



Developing a Lightweight Edge-AIoT Model for Real-Time Prediction of Mechanical Failures to Improve Maintenance

Prepared By

Name: Abdulaziz Saleh Mohammed Alhasil ID No.: 261002308

Name: Majed Humud Alhumuda ID No.:253002679

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

Collage of Informatics

Midocean University

Supervisor Name: Dr Sherif Mostafa

Collage of Informatics, Midocean University, Comoros

Email:.....

Acknowledgements

Special thanks to Dr. .

Abstract

List of Publications

Table of Contents

List of Abbreviations	6
List of Figures.....	7
List of Tables	8

Chapter 1: Introduction

1.1 Overview	1
1.2 Problem Definition	2
1.3 Thesis Objectives	3
1.4 Thesis Contribution	4
1.5 Thesis Organization.....	5

Chapter 2: Background and Related Work

.....	31
-------	----

Chapter 3: Proposed Approach

78

4.1 Introduction	78
------------------------	----

Chapter 4: Experimental Results and Discussion

86

Chapter 5: Conclusion and Future Work

127

7.1 Important Features	127
------------------------------	-----

7.3 Future Work	129
-----------------------	-----

References	130
------------------	-----

List of Abbreviations

- **Acc.:** Accuracy
- **ANN:** Artificial Neural Network
- **ANOVA:** Analysis of Variance
- **ROC:** Receiver Operating Characteristic

List of Figures

2.1 Educational Data Mining Process	7
2.2 Applications of Educational Data Mining	8
2.3 Distribution of Papers that used different attributes for PSP problem	19
2.4 Student Performance Data Set	21
2.5 Educational Process Mining (EPM): A Learning Analytics Data Set ...	23

List of Tables

2.1	Student Performance Data Set Description	21
2.2	Students' Academic Performance Dataset Description	23

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Overview

There is no doubt that today's industrial outputs are an essential part of human life and the development of nations across all aspects of life (such as food, health, sports, and education). Without industry, many resources would not have been available. Since the First Industrial Revolution, and passing through the Second, industry has become a top priority for countries, completely reshaping the human lifestyle.

Driven by the significant impact of these factories—where any downtime directly affects the availability of resources and leads to substantial profit losses—factories have been placing a paramount importance on maintenance since the Fourth Industrial Revolution (4IR). Maintenance has shifted from traditional periodic maintenance or reactive maintenance (after a breakdown occurs) to predictive maintenance. These factories have begun integrating Artificial Intelligence (AI) with the Internet of Things (IoT) to provide a powerful tool that collects and analyzes data in real-time, predicting the lifespan of machinery before they stall or fail.

Therefore, the project will be: "**Developing an AIoT model to predict mechanical failures and improve maintenance.**" This project aims to help factories ensure business continuity without disruptions or incurring heavy losses, while incorporating the latest research and experiments in this field to deliver a project that enhances factory efficiency and seamless operations without downtime

1.2 Problem Definition

First: The Mechanical and Economic Problem

Sudden breakdowns of mechanical machinery cause massive financial losses and disrupt production lines. This, in turn, directly impacts both manufacturers and end-consumers regarding the availability of goods and services.

Second: The Current Software Problem

Current solutions rely either on connecting Artificial Intelligence (AI) models to the cloud—which introduces data transmission delays (latency)—or on using compressed models that compromise accuracy, as well as dealing with noisy or distorted data.

1.3 Thesis Objectives

Objective 1: Design a lightweight AIoT model optimized for edge computing devices to ensure real-time mechanical failure prediction.

Objective 2: Develop a fault-tolerant data fusion algorithm capable of handling noisy or missing sensor data (vibration, temperature).

Objective 3: Implement an event-driven mechanism to balance between IoT device energy efficiency and high-fidelity transient signal capture.

