as energy-efficient [5]. Innovations in Predictive Maintenance in recent years have been based primarily on DL models: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRU) and Transformer based models. Such approaches have been incredibly precise at finding patterns using sensor data [6], [7].

Their computation complexity, large memory footprint and energy demands are however a limitation to their use in resource constrained edge computing environments [8]. In addition, such models tend to be sensitive to noisy or imbalanced data typically occurring in the industrial setting, which causes lower prediction accuracy and ineffective energy consumption [9]. Although it is expected to be excellent in cloud-based configurations, the latency and overhead of the communication restrict its practical use at the industrial level in real-time [10]. In order to address these issues, the current study proposes an Energy-Aware Predictive Maintenance framework based on the Pruned LSTM Networks on Sensor-Based Time-Series Data. The model proposed combines the systematic pruning that removes redundant network parameters in order to greatly reduce computational costs and power usage, without affecting the predictive performance. When it is deployed at the edge with the optimized model, it can be used to provide real-time fault detection, efficient energy management, and adaptive sensor data stream learning. This will provide a trade-off between predictive accuracy, energy-saving, and operational-responsiveness, which are important to the modern industrial systems.

1.1 Problem Statement

Conventional predictive maintenance models based on DL are computationally intensive and power-intensive, accurate, but impractical to use on edge computing devices. The majority of the current models, including CNN, GRU, and hybrid models, are computationally and memory intensive and thus expensive to run and slow in detecting faults in real time. Also, these models do not fit well to different machine conditions and are overfitting when trained on unbalanced industrial data [15]. The deficiency of optimized architectures with a balance between predictive accuracy and energy efficiency provides a gap in the critical research in the design of lightweight and reliable edge-based predictive maintenance models [17]. This study will address this gap by suggesting a time-conscious pruned LSTM model that can learn temporal relationships on multi-sensor time-series data with a low level of computational complexity. The strategy fills the performance verses efficiency gap, guaranteeing quicker inference and sustainable edge dispensation of industrial systems.

1.2 Research Motivation

The impetus behind this study is the increased desire to have smart, energy saving predictive systems that can operate in edge scenarios with minimal resources. In fact, the complex time-series information produced by industrial machines are in constant need of real-time analysis without depending on cloud infrastructures with high power. To fulfil this requirement, the development of a pruned LSTM-based predictive maintenance model, with its capacity to combine computational efficiency and a high level of accuracy, is encouraged to make a shift to autonomous, low-power, and adaptive maintenance models that can revolutionize industrial reliability and sustainability.

1.3 Research Significance

The proposed study has a great industrial importance because it can solve the twofold problem of precise fault forecasting and power-saving model execution on peripheral devices. The suggested pruned LSTM architecture is better at achieving higher predictive performance by consuming less power and inference latency, which facilitates real-time decision-making. The research will lead to sustainable intelligent manufacturing, enhancing equipment life, reducing unplanned downtime, and further practical implementation of industry 4.0-ready energy-efficient predictive maintenance systems by making it cost-effective, scalable, and intelligent.

1.4 Key Contributions

- A new predictive maintenance framework is proposed that integrates pruned LSTM networks with energy-aware optimization for efficient fault detection and condition monitoring in industrial edge systems.
- The study introduces a structured pruning mechanism to remove redundant neurons and connections in the LSTM network, significantly reducing computational overhead, energy usage, and inference time without compromising prediction accuracy.
- The proposed model is specifically designed for industrial edge environments, enabling real-time fault detection and maintenance prediction on low-power devices while maintaining high model fidelity.
- The framework is implemented using Python and TensorFlow, demonstrating practical applicability and scalability across different industrial domains with varying sensor configurations.
- This study establishes a strong foundation for future advancements by enabling the integration of transfer learning, self-supervised adaptation, and federated edge intelligence for scalable and sustainable predictive maintenance in Industry 4.0 environments.

The rest of the paper is organized as follows. Section 2 review the related works, Section 3 detailed about the proposed methodology, Section 4 describes the results and discusses about the study, and finally Section 5 concludes the study and direction for future work.

2. Related Work

Chen et al. [11] suggested a Low-Power On-Device Predictive Maintenance (LOPdM) system to combine Self-Powered Sensors (SPS) and Tiny Machine Learning (TinyML) methods to make real-time fault detection possible and energy-efficient. The aim of the study was to address the high power and cost requirements of the conventional AI-based PdM systems. There were 6 ML models tested where it was found that both the Random Forest and Deep Neural Network models performed as well as 99% accuracy even in low conditions of data and sampling. The system saved on energy by a rate of 66.8 as opposed to IMU-based systems. Nevertheless, the method might be constrained in terms of addressing complicated industrial statistics as well as scalability in various settings.

Rahman et al. [12] have performed a review article on how to combine Machine Learning (ML) and Digital Twin (DT) with Edge AI to improve intelligent industrial automation. The objective of the study was to enhance predictive maintenance, quality control, and optimization of processes with the help of real-time data-driven insights. Through the analysis of different ML models, datasets and industrial platforms, the review reflected on the emerging role of

deep learning, especially convolutional and recurrent architecture, in the industrial systems. The study has successfully realized a definite mapping of the transformational role of ML in the automation. Nevertheless, the lack of model generalization, real-time deployment, interpretability, scalability, and safety in autonomous decision-making are some of the weaknesses.

Rosca and Stancu [13] presented a bibliometric and thematic review of the literature on the topic to discuss the incorporation of Artificial Intelligence (AI) and specifically ML into self-powered IoT sensors. The goal of the study was to categorize the areas of IoT and evaluate the uptake of AI in such sectors as healthcare, industry, and smart cities. The authors conducted literature analysis in 2020 to 2025 and found the major sensors and the most efficient ML models such as CNN, LSTM, SVM and RF with accuracy as high as 99.92. The study made a clear presentation of AI-IoT developments. It, however, pointed at shortcomings in its form of inadequate standardization, asymmetrical AI usage, energy usage and insufficient study in underrepresented sectors like agriculture.

Ang et al. [14] have suggested a new way of detecting early anomaly in sensor-based Multivariate Time Series (MTS) through a technique known as Correlation Analysis based Detection (CAD). The aim of the study was to address the weaknesses of conventional and DL techniques, which need large data sets or generate volatile outcomes. MTS data are transformed into Time-Series Graphs (TSGs) by CAD as a means of measuring correlations between sensors as well as detecting anomalies by analysing how much correlations vary. The method attained more than 85% accuracy on big data sets and surpassed nine state-of-the-art tasks. Nevertheless, it is more deterministic and might not be able to be flexible to nonlinear relationships and hidden dynamic industrial conditions.

Rojas et al. [15] performed a systematic literature review on the topic of AI, the IoT, and DT application to predictive maintenance in the mining sector. The article examined 166 articles in Scopus and Web of Science that are concerned with fault detection, hybrid AI models, and real-time monitoring. The results obtained indicated that deep and reinforcement learning are very effective in predicting fault at the early stage and efficiency of operations. Nevertheless, there are still constraints in the standardization of data, scalability of models, interoperability and explainability, which do not allow the realization of large-scale application and real-time flexibility in complex mining conditions.

Achouch et al. [16] represents the extensive review of intelligent predictive maintenance approaches in Industry 4.0 as the means of enhancing the uptime of machines, lifecycle management, and the quality of production. The methods identified in the study included condition-based maintenance (CBM), prognostics and health management (PHM), and remaining useful life (RUL), and suggested a new multimodal predictive maintenance system consistent of varying sensors and prescriptive prognostic models. A case study of an industry based on the centrifugal compressor revealed proper prediction of defects and breakdowns which matched the actual maintenance schedules. Nevertheless, the downsides consist of inability to standardize models across various equipment, reliance on high quality data, complicated instrumentation of the system, and demand of a thorough validation and cybersecurity measures.

Bermeo-Ayerbe et al. [17] suggested an adaptive, data-driven energy modelling methodology to the industrial machinery to improve energy efficiency and sustainability with the integration of digital twins. The study was aimed at developing dynamic models to detect behaviour changes in machines with a concept drift detector, which is able to adapt to the degeneration and uncharacteristic energy behaviours. The method, tested on an industrial testbed with simulated drifts, outperformed non-adaptive models by at least a factor of two in terms of prediction accuracy as the method gave an 82.81% fit rate. Nevertheless, the method has weaknesses including temporary delays when a drift is detected and false drift detection and poor robustness that needs additional input characterization and automation enhancement.

Moleda et.al [18] conducted a review of maintenance strategies in the power industry with emphasis on the shift of classical corrective solutions to the predictive and prescriptive solutions based on Industry 4.0 solutions. The study focused on the analysis of the available practices, AI-based analytics, Big Data, and IoT applications in equipment monitoring, fault detection, and maintenance planning. The authors have also compared the traditional and the latest methods by outlining the strengths, weaknesses, and the integration difficulties. The review managed to map the state-of-the-art predictive maintenance methods comprehensively and thus, researchers can be guided on how to improve it. Nevertheless, there are disadvantages like challenges in real-life applicability because of the safety laws, cyber-security needs, operator limitations, and the expensive nature of industrial applications.

