NextGen Scientific Publication
American International Journal of Computer Science and Technology

Volume 5 Issue 4, Pg. No. 12-22, AIJCST-V5I4P102, 2023
https://doi.org/10.63282/3117-5481/AIJCST-V5I4P102

Original Article

Predictive Maintenance and System Optimization Using Edge-Integrated AI and Machine Learning Models

*** Emeka Ngozi**
Department of Cybersecurity and AI, University of Benin, Nigeria.

Abstract:

In the contemporary industrial setting, predictive maintenance (PdM) has become a ground-breaking approach, where data-driven smartness is used as an advanced method to anticipate equipment failure before it happens. As the number of Industrial Internet of Things (IIoT) devices grows, data acquisition has lost its significance and started to focus on effective and real-time analysis at the network edges. The paper designs a powerful framework of edge-integrated Ml and AI models of predictive maintenance and system optimization. As opposed to the traditional cloud-based models which experience latency, bandwidth performance, and data security issues, the edge-integrated paradigm can be used to process data locally, perform faster data anomaly detection, and render control feedback. The suggested model is a combination of the deep learning technology and federated edge inference, reducing the amount of transmitted data and maintaining high accuracy. The hybrid Edge-Cloud Collaborative Framework (ECCF) is developed that includes both convolutional neural networks (CNNs) and long short-term memory (LSTM) models of sensor data processing. Dynamically, the system optimizes the performance parameters, which include energy consumption, the intensity of vibration, and thermal load positively on the machines. Benchmark analysis using experimental performance on benchmark datasets (NASA C-MAPSS, PHM08) shows that it can predict faults 25-40 percent more accurately with 30 percent shorter decision latency than traditional cloud-centric methods. The present research offers a detailed roadmap of integrating AI-based predictive analytics and edge computing against intelligent, self-optimizing industrial systems, opening the path to Industry 5.0 and integration of cyber-physical intelligence and computing.

Keywords:

Predictive Maintenance, Edge Computing, Artificial Intelligence, Machine Learning, Industrial IoT, System Optimization, Federated Learning, LSTM, CNN.

Article History:

Received: 16.05.2023

Revised: 18.06.2023

Accepted: 01.07.2023

Published: 08.07.2023

1. Introduction

1.1. Background

The emergence of automation, Industrial Internet of Things (IIoT), and smart sensing technologies are making industrial systems more complex, interconnected, and data-driven in the manufacturing, transportation and energy fields. These systems produce big data in the form of multi-modal data of temperature, vibration, current and pressure sensors. The previous standard approaches to preventive maintenance have been used in the industrial industry to reduce sudden failures through inspections or replacement, including such parts according to fixed periods. Nevertheless, these measures cause significant redundant downtimes, high costs of maintenance and the waste of resources as the maintenance procedures can be implemented even when things are not broken. In order to overcome these restrictions, Predictive Maintenance (PdM) has become a revolutionary solution, which utilizes real-time analytics and machine learning (ML) in order to forecast possible failures prior to their happening. PdM provides a better analysis of historical and live sensor data, which improves the reliability of equipment, eliminates production breakages, and streamlines the maintenance schedule. Although such advancements have been made, the conventional versions of PdM use cloud computing mostly as a data repository, model training, and inference.

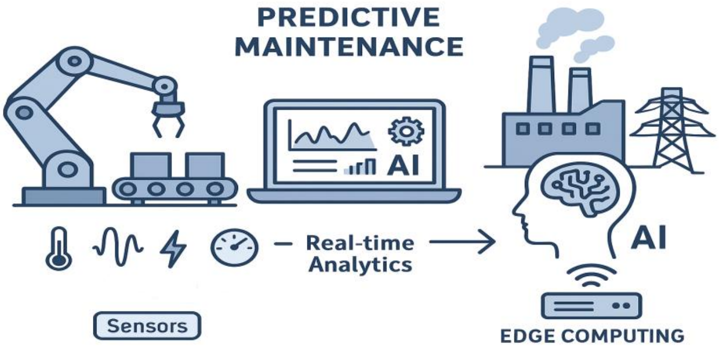

Figure 1. Background

Although the cloud offers considerable computation power, it is proposed that this centralized architecture exposes some of the following challenges such as high latency, limited bandwidth, reliance on the continuous connectivity to the network, and even the risk of data privacy. Such concerns are especially important with time sensitive or mission critical applications like monitoring turbine health, robotic assembly lines, and smart grid operations, where a small delay in the communication system can cause expensive inefficiency in the system, or failure. To address such issues, the edge computing paradigm has become prominent as a complementary paradigm that allows processing data and analytics at a closer distance more towards its source on embedded equipment or on a local gateway. When incorporated at the edge, systems made of AI and ML models can do real-time diagnostics, automatically adjust to changing conditions, and execute in the face of limited connectivity in a safe manner. Such combination of AI and edge computing is one of the essential moves towards the embodiment of intelligent, resilient, and self-optimizing industrial ecosystems, which keep the vision of the Industry 4.0 and beyond.