Objective 4: Evaluate the proposed framework using real-world industrial datasets to benchmark accuracy and processing latency

1.4 Thesis Contribution

- **Contribution 1:** Introducing an **Adaptive Edge-AI framework** that self-updates locally on the IoT gateway without relying on cloud retraining.
-
- **Contribution 2:** A novel **Fault-Tolerant multi-sensor fusion architecture** that maintains predictive accuracy even during individual sensor failure.
-
- **Contribution 3:** An optimized **Asynchronous XAI mechanism** (like lightweight TreeSHAP) that explains failure root-causes to mechanical engineers without violating real-time execution constraints.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

- **Chapter 2 (Background and Related Work)**
provides the theoretical background of AIoT and mechanical fault analysis, alongside a critical review of related work.
- **Chapter 3 Proposed Approach (Methodology)**
details the proposed adaptive AIoT model architecture, sensor fusion, and edge deployment strategies.
- **Chapter 4 (Experimental Results and Discussion)**
presents the software implementation setup, dataset description, and performance evaluation metrics.
- **Chapter 5 (Conclusion & Future Work)**
summarizes the core findings of this thesis and outlines potential directions for future research".

Chapter 2

Background and Related Work

Chapter 2

Background and Related Work

2.1 Research Background

2.1.1 Industrial Maintenance Paradigms

In modern industrial engineering and infrastructure management, maintaining the structural and operational integrity of mechanical assets is paramount. Historically, facilities relied heavily on Reactive Maintenance (Run-to-Failure), which entails repairing equipment only after a breakdown occurs. While structurally simple, this paradigm incurs severe financial deficits due to unexpected downtime, safety hazards, and catastrophic secondary damage to adjacent machinery.

To mitigate these risks, industries transitioned to Preventive Maintenance, which implements schedule-based servicing. Although more structured, preventive schedules often result in "over-maintenance," causing premature replacement of functional components and unnecessary operational overhead. This economic and logistical imbalance has accelerated the global adoption of Predictive Maintenance (PdM)—a proactive approach that evaluates the actual real-time condition of an asset to forecast anomalies before failure occurs.

2.1.2 IoT Evolution and the Bandwidth Bottleneck

The realization of Predictive Maintenance was significantly advanced by the Internet of Things (IoT). Deploying specialized wireless sensory networks allowed engineers to stream physical machinery telemetry—such as fluid pressure oscillations, mechanical vibration amplitudes, and heat gradients—to centralized storage frameworks.

However, classic cloud-centric IoT architectures suffer from critical constraints when scaled across extensive facilities. Streaming

continuous, high-frequency time-series data to remote cloud servers demands immense network bandwidth, exposes data to privacy vulnerabilities, and introduces transmission latencies. In high-pressure hydraulic setups or high-speed rotational systems, a delay of milliseconds in anomaly detection can mean the difference between a minor leak and a catastrophic rupture.

2.1.3 Edge Intelligence (AIoT) and Hardware Optimization

To resolve cloud over-dependence, the modern engineering paradigm introduces AIoT (Artificial Intelligence of Things), which embeds cognitive machine learning capabilities directly into localized hardware layers, establishing Edge Intelligence. By migrating model inference to the edge, sensory nodes can interpret raw physical signals right at the point of measurement.

Recent breakthroughs in deep learning optimization—such as Model Quantization (converting weights from Float32 to INT8 precision) and embedded runtimes like TensorFlow Lite Micro—have made it possible to deploy complex mathematical models onto low-cost, resource-constrained microcontrollers (e.g., the ESP32 platform). This ensures that heavy computational diagnostics no longer require expensive industrial computers or active internet dependencies.

2.1.4 Multi-Class Fault Analytics and Localized Anomaly Detection

The ultimate efficacy of an AIoT predictive framework relies on its algorithmic capacity to not only flag general anomalies but to accurately map and classify specific degradation behaviors concurrently (Multi-Class Fault Classification). In mechanical and hydraulic domains, raw sensory signatures (such as fluid pressure telemetry or vibration frequencies) must be classified at the edge node to distinguish between a "benign transient pressure drop," a "structural line leak," and a "critical valve failure."