Table 1. Summary of Existing Studies

Reference	Method	Advantages	Limitations

Chen et al. [11]	Low-Power On-Device Predictive Maintenance (LOPdM) integrating SPS and TinyML	Achieved 99% accuracy using Random Forest and Deep Neural Network even under low data/sampling; 66.8% energy savings over IMU-based systems	Limited scalability and ability to handle complex industrial data
Rahman et al. [12]	Review on combining ML, DT, and Edge AI for intelligent industrial automation	Mapped the transformational role of ML in automation; improved predictive maintenance, quality control, and process optimization	Lack of model generalization, real-time deployment, interpretability, scalability, and safety in autonomous decision-making
Rosca and Stancu [13]	Bibliometric and thematic review of AI and ML integration in Self-Powered IoT Sensors	Identified key AI-IoT applications (healthcare, industry, smart cities); high model accuracy (up to 99.92%)	Lack of standardization, uneven AI adoption, energy inefficiency, and limited research in sectors like agriculture
Ang et al. [14]	CAD for anomaly detection in Multivariate Time Series (MTS)	Achieved >85% accuracy, outperforming nine state-of-the-art methods; efficient correlation-based detection	Deterministic approach; poor adaptability to nonlinear relationships and dynamic industrial environments
Rojas et al. [15]	Systematic review of AI, IoT, and DT applications in predictive maintenance for mining	Demonstrated deep and reinforcement learning effectiveness in early fault detection and operational efficiency	Data standardization, model scalability, interoperability, and explainability remain unresolved
Achouch et al. [16]	Review of intelligent predictive maintenance approaches in Industry 4.0 using CBM, PHM, and RUL models	Enhanced uptime, lifecycle management, and production quality; validated through a real industry case study	Lack of model standardization, dependency on high-quality data, and cybersecurity challenges
Bermeo-Ayerbe et al. [17]	Adaptive, data-driven energy modeling with DT and concept drift detector	Improved energy efficiency and adaptation to machine behavior changes; achieved 82.81% fit rate	Issues with false drift detection, temporary delays, and low robustness
Moleda et al. [18]	Review of AI-based predictive and prescriptive maintenance strategies in the power industry	Comprehensive mapping of state-of-the-art PdM practices; facilitates guidance for future industrial applications	Implementation challenges due to safety, cybersecurity, operator skills, and high costs

Table 1 contains a summary of the existing research works that were consolidated regarding the topic of AI-driven and energy-aware predictive maintenance systems in industrial environments in a more condensed form. The surveyed studies prove that there is a significant progress in uniting MLand DT technologies as well as IoT-based data collection to improve the efficiency of maintenance and operational stability. Such works have had astounding levels of accuracy, energy savings and enhanced automation performance in a range of industrial uses. Nonetheless, these developments are associated with several significant shortcomings in the literature including inadequate scalability, reliance on high-quality datasets, absence of real-time implementation, interoperability and inability to operate successfully in complex, nonlinear and dynamic industrial systems. The issues with many current methods are also optimization of energy, generalization of models, and explainability, which makes implementation of such methods in large-scale industries challenging. The current study will overcome these difficulties by presenting an energy-gauge predictive-maintenance framework, which combines adaptive feature learning, and an LSTM-based deep-learning device in order to effectively process time-series sensor data. The proposed system is more energy efficient, scalable, and predictive of faults and is robust to a wide range of industrial environments- in effect addressing the limitations that were found in earlier research works as summarized in the table.

3. Proposed Pruned LSTM for Energy-Aware Predictive Maintenance in Industries

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

The proposed study introduces an energy-conscious predictive maintenance architecture of industrial edge systems based on Pruned LSTM networks on sensor-based time-series. This method starts by obtaining multivariate sensor data like temperature, vibration, current and energy consumption data of industrial equipment. The data are preprocessed through such steps as noise filtering, normalization and sliding-window segmentation to organize sequences in a manner that can be modeled in time. A Pruned LSTM model is then constructed to encompass temporal dependencies and one of these is to eradicate unnecessary neurons and connections, hence lowering the computation cost and enhancing the speed of inference to deploy edges. To improve performance, attention mechanisms are added to emphasize significant sensor signals that make the most contribution to failure prediction. The model quantization is also used to reduce the energy consumption further without compromising on the accuracy. The trained model forecasts possible failures or deterioration conditions in real-time, which allows to plan the maintenance proactively. In general, the presented approach offers a scalable, low-power, and intelligent predictive maintenance solution in real-time in smart manufacturing and an industrial IoT setting. The workflow of the proposed framework is illustrated in Fig 1.

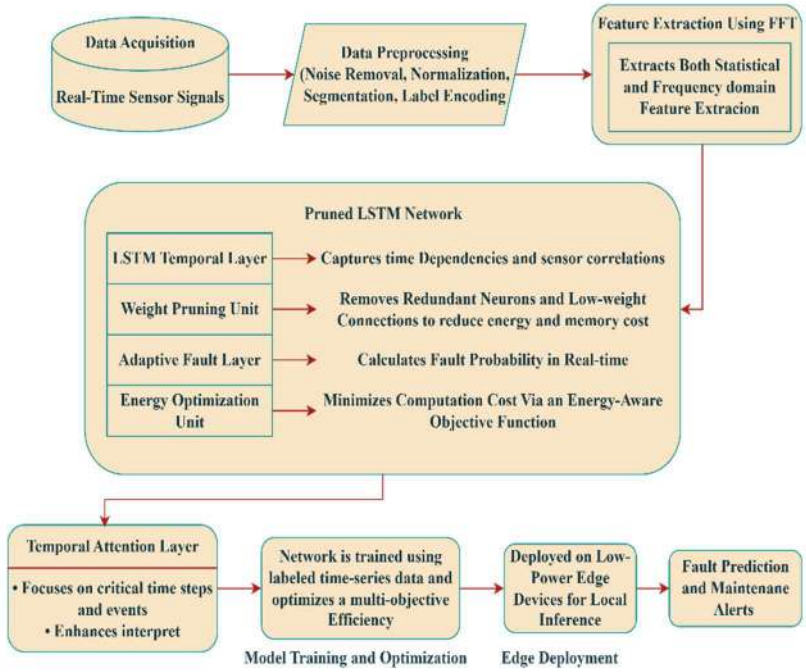

Fig 1. Workflow of the Proposed Framework

3.1 Data Collection

The publicly available Kaggle dataset, a smart manufacturing process dataset used in this study that included real-time multivariate sensor measurements taken in industrial equipment [19]. Among the most important parameters like temperature, vibration, current, speed, and energy consumption are constantly measured at any time that the machine is running. These time-series observations both observes normal and faulty states allowing the model to acquire information on degradation patterns and forecast failures. Timing Data is collected at regular sampling rate and coordinated between all the sensors to achieve timing consistency. The dataset is realistic as it offers an industrial setting to assist in the development and testing of the proposed energy-aware pruned LSTM model in predictive maintenance of edge systems.

Table 2. Dataset Description

Timestamp	Temperature (°C)	Machine Speed (RPM)	Production Quality Score	Vibration Level (mm/s)	Energy Consumption (kWh)	Optimal Conditions
2025-04-01 08:00:00	78.92	1461	8.49	0.07	1.97	0
2025-04-01 08:01:00	71.83	1549	8.97	0.04	1.01	0

2025-04-01 08:02:00	74.88	1498	8.52	0.08	1.60	0
2025-04-01 08:03:00	77.27	1478	8.28	0.09	1.87	0
2025-04-01 08:04:00	76.50	1524	8.07	0.04	1.53	0

Table 2 shows a sample representative to the manufacturing data utilized in the formation and confirmation of the suggested Energy-Aware Predictive Maintenance design. The data set is a time-series of multivariate data about industrial machinery that is captured by the data and the data is a reflection of real-time operational and environmental parameters. The LSTM-based predictive model is used to extract features, detect anomalies and predict faults on these data. The organized data allows to evaluate the health of machines, detect the deviations in performance and estimate what maintenance should be performed in different operating conditions.

3.2 Data Preprocessing

Preprocessing of data phase involve the preparation of raw multivariate sensor data to do effective modelling and analysis. It is to guarantee the quality of data, uniformity, and preparedness to time-series learning of the proposed pruned LSTM model. The preprocessing of the signals measured by the industrial sensors converts the signals into structured temporal sequences that indicate the behaviour of the machine in normal and abnormal operating conditions. This clean data increases the sensitivity of the model to subtle signs of degradation and also manages to better predict. The result of this phase is a clean, well-organised and balanced dataset that can be used to predictively maintain the industry in an energy-efficient manner.

3.2.1 Data Cleaning

This will eliminate incomplete, clustered, and noisy sensor readings which may alter time-series patterns. In the industrial settings, the noise can be due to faulty sensors or transmission errors. Missing data are either interpolated or deleted. This will guarantee good quality and consistent data to model and this will assist the pruned LSTM to learn actual equipment behaviour patterns.

3.2.2 Normalization

The normalization of all sensor properties (e.g., temperature, vibration, energy) is performed in order to stabilize gradient updates and speed up the training of LSTM. It avoids the large features with large numeric ranges to prevail over the smaller features. Data is scaled using the min-max scale to the range [0, 1].

$$x_i^{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Here x_i is the original sensor reading, x_{max} and x_{min} are the maximum and minimum values of that sensor feature respectively.

