Real-Time Predictive Maintenance for Industrial IoT Using Edge Intelligence

Juan Neyra*, Avinash Kankari*, Zakir Afzal Khan *
* University of Bremen, Bremen, Germany
Emails: jneyra@uni-bremen.de, kankari1@uni-bremen.de, zakkhan@uni-bremen.de

***Abstract*—Machinery downtime in industrial settings can lead to significant productivity losses and revenue reduction. Unplanned maintenance and sudden equipment failures pose major operational challenges. This project focuses on developing a predictive maintenance system using edge computing to monitor machine health, forecast maintenance needs, and mitigate unexpected failures.**

Edge computing is well-suited for predictive maintenance as it facilitates real-time monitoring and local data analysis, enabling immediate anomaly detection and prompt action. The system features real-time data collection through sensors, predictive maintenance alerts to schedule repairs before critical breakdowns, and automated logging of maintenance activities.

Machine learning techniques play a pivotal role in the system, utilizing unsupervised learning methods such as Autoencoders and Isolation Forests for anomaly detection, deep learning models like Convolutional Neural Networks (CNN) and classification algorithms like Support Vector Machines (SVM) to categorize faults by type and severity.

The system is deployed on edge computing boards, integrated with temperature and acoustic sensors for comprehensive machinery condition monitoring. By leveraging edge computing and advanced machine learning algorithms, this system aims to minimize unplanned downtime, enhancing operational efficiency in industrial environments.

Keywords- Edge computing, predictive maintenance, anomaly detection, machine learning, Long Short-Term Memory (LSTM), Support Vector Machines (SVM), remaining useful life (RUL), ESP-EYE-S3, sensors.

I. Introduction

The Internet of Things(IoT) is meant to improve our quality of life. Since its introduction, almost all businesses have employed it in some capacity. For instance, the IoT is extensively utilized. For smart industrial, smart agriculture, smart healthcare, smart grid, and other domains. Information and communication technology is used by the smart industry, also known as Industry 4.0, to facilitate efficient production. IoT has attracted a lot of attention from academics lately. It is also referred to as Industrial IoT (IIoT) in the context of industry. The IIoT makes use of a variety of sensors to keep an eye on how machinery or even entire industrial processes are operating. Predictive Maintenance (PM), an IIoT method, has garnered attention recently.

Multiple sensors are utilized in IIoT to monitor the operation of machines or a whole production cycle. A recent development in IIoT is the concept known as Predictive Maintenance (PM) support from sensor data to forecast future equipment deterioration or failure. PM uses machine learning (ML) to provide forecasts based on the collected information. The accuracy of the ML models is determined by the acquired data. One common technique for gathering data in the IIoT is streaming data from sensing devices to the cloud, where it is analyzed and modeled. Large volumes of data are often produced by sensing equipment, either constantly or intermittently, and often in a very short time. In little than a second, for example, a computer may generate thousands of records. To do this, a variety of techniques including filtering, compression, and sampling are used. These techniques can reduce the quantity of data sent to the cloud. There can be trade-offs in accuracy for machine learning models that employ fewer datasets [1].

The way that individuals get information has changed significantly in recent decades due to technological developments. Numerous sources are now being used to collect data. Smartphones, wearable technology, and other smart gadgets are becoming necessities. Data is being collected faster than ever before, yet its structure is getting more intricate. The primary difficulty lies in swiftly assessing large amounts of data collected from diverse sources and contexts to use it for better decision-making in critical domains such as real-time operations management, device data processing and evaluation, and fast defect and risk detection [2], [3].

The production environment cannot operate if the equipment is not operating correctly. A single system or component failure might bring the production line to a complete halt. Production stops have significant costs associated with them, such as lost output and downtime, time and effort lost in identifying and resolving the issue, wasteful products produced due to low quality as soon as the system was operational again, repair expenses, and deteriorating equipment [1].

There might be several reasons for the stoppage, such as malfunctioning machinery, errors made by people, or unfavorable weather. Although it may not be feasible to completely eradicate system malfunctions, a significant portion of production halt might be linked to preventable mistakes. Early identification of likely issues enables prompt resolution without needing to halt production, saving money on associated costs and time [1].