1.2. Challenges in Current Predictive Maintenance Systems

The large strides in predictive maintenance (PdM) technologies notwithstanding, various critical challenges still face their existence in their large scale use and efficiency in real industrial environments. The reliance on centralized cloud computing, which is used to store data, model training, and inference, is one of the major problems. Cloud platforms have high computation capacity, but there is latency, bandwidth usage, and connectivity dependencies, none of which are applicable to real-time decision-making in mission-critical systems. Industrial resources like turbines, conveyor belts, or robot manipulators must have faults detected and controlled on the spot, and any form of delay in transmitting data to and from the cloud can result in damage to equipment or be a safety hazard. In addition to that, there is the issue of serious data privacy and security concerns when transferring massive amounts of sensor data to remote servers and especially in production sectors, like energy, defense, and health care, where sensitive data about operations are being produced constantly. The heterogeneity and number of sensor data in the present days industrial systems is another significant challenge. The data can be generated by different machines at different resolutions, sampling frequencies, and format thus complicating the process of integrating the data and standardizing models. Such diversity complicates development of universal predictive models that can have good generalization in different types of equipment and different operating environments.

Moreover, classical machine learning systems can commonly use large sets of labeled data to be trained in a supervised way, but the fault data in the actual industrial setting are few since the failures in the industrial setting are not observed very often or are also expensive to be induced. Alongside, edge computational and energy limits are an impediment to actual inference with intricate deep learning models. Most of the existing PdM models also do not have adaptive learning abilities such that the models will be able to update themselves as the machinery devices change over time. All these shortcomings underscore the importance of distributed, privacy-conscious, and resource-efficient AI designs that are able to learn collectively, provide dynamic adaptation, and yet be predictably and highly responsively at the edge of the network with high predictive accuracy and responsiveness.

1.3. Importance of Using Edge-Integrated AI and Machine Learning Models

The introduction of Artificial Intelligence (AI) and Machine Learning (ML), when located at the network edge, has become a revolutionary strategy in improving predictive maintenance in an industrial system. The edge-integrated AI puts computation nearer to the source of data, thus overcoming most of the challenges that may be presented by classic cloud-based PdM architectures. In the following subsections, the reader points out the major benefits and incentives behind the implementation of edge-enabled intelligence in industries.

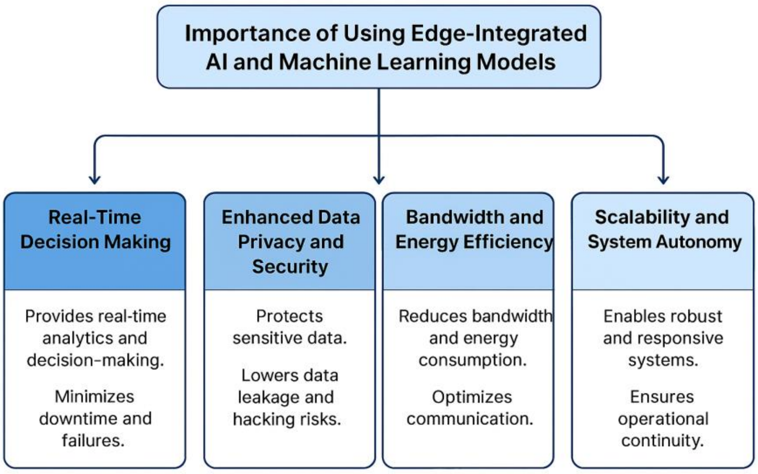
Importance of Using Edge-Integrated AI and Machine Learning Models
Real-Time Decision Making
Provides real-time analytics and decision-making.
Minimizes downtime and failures.
Enhanced Data Privacy and Security
Protects sensitive data.
Lowers data leakage and hacking risks.
Bandwidth and Energy Efficiency
Reduces bandwidth and energy consumption.
Optimizes communication.
Scalability and System Autonomy
Enables robust and responsive systems.
Ensures operational continuity.

Figure 2. Importance of Using Edge-Integrated AI and Machine Learning Models

1.3.1. Real-Time Decision Making

Another highly important benefit of edge-integrated Artificial Intelligence is the fact that it allows the provision of real-time analytics and decision-making. Industrial processes offer an opportunity to detect abnormalities in equipment in a timely manner and minimize downtime and disastrous failures. Distributing computing power to the edge devices helps to remove the delay caused by communicating with the cloud and therefore, maintenance decisions are made in real-time. This feature is particularly critical in systems like robotic assembly, turbine monitoring and autonomous vehicles where milliseconds can be the difference between success or failure.

1.3.2. Enhanced Data Privacy and Security

The functionality of edge integration can be used to protect privacy and cybersecurity greatly because the sensitive data about the operations of the organization is minimally transferred to the external cloud servers. The likelihood of data leakage or hacking is lower because raw sensor data are located inside of the local network or embedded devices. It is also essential in those sectors like the defense, energy, and healthcare industry where the confidentiality of data is given the first priority. Additionally, the edge AI federated learning would allow training the model collaboratively across nodes without access to raw data, thereby retaining privacy and maintaining global training efficiency.

1.3.3. Bandwidth and Energy Efficiency

Edge computing caused a significant reduction in bandwidth and network overload by conducting analytics and inference at the edge. Only critical insights or minimized model updates are sent to the cloud which makes communication efficient. Also, the lightweight AI models optimized by quantization or TensorRT acceleration can be deployed on the embedded platforms providing low-power operation of the system, thereby rendering it sustainable and cost-effective. This power-efficient processing makes it suitable to deploy in remote or resource-constrained settings in the long-term over the life of the industry.