By executing lightweight, highly optimized on-device deep learning

algorithms, a standalone microcontroller can analyze streaming time-series data locally and isolate specific failure roots in real-time. This localized processing enables the system to issue instant diagnostic responses or initiate safety overrides—such as shutting off local valves or pumps—within milliseconds. Crucially, this advanced diagnostic protection remains fully operational even during complete communication network dropouts, ensuring robust, low-power, and continuous hardware asset reliability.

2.2 Related Work And Literatures Review

Citation / Author & Year	Objective & Problem Statement	AI / Algorithmic Methodology	IoT Environment & Edge Deployment	Key Results & Performance Metrics	Identified Research Gap
Montes-Sánchez et al. (2025)	• Early detection of tube degradation and mechanical wear in peristaltic pumps. • Preventing sudden runtime structural failures in highly time-sensitive medical and industrial fluid operations.	• Evaluated 939 hybrid architectural variations. • Utilized Recurrent Neural Networks, specifically Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). • Optimized hidden unit tuning (scaled 2 to 64 nodes) to discover the lightest execution architecture.	• Raw multi-modal signals captured via 6 physical sensors (3 accelerometers, 1 gyroscope, 1 magnetometer, 1 microphone). • Local target model optimization downsampled to 12 Hz for low-footprint execution across 4 different embedded microcontrollers (MCUs).	• Excellent predictive capability achieving F1-scores between 0.99 and 1.00 for top candidate networks. • Demonstrated solid real-time inference viability regarding edge SRAM consumption and ultra-low latency.	• Constrained to rigid, static operational environments (unchanging fluid thickness and motor speed vectors). • Your Opportunity: Developing a resilient Edge AIoT system capable of leak forecasting under complex multi-tier dynamic pressure variations and unpredictable flow environments.
Reis (2025)	• Solving the critical overhead complexity and surging fuel/labor costs inherent to last-mile commercial delivery logistics.	• Structured Systematic Literature Review executed via PRISMA 2020 protocol. • Empirical expert validation	• Architectural design of an Edge Intelligence Ecosystem integrating edge sensory layers with low-latency	• Measurable enhancements in dynamic vehicle route optimization, delivery latency compression, and user trust. • Highlighted regional European SME hesitancy due to steep	• Heavily exploratory and methodological framework lacking concrete edge خوارزمية development or physical silicon bench tests. • Your Opportunity: Bridging the gap by embedding these high-level managerial strategies directly into a compact,

	Formalizing the structural deployment pillars of Edge Intelligence (EI) for decentralized edge stream handling.	using a qualitative 2-round Delphi investigative approach to finalize strategic adoption priorities.	network processing. Map decentralized computing clusters near end-users, outlining edge-cloud transactional offloading topologies.	initialization costs and deep specialized skill shortages.	budget-conscious Edge AI embedded node layout.
Razaq et al. (2026)	- Tackling degraded energy efficiency and high operational maintenance expenses (OpEx) caused by grid inverter failures. - Bridging disconnected edge data silos that relied solely on distant centralized cloud backends.	- Formulated an efficient deep hybrid Convolutional LSTM (CNN-LSTM) topology for immediate multi-class electrical fault classification. - Implemented a Health Index Prognostic Model to approximate Remaining Useful Life (RUL) limits.	- Native local filtering of volatile raw transient thermal and electrical signals right at the generation point. - Employed model compression (Pruning and Quantization) using TensorFlow Lite to port models onto ARM-based nodes (Jetson Nano, Raspberry Pi) via MQTT.	- Attained diagnostic classification accuracy up to 96.8%. - Compressed processing delays sharply from 85 ms (cloud) down to 18 ms (edge device), eliminating communication backhauls.	- Relied partially on synthetic asset generation due to sparse real-world life-cycle fault records; validated mostly in enclosed pilot setups. - Your Opportunity: Extending embedded analytics to rugged multi-point deployment settings or configuring low-power distributed frameworks (like localized Master/Slave nodes).
Kumar et al. (2026)	Eliminating structural vulnerabilities linked with cloud-reliant predictive maintenance, such as deep packet latency and network downtime risk.	Configured an optimized INT8-quantized edge classifier for parallel multi-fault prediction alongside Time-to-	Implemented a resilient hybrid Edge-Cloud orchestration layout featuring a smart confidence-based fallback mechanism.	- Locked a predictable, local real-time inference latency averaging just 42.3 ms. - Retained 97.3% of premium cloud model precision boundaries when running at a 0.70 confidence threshold. - Kept fault recall	- Confined exclusively to standalone execution on a isolated single MCU unit (ESP32-S3) inside a static environment. - Your Opportunity: Utilizing this ultra-lightweight architecture to build a coordinated Master/Slave smart multi-point distributed sensory grid to monitor wide-area building networks.