3.2.3 Segmentation

Segmentation breaks continuous sensor streams into fixed length overlapping windows that represent the temporal dependencies that would be used in predictive maintenance. The individual segments are used as inputs of the LSTM model, which can learn time-related trends of degradation that results in faults.

3.2.4 Label Encoding

Each sequence that has been segmented is given a label that shows whether it is in a normal state or faulty state. Binary encoding is used in which 0 is normal operation and 1 failure or anomaly. This enables the LSTM to go through supervised prediction learning of maintenance.

3.2.5 Data Balancing

The industrial data have a small number of faulty samples, balancing provides equal contribution to model learning by normal and fault classes. Hybrid resampling is used to eliminate bias by oversampling rare segments of faults and under sampling normal segments.

3.3 Pruned LSTM Feature Extraction

The Pruned LSTM Feature Extraction step is important in realizing an energy efficient predictive maintenance in industrial edge system. The Long Short-Term Memory (LSTM) network is selected in particular since it is a good demonstration of temporal correlations in sensor-based time-series, including vibration, temperature, current, and energy consumption data, incurred by the industrial equipment. Nonetheless, traditional LSTM designs tend to have superfluous neurons and parameters which add to the computational cost, memory and energy cost of the system and thus cannot be deployed to resource-constrained edge devices. To overcome these difficulties, LSTM network is pruned in a systematic way to maximize the efficiency of the model and its prediction efficiency. Once the first LSTM model is trained with the preprocessed sensor data it starts the pruning process. Weight magnitude is used and together with activation-based sensitivity analysis, neurons, gates and connections are identified which contribute insignificantly to the output of the model. These unimportant elements are removed systematically in effect downsizing and simplifying

the LSTM model. The LSTM Cell Computation is represented in (2) – (6).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$
$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (4)$$
$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (5)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (6)$$

Where x_t is the input sensor vector at time t (e.g., temperature, vibration, energy); h_{t-1} is the hidden state output at time t; C_t is the cell state; i_t, f_t and o_t are the input, forget, and output gates respectively; $\widetilde{C_t}$ is the candidate memory; W_i, W_f, W_o and W_C are the weight matrices for gates; b_i, b_f and b_o are the bias terms; σ denotes the Sigmoid activation, and $\odot$ denotes the element-wise multiplication. Structured pruning is more similar to unstructured pruning, except that the weights are removed more densely and the entire neurons or even complete LSTM layers are eliminated, preserving the computational regularity of the model and allowing the model to be run on edge hardware accelerators. The weight pruning function is denoted in (7).

$$W_{pruned} = W \cdot I(|W| > \tau) \quad (7)$$

Where W is the original weight matrix of the LSTM layer, τ is the pruning threshold (set based on weight magnitude or sensitivity), I $(\cdot)$ is the indicator function that retains weights greater than threshold τ, and W_{pruned} is the pruned weight matrix after eliminating low-importance connections. After pruning, a fine-tuning stage is performed in order to regain the small loss in prediction ability. The Fine-Tuning Loss Function is given in (8).

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (8)$$

Where N is the number of samples, y_i is the true label (0 = normal, 1 = fault), $\hat{y}_i$ is the predicted fault probability from pruned LSTM. This retraining helps to make sure that the rest of the network parameters are adjusted to the smaller architecture without damaging the model that can be used to predict possible faults of the equipment and the patterns of its degradation. Optimal trade-off is generated between the size of the model and the accuracy of fault prediction by successively pruning and fine-tuning the model. The pruning is guided not only by accuracy but also by energy efficiency. The objective minimizes loss while penalizing high energy consumption is represented as E_{comp} in (9).

$$\mathcal{J} = \mathcal{L} + \lambda E_{comp} \quad (9)$$

Where J is the joint optimization objective, L is the prediction loss, E_{comp} is the computational energy consumption of the model, λ is the regularization factor balancing accuracy and energy. The final decision layer outputs the probability of fault occurrence is given in (10).

$$\hat{y}_t = \sigma(W_y h_t + b_y) \quad (10)$$

The pruned LSTM makes active neurons, parameters, much fewer, and thereby, inference latency and energy consumption are reduced by a significant margin, which is essential when continuous monitoring is needed in the industry. The lightweight design enables real time execution on edge computing devices to reduce the frequency of cloud communication hence minimising network overhead of energy consumption. Also, it is designed to reduce the thermal and power footprint of the hardware, which reduces the lifespan of the device itself and enhances the sustainability of the entire system. Finally, the Pruned LSTM Feature Extraction step will be used to convert the standard LSTM model into an energy and computationally friendly edge-based predictive maintenance LSTM model. This allows fault detection and prediction of anomaly in real-time with the minimum amount of power consumption, which fits within the objective of this study of designing an intelligent and low power and scalable solution to contemporary industrial systems. The architecture of the proposed system is illustrated in Fig 2.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

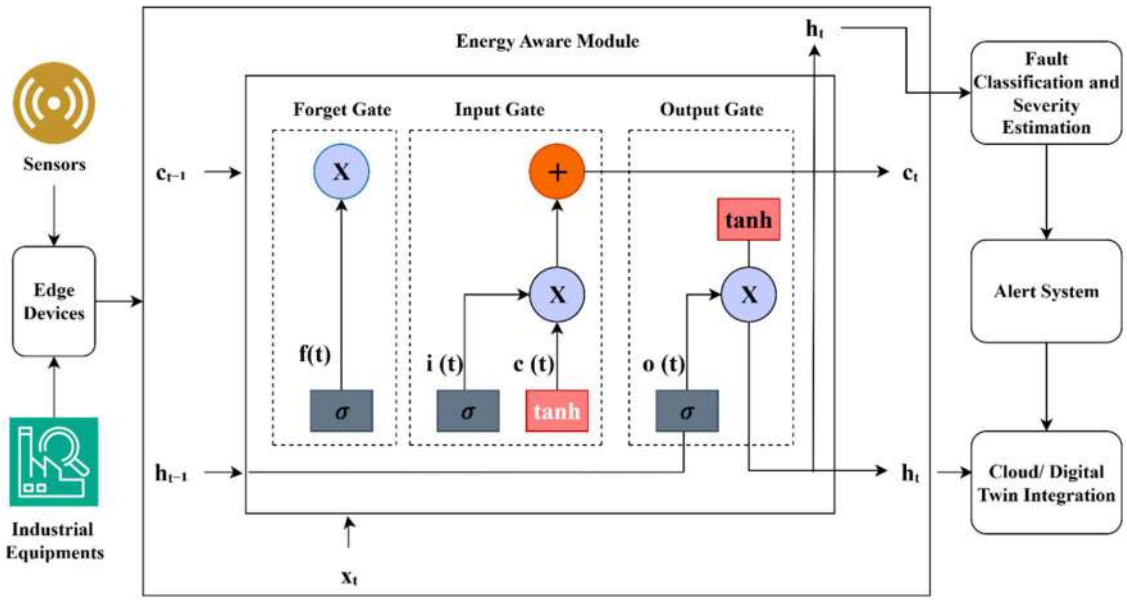

Fig 2. Architecture of the Proposed System

3.4 Temporal Attention Layer

The Temporal Attention Layer is an important addition to the pruned LSTM network, which allows the model to selectively pay attention to the most informative time steps in the sensor-based time-series data. Although the LSTM is effective in the long-term capturing of sensor reading dependencies over time, it equally considers all the time steps equally when processing a sequence. But in the world, industrial system, some events, such as sudden spikes in vibration, sudden temperature surges or sudden shifts in energy consumption, have much stronger hints of possible faults than others. The temporal attention mechanism resolves this shortcoming by dynamically weighting the importance of each hidden state output of the LSTM, and as a result, which enables the model to focus on important temporal features that are most significant to predictive maintenance. The attention layer will receive the series of the hidden states produced by the pruned LSTM which is denoted as $h_1, h_2, .., h_T$ and calculates a set of attention scores, which are indicators of the contribution that each time step makes to the prediction of the fault. Attention weights are constructed using these scores with the help of a SoftMax function. Ensuring that the overall significance of all the time steps is one. The context is the sum of these hidden states weighted α_t which is a useful summary of the sequence and highlights the most important patterns associated with machinery degradation or failure c. Mathematically, the attention mechanism is represented as in (11) – (13).

$$e_t = v^T \tanh(W_a h_t + b_a) \quad (11)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^{T} \exp(e_k)} \quad (12)$$

$$c = \sum_{t=1}^{T} \alpha_t h_t \quad (13)$$

Where e_t is the attention score for time step t , v is the context vector (trainable parameter), W_a is the weight matrix for attention, h_t is the hidden state output from the pruned LSTM at time step t , b_a is the bias term, α_t is the normalized attention weight at time step t , e_k is the attention score for each time step k , T is the total number of time steps, c is the context vector representing the weighted sum of hidden states. This context vector can then be sent to the last prediction layer which will give a probability of an impending fault or an anomaly. The attention layer increases interpretability of the model by prioritizing key temporal areas, which enables engineers to know which periods of time or sensor patterns gave a fault prediction. Moreover, the attention computation is lightweight, unlike convolutional or dense layers, and thus, keeps the energy consumption needed to deploy edges. The Temporal Attention Layer is a natural extension of the pruned LSTM architecture that enhances the prediction accuracy, interpretability, and responsiveness without adding too much to the computational and energy expenses, which means that it fits the study objective, that is, energy-aware predictive maintenance in industrial edge systems, perfectly.