Predictive maintenance relies on continuous monitoring of machine performance through sensors that collect real-time data. By leveraging machine learning (ML) algorithms, these systems can analyze historical and real-time data to predict the remaining useful life (RUL) of equipment and detect anomalies in the machine's operation. This proactive approach to maintenance not only reduces unexpected downtime but also extends the lifespan of industrial assets by optimizing maintenance schedules.

Traditional cloud-based predictive maintenance systems, while effective, have latency issues due to data transmission times and dependence on network connectivity. These limitations can delay the identification of critical failures, resulting in operational inefficiencies. To address this, edge computing is increasingly being integrated with PdM solutions. Edge computing enables data processing closer to the source, allowing real-time analysis and decision-making without the need for constant communication with a centralized cloud server

In this project, we develop an edge computing-based predictive maintenance system that processes machine data in real time using a combination of temperature and acoustic sensors. The system integrates machine learning models, such as Autoencoders for anomaly detection, Convolutional Neural Networks (CNN) and Support Vector Machines (SVM) for fault classification. The use of edge devices, specifically ESP−EYE-S3 boards, ensures low latency and high reliability in detecting faults before they become critical. This solution addresses the key challenges in industrial maintenance by minimizing downtime, optimizing repair schedules, and reducing maintenance costs.

II. RELATED WORKS

The authors of the study discuss how predictive maintenance for an industrial fan can be implemented through the integration of the Internet of Things (IoT) and machine learning. They outline the methods employed for gathering, interpreting, and analyzing real-time sensor data from the fan to anticipate maintenance requirements. The research demonstrates the application of these technologies within a particular industrial context. The study also defines predictive maintenance, explores its industrial applications, and explains how IoT and machine learning can enhance maintenance processes. It provides a rationale for incorporating IoT and machine learning to develop data-driven, preventive maintenance strategies. Additionally, it explores IoT-based condition monitoring, focusing on IoT sensor technologies for real-time data collection and monitoring. Various types of condition monitoring sensors, such as vibration, temperature, and audio sensors, are examined. The study also covers efficient data collection and analysis using wireless sensor networks and edge computing [4].

Data analytics and feature extraction: In predictive maintenance, machine learning techniques are evaluated for their effectiveness in data analysis and pattern recognition. The study compares methods for anomaly detection and defect identification using both supervised and unsupervised learning approaches. It also provides an overview of feature extraction techniques used to identify key features from sensor data. Fault diagnosis methods based on machine learning are explored, focusing on error detection and prediction. Prognostic models are investigated to estimate the remaining useful life (RUL) of industrial equipment, with a discussion on combining time-series analysis and anomaly detection for fault detection and prediction.

Decision-Support Technologies: The analysis covers decision-support technologies for preventive maintenance, driven by machine learning and IoT. Techniques to improve equipment uptime and optimize repair scheduling are summarized, along with an assessment of the return on investment (ROI) and cost-effectiveness of predictive maintenance systems.

Problems and Suggested Action Plans: Issues such as cybersecurity, scalability, and data quality are highlighted, along with other challenges in implementing IoT and machine learning for predictive maintenance. The study discusses future directions, including edge intelligence, hybrid modeling, and explainable AI. Privacy and security concerns, especially in healthcare, are emphasized due to the sensitive nature of healthcare data in predictive maintenance systems.

This study explores the risks to patient privacy when medical records are collected and used for predictive maintenance. It focuses primarily on proactive maintenance techniques rather than just failure prediction. The research examines prediction models for adaptive maintenance strategies that continuously learn and adjust to changing equipment conditions. It offers insights into practical challenges, implementation methods, and lessons learned from applying predictive maintenance in healthcare settings. The reviewed studies provide a solid foundation for future research, offering valuable information on developing effective predictive maintenance algorithms, integrating them with current systems, addressing privacy and security issues, and utilizing proactive maintenance approaches.