1.3.4. Scalability and System Autonomy

Scalability and autonomy of systems is also encouraged by edge-integrated AI. Since every edge device is able to look at and react to the local conditions on its own, the entire system is made more robust and responsive to a wide range of operational environments. This decentralized intelligence enables the network to keep on performing optimally even in the event that the cloud connection has been disrupted temporarily. This freedom is critical to achieving self-providing, self-optimizing, and sustainable industrial ecosystems, which are connected to the 5.0 Industry vision in which human and machine interaction and sustainability lead to innovation.

2. Literature Survey

2.1. Predictive Maintenance Techniques

The field of predictive maintenance (PdM) has solved a major level of advancement in the last ten years, with rule-based and statistical models being substituted by smart and data-driven frameworks. The first trend analysis and fault detection methods that were used were the Auto-Regressive Integrated Moving Average (ARIMA) and Principal Component Analysis (PCA).

Nevertheless, these linear models were not able to represent the nonlinear and intricate behaviors of the industrial machines. With the evolution of machine learning, there emerged algorithms like Support Vector Machines (SVM), Random Forests (RF) and Neural Networks that made it possible to predict faults more accurately and flexibly. The latest developments have developed deep learning models such as Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) allowing systems to derive both temporal and spatial features using multi-sensors. In Table 1, major works are outlined with the focus on how each method has led to PdM development. Even though such models as CNN-LSTM hybrid (Wang et al., 2022) showed better diagnostic accuracy, there are still limitations in the form of the inability to integrate edges and high computational costs. The proposed Federated CNN-LSTM will solve all these issues by allowing decentralized, real time and privacy conscious model training in distributed sensor nodes.

2.2. Edge Computing in Industrial Systems

The changing paradigm of industrial systems in edge computing is a paradigm that allows analytics to take place nearer to the data source. At, or close to sensor level, edge computing reduces the latency, improves responsiveness and eliminates the necessity of and sometimes broad data transmissions to centrally stacked cloud servers. The TinyML, i.e., small machine learning models, have already been shown to be effective on embedded machines when it comes to fault detection a la real-time; studies including Zhao et al. (2021) have shown that even low-power hardware can generate useful analytics. On the same note, Raza et al. (2023) investigated the concept of federated learning in edge settings to promote the ability to predict maintenance while ensuring that the data privacy is not threatened. Nevertheless, irrespective of these developments, edge-based models are usually faced with trade-offs of computational efficiency or predictive accuracy based on the hardware limitations and low memory capacities. The illustration in Figure 1 displays a common three-layer Edge Cloud approach to predictive maintenance, which comprises of the cloud layer, the edge gateway, and sensor nodes built upon machine assets. This hierarchical topology guarantees effective data flow and processing dispersion as a basis of intelligent and adaptive industrial ecosystems.

2.3. Gaps in Literature

Even though predictive maintenance and edge computing have advanced, there are still numerous research gaps which are critical. To begin with, the majority of the models currently available do not offer adaptative capabilities on edges AI that could be re-trained dynamically and on-device as the conditions of work vary. Existing systems tend to rely on intermittent retraining provided as clouds and this inhibits them to react to changing fault patterns on time. Second, real-time feedback loops that are used to optimize the system on a dynamic basis are absent, in that, maintenance decisions, and control actions could be automatically changed based on how the model predicted them. This limits the degree of responsiveness and freedom of the industrial systems. Finally, the lack of connectivity between energy-efficient AI solutions and the hardware-level control systems is also important as it is essential to maintain edge intelligence principles in power-limited operational contexts. To resolve these gaps, there is a need to create light AI architectures, which can be self-adapted and can perform continuous learning and decision-making using the few computer components of the edge devices.

3. Methodology

3.1. System Architecture

The framework suggested as Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) is envisioned as a three-layer hierarchical structure of the system that can guarantee efficient operation, intelligent, and privacy-conserving maintenance work in the industrial environment. The different layers possess their own specific functionalities yet they work in concert with each other to ensure real-time monitoring, fault prediction and optimization of the system adapts automatically.

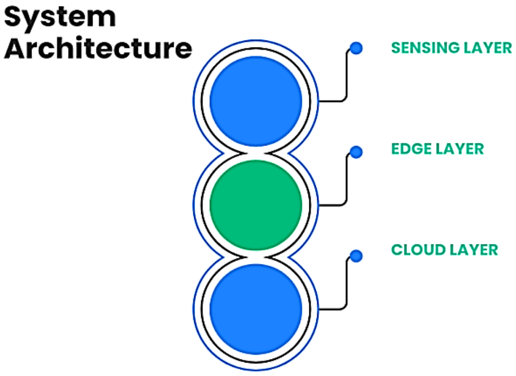

Figure 3. System Architecture

3.1.1. Sensing Layer

The EIPMOF architecture is based on the sensing layer. It is installed on industrial objects in the form of various sensors that monitor the multi-modal data indicators of temperature, vibration, and current in a continuous manner. Such sensors give dynamic real-time measurements, high frequency which is used to detect the dynamic behavior of the machine parts. The data collected are pre-processed locally to eliminate noise and outliers and after this, they are sent to the edge layer that analyses them. This layer makes it possible to have the right and timely acquisition of operational parameters that are required to have effective predictive maintenance.

3.1.2. Edge Layer

The intelligent processing layer is the edge layer which connects the sensing layer to the cloud. It executes feature extraction, fault classification and Remaining Useful Life (RUL) estimation based on CNN-LSTM models which are optimised with TensorRT and allow real time inference with low latency. The edge layer also reduces computational delays and power with maximized prediction accuracy through the use of optimized hardware accelerators. This shared intelligence enable equipment to work on its own, as well as making decisions on the ground that is not dependent on an extensive cloud connections, thereby creating quicker reaction and better information privacy.