3.5 Quantization for Edge Deployment

The Edge Deployment, quantization, is critical towards an energy-efficient predictive maintenance system to be deployed in industrial settings. The

optimally pruned LSTM model is then quantized to reduce the computational complexity, memory footprint and energy consumption downstream even more. Quantization The high-precision 32-bit floating-point weights and activations of a model are converted into lower-bit integer representations, usually 8-bit quantization without a markedly different impact on the predictive accuracy of the model. This will be the necessary step to make sure that the trained model can be successfully deployed to industrial edge devices with small computational and power capabilities including embedded controllers or IoT gateways. The quantization process can be expressed mathematically as in (14) and (15).

$$Q(w) = \text{round}\left(\frac{w - w_{\min}}{s}\right) \quad (14)$$

$$s = \frac{w_{\max} - w_{\min}}{2^b - 1} \quad (15)$$

Where Q (w) represents the quantized weight, $w_{\min}$ and $w_{\max}$ are the minimum and maximum weight values, s is the scaling factor, and b is the bit precision (e.g., 8 bits). The formula reduces the continuous axis of the weight values into discrete levels of integer values hence making the computation simplified and easier to be handled using hardware. There are two general approaches to quantization, namely Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT). To preserve high accuracy post-compression, Quantization-Aware Training is applied in the current research. QAT models the consequences of quantization when training a model and enables the pruned LSTM to train distributions of weights that is resistant to the loss of numeric precision. This method will make sure that the accuracy of the model at predicting fault and its sensitivity to minor patterns of machine degradation is not lost once deployed. The use of the quantized model deployed to edge devices has a number of benefits in operation. It minimizes model size resulting in a reduction in loading and inference time and reduces energy consumption which is vital in continuous monitoring in remote industrial settings. Also, quantized computation can be compatible with integer-based hardware accelerators such as Tensor Processing Units (TPUs) or low power microcontrollers, which can be real-time. Altogether, the quantization step is the step connecting DL studies with the practical industrial use. The quantized pruned LSTM model offers a scalable and efficient predictive maintenance solution in the industrial edge systems (with significantly lower computational costs) that would be a perfect fit to the objective of the study of sustainable and intelligent manufacturing processes.

3.6 Energy-Aware Adaptive Inference

The Energy-Aware Adaptive Inference mechanism is created to guarantee that the suggested pruned LSTM-based predictive maintenance system can work effectively, even with the different energy and computational issues, which are inherent to industrial edge conditions. The edge devices of industries usually vary over time in the amount of energy and processing power with several simultaneous tasks or volatile power supply. Consequently, ensuring quality fault detection performance, but at the same time reducing the energy usage is important to ensure the continuous operation. This module presents a dynamically controlled energy monitoring controller that dynamically switches between the inference and prediction modes of the model depending on the dynamically available energy and workload on the system. There are two major operational modes in the inference framework which include high-energy mode and low-energy mode. Under energetically demanding conditions, on the event that sufficient power and computational resources are available to the edge device, the system switches to the full pruned + attention LSTM model, providing the optimal accuracy and fault detection limit. This version incorporates all the learned parameters and attention mechanism to concentrate on the important sensor time steps so that the early indications of machine degradation or malfunction are correctly identified. Conversely, when operating in low-energy states, the controller automatically scales to either an ultra-pruned or lightweight version of LSTM, where additional unnecessary neurons and parameters are further eliminated, and attention layers can be bypassed at least partially, to reduce the complexity of the computation. The arrangement requires a much lower energy consumption without reducing prediction accuracy to be acceptable. The policy of switching these two modes is informed by an energy threshold policy.

This dynamic adaptation guarantees stable and uninterrupted operation without human intervention providing a balance between accuracy of predictions and energy efficiency. The controller maintains real-time control over system measurements (CPU load, power draw, temperature, etc.) by using them to make decisions in order to maximize inference performance. The system through this adaptive mechanism can provide sustainable and autonomous predictive maintenance and can effectively work even under the energy-restricted industrial environment. On the whole, the Energy-Aware Adaptive Inference approach is appropriate to the aim of the study because it seeks to develop intelligent, low-power, and resilient edge-based maintenance systems in future smart manufacturing settings.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 4, 2025

Algorithm 1: Energy-Aware Predictive Maintenance using Pruned LSTM

Input:
X = {x_1, x_2, ..., x_n} // Multivariate sensor time-series data
Y = {y_1, y_2, ..., y_n} // Corresponding fault/normal labels
θ = pruning threshold
λ = energy regularization coefficient
Data Preprocessing
Normalize sensor readings in X
Segment X into time windows W_1, W_2, ..., W_k
Extract statistical and frequency-domain features F from each window
Model Initialization
Initialize LSTM network parameters W = {W_i, W_f, W_o, W_C }
Initialize bias terms b = {b_i, b_f, b_o, b_C }
Training Phase
For each epoch do
For each mini-batch (X_b, Y_b) in training data do
Compute hidden states h_t and cell states $C_t$using LSTM equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$
$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

Compute fault prediction: P_{fault} = σ (W_o * h_t+ b_o)
Compute total loss: L_{total} = $L_{predict}$+ λ * L_{energy}
Update weights W using gradient descent
Energy-Aware Pruning
For each weight w in W do
If |w| < θ then
w ← 0 // Prune low-importance connections
End If
End For
End For
Model Deployment on Edge Device
Compress pruned model for lightweight deployment
Upload optimized model to edge node
Real-Time Fault Prediction
For each new sensor input sequence X_t do
Compute P_{fault}= model (X_t)
If P_{fault}≥ 0.5 then
Trigger maintenance alert A = 1
Else
A = 0 // Normal operation
End If
End For
Output:
Predicted fault probability P_{fault} and maintenance alert signal A

Algorithm 1 presents a step by step outline of the proposed Energy-Aware Predictive Maintenance System. It initially preprocesses sensor time scientific data and eliminates pertinent characteristics. Afterwards, an energy-regularized loss function is used to train a Pruned LSTM to trade-off accuracy and computation efficiency. The training process is then performed with low-importance weights that are removed sequentially to minimize the energy usage. Lastly, the lightweight model is implemented in industrial edge devices where the lightweight version predicts faults in real-time and sends maintenance

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

notifications whenever the probability of anomalies surpasses a pre-determined threshold.

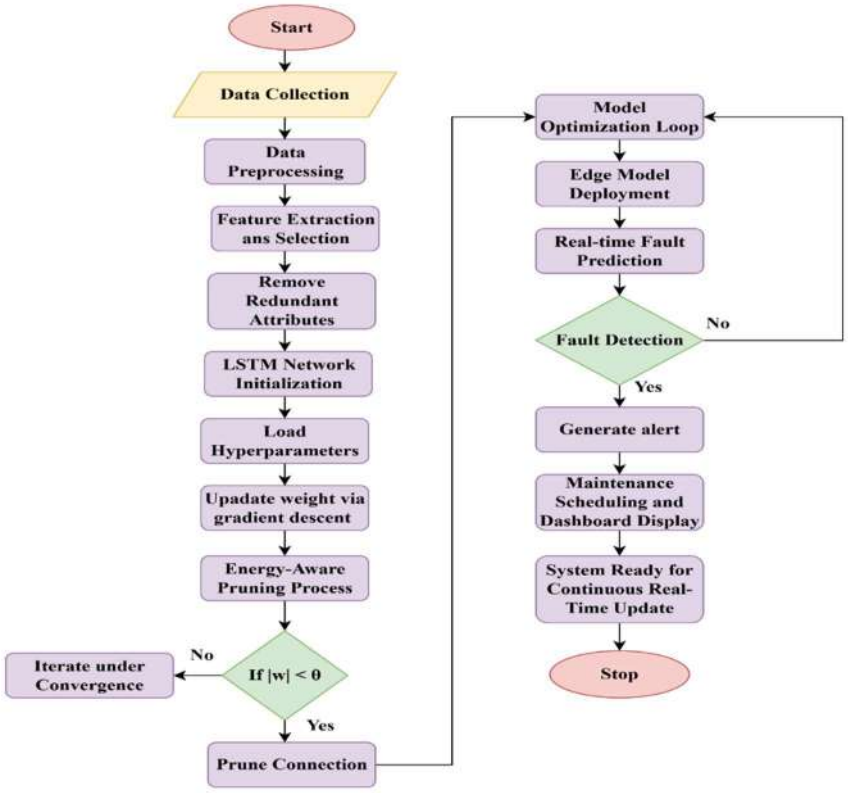

Fig 3. Flowchart

Fig 3 summarizes all the computational steps starting with the sensor input to the fault alert. It also highlights interaction of the multi-stage between data processing, LSTM learning, pruning and energy optimization. Multiple levels provide decision points that make sure of dynamic adaptability data integrity, pruning thresholds and model stability are all under iterative control. The closed self-improving predictive maintenance ecosystem is linked to the real-time deployment loop through the pruned model and the industrial edge hardware.