	• Preserving continuity in critical time-bounded smart machinery setups.	Failure (TTF) tracking. • Formulated a conservative asymmetric loss function to avoid risky asset life overestimations. • Infused Explainable AI (XAI) using SHAP via Natural Language descriptions.	• Deployed natively via TensorFlow Lite Micro directly onto a low-cost, resource-restricted commercial ESP32-S3 microcontroller without custom hardware accelerators.	above 0.924 during extreme 50% simulated network dropouts.	
Garg et al. (2026)	• Resolving network processing lag, massive bandwidth costs, scalability bottlenecks, and data privacy vulnerabilities inherent to cloud-based streaming IoT telemetry. • Enabling low-latency local decision-making in Cyber-Physical Smart Environments.	• Engineered a collection of compact, high-efficiency CNN and RNN lightweight deep learning models tailored for native edge compilation. • Developed adaptive stream handling pipelines for local Anomaly Detection and localized real-time risk scoring.	• Devised a distributed 4-tier structural architecture (Sensors, Intelligent Edge, Cloud Coordination, and Applications). • Integrated MQTT for lightweight communications. • Compiled models via TensorFlow Lite for concurrent execution across Raspberry Pi and NVIDIA Jetson Nano boards.	• Captured an outstanding local anomaly identification precision of 97.1%. • Slashed computing latency from 180 ms (cloud telemetry) to a mere 45 ms local execution response. • Reduced operational power draw to 7.2W while maintaining a high stream throughput of 1620 transactions per second.	• Targeted generic multi-sensor data patterns on relatively heavy edge compute nodes (Jetson Nano/Raspberry Pi) lacking extreme resource micro-optimization. • Your Opportunity: Customizing this structural 4-tier approach down to bare-metal microcontrollers (such as your ESP32-driven architecture in the 'Mansob' leak detection system) for specialized hydraulic pressure anomaly filtering right at the sensor head.