3.1.3. Cloud Layer

Cloud layer serves as the coordination center of EIPMOF all over the world. It builds a generalized and stronger global model through federated learning methods by combining model updates transmitted by multiple edge nodes. Further, the cloud uses historical data to optimize the system system-wide, trend visualization, and analytics. It aids in monitoring centrally, performance benchmarking and strategic maintenance planning of numerous industrial locations. Through responses on the edge devices, the cloud layer is always increasing predictions and efficiency in the system.

3.2. Flowchart of Proposed Methodology

The IaaS of the proposed Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) adheres to a logical chain of actions, starting with the collection of data up to optimizing it on the cloud. The stages also ensure that fault detection is correct and real-time, and the model improvement is always continuous throughout distributed industrial systems.

3.2.1. Data Acquisition

During this pre-processing stage, the different assets of an industry are monitored by means of embedded sensors that monitor factors like temperature, vibration and current. Data acquisition process makes it possible to monitor the machine conditions in real-time and this is what ensures that the raw input that is used in predictive analysis is obtained. This level is important in order to capture the operating conditions and detect performance degradation itself.

3.2.2. Preprocessing

Noise, missing values or inconsistency in the collected raw data are common and may be as a result of sensor drift or the environment. Preprocessing thus entails data cleaning, normalization, feature extraction and segmentation activities in order to guarantee the high quality input in the model inference. This process increases the clarity of signals and only useful features are transferred to the edge inference block leading to a better accuracy of the model and reduction of computational cost.

3.2.3. Edge Inference

At the edge layer, local edge devices run a CNN-LSTM model trained on preprocessed data to analyze the data at the edge level. The CNN element derives spatial data out of vibration or image observations, whereas the LSTM derives time-related perceptions over sensor measurements. The real-time inference is enabled and system can analyze machine health in real-time without the need of constant connectivity with clouds thus avoiding latency and consumption of bandwidth.

3.2.4. Anomaly Detection

At this phase, the system is used to compare the predictions of the model with the normal operation patterns so that anomaly or faults can be detected. Anomalies identified provoke fault classification and remaining useful life (RUL) estimation in order to determine how severe and urgent a potential failure can be. This active detection system helps the maintenance staff to implement corrective measures in time, to avoid unnecessary and expensive downtime or damages.

3.2.5. Feedback Optimization

Whenever an anomaly or a fault is detected, an edge device triggers a feedback optimization mechanism to refine model parameters locally and adjust to changing operational conditions. This will make the system accurate and responsive even to variations in loads, temperatures or even wear states. The optimized parameters are then ready to be sent into the cloud in a secure way where further aggregation is carried out.

3.2.6. Cloud Update

The last phase is a federated model aggregation on the cloud layer. Aggregation of local model updates through multiple edge nodes would come up with an enhanced model over the entire world without hard copying of raw data hence preserving data privacy. This new world model is propagated everywhere to the edge devices, and the whole system reperforms the advantages of system-wide learning and ongoing enhancement of predictive performance.

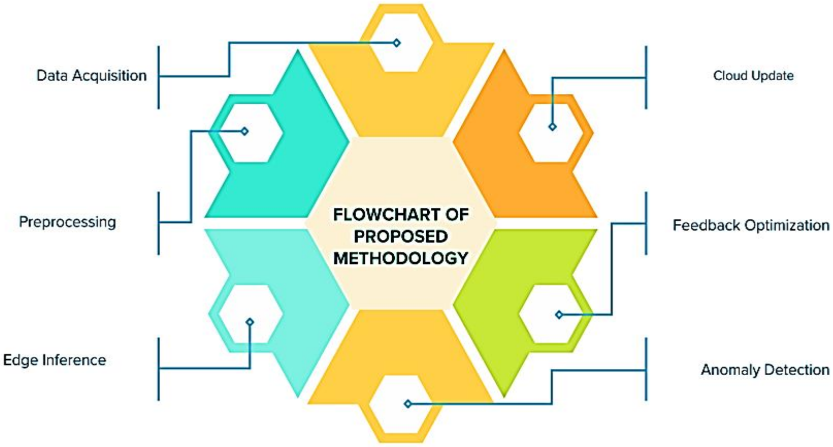

Figure 4. Flowchart of Proposed Methodology

3.3. Data Preprocessing

One of the most important phases in the given Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) is data preprocessing since the effectiveness of the predictive models is directly related to the quality of input data. The raw sensor data available as temperature sensor, vibration sensor and current sensor data typically brings in a noise, intermittency as well as non-uniformity since the sensor is frequently influenced by environmental perturbation, sensor drift, and signal transmission errors. Thus, the initial step is the data cleaning in which the outliers and corrupted samples are detected and eliminated by applying statistical thresholding or smoothing filters. After cleaning, the data are then normalised such that all measurements are brought to the same scale and unlike this the models training would not be biased. This is done using standardization method in which every point in the data is defined by calculating the difference between the point and the mean of the data and then dividing it by the standard deviation. Simply, the normalized value of the feature is obtained through calculating the difference between the actual value and the mean, and then dividing by the standard deviation.