The proposed approach presents a new type of energy-aware predictive maintenance framework that describes a unique combination of model pruning, temporal attention, quantization, and adaptive inference into a single LSTM-based model that is optimized to be deployed to industrial edges. This research touches on both performance and sustainability unlike the traditional predictive maintenance models which focus on the accuracy at the expense of the computational and energy efficiency. Pruned LSTM allows the removal of unnecessary neurons without losing the ability to learn temporal features, which leads to light-weight computation and does not deteriorate the accuracy. The model has a temporal attention mechanism integrated into it, which allows it to dynamically focus on important sensor time points, increasing interpretability and fault detection accuracy. Moreover, quantization reduces the power and memory cost of high-bit model versions by converting high-precision models into low-bit integer versions that are friendly to edges. The most exciting feature is the Energy-Aware Adaptive Inference module which enables the switching between full and lightweight inference modes in real-time depending on the available energy so that it can be used even when it is constrained by power. This self-optimizing mechanism is a major improvement against the use of the static models in the previous research. In general, the proposed solution is a balanced trade-off between accuracy, latency, and energy consumption, and it is very appropriate in the case of real-world industrial edge scenarios.

4. Result and Discussion

The findings of this study indicate that the proposed energy-conscious pruned LSTM model with temporal attention and adaptive inference has a better performance in predictive maintenance under industrial edge conditions. It is successful in acquiring temporal patterns and degradation trends based on

complex sensor data to enhance effective prediction of faults. By organized pruning and quantization, the system dramatically lowers the amount of computational overhead and energy usage making the system usable in real-time operation on low-power industrial edge devices. The model also introduces the temporal attention mechanism that enables the model to focus on the most informative sensor intervals resulting in more credible fault recognition despite Imbalanced or noisy datasets. The adaptive inference mechanism is also successful in controlling variations in power by dynamically changing the model complexity based on the energy availability to provide continuous functioning in the limited environments. The comparative analysis of this hybrid approach with traditional DL models confirm that this hybrid approach has high reliability in predictions and also attains high efficiency improvements. On the whole, the findings support the feasibility and strength of the suggested framework, and prove that it is a perfect choice as it considers predictive accuracy, cost of computation and energy efficiency, which makes it the optimal answer to sustainable and intelligent and continuous predictive maintenance in the contemporary industrial systems.

Table 3. Simulation Parameter

Parameter	Value
Dataset	Multivariate Time-Series Data
Sampling Frequency	1 Hz
Window Size	50-time steps
Overlap Ratio	25%
Training–Testing Split	80% – 20%
Batch Size	64
Learning Rate	0.001
Optimizer	Adam
Loss Function	Binary Cross-Entropy
LSTM Layers	2
Hidden Units	128
Dropout Rate	0.3
Pruning Ratio	30%
Quantization Precision	8-bit integer
Attention Mechanism	Enabled
Epochs	100
Deployment Platform	ARM Cortex-A57, 4 GB RAM
Energy Monitoring Interval	Every 10 seconds

Table 3 summarises the simulation parameters of the experiment to be used in the assessment of the proposed energy-conscious predictive maintenance framework. The structure is to be balanced in terms of computational capability and model precision and to allow real time fault detection on low power embedded systems. The parameters make the system to capture the temporal dependencies of the multivariate time-series data in an optimized sequence processing as well as controlled training dynamics. The model training system focuses on a robust learning but with sufficient regularization and optimization policies, so that the learning process would be stable as it approaches convergence. Moreover, model compression methods including pruning and quantization can be used to run on hardware with resource limits with minimal performance loss. Attention mechanism promotes prioritization of features to be more precise in fault recognition whereas monitoring of energy at a predefined period aids in evaluation of efficiency and sustainability of the system in the real-time running.

4.1 Predictive Performance Analysis

The predictive performance assessment determines the success of the suggested pruned LSTM model in predicting possible faults and patterns of degradation in industrial equipment using sensor-based time-series data. The temporal dependency and the ability to identify important patterns in the model by use of the attention mechanism makes the model very reliable when it comes to fault detection. Its performance is evaluated by comparing its predictions to actual fault occurrences to determine its performance in terms of detecting faults and consistency. The proposed method has a higher convergence rate, lower false alarms, and better fault detection during different operational conditions than the classic models, which proves its rigor and applicability in real-time predictive

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

maintenance in the industrial setup with limited energy resources.

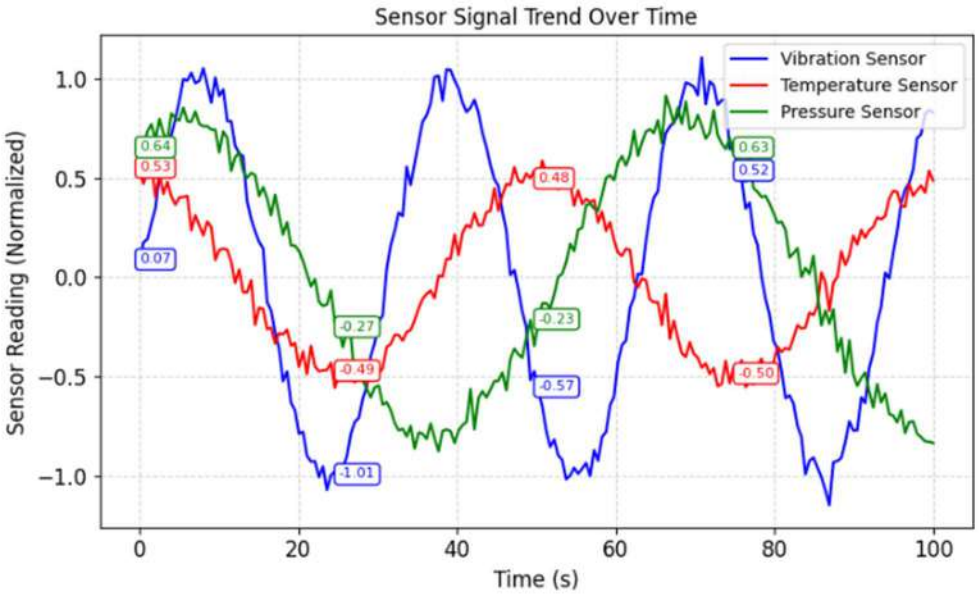

Fig 4. Sensor Signal Trend Over Time

Fig 4 shows the time dynamics of three important industrial sensors, such as vibration (blue), temperature (red), and pressure (green), which were recorded during consecutive processes and operations. The blue curve would signify the amplitude of vibrations that would represent oscillation of the machine, the red curve would denote the stability of temperature when the load is altered and the green curve would denote how the pressure varies over changes in working periods. These values are boxed values indicating sampled sensor values at particular timestamps to validate the trends. A combination of such time-series trends points to possible deteriorating pattern, and thus the proposed energy-conscious predictive maintenance framework will be capable of learning the temporal pattern of fault evolution.

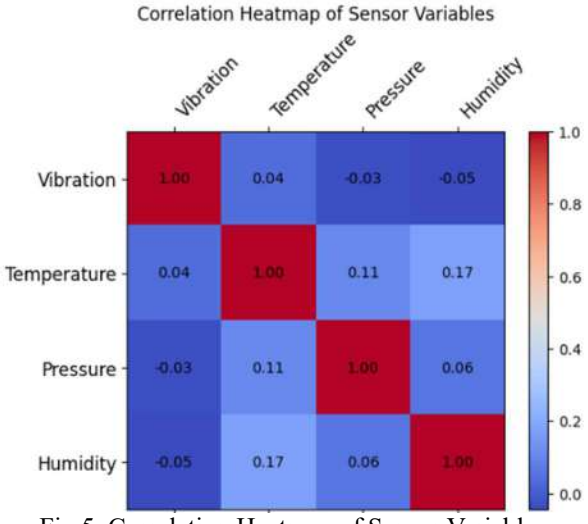

Fig 5. Correlation Heatmap of Sensor Variables

Fig 5 shows the relationship of the various sensor indicators: vibration (red squares), temperature (orange tones), pressure (blue gradients), and humidity (pale tones). Areas in red are strong positive correlations and deep blue colour depicts inversive relationships. The diagonal line is an indication of ideal correlation of each sensor to itself. The numerical values in boxes provide the actual strength of correlation between sensor pairs. The visualization is useful in revealing redundant features and inter senor dependence critical in pruning and training of models with low-energy consumption.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

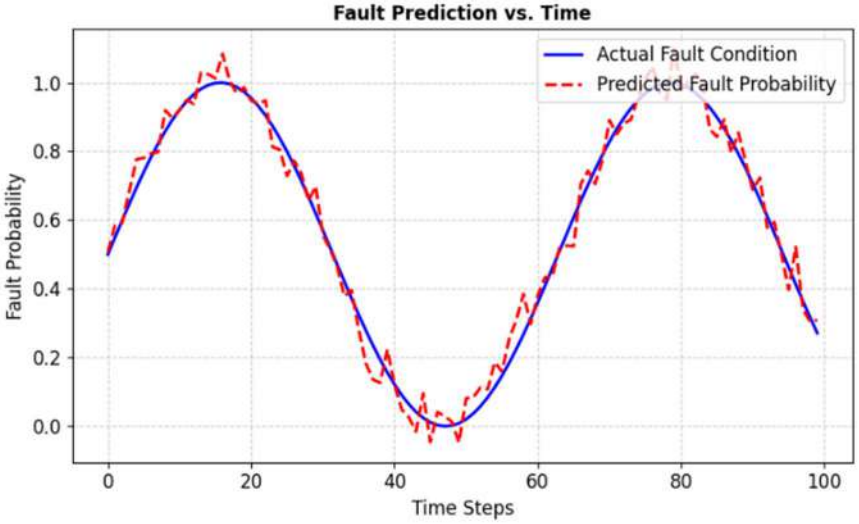

Fig 6. Fault Tolerance Versus Time

Fig 6 depicts the fault prediction verses time line graph. The solid line of blue colour shows the real condition of the machine based on sensor data, and the red line with a dotted form is the forecasted probability of a fault made by the pruned LSTM model. The high correlation between the two lines implies that the model is able to follow the trends of temporal degradation of the equipment and this proves that it is reliable in undertaking any predictive maintenance exercise.