Kang, J. S., et al. (2021)	To reduce reliance on cloud computing and mitigate data transmission latency that delays real-time failure prediction.	Lightweight and compressed Convolutional Neural Networks (Quantized CNNs).	Industrial IoT setup with wireless vibration sensors deployed on an Edge Gateway (e.g., Raspberry Pi) for local inference.	Reduced response latency by 40% , minimized network bandwidth, and achieved an initial fault classification accuracy of 89% .	Software/Algorithmic Gap: Model quantization led to a loss of accuracy in predicting complex or overlapping mechanical faults, and the model is static with no online learning capabilities to adapt to changing environments.
Zhao, M., et al. (2020)	To address the difficulty of predicting the Remaining Useful Life (RUL) of machinery when relying on a single sensor type.	Multi-sensor Data Fusion algorithm coupled with a Long Short-Term Memory (LSTM) recurrent neural network to track degradation over time.	A heterogeneous IoT sensor network (vibration, temperature, and electrical current sensors) connected to a centralized control system.	High accuracy in RUL estimation with a 15% reduction in Mean Squared Error (MSE) compared to single-sensor models.	Software/Algorithmic Gap: The system programmatically assumes 100% sensor reliability; it experiences total breakdown or yields erratic outputs when receiving noisy or missing data (No Fault-Tolerance built into the AI).
Berghout, T., et al. (2021)	To tackle severe Data Imbalance in smart factories, where actual failure data accounts for less than 1% of the collected dataset.	Generative Adversarial Networks (GANs) to generate synthetic fault data to balance the training dataset.	An IoT framework capturing operational data from gears and motors, streaming it to a Local Server for data balancing and model training.	Improved the prediction rate of rare mechanical failures (Recall/Sensitivity) from 60% to 92% .	Software/Algorithmic Gap: The synthetic data generated by GANs consists of isolated, snapshot instances; it programmatically fails to replicate the continuous, time-dependent degradation trend of physical machinery.
Zhang, L., et al. (2022)	To solve high battery consumption in wireless IoT sensors caused by the continuous sampling and broadcasting	An intelligent scheduling algorithm utilizing Adaptive Sampling combined with	Battery-powered smart sensors utilizing low-power wide-area networks (LoRaWA)	Decreased IoT device energy consumption by 35% , significantly extending the operational battery life of the sensors.	Software/Algorithmic Gap: To save energy, the algorithm reduced the data sampling rate, causing the ML model to miss transient signals and sudden microscopic shocks that precede major mechanical failures.

	of high-frequency data.	lightweight embedded ML models.	N) attached to mechanical assets.		
Silva, R., et al. (2023)	To overcome engineers' lack of trust in deep learning models due to their " Black Box " nature, which conceals the reasoning behind failure alerts.	Integrating Explainable AI (XAI) frameworks, specifically SHAP and LIME , on top of the deep learning layers.	An integrated AIoT platform mapping machine sensors to an interactive Dashboard that displays failure predictions alongside their root causes.	Enhanced engineering trust (User Trust) and streamlined maintenance workflows by pinpointing the exact anomalous sensor attributes.	Software/Algorithmic Gap: Calculating SHAP values is highly computationally intensive, introducing severe algorithmic latency that violates real-time constraints required for emergency shut-offs.

Chapter 3 Proposed Approach (Methodology)

Chapter 4 :Experimental Results and Discussion

Chapter 5 Conclusion and Future Work

References

[Abiodun, Oludare Isaac, et al., 2018]

Abiodun, Oludare Isaac, Aman Jantan, Abiodun Esther Omolara, Kemi Victoria Dada, Nachaat AbdElatif Mohamed, and Humaira Arshad. “State-of-the-art in artificial neural network applications: A survey”. *Heliyon* 4. 2018.

[Abu Saa, Amjed, et al., 2019]

Abu Saa, Amjed, Mostafa Al-Emran, and Khaled Shaalan. “Factors Affecting Students’ Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques”. *Tech Know Learn* 24, 567–598., 2019.

[Adewale, Amoo M., et al., 2018]

Adewale, Amoo M., Alaba O. Bamidele, and Usman O. Lateef. “Predictive modelling and analysis of academic performance of secondary school students: Artificial Neural Network approach”. *International Journal of Science and Technology Educational Research* 9(1), 1-8, 2018.

[Ali S, Rasheed Mansoor, and S Perumal, 2022]

Ali S, Rasheed Mansoor, and S. Perumal. “Multi-class LDA classifier and CNN feature extraction for student performance analysis during Covid-19 pandemic”. *International Journal of Nonlinear Analysis and Applications*, 13(1), pp. 13291339., 2022.

[Alpaydin, Ethem, 2010]

Alpaydin, Ethem (2010). *Introduction to Machine Learning*. London: The MIT Press. ISBN 978-0-262-01243-0