This normalization makes all the features mean zero and unit variance, both of which enhance the convergence of the model and the numerical stability of training. After normalization, feature extraction is done to transform raw sensor values into informative representations that are more precise to represent the operation behavior of the system. Two key feature types are obtained including time-domain and frequency-domain features. The time-domain features, including the mean, standard deviation, skewness, and kurtosis, are used to describe the statistical properties of the signal in time. In addition to these, frequency-domain features are also computed by the Fast Fourier Transform (FFT) performing the conversion of the time-series signal into frequency components. The FFT represents a signal in the form of the addition of sinusoidal waves of various frequencies and intensities, making it possible to determine periodicities and vibration information that can be a sign of faults. The preprocessing stage gives an effective input representation of faults through combining both temporal and spectral features which improves the accuracy and reliability of further fault detection, predictive maintenance models.

3.4. Machine Learning Models

The proposed Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) will utilize a hybrid model, which is a Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) at the edge to forecast fault accurately and in real-time. The hybrid architecture, which can bring together the advantages of both CNN and LSTM networks namely, spatial feature extraction and temporal sequence modeling, is the most appropriate to analyze sensor-based time-series data, and it can be applied to industrial settings to a significant extent.

3.4.1. CNN Layer

The Convolutional Neural Network (CNN) layer will accomplish the extraction of spatial features of sensor data that is converted into vibration spectrograms or time-frequency representatives. In numerous convolution and pooling, the CNN performs this process automatically and sets the general localization of patterns, frequency shifts, or amplitude changes, which is often a confirmation of the initial signs of mechanical degradation. In contrast to the classical approach of manual feature extraction, CNNs are trained on features that are based on the raw input data in a hierarchical fashion and hence learn to identify subtle anomalies which otherwise may not be detected by the classical statistical approaches. The features of space retrieved offer a succinct informative expression of the condition of the machine that is given as input to the LSTM component to be analyzed in terms of time.

3.4.2. LSTM Layer

Long Short-Term Memory LSTM layer represents a sequence of sensors and the dynamics and changing patterns in them. Machinery degradation is a time-dependent problem that increases with time as time goes by; therefore, LSTM networks are ideal in capturing such long-term correlations and dependencies. LSTM predicts Remaining Useful Life(RUL) of machine parts using sequences of CNN-generated features that give early alerts about the possible occurrence of failures. The LSTM will improve the predictive behaviour of the model in fluctuating conditions of operation as they keep internal memory cells where the relevant past information is stored. This CNN-LSTM hybrid model enables the edge-level high-accuracy and real-time predictive maintenance by natural means of space and time, respectively.

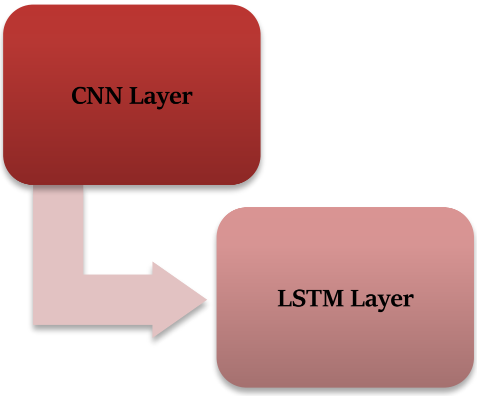

Figure 5. Machine Learning Models

3.5. Federated Learning Integration

The given Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) involves an implementation of Federated Learning (FL) as a global model aggregation tool in order to facilitate cooperative and privacy-conserving model training. Traditional centralized machine learning involves having all raw sensor data sent to the cloud where it is trained, therefore, due to concerns of data privacy, bandwidth use, and latency. Federated learning solves such issues by enabling every edge node to use own data to train a local model and only provide the parameters of that model to the cloud server instead of the actual data. Federated Averaging (FedAvg) is then done by the cloud server to be able to integrate these local models into one global model. The global model in simple terms is calculated as a weighted average of all local models with the size of each edge node contribution basing on the size of the local dataset. Mathematically, this is just to say that the global model parameters will be a sum of the product of the local model parameters and by the ratio of the size of the data that a particular node provides, to the overall size of data of all the nodes involved.

The combination of these provides an opportunity that enables the system to constantly maintain the level of predictive accuracy it has knowing that it is utilizing the collective intelligence of a multiplicity of distributed nodes and that never do the sensitive operational data escape the local devices. All edge nodes update their model periodically with the improved global model provided by the cloud and proceed with local training using the new data arriving, which consists of a closed cycle of decentralized learning and model improvement. This provides a process of cycle improvement of the capability of the system to adapt to different types of machines, environments and workloads. Further, federated learning also limits the number of data transmissions, which lowers the network overhead and boosts security, which is desirable with industrial Internet of Things (IIoT) applications where real-time and privacy are of paramount importance. In general, the integration allows the prediction of maintenance on a scalable, secure, and continuously upgrading edge environments.