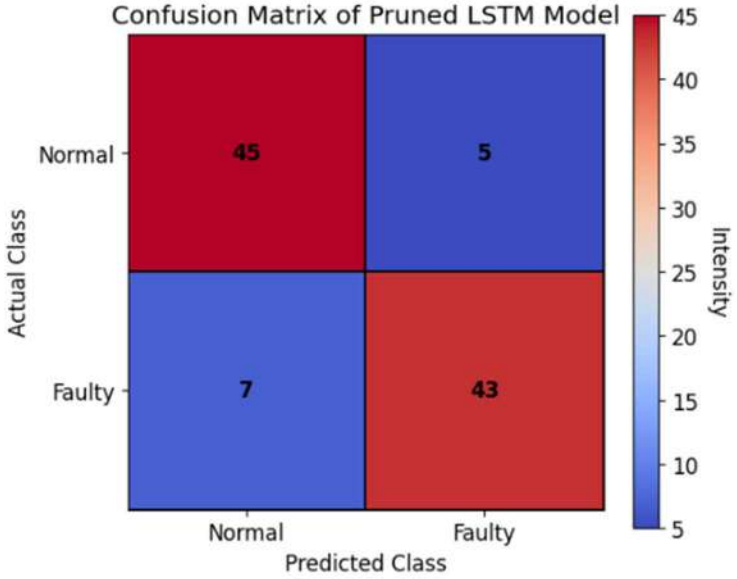

Fig 7. Confusion Matrix of Pruned LSTM Model

Fig 7 represents the results of the classification between the Normal and Faulty machine conditions based on sensor-based time-series data. The rows reflect real equipment states whereas columns reflect the forecasted states. The dark red diagonal boxes represent the correct predictions in both normal and faulty states and the light parts outside the diagonal denote the misclassifications. The number of samples in each category is represented in the data, thus giving a clear indication on how the model works in terms of categories. The large scores on the diagonal indicate that the proposed Pruned LSTM is very effective in degradation trends in the time lag and reducing false alarms. This validates the fact that it can be used in energy-efficient predictive maintenance in the industrial frontline to ensure that both computational economy and strong fault recognition are achieved.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

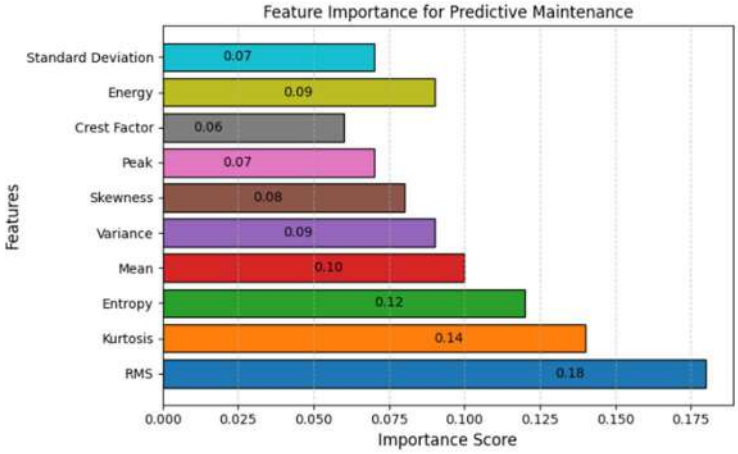

Fig 8. Feature Importance of Predictive Maintenance

Fig 8 shows the importance of statistical features that are extracted to predict equipment faults. Every colour represents specific characteristic, namely blue-RMS, orange-Kurtosis, green-Entropy and red-Mean values. The values in the box are the calculated values of the importance scores of the model analysis. The more the importance (e.g., RMS and Kurtosis) the more correlated with fault progression patterns. The colour diversity helps to visual differentiation, with the feature selection making the interpretability better and the energy-aware pruning to retain only the influential variables.

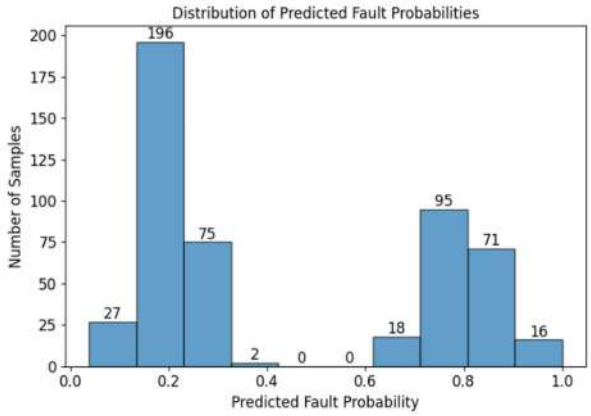

Fig 9. Distribution of Predicted Fault Probabilities

Fig 9 indicates the predicted fault distribution of the pruned LSTM model. The blue bars are the sample size within various probability ranges. The samples around 0.8 are strong indications of a lot of confidence of the model in its ability to detect faults whereas the samples around 0.2 are good signatures of healthy operation states. The visualization contributes to identifying the threshold at which predictive maintenance alerts will be activated and will allow following the necessary calibration of the model.

4.2 Energy Consumption Analysis

The energy consumption analysis aims at assessing the effectiveness with which the proposed pruned and quantized LSTM model is power-conserving in edge-based predictive maintenance. The model is much more energy efficient with less redundant computations as it uses low-bit quantization to achieve high prediction reliability. The adaptive inference process allows dynamically changing the complexity of the models depending on the real-time power availability, which guarantees its continuous operation even when operating with low energy conditions. Relative comparison with non-optimized models shows that this method can result in significant energy savings without compromising the fault detection accuracy, which proves the appropriateness of the method in sustainable, real-time predictive maintenance of industrial edge systems where power efficiency is paramount.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

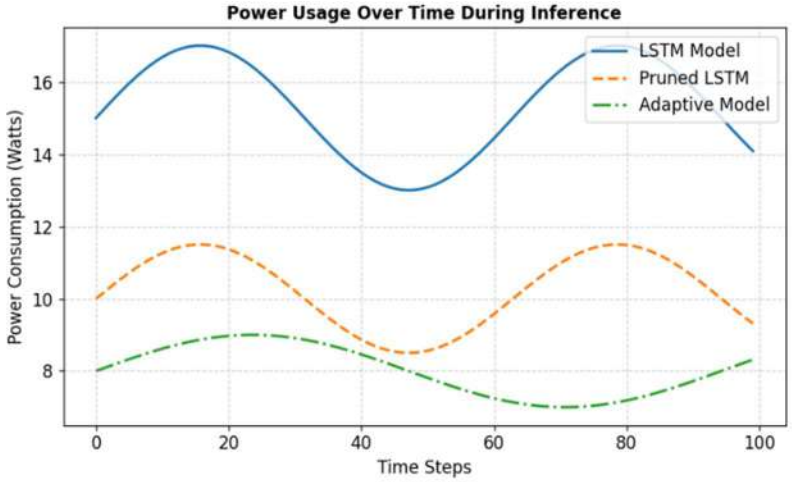

Fig 10. Power Usage Over Time During Inference

Fig 10 shows the variation in power consumption with inference in various model configurations. The blue solid line depicts the baseline LSTM model which is always consuming more power because of the dense computation and large number of parameters. The orange dashed line is associated with the Pruned LSTM model which attains a significant decrease in energy expenditure by pruning the parameters and structure optimization. Green dash-dot line represents the Adaptive Energy-Aware Model, which dynamically scales its complexity with regards to the inference to the sensor activity, leading to the smoother and lower power consumption trends over the duration. Generally, the diminishing trend of power between blue to green indicates the benefit of the suggested pruned and adaptive solution in providing energy efficient predictive maintenance that can be used in industrial edge systems

.