As the availability of data increases, the use of machine learning and reasoning for predictive maintenance is becoming more widespread. Machine learning experts, in particular, are essential in addressing several challenges, such as the need for precise prediction models, real-time demand monitoring while handling latency, scalability, and integrating data from various sources (source heterogeneity). Consequently, the key benefits of machine learning models in this context can be explained in terms of interpretability, prediction accuracy, and computational efficiency [5].

Multiple Instance Learning (MIL) is a type of supervised learning where a set of instances, known as a bag, is assigned a single class label, either positive or negative. The objective is to develop a classifier that can accurately classify unseen bags or individual instances, using bags obtained from various sources and at different times. However, to the best of our knowledge, no research has provided similar guidance on using data logs from CNC machine tools to achieve predictive maintenance (PDM) or to predict the failure of specific mechanical components (such as roller bearings, motor inverters, etc.) [6].

III. Methodology

The proposed methodology integrates edge intelligence with predictive maintenance strategies in Industrial IoT (IIoT) environments, utilizing deep learning models on resource-constrained devices. The core of the system is built around the ESP32 microcontroller, chosen for its low power consumption and capability to support real-time data processing. The ESP32 is leveraged to perform on-device inference using deep learning techniques, enabling real-time detection of anomalies in industrial equipment.

The methodology is designed to address the unique challenges of IIoT systems, where network latency, bandwidth limitations, and real-time constraints must be managed efficiently. The deep learning model is trained to classify normal and anomalous behaviors of industrial machinery, using audio signals as input. These signals are processed locally on the ESP32, which ensures timely prediction and minimizes reliance on cloud-based infrastructure. In this section, we detail the processes involved in data collection, model development, deployment on the ESP32, and integration with edge intelligence techniques to achieve real-time predictive maintenance.

As previously mentioned, the ESP32 was selected as the device for executing the model. Specifically, the ESP32-S3-EYE variant was chosen due to its integrated digital microphone. Additionally, it is equipped with 8 MB of Octal PSRAM and 8 MB of flash memory, which provides sufficient resources for real-time neural network execution. The device also benefits from support provided by the ESP-IDF libraries, real-time operating system (RTOS) capabilities, and compatibility with the Edge Impulse platform, facilitating the deployment of the model on the device. A small summary of the device's capabilities are shown on Table I.

Feature	Value
SoC	ESP32-S3 - dual core 32 bit 240 MHz
PSRAM	8MB Octal PSRAM
Flash	8MB
Integrated sensors	camera, internal microphone, accelerometer
Graphical interface	LCD screen
Available Network Options	WiFi and bluetooth
Internal microphone	61 dB SNR and -26 dBFS sensitivity

TABLE I
ESP32-S3-EYE

The anomaly detection algorithm is illustrated in Figure 1. Audio signals are captured from the microphone and classified by a deep learning model into two categories: normal behavior (high speed) and abnormal behavior (low speed). In conjunction with temperature readings, a decision-making algorithm further categorizes the signals into four conditions:

- normal behavior: high speed and temperature below threshold (T max), display color: green
- abnormal temperature: high speed and temperature above threshold (T max), display color: yellow
- abnormal speed: low speed and temperature below threshold (T max), display color: blue
- abnormal behavior: low speed and temperature above threshold (T max), display color: red

normal, temperature alert, noise alert, and abnormal. For each category, the device will adjust the color of the LCD display to indicate the corresponding alert status.