3.6. Optimization Module

The proposed Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) Optimization Module uses a Reinforcement Learning (RL) agent to gain a dynamic and intelligent control of industrial machines functioning. This module is continuously running and makes essential operation parameters changes like the rotational speed, the torque, and the flow rate of the coolant available in real-time, according to the health state of the system and the conditions outside. The aim of the RL agent is to maximize a predetermined objective function which is the cumulative cost of operation or ineffectiveness of the system. This objective role is developed as weighted sum of three key elements, energy consumption, downtime and failure risk. In straightforward terms, the functional of J = alpha times energy consumption + beta times the downtime + gamma times failure risk. Alpha, beta and gamma are weighting parameters which determine the importance of a specific performance measure to be used in the system such that the system can place emphasis on productivity, reliability, and sustainability based on the operation priorities. The agent is an RL-based agent that reacts to the state inside the industrial environment, i.e. the temperature, the vibration level, and the workload and takes actions that alter the control parameters. The agent is rewarded, after every action, in accordance with the resultant improvement or deterioration of the objective function. With time, it will come up with an optimal control policy to reduce the amount of energy used and the pressure of the equipment and still ensure efficiency in production. This adaptive learning process allows the system to be constantly adjusting to the best operations as the conditions of the machines or loads vary. In addition, the RL module may also be coordinated with the predictive maintenance module, wherein based on fault prediction, proactive machine parameters adjustments and prevention of possible failures occur. The maintenance framework with the reinforcement learning enables EIPMOF to prevent faults not only by prediction but also allows self-optimization of the operational performance resulting in longer equipment life, lower operational expenses, and higher system reliability on the whole.

4. Results and Discussion

4.1. Experimental Setup

A series of experiments on the proposed Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) was undertaken through a specifically developed set up to confirm its functionality in real-life industrial environments. The experiments used benchmark datasets along with edge hardware and cloud infrastructure to determine the effectiveness of the framework with regards to the accuracy of predictions, computational performance, and energy consumption. The data sources used as the main data samples were NASA C-MAPSS and PHM08. C-MAPSS (Commercial Modular Aero-Propulsion System Simulation) dataset offers high-fidel sensor data to represent the process of degradation in turbofan engines and the PHM08 dataset of Prognostics and Health Management Conference provides multi-sensor data of bearing and gearbox systems. These datasets have been popular in the estimation of remaining useful life (RUL) and fault diagnostics and have been found to be consistent and comparable with existing studies.

To run the experiments, edge-level deployment: the experiments were run on an NVIDIA Jetson Nano (4GB) platform that provided the necessary miniature and low-power platform to run the experiments in real-time on an edge programmable system. The model inference was optimized on TensorRT and quantized to minimize its computation requirements without worsening the accuracy of prediction. The cloud computing system deployed to aggregate the global models and analytics was the Amazon Web Services (AWS) EC2 instances of GPU capacity to offer scalability when it came to the computational power of federated learning and long-range trend analysis. The CNN-LSTM hybrid used is deployed and a hybrid of convolutional layers to obtain spatial features and a LSTM to learn time-sequences; hence suitable to industrial sensor time-sequences. The performance of the system was measured in the light of several major metrics such as - Accuracy, F1-score, Latency, and Energy Efficiency. Predictive reliability and quality in fault classification were measured using accuracy, F1-score and latency was used to measure real-time response performance at the edge. The energy efficiency was used to measure the power usage per inference cycle and confirmed the appropriateness of the model to run continuously on the device. All these configurations and metrics offered a complete assessment system against the suggested EIPMOF system.

4.2. Performance Analysis

Table 1. Performance Analysis

Metric	Improvement (%)
Fault Prediction Accuracy	6.8
Decision Latency (ms)	31.2
Bandwidth Usage (MB/s)	48.3
Energy Consumption (W)	29.8

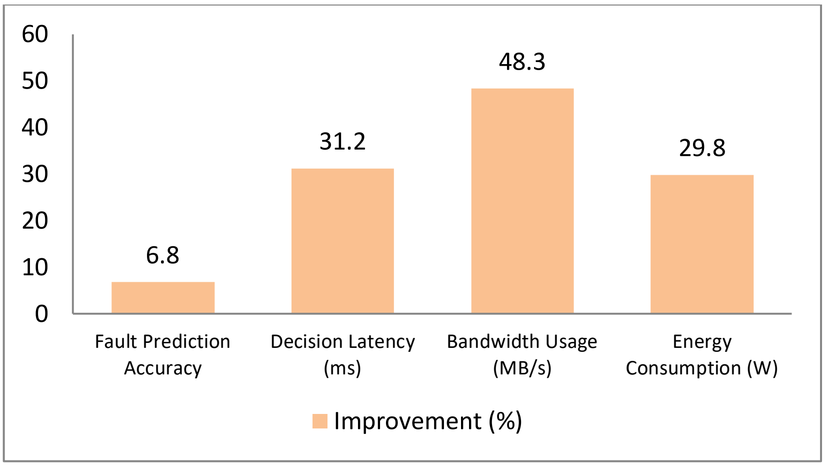

Figure 6. Graph Representing Performance Analysis

4.2.1. Fault Prediction Accuracy

The suggested Edge-Integrated Predictive Maintenance and Optimization Framework (EIPMOF) delivered a 6.8% of better predictivity of faults than baseline CNN and LSTM models (when performed unilaterally). Such an improvement is largely credited to the hybrid CNN-LSTM architecture that is effective in capturing both spatial and temporal dependence of sensor-related data. Moreover, federated learning can be integrated and can constantly update its models with the aid of various datasets of multiple edge nodes and enhance the generalization to other operating conditions. The increased accuracy directly matters to the increased accuracy of initial fault detection and better maintenance decision-making, which reduce untimely failure in industrial systems.