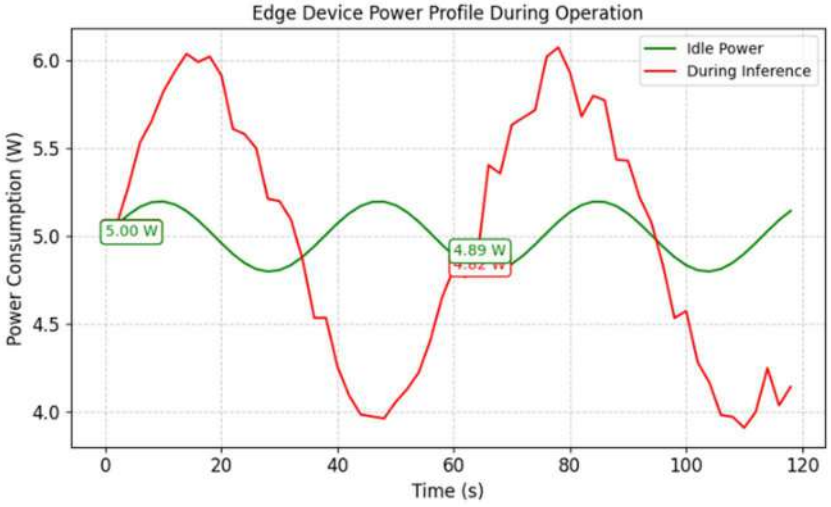

Fig 11. Edge Device Power Profile During Operation

Fig 11 indicates the behaviour of the power consumption of the edge device over time when it is performing predictive maintenance tasks. The green line is where the idle power is, whereas the red line is power consumption during model inference cycles. The readings that are boxed show real-time energy consumption in watts. The red curve peeks indicate computationally expensive periods of time when LSTM model is handling incoming sensor data. The stable green floor gives energy stability, which proves that the pruning and quantization strategies can greatly lower the operational energy requirement at the edge.

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

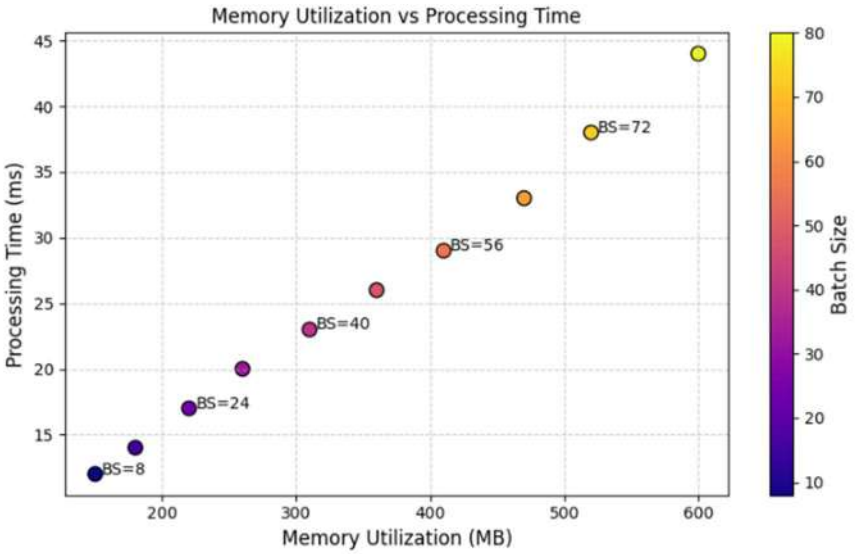

Fig 12. Memory Utilization Vs Processing Time

Fig 12 represents the trade-off between memory usage and inference processing time introduced by different model configurations that were run on the edge system. Every point corresponds to a specific configuration, and the intensity of the colors (yellow to dark purple) would be related to the batch size - lighter colors would be associated with smaller batch sizes, and dark purple colors with larger batches. Boxed labels have reference batch sizes. The rising pattern shows that as memory allocation is increasing, the processing time tends to increase which means there is a resource-performance tradeoff. This observation is in line with the presented energy-aware pruning method that reduces memory footprint without affecting responsiveness - guaranteeing efficient predictive maintenance under limited edge conditions.

4.3 Edge Deployment Validation

The edge deployment validation is an analysis done to determine the viability and fitness of the suggested pruned and quantized LSTM model as it is implemented on actual industrial edge devices. The model uses low-power hardware and is able to achieve very steady performance in real-time conditions. Resource consumption tests verify low CPU, GPU and memory consumption by pruning and quantization and allow operation to run smoothly with continuous operation without overheating and latency problems. The mechanism of adaptive inference guarantees the smooth transition between modes depending on the energy levels, which proves the resilience and scalability of the model. As a whole, the deployment validation states that the system is completely geared towards real-world, is energy efficient industrial edge applications.

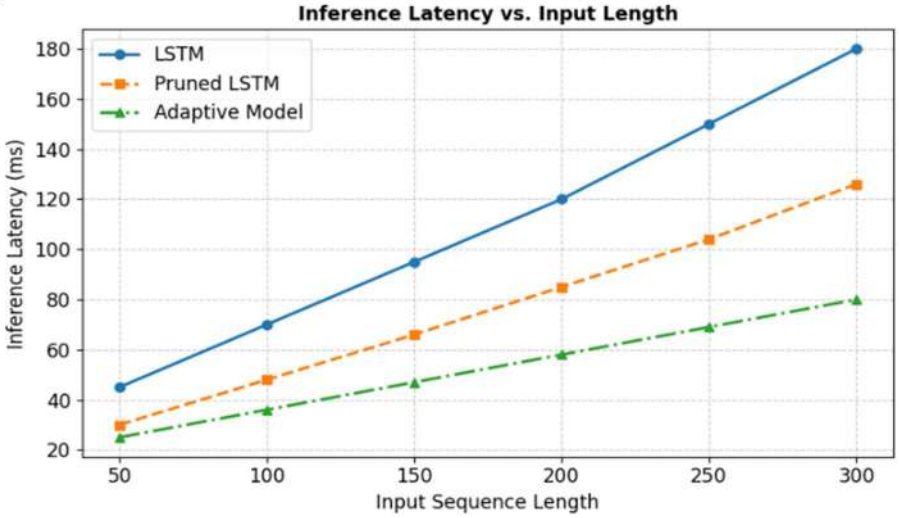

Fig 13. Inference Latency Versus Input Length

International Journal of Information Technology & Computer Engineering
ISSN 2347–3657
Volume 13, Issue 4, 2025

Fig 13 shows the latency of inference as a function of length of input sequence over the various model architectures. The blue circular curve is the blue line that denotes the bottom of LSTM model since it has the highest latency increase because of its dense computation and huge memory access constraints. The square marker of an orange dashed line shows the Pruned LSTM which significantly cuts down the latency by eliminating the unnecessary neurons, and by cutting down on the path of computation. The green dash-dot curve with the triangular markers corresponds to the Adaptive Energy-Aware Model which shows the lowest latency of any input length due to the ability to dynamically increase computational resources as the complexity of real-time data changes. The blue to green colour flow clearly shows how every optimization step, pruning and adaptive scheduling, cuts down on the computational delay to enhance real-time responsiveness which is an important feature of industrial edge predictive maintenance systems.

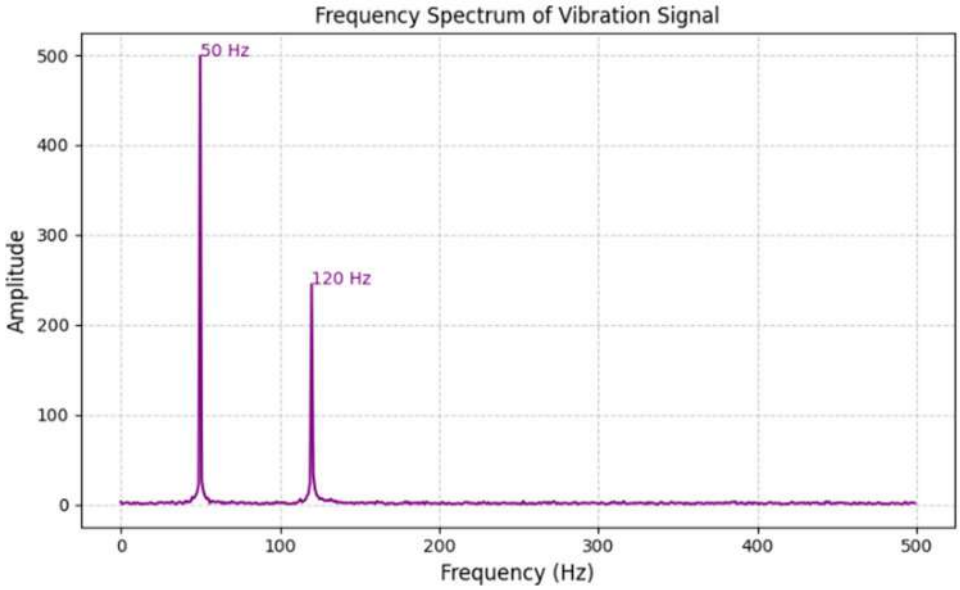

Fig 14. Frequency Spectrum of Vibration Signal

Fig 14 demonstrates the frequency-domain representation of the sensor of vibrations data with Fast Fourier Transform (FFT). The purple curve shows the distribution of amplitudes of various frequencies. The frequency markers are boxed as frequency values of 50 Hz and 120 Hz which are the dominant peaks and represent machine rotation and harmonic vibrations. Peak values at certain frequencies signify potential mechanical imbalance or wear of bearings and this would enable the predictive maintenance system to identify early fault trends in the piece of equipment.

