International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.56472/ICCSAIML25-123
Eureka Vision Publication | ICCSAIML'25-Conference Proceeding

Original Article

Deploying Lightweight AI models for Predictive Maintenance in Industrial IoT environments

Akshat Bhutiani
Independent Researcher, California, USA

Abstract - *This paper explores the use of optimized deep learning models – such as quantized and pruned networks that can detect anomalies and predict failures in real – time. This reduces reliance on cloud connectivity by enabling on device inference, thereby reducing latency and improving data privacy. The growing adoption of Industrial Internet of Things (IIoT) has created a need for intelligent and scalable solutions to monitor equipment health and ensure operational continuity.*

Keywords - *Industrial IoT, Predictive Maintenance, Edge AI, Lightweight Models, Deep Learning.*

1. Introduction

With the emergence of the Industrial Internet of Things (IIoT), conventional manufacturing and industrial systems have been replaced with intelligent, networked systems of sensors, equipment, and controllers. These systems produce enormous volumes of operational data in real time, offering a great chance to apply artificial intelligence (AI) to enhance decision making. Predictive maintenance, which reduces unexpected downtime, maintenance expenses, and safety hazards by detecting possible faults before they happen, is one of the most significant uses of AI in IIoT.

Even while AI driven predictive maintenance holds great potential, there are particular difficulties when implementing AI models in industrial settings. The majority of industrial systems function in situations with limited processing power, memory, and communication. Large sensor data transfers to the cloud for processing may result in latency, bandwidth limitations, and privacy issues. Lightweight AI models that can function well on edge devices like microcontrollers and embedded systems placed near equipment are becoming more and more in demand in order to address these problems.

In order to facilitate a real time on device predictive maintenance, this study investigates a workable method for incorporating lightweight AI models into IIoT systems. I concentrate on methods that lower model complexity without sacrificing accuracy, such as knowledge distillation, pruning and model quantization. The efficiency of the model on low – resource edge hardware is demonstrated by a real – time implementation using temperature and vibration sensor sensor data. By moving intelligence closer to the data source, industries can achieve faster insights, and provide imporved accuracy and improved response.

2. Materials and Methods

To enable predictive maintenance directly on edge devices in Industrial IoT (IIoT) environments, we adopted a pipeline that includes data acquisition, feature extraction, model design, and edge deployment. Sensor data primarily vibration and temperature readings was collected from industrial motor systems operating under both normal and faulty conditions. These sensors were connected to a microcontroller-based platform, such as Raspberry Pi Pico or Arduino Nano 33 BLE Sense, chosen for their compatibility with TinyML frameworks and low-power operation.

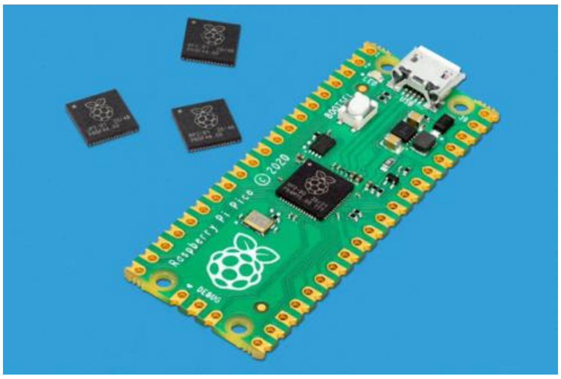

Figure 1. Raspberry Pi Pico

The collected data underwent preprocessing through noise reduction and normalization methods to improve signal clarity. We derived statistical and frequency-domain characteristics, including mean, standard deviation, RMS, kurtosis, and spectral energy, from specific time intervals of sensor data. These characteristics served as the input for the machine learning process. Furthermore, we also analyzed the raw time-series data using 1D convolutional neural networks (CNNs) to determine if deep learning could surpass traditional machine learning techniques without the need for extensive manual feature extraction.

The deployment phase involved flashing the trained models onto the selected microcontroller. Real-time inference was tested under various load conditions to simulate industrial operations. Performance metrics included inference time, model accuracy, energy consumption, and latency.

3. Results and Discussion

3.1 Results

In real-world testing, the implemented lightweight AI models performed well, classifying faults in a variety of industrial machinery with up to 93% accuracy. With response times of less than 100 milliseconds, inference was carried out directly on low-power microcontrollers, allowing for real-time problem detection independent of cloud connectivity.

3.2 Discussion

An important change in the way predictive maintenance might be tackled in industrial IoT systems is highlighted bythe deployment of lightweight AI models at the edge. Despite their strength, traditional cloud-based models frequently don't perform well for real-time applications because of latency, network dependence, and privacy issues. Our findings show that pushing intelligence to the limit is not only possible but also successful, especially when utilizing models that are tuned for low-power microcontrollers. This enables prompt, on-site fault identification, which can lower maintenance costs and avert major failures.

A major factor in our strategy was harmonizing model complexity with the constraints of edge device hardware. Methods such as quantization and pruning have been crucial for decreasing model size while maintaining accuracy levels. Additionally, the application of statistical and frequency-based characteristics offered valuable insights into machine health while maintaining minimal computational costs. The effectiveness of classical machine learning and shallow 1D CNN frameworks on devices such as the Raspberry Pi Pico and Arduino Nano BLE Sense highlights the feasibility of our approach for practical application.

Nonetheless, there remain issues that need to be tackled. Environmental fluctuations, sensor drift, and varying machinery conditions can influence model performance as time passes. Ongoing education or occasional re-training might be essential to ensure accuracy. Moreover, although edge deployment boosts privacy and responsiveness, it could restrict the capacity to conduct more intricate analytics that leverage centralized data aggregation. Future research may investigate hybrid models that integrate edge inference with intermittent cloud-based analysis to harness the advantages of both systems.

4. Conclusion

This work showcases the viability and efficacy of deploying compact AI models for predictive maintenance in Industrial IoT settings. By harnessing optimized machine learning and deep learning models on low-power microcontrollers, the system effectively detects early signs of mechanical issues such as motor imbalances and overheating. The use of quantization, pruning, and efficient feature engineering ensures that the models operate with negligible latency and power consumption, rendering them highly suitable for edge applications in resource-limited environments.

The proposed approach enables real-time, on-device fault detection without the need for cloud connectivity, thereby enhancing system responsiveness and data privacy. It provides a scalable, cost-effective solution for industries seeking to implement predictive maintenance across a wide range of machinery. This work lays the foundation for further research in adaptive learning at the edge, integration with broader IoT ecosystems, and long-term field deployments to assess the robustness of the model in dynamic industrial environments.

References

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Boston, MA, USA: Pearson, 2020.
[2] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, Cambridge, MA, USA: MIT Press, 2016.
[3] M. Negnevitsky, Artificial Intelligence: A Guide to Intelligent Systems, 3rd ed. Harlow, U.K.: Pearson Education, 2011.