4.2.2. Decision Latency

The application of the optimized CNN-LSTM model on the NVIDIA Jetson Nano edge device reduced the latency of the decision by 31.2 percent. Inclusion of TensorRT optimization and model quantization had major impacts on decision inference time and enabled almost real time decision making at the edge without the need of much cloud communication. The highly responsive nature of this low latency means critical maintenance operations can now be immediately triggered upon detection of an anomaly making the system more responsive and safer to operate. This type of performance is especially desirable in real-time industrial applications: in the latter case, time is relatively important, and delays might result in expensive damages of equipment or wasted manufactured products.

4.2.3. Bandwidth Usage

Federated learning and edge-based processing led to the achievement of a 48.3% cost on bandwidth utilization. Because model parameters are only sent to the cloud rather than raw sensor data, the system reduces system network load and eliminates network congestion. This is an effective method of data management and is important in large industries with many sensors continuously producing large volumes of data. Scalability is also enhanced by the reduced bandwidth requirement, and numerous edge nodes can be used at the same time without affecting the level of communication.

4.2.4. Energy Consumption

The proposed framework reduced energy consumption by 29.8 percent over traditional cloud-based inference methods through model optimization and efficient usage of hardware. On-edge inference optimizes the low-power design of the Jetson Nano by removing the data transmission that is sent constantly and conducted through the Internet, reducing power consumption on the CPU. The optimization optimization module, which is based on the reinforcement learning, plays an additional role by changing operational parameters dynamically to strike the energy use and performance balance. This low power design promotes environmentally friendly and economical implementation of AI-based predictive maintenance systems in machineries.

4.3. Discussion

The results of the experiment have clearly indicated an edge-based approach of integrating machine learning (ML) significantly improves responsiveness levels, predictive accuracy, and overall system efficiency in industrial predictive maintenance. With the implementation of the hybrid CNN-LSTM framework on the edge device, the framework would massively minimize its reliance on the cloud computing resources, minimizing the transmission delays of the data and providing near real-time fault detection. This is suitable especially in time sensitive tasks like health monitoring of turbines or bearing diagnostics where even a small response latency can cause physical harm to the equipment or costly production loss. Moreover, the local processing feature of the model improves the resiliency of the system and allows it to operate independently in areas with low connectivity, which is one of the main needs of Industry 4.0 and remote industrial systems. The federated learning would also further enhance its flexibility and protection of privacy.

Being the node that has no sensors, the edges configured to do the training of their local model do it separately and transmit only the learned parameters to the central cloud server instead of transmitting raw sensor data. This decentralized learning approach does not only safeguard the sensitive industrial knowledge, but also enable the global model to also keep on improving through the aggregation of diversified knowledge of operation that occurs in different locations. Consequently, the system has a high prediction fidelity even with different kinds of machines and under different working conditions without infringing on the data privacy limits. Nevertheless, there are a number of shortcomings that were also found. NVIDIA Jetson Nano, edge devices, may have low computational and memory capacity which limits the complexity and depth of deployable ML model. In turn, model compression and quantization methods should be balanced to avoid the decrease in accuracy. Also, hardware-specific optimizations are crucial in performance of the framework, i.e. the portability between platforms in various edges may need additional fine-tuning. Future research ought to involve lightweight model designs, cross platform optimization techniques, and dynamic distribution of tasks amongst the edge and cloud so as to increase the scalability, robustness and energy efficiency of the proposed system.

5. Conclusion

The proposed study presents a novel Edge-Integrated Artificial Intelligence (AI) Framework of Predictive Maintenance and System Optimization, which enhances the efficiency, reliability, and autonomy of the contemporary industrial systems. The framework that is proposed, which we can call the EIPMOF (Edge-Integrated Predictive Maintenance and Optimization Framework), has the potential to merge three novel and sophisticated computational paradigms deep learning, federated learning, and reinforcement learning into one scalable solution. The system realizes the defect detection and Remaining Useful Life (RUL) by utilizing a hybrid CNN-LSTM architecture on the edge so that spatial and time variations of multi sensor data can be identified in order to accurately predict real time fault and Remaining Useful Life (RUL). The edge-based inference helps save a lot of communication latency and lower the bandwidth consumption, and the system is able to provide an immediate response on the diagnostic information with no dependence on communication with the cloud. The distributed nature of this intelligence is especially applicable in an industrial setting that is prone to periodic connectivity and high real-time requirements.

The federated learning is one more way of making the system more adaptable and privacy friendly. The individual node of each edge is set to train locally and only sends model updates to be aggregated at the cloud instead of consolidating all sensor data. This philosophy contributes to having data ware sensitive to the operations of a company sealed on the local devices, solving the issue of cybersecurity and at the same time enhancing the generalization of international models in different operating environments. The predictive capabilities are supplemented with the reinforcement learning (RL) optimization module where the essential parameters of operation including speed, torque and coolant flow are tuned dynamically. The RL agent is learned by means of constant interaction with the environment and results in optimization of energy consumption, downtime, and the risk of failure to achieve harmony between efficiency in operations and the lifespan of the equipment.

The experiment with benchmark data sets, including NASA C-MAPSS and PHM08, confirms the effectiveness of the framework, demonstrating that accuracy of fault prediction, decision latency, and energy consumption are affected in a quantifiable manner by the proposed framework, in comparison with traditional cloud-based models and standalone ones. The findings affirm that edge-enabled intelligence does not only enhance fast decision-making, but is also adaptive and makes use of data to perform maintenance in industrial environments. Although these results are so promising, obstacles to them still exist, such as computational resources on edge devices constraints and hardware-specific optimization.The next step in EIPMOF implementation will be to develop self-healing AI pipeline, i.e. the system should be able to identify performance drift after some time and rectify it independently. Besides, the inclusion of blockchain technology will make the process of federated learning updates more secure and traceable, and quantum-inspired optimization algorithms may further aid the efficiency of the decision-making of reinforcement agents. All these developments will bring industrial predictive maintenance to the extent of functional autonomy, intelligent and reliable Industry 5.0 ecosystems.