4.4 Performance Comparison

The performance comparison in this study shows that the presented Pruned LSTM-based predictive maintenance model is much more effective than the conventional DL procedures, including standard LSTM, GRU, and CNN models. The pruning method is useful in minimizing computational cost and memory usage with predictive accuracy not being affected. The proposed system in comparison with the baseline models has a shorter inference time, consumes less energy, and is more stable in real-time. Accuracy, efficiency, and responsiveness are all what make this balance suit the model to the industrial edge environments. The findings indicate that energy mindful pruning and optimization yield a sparse but very robust predictive maintenance framework of sensor-based systems.

Table 4. Performance Comparison Across Various Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Energy Consumption (W)
CNN-Based Fault Detection [20]	98.3	98.7	98	98.3	4.8
GRU-Based Predictive Model [21]	92.2	92.2	92.2	92.06	4.3
CNN-RNN [22]	97.7	98.7	97.3	96.5	4.1

LightGBM [23]	93	94	93	93	3.9
XGBoost [24]	95.9	96.1	95.9	95.9	3.7
Proposed Pruned LSTM	98.8	98.9	98.8	98.9	2.5

Table 4 offers a comparative study of different predictive maintenance models that were tested under the same experimental conditions. Although basic DL networks like CNN, GRU, and hybrid CNN-RNN networks have proven to be quite effective in fault detection, they are slightly more energy-consumptive and, therefore, are not optimal when it comes to edges deployment. LightGBM and XGBoost are gradient-boosted approaches which demonstrate moderate accuracy, but are computationally light. On the contrary, Pruned LSTM model proposed has a better predictive performance with highest accuracy, precision, recall, and F1-score and consumes only a small power. This advancement underscores the effectiveness of the model in terms of efficiency and energy-consciousness in predictive accuracy, and makes it an appropriate choice when using real-time industrial settings on edge devices that are resource-constrained.

4.5 Discussion

The experimental analysis shows that the proposed Energy-Aware Predictive Maintenance framework based on Pruned LSTM Networks is a successful method of improving the prediction quality and energy consumption in industrial edge systems. The model is able to learn the complicated time-dependent relationships using multi-sensor time-series data, and at low computational cost due to pruning. The proposed architecture outperforms such traditional architectures as CNN, GRU, and XGBoost in terms of detecting early fault patterns, reducing false alarms, and being able to maintain stable operation at varying sensor conditions. The lower power usage highlights the fact that the method is suitable to be implemented on low power industrial edge devices and thus best suited to the implementation of real time monitoring and scheduling of maintenance. Besides, the trade-off between the performance and the computational efficiency is balanced to confirm the flexibility of the proposed model to different setups in industries. Nevertheless, one of the main shortcomings of this research is the fact that the model is based on labelled data and is limited to generalizing to various types of machines. Further research will be directed to include the element of self-supervised learning and adaptive transfer mechanisms in order to improve the scalability and performance of the model in unseen industrial settings.

5. Conclusion and Future Work

The proposed study introduced an Energy-Aware Predictive Maintenance model based on Pruned LSTM Networks that would suit sensor-based industrial edge systems. The strategy is a successful combination of intensive time modelling and pruning optimization to provide high prediction accuracy and reduction of computing and energy costs. Experimental evidence shows that the suggested approach largely excels over the standard ML and DL models in detecting early equipment failures, minimizing the inference latency, and boosting its energy consumption. By integrating pruning methods, the deployment can be done real-time on edge devices, enabling the operation to be cost-effective, intelligent, and sustainable to the maintenance operations, in the smart manufacturing environment. In spite of these promising results, imbalance in data, noises, and machine specific variations can affect the performance of the model. Therefore, the future studies will be aimed at the creation of adaptive transfer learning mechanisms, federated edge intelligence mechanisms, and hybrid Transformer-LSTM models to enhance cross-domain generalization and resilience. More so, using self-managed learning and power conscious scheduling systems will also increase the efficiency of fault prediction and allow deployment in large industrial settings with minimum human intervention.

References

[1] G. Tsaramirsis *et al.*, “A modern approach towards an industry 4.0 model: From driving technologies to management,” *Journal of Sensors*, vol. 2022, no. 1, p. 5023011, 2022.

[2] D. P. Yedurkar, T. Schlech, and M. G. Sause, “A Systematic Review on Smart and Predictive Maintenance in Tool Condition Monitoring,” *IEEE Access*, 2025.

[3] A. Hamdan, K. I. Ibekwe, V. I. Ilojianya, S. Sonko, E. A. Etukudoh, and others, “AI in renewable energy: A review of predictive maintenance and energy optimization,” *International Journal of Science and Research Archive*, vol. 11, no. 1, pp. 718–729, 2024.

[4] H. Nizam, S. Zafar, Z. Lv, F. Wang, and X. Hu, “Real-time deep anomaly detection framework for multivariate time-series data in industrial IoT,” *IEEE Sensors Journal*, vol. 22, no. 23, pp. 22836–22849, 2022.

[5] S. Bello, I. Wada, O. Ige, E. Chianumba, and S. Adebayo, “AI-driven predictive maintenance and optimization of renewable energy systems for enhanced operational efficiency and

longevity," *International Journal of Science and Research Archive*, vol. 13, no. 1, pp. 2823–2837, 2024.

[6] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU," *arXiv preprint arXiv:2305.17473*, 2023.

[7] M. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient acceleration of deep learning inference on resource-constrained edge devices: A review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, 2022.

[8] Z. Wei, X. Yu, and L. Zou, "Multi-resource computing offload strategy for energy consumption optimization in mobile edge computing," *Processes*, vol. 10, no. 9, p. 1762, 2022.

[9] S. Kapp, J.-K. Choi, and T. Hong, "Predicting industrial building energy consumption with statistical and machine-learning models informed by physical system parameters," *Renewable and Sustainable Energy Reviews*, vol. 172, p. 113045, 2023.

[10] B.-Y. Ooi, W.-K. Lee, M. J. Shubert, Y.-W. Ooi, C.-Y. Chin, and W.-H. Woo, "A flexible and reliable internet-of-things solution for real-time production tracking with high performance and secure communication," *IEEE transactions on industry applications*, vol. 59, no. 3, pp. 3121–3132, 2023.

[11] Z. Chen, Y. Gao, and J. Liang, "Lopdm: A low-power on-device predictive maintenance system based on self-powered sensing and tinyml," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[12] M. A. Rahman, M. F. Shahrior, K. Iqbal, and A. A. Abushaiba, "Enabling Intelligent Industrial Automation: A Review of Machine Learning Applications with Digital Twin and Edge AI Integration," *Automation*, vol. 6, no. 3, p. 37, 2025.

[13] C.-M. Rosca and A. Stancu, "Integration of AI in Self-Powered IoT Sensor Systems," *Applied Sciences*, vol. 15, no. 13, p. 7008, 2025.

[14] Y. Ang, Q. Huang, A. K. Tung, and Z. Huang, "A stitch in time saves nine: Enabling early anomaly detection with correlation analysis," in *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, IEEE, 2023, pp. 1832–1845.

[15] L. Rojas, Á. Peña, and J. Garcia, "AI-driven predictive maintenance in mining: a systematic literature review on fault detection, digital twins, and intelligent asset management," *Applied Sciences*, vol. 15, no. 6, p. 3337, 2025.

[16] M. Achouch *et al.*, "On predictive maintenance in industry 4.0: Overview, models, and challenges," *Applied sciences*, vol. 12, no. 16, p. 8081, 2022.

[17] M. A. Bermeo-Ayerbe, C. Ocampo-Martinez, and J. Diaz-Rozo, "Data-driven energy prediction modeling for both energy efficiency and maintenance in smart manufacturing systems," *Energy*, vol. 238, p. 121691, 2022.

[18] M. Molęda, B. Małysiak-Mrozek, W. Ding, V. Sunderam, and D. Mrozek, "From corrective to predictive maintenance—A review of maintenance approaches for the power industry," *Sensors*, vol. 23, no. 13, p. 5970, 2023.

[19] "Smart Manufacturing Process Data." Accessed: Oct. 14, 2025. [Online]. Available: https://www.kaggle.com/datasets/programmer3/smart-manufacturing-process-data

[20] P. Iswarya and K. Manikandan, "TransCNN: Deep Hybrid Model for Effective Intermittent Fault Diagnosis in Wireless Sensor Networks," *IEEE Access*, 2025.

[21] A. A. Adiputra, J. Yun, H. Kim, and others, "Anomaly Detection in Industrial Machine Sounds Using High-Frequency Features and Gate Recurrent Unit Networks," *IEEE Access*, 2025.

[22] C. Eang and S. Lee, "Predictive maintenance and fault detection for motor drive control systems in industrial robots using CNN-RNN-based observers," *Sensors*, vol. 25, no. 1, p. 25, 2024.

[23] A. Hosseinzadeh, F. F. Chen, M. Shahin, and H. Bouzary, "A predictive maintenance approach in manufacturing systems via AI-based early failure detection," *Manufacturing Letters*, vol. 35, pp. 1179–1186, 2023.

[24] R. H. Hadi, H. N. Hady, A. M. Hasan, A. Al-Jodah, and A. J. Humaidi, "Improved fault classification for predictive maintenance in industrial IoT based on AutoML: A case study of ball-bearing faults," *Processes*, vol. 11, no. 5, p. 1507, 2023.