

MASTER OF SCIENCE IN ARTIFICIAL INTELLIGENCE

Developing a Lightweight Edge-AIoT Model

For Real-Time Prediction of Mechanical Failures to Improve Maintenance

Prepared By: Abdulaziz Saleh Alhasil & Majed Humud Alhumuda

Supervisor: Dr. Sherif Mostafa

Presentation Agenda

1. Introduction & Background
2. Problem Definition & Objectives
3. IoT & Edge Architecture
4. Proposed Methodology & Hardware
5. Expected Results & Conclusion

Introduction: The Industrial Shift

The Fourth Industrial Revolution

Industry is a top priority globally, completely reshaping human lifestyle. The availability of vital resources depends heavily on the uninterrupted outputs of modern factories. Any downtime leads to substantial profit losses and resource scarcity.

Shift to Predictive Maintenance

Since the advent of 4IR, maintenance strategies have shifted drastically. Factories are moving away from traditional reactive approaches (fixing after breakdown) to **Predictive Maintenance (PdM)** by integrating AI and IoT.

Problem Definition



Economic Impact

Sudden breakdowns of mechanical machinery cause massive financial losses and disrupt entire production lines, directly impacting manufacturers and consumers alike.



Cloud Dependency

Current solutions heavily rely on connecting AI models to remote cloud servers. This introduces critical data transmission delays (latency) and privacy vulnerabilities.



Software Trade-offs

Conversely, using compressed edge models often compromises diagnostic accuracy, struggling to deal with noisy or distorted mechanical sensor data effectively.

Thesis Objectives

- 🎯 **Objective 1: Lightweight Design.** Design a lightweight AIoT model optimized for edge computing devices to ensure real-time mechanical failure prediction.
- 🎯 **Objective 2: Fault-Tolerant Fusion.** Develop a data fusion algorithm capable of handling noisy or missing multi-modal sensor data (e.g., vibration, temperature).
- 🎯 **Objective 3: Event-Driven Mechanism.** Implement a mechanism to balance between IoT device energy efficiency and high-fidelity transient signal capture.
- 🎯 **Objective 4: Real-World Evaluation.** Evaluate the proposed framework using real-world industrial datasets to benchmark accuracy and processing latency.

Key Contributions

Adaptive Edge-AI Framework

Introducing a novel framework that self-updates locally on the IoT gateway, eliminating the need to rely on continuous cloud retraining and reducing bandwidth consumption.

Fault-Tolerant Architecture & XAI

A multi-sensor fusion architecture that maintains predictive accuracy during individual sensor failure, coupled with an optimized **Asynchronous XAI mechanism** to explain root-causes without violating real-time constraints.

Evolution of Industrial Maintenance

1. Reactive Maintenance (Run-to-Failure):

Repairing equipment only after breakdowns. Incurs severe deficits due to unexpected downtime and collateral damage.

2. Preventive Maintenance:

Schedule-based servicing. Often results in "over-maintenance" and premature replacement of perfectly functional components.

3. Predictive Maintenance (PdM):

A proactive approach evaluating actual real-time condition to forecast anomalies. Highly dependent on advanced sensing and AI.



The Bandwidth Bottleneck

Cloud-Centric IoT Limitations

Streaming continuous, high-frequency time-series data to remote cloud servers demands immense network bandwidth and exposes operational data to privacy vulnerabilities.

More critically, in high-speed rotational systems, transmission **latencies** of hundreds of milliseconds can mean the difference between a minor issue and a catastrophic rupture.

The Edge Solution

Migrating cognitive machine learning directly to localized hardware layers ensures raw physical signals are interpreted right at the point of measurement.

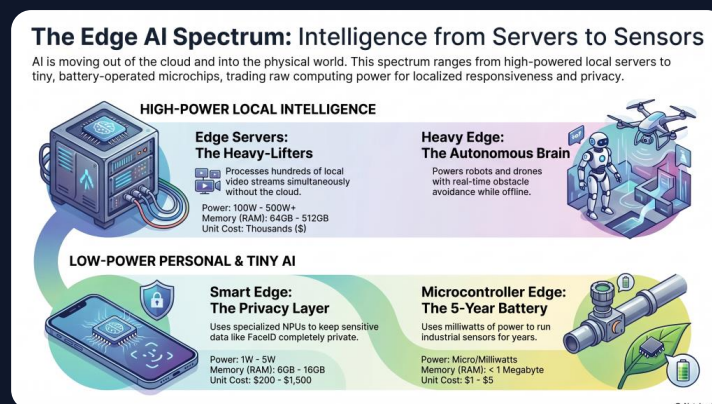


Four-Tier AIoT Reference Architecture



1. Perception Tier

Acquires raw physical signals directly from machinery, performing high-frequency sampling and initial anti-alias filtering.