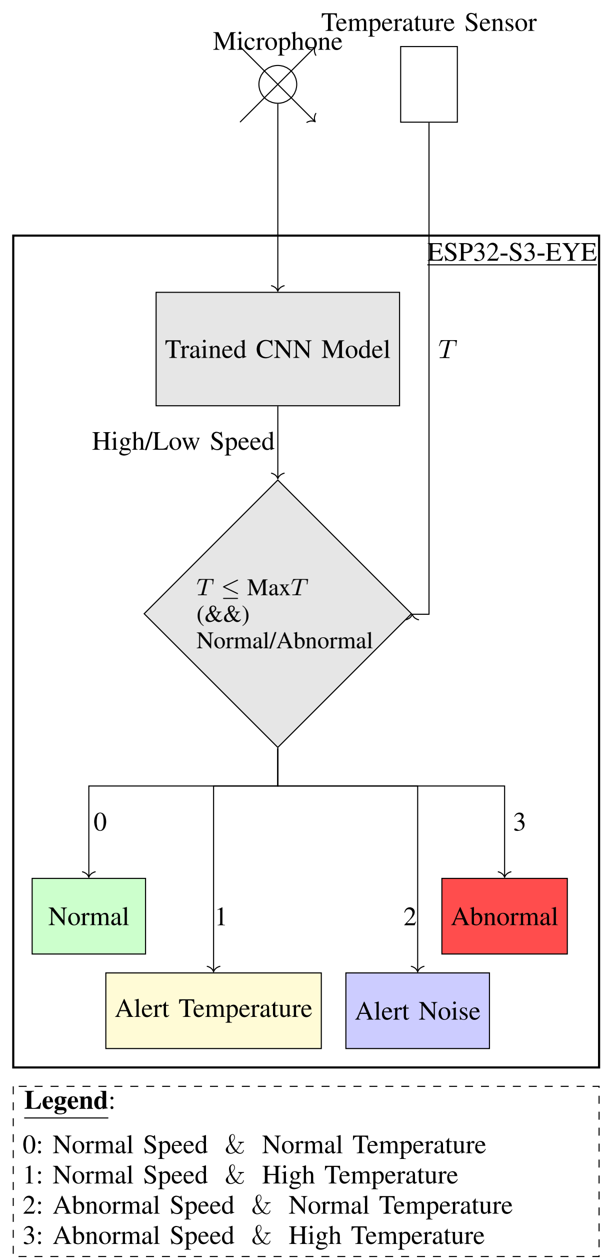


Fig. 1. Block Diagram

Samples were obtained by recording the sound of a ventilator (within a hair dryer) over a 30-second period, through a simple smartphone's recorder. Two classes were designated for the classifier: normal behavior and anomalous behavior. The machine's low-speed mode was employed to model anomalous behavior, while the high-speed mode represented normal behavior. A total of 10 samples were recorded for the anomalous behavior and 9 for the normal behavior. These samples were subsequently augmented through resampling, resulting in 14 samples for anomalous behavior and 16 samples for normal behavior.

The audio signals were captured by the device at a sampling frequency of 16 kHz and a resolution of 16 bits per sample. Signal processing was carried out in a separate thread, running concurrently with the main thread. The processed data were stored in a ring buffer, which was accessible to the main program.

Due to the memory limitations of the device, the proposed model was adapted to utilize input segments of 1 milliseconds and window increases of 512 ms. The model was developed using the Edge Impulse platform, enabling the creation of a library that could be seamlessly integrated into the core software. Mel Frequency Cepstral Coefficients(MFCC) were calculated using the following parameters:

- number of coefficients: 13
- FFT length: 256
- number of filters: 32
- frame length: 0.02

Following the implementation on the device, the proposed model will be further refined to adhere to constraints related to time and accuracy, culminating in the development of the final model as outlined below:

- Input Layer (663 features)
- Reshape layer (13 columns)
- 1D conv layer (8 neurons, 3 kernel size)
- Dropout (0.25 rate)
- Flatten layer
- Output layer (2 classes)

The 663 features in the input layer correspond to the number of features extracted through MFCC processing. The model and the preprocessing algorithm can then be quickly deployed into a library by a tool offered by Edge-Impulse platform.

A maximum temperature threshold of 30° was established. Temperature readings exceeding this threshold were classified as anomalies. Additionally, to display the results on the LCD screen, the ”lvgl” library, integrated within the ESP32 framework, was utilized.

IV. Results

The following section presents the outcomes of the real-time predictive maintenance system implemented on the ESP32 platform, utilizing edge intelligence techniques. The results focus on the system's ability to process audio signals, detect anomalies, and perform predictive maintenance tasks with minimal latency. Performance metrics such as model accuracy, inference time, and memory usage are analyzed to assess the viability of deploying deep learning models on resource-constrained devices for industrial IoT (IIoT) applications.