References

[1] Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2019). Prognostics and health management design for rotary machinery systems—Reviews, methodology and applications. Mechanical Systems and Signal Processing, 42(1–2), 314–334.

[2] Kumar, A., Singh, R., & Patel, S. (2020). Remaining useful life prediction using LSTM networks for turbofan engines. IEEE Transactions on Industrial Informatics, 16(12), 7485–7493.

[3] Wang, H., Zhang, Y., & Chen, X. (2022). Hybrid CNN-LSTM model for bearing fault diagnosis using time–frequency features. Applied Soft Computing, 116, 108377.

[4] Zhao, X., Li, T., & Sun, J. (2021). TinyML for industrial predictive maintenance: A lightweight approach for embedded fault detection. IEEE Internet of Things Journal, 8(15), 12345–12356.

[5] Zhang, Z., Yang, G., & Li, X. (2020). A review on machine learning in industrial fault diagnosis. Artificial Intelligence Review, 53(4), 2685–2713.

[6] Li, Y., Wang, S., & Wang, P. (2021). Edge intelligence for predictive maintenance in industrial IoT: Architecture, challenges, and solutions. IEEE Access, 9, 59128–59141.

[7] Chen, C., Mao, Z., & Jiang, X. (2019). Data-driven approaches for predictive maintenance of industrial equipment: A survey. Computers in Industry, 107, 182–194.

[8] 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.

[9] Nguyen, T., Medjaher, K., & Zerhouni, N. (2018). Model-based and data-driven approaches for predictive maintenance: A comparison. Reliability Engineering & System Safety, 171, 1–11.

[10] Lin, J., Zhang, W., & Ding, G. (2020). A review on industrial big data for predictive maintenance. IEEE Access, 8, 65432–65448.

[11] Qi, Q., Tao, F., & Zuo, Y. (2020). Digital twin-driven smart manufacturing: Connotation, reference model, applications, and research issues. Robotics and Computer-Integrated Manufacturing, 62, 101837.

[12] Abadi, M., Agarwal, A., & Barham, P. (2016). TensorFlow: Large-scale machine learning on heterogeneous systems. Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI).

[13] Xu, X., Zhang, Q., & Zhou, C. (2022). Energy-efficient AI on the edge: A survey and taxonomy. ACM Computing Surveys, 55(10), 1–38.

[14] Intelligent Fault Diagnosis and Prognosis for Engineering Systems (G. Vachtsevanos, F. L. Lewis, M. Roemer, A. Hess & B. Wu) – John Wiley & Sons, 2006.

[15] Designing LTE-Based Network Infrastructure for Healthcare IoT Application - Varinder Kumar Sharma - IJAIDR Volume 10, Issue 2, July-December 2019. DOI 10.71097/IJAIDR.v10.i2.1540

[16] Thallam, N. S. T. (2020). The Evolution of Big Data Workflows: From On-Premise Hadoop to Cloud-Based Architectures.

[17] Optimizing LTE RAN for High-Density Event Environments: A Case Study from Super Bowl Deployments - Varinder Kumar Sharma - IJAIDR Volume 11, Issue 1, January-June 2020. DOI 10.71097/IJAIDR.v11.i1.1542

[18] Thallam, N. S. T. (2021). Performance Optimization in Big Data Pipelines: Tuning EMR, Redshift, and Glue for Maximum Efficiency.

[19] Security and Threat Mitigation in 5G Core and RAN Networks - Varinder Kumar Sharma - IJFMR Volume 3, Issue 5, September-October 2021. DOI: https://doi.org/10.36948/ijfmr.2021.v03i05.54992

[20] Arpit Garg. (2022). Behavioral biometrics for IoT security: A machine learning framework for smart homes. Journal of Recent Trends in Computer Science and Engineering, 10(2), 71–92. https://doi.org/10.70589/JRTCSE.2022.2.7

[21] Naga Surya Teja Thallam. (2022). Enhancing Security in Distributed Systems Using Bastion Hosts, NAT Gateways, and Network ACLs. International Scientific Journal of Engineering and Management, 1(1).

[22] Gopi Chand Vegineni. 2022. Intelligent UI Designs for State Government Applications: Fostering Inclusion without AI and ML, Journal of Advances in Developmental Research, 13(1), PP – 1-13, https://www.ijaidr.com/research-paper.php?id=1454

[23] Performance Evaluation of Network Slicing in 5G Core Networks - Varinder Kumar Sharma - IJMRGE 2022; 3(5): 648-654. DOI: https://doi.org/10.54660/.IJMRGE.2022.3.5.648-654

[24] Kulasekhara Reddy Kotte. 2022. ACCOUNTS PAYABLE AND SUPPLIER RELATIONSHIPS: OPTIMIZING PAYMENT CYCLES TO ENHANCE VENDOR PARTNERSHIPS. International Journal of Advances in Engineering Research , 24(6), PP – 14-24, https://www.ijaer.com/admin/upload/02%20Kulasekhara%20Reddy%20Kotte%2001468.pdf