2. Edge Intelligence

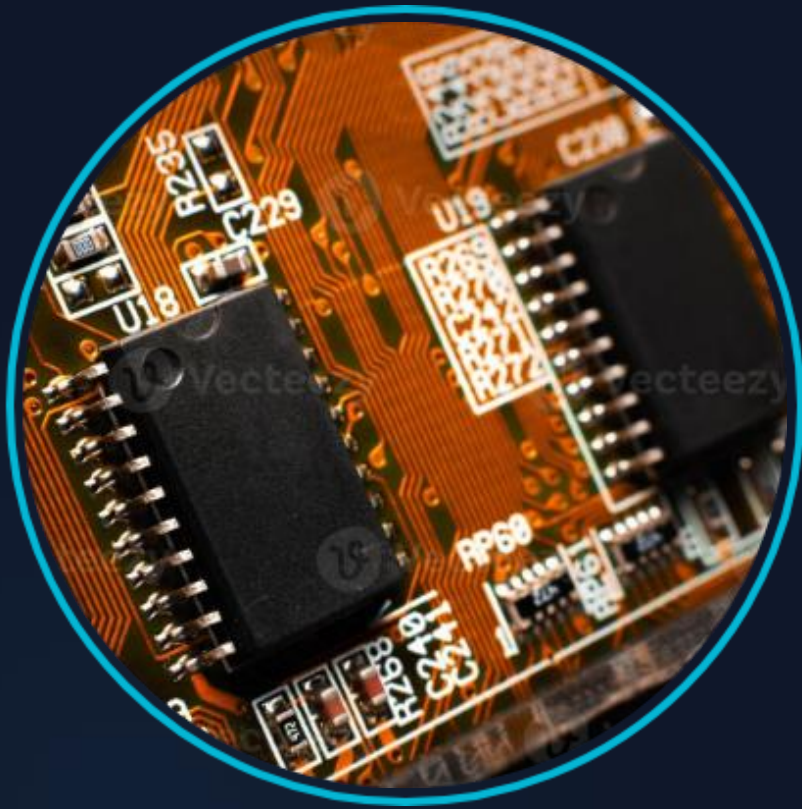
Executes feature extraction, quantized inference, and lightweight explainability directly on the microcontroller.



3. Cloud & App Tier

Provides fleet-level dashboards, historical archiving, and periodic model retraining using highly aggregated events.

Hardware Optimization (TinyML)



Empowering Resource-Constrained Devices

Recent breakthroughs in deep learning optimization have made it possible to deploy complex mathematical models onto low-cost microcontrollers (e.g., the ESP32 platform).

Model Quantization: Converting neural network weights from Float32 to INT8 precision significantly reduces flash footprint and RAM usage, while maintaining ~97% of original accuracy. Using embedded runtimes like **TensorFlow Lite Micro**, heavy computational diagnostics no longer require expensive industrial computers or active internet dependencies.

Multi-Class Fault Analytics



1. Signal Acquisition

A sliding-window isolates transient events in continuous telemetry streams (e.g., vibration frequencies, fluid pressure), transforming them into compact spectral descriptors.



2. Quantized Inference

The descriptor is submitted to the quantized INT8 model on the edge device, outputting confidence values across multiple distinct fault classes rather than a simple binary flag.



3. Localized Override

Diagnosis is graduated across severities. A critical valve failure triggers an immediate autonomous safety override (shut-off) in less than a millisecond, completely independent of network connectivity.

Literature Gap Analysis

Previous Work Focus	Identified Limitation / Gap	Our Proposed Solution
Cloud-reliant AI Algorithms	High data-transmission latency; vulnerable to network outages.	Fully Localized Edge AI acting in milliseconds.
Multi-sensor Data Fusion (LSTM)	Assumes 100% sensor reliability; fails if a sensor degrades.	Fault-Tolerant Fusion maintaining accuracy despite sensor failure.
Trust & Explainable AI (SHAP/LIME)	Requires high computational power; causes prolonged execution times.	Asynchronous Lightweight XAI to prevent execution delays.
Model Compression / Pruning	Models validated only in static environments or on heavy compute nodes.	Adaptive Framework on bare-metal microcontrollers (ESP32).

Proposed Methodology & Architecture

Bridging the gap between raw hardware constraints and intelligent industrial safety.

Core Architecture Features

1. Fault-Tolerant Sensor Fusion

Industrial environments are harsh; sensors degrade. The architecture utilizes a dynamic weighting mechanism. If the vibration sensor generates anomalous noise, the model automatically shifts reliance to the temperature or acoustic sensor, preserving predictive accuracy without triggering false alarms.

2. Event-Driven Optimization

To preserve energy in battery-operated IoT nodes, the system idles during normal operations. It only triggers high-frequency, energy-intensive sampling algorithms when statistical anomalies cross a pre-defined threshold, capturing transient signals perfectly.