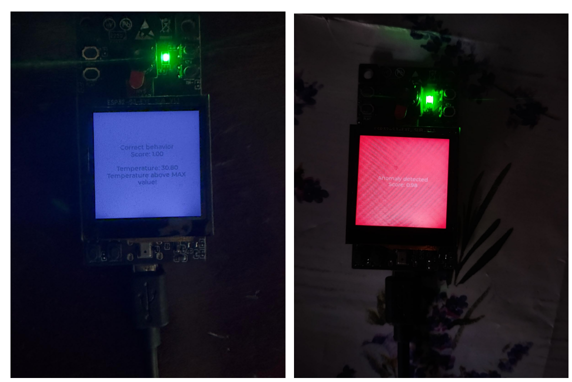
Specifically, we evaluate the model's classification accuracy on both normal and anomalous behavior, as well as its ability to operate in real-time under practical conditions. These results demonstrate the system's potential for improving equipment reliability while maintaining the constraints of real-time industrial environments.

The model was trained on the Edge Impulse platform, achieving a training accuracy of 100% and a test accuracy of 98.76%. The model successfully classified all normal behavior samples, however, it exhibited some challenges in identifying anomalous behavior, correctly classifying only 97.5% of the anomaly samples (confusion matrix shown in Table II).

	High Speed	Low Speed	Uncertain
High Speed	100%	0%	0%
Low Speed	1.7%	97.5%	0.8%
F1 Score	0.99	0.99	-

TABLE II
CONFUSION MATRIX FOR TEST SET

(a) Normal (b) High Temperature



(c) High Speed (d) Abnormal

Fig. 2. Detection of anomalies on device

Figure 2, illustrating the results of model's ability to classify normal behavior and detect anomalies based on real-time data. Alongside, continuously monitoring temperature of the module and helps to make a decision and displayed on the ESP32-S3-EYE LCD screen. For the purpose of testing, the device was artificially heated to assess its response to elevated temperatures. Furthermore, when applied in a real-world scenario, the model could demonstrate successful performance in Industries. Each of the four predefined categories was evaluated on the device. The LCD screen displayed corresponding messages based on the predicted outcome, along with the inference score and the recorded temperature value for each instance.

- Normal speed and temperature: Correct behavior
- Anomalous speed and normal temperature: Anomaly detected
- Normal speed and abnormal temperature: Correct behavior. Temperature above MAX value!
- Anomalous speed and abnormal temperature: Anomaly detected. Temperature above MAX value!

To further evaluate the model's performance, ten real-time tests were conducted for each category (normal and abnormal speed). The classifier, implemented using the Edge Impulse library, returns a score representing the model's confidence in classifying each instance. These confidence scores were utilized to generate the bar graph shown in Figure 3.

The bar graph presents the detection scores for both normal and abnormal speed conditions. Each test is represented along the x-axis, while the y-axis indicates the detection scores, ranging from 0 to 1.

Normal Speed Scores (Blue Bars): The blue bars represent the detection scores for normal speed conditions. As observed, the scores are consistently high, with most readings nearing 1. This indicates that the model accurately identified the normal speed conditions, demonstrating its effectiveness in detecting proper machine behavior.

Abnormal Speed Scores (Red Bars): In contrast, the red bars depict the detection scores for abnormal speed conditions. The scores vary but tend to be lower than those for normal speed. This suggests that while the model is still able to detect abnormal speeds, it is less reliable compared to its performance on normal speeds. The variability in these scores indicates that there may be challenges in correctly identifying all instances of abnormal behavior.

Overall, this graph visually illustrates the model's strong performance in detecting normal machine speeds while highlighting areas for improvement in accurately identifying abnormal conditions. The results underscore the model's potential utility in industrial applications where precise machine condition monitoring is critical.

Edge-impulse's library also has a tool 'edge-impulse-sdk' to read the overall time for signal processing MFCC and inference (model's prediction). The processing and inference times for the model were 489 ms and 2 ms respectively. By utilizing audio capture in a separate thread, real-time prediction was achieved. Specifically, the device required an average of 1 second per prediction. This is can read from the same library.