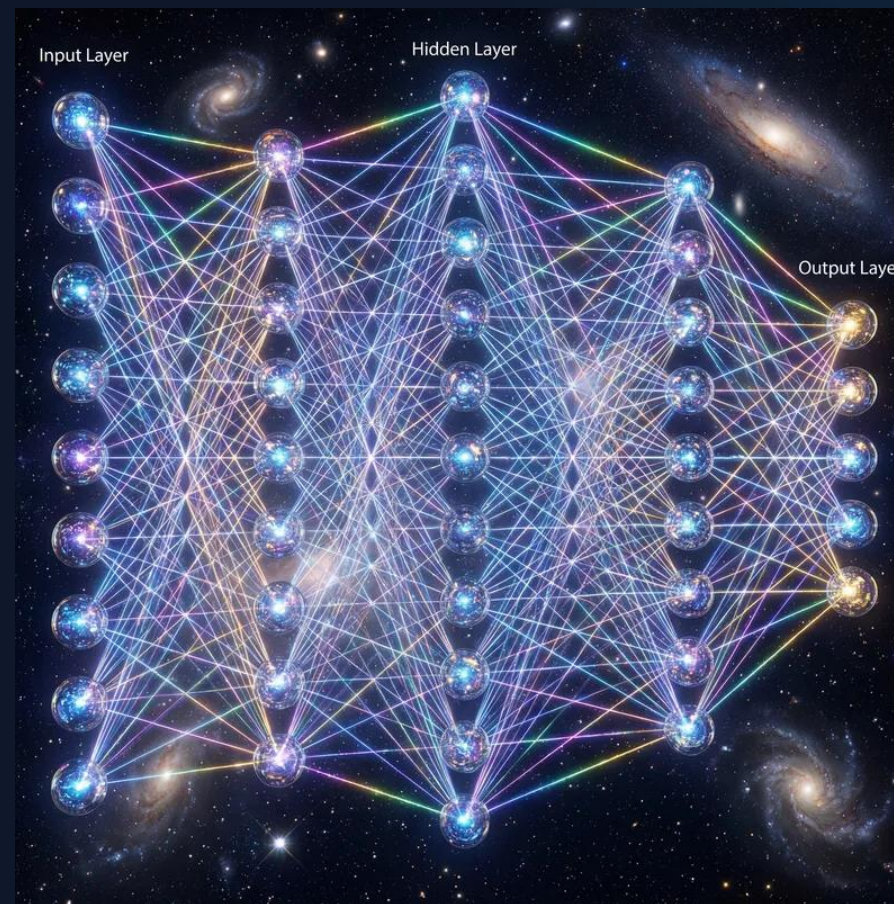
Explainable AI (XAI) at the Edge

Building Engineer Trust

Black-box AI models hinder industrial adoption. Engineers need to know *why* a machine was shut down.

While traditional Explainable AI methods (like standard SHAP or LIME) demand excessive compute power and time, our framework introduces an **Optimized Asynchronous XAI mechanism**.

It generates a lightweight ranking of the features that drove the decision (e.g., "High Frequency Vibration peak at 2kHz"), supplying maintenance engineers with clear root-cause evidence without stalling the real-time emergency override.



Conclusion & Future Work

A new paradigm for localized industrial
reliability.

Summary & Path Forward



Achieved Autonomy

Developed a framework where edge nodes diagnose machinery and trigger safety overrides in milliseconds, completely independent of cloud connectivity.



Ensured Reliability

Implemented fault-tolerant fusion and lightweight explainability, ensuring that hardware constraints do not compromise diagnostic accuracy or user trust.



Future Research

Future work will focus on expanding the multi-class taxonomy to broader asset types and implementing federated learning across multiple factory deployments.

Questions?

Thank you for your time and attention.

Abdulaziz Saleh Alhasil & Majed Humud Alhumuda
Master of Science in Artificial Intelligence
Midocean University

Image Sources



<https://8962390.fs1.hubspotusercontent-na1.net/hubfs/8962390/digital%20twins%20article%20header%20-%20optimized%20%20%281%29.jpg>

Source: www.cyngn.com



https://png.pngtree.com/thumb_back/fh260/background/20260416/pngtree-abstract-digital-network-glowing-with-many-connected-points-and-lines-in-image_21680575.webp

Source: pngtree.com



https://www.azorobotics.com/images/Article_Images/ImageForArticle_355_15825580285584654.png

Source: www.azorobotics.com



https://blog.nashtechglobal.com/wp-content/uploads/2026/07/The_Edge_AI_Spectrum-1024x572.png

Source: blog.nashtechglobal.com



<https://www.justinmind.com/wp-content/uploads/2018/12/6-best-practices-for-Dashboard-Design-Justinmind-header.png>

Source: www.justinmind.com



https://static.vecteezy.com/system/resources/previews/067/973/741/non_2x/printed-circuit-board-close-up-small-microprocessor-microcontroller-on-the-motherboard-of-a-modern-desktop-pc-macro-of-an-object-close-up-shallow-depth-of-field-nobody-photo.jpg

Source: www.vecteezy.com

Image Sources



https://media.easy-peasy.ai/7cd52b37-a38b-4e6c-9ae7-91d3588a7d76/f05cff9d-bc1d-47d7-9908-ea4d5cc10a8b_medium.webp

Source: easy-peasy.ai