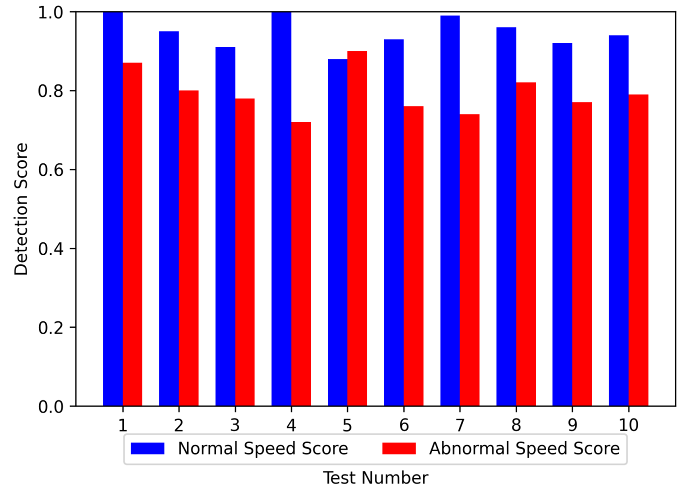


Fig. 3. Detection Scores for Normal and Abnormal Speeds. Blue bars represent normal speed score and red bars abnormal speed scores

V. Discussion and Conclusion

In this project, we utilized two datasets: Normal Speed and Abnormal Speed recordings, with temperature being measured alongside but not directly inputted into the model. Despite this, we observed that temperature influences the device's decision-making process. The model outputs four combinations: normal speed and normal temperature, normal speed and high temperature, high speed and normal temperature, and high speed with high temperature. This reveals the potential for multi-dimensional assessment in machine condition monitoring, where sound acts as the primary indicator, but additional parameters like temperature can indirectly affect decision-making. This approach provides several benefits, especially in industrial applications where machine sound data can indicate potential wear, faults, or malfunctions. With the implementation of edge computing, this model generally reduces latency because real-time monitoring and localized processing enable quicker fault detection and response, which is critical for predictive maintenance, helping prevent costly machine downtimes or failures. The model, trained on industrial machine sound data, has shown its ability to classify speed conditions accurately as either normal or abnormal. However, potential improvements can be made by incorporating real-time temperature measurements directly into the model, enhancing its accuracy and offering a more comprehensive assessment of machine health. Furthermore, integrating more powerful processors and additional sensors could optimize performance, paving the way for more advanced edge computing solutions in predictive maintenance. This device could be installed on critical machines in process or power industries, such as gas compressors, pumps, or other vital equipment, where reducing machine downtime is essential. This solution would have a considerable impact on industrial operations by enabling faster, localized decisions and preventing failures before they occur.

References

[1] A. Jadhav, R. Gaikwad, T. Patekar, S. Dighe, B. Shaikh, and N. S. Patankar, "Predictive maintenance of industrial equipment using iot and machine learning," *IEEE Automation and Knowledge Management (ICCAKM 2023)*, 2023.

[2] S. R., V. A., , and S. V., “Predictive maintenance of industrial equipment using machine learning and iot,” *In 2019 International Conference on Advances in Electrical Engineering (ICAEE)*, pp. 1–6, 2020.

[3] M. F. H. Khan, M. R. Hasan, and M. S. Islam, “Predictive maintenance of industrial machines using iot and machine learning,” *IEEE Automation and Knowledge Management (ICCAKM 2023)*, pp. 110–115, 2019.

[4] S. Sharma, S. Rani, and K. Singh, “A survey on predictive maintenance using machine learning technique,” *International Journal of Computer Applications*, pp. 13–18, 2019.

[5] Y. Liu, M. Pasquier, and Y. Gousseau, “Deep learning for time-series analysis,” *arXiv:1809.04356*, 2019.

[6] J. Lin, G. Wei, X. Sun, Z. Yan, and B. Luo, “Predictive maintenance of industrial machines using iot and machine learning,” *IEEE Access*, pp. 69 029–69 037, 2018.