

A Generalizable Machine Learning Framework for Performance Prediction and Fault Detection Across Power Plant Units

Salem Ahmed Buhluli

Student ID: 262000880

Course: Research Methodology in Computer Science (AI800)

Lecturer: Dr. Mohammed Anbar

Table of Contents

Abstract	3
Chapter 1: Introduction.....	4
1.1 Introduction	4
1.2 Background	4
1.3 Problem Statement.....	5
1.4 Research Objectives	7
1.5 Research Scope	7
Chapter 2: Literature Review	9
2.1 Machine Learning in Power Plants.....	9
2.2 Studies That Address the Weak Points	10
2.3 Gap Analysis	12
Chapter 3: Research Methodology	14
3.1 Phase 1: Literature Review14	
3.2 Phase 2: Data	14
3.3 Phase 3: Building the Model	15
3.4 Phase 4: How Results Are Measured.....	15
3.5 Phase 5: Experiments and Comparison.....	15
3.6 Ethical Considerations	16
3.7 Limitations	16
3.8 Expected Outcomes	16
3.9 Gantt Chart	17
References	18

Abstract

Every day power plants collect huge amounts of sensor data. It is now being applied to turn that data into two useful things: predicting how the plant will perform and catching equipment faults early. The published results are very good but almost all are from one plant, one machine or a few days of data. If you take the same model and apply it to a different machine, or a different season, the accuracy will usually drop. Three other problems:

Real fault examples are rare, most papers are based on private data that nobody else can check, and the way results are measured is not always honest.

This proposal outlines a study that takes this on as the main problem rather than a side issue. The work involves five phases: revision of the existing studies and listing their gaps; preparation of public datasets containing more than one machine; building a model involving sensor selection, a deep time-series network, few-shot learning and a few simple physical rules; fixing the measuring rules before running anything; and finally testing the model on a machine that has never been seen during training.

The expected result is a model which preserves most of its accuracy on a new machine, an unambiguous answer to how many labeled faults are actually needed and a simple set of reporting rules which other students and researchers can re-use. At the end there is a Gantt chart of the whole plan and it takes twelve months.

Keywords: Machine Learning, Power Plants, Fault Detection, Predictive Maintenance, Generalization, Few Shot Learning.

Chapter 1: Introduction

1.1 Introduction

People think of electricity only when it fails. Every socket is supported by a power station with hundreds of sensors recording temperature, pressure, flow, vibration and load every few seconds. For years this data was only used for alarms and reports that were written after something already had broken. That's changed. Storage is cheap, computers are fast and machine learning can now use the same data to predict what the plant will do next and to notice a fault before it is too late to fix it.

Most of the attention goes to two jobs. The first is prediction: predicting the output or the emissions of a unit from the conditions under which it is running. Tufekci (2014) predicted the output of a combined cycle plant using four ambient variables only. Tan et al. (2019) applied an LSTM network to monitor NOx emissions from a 660 MW boiler. The second job is fault detection: spotting the early signs of a boiler tube leak, a turbine problem or a bad bearing. Khalid et al. (2021) carried out sensor selection on real plant data, and Fahmi et al. (2024) used a deep autoencoder for vibration data from an operating gas turbine.

These results are good and are based on real industrial data, thus more credible than simulation studies. But there is one problem that is common in all of them. Each model is trained and tested on a single plant, single machine or short time period. Nobody has really shown that if you train a model on machine A, it will still work on machine B. That's the first question an engineer has to trust the model with a maintenance decision, and the field doesn't have a good answer yet.

This research targets this issue specifically. The focus is shifted from dataset-specific optimization toward designing a framework explicitly intended for generalization across different machines, supported by realistic and dependable evaluations.

1.2 Background

1.2.1 From Fixed Schedules to Data-Driven Maintenance

Calendar-based maintenance often sacrifices good components or misses early breakdowns due to fixed schedules. Predictive maintenance resolves this by using live machine data to guide decisions. As Carvalho et al. (2019) observed, machine learning algorithms like Random Forest, SVMs, neural networks, and k-means form the backbone of this approach, with multi-algorithm frameworks generally delivering superior performance.

1.2.2 Machine Learning in Plant Operations

In plant operations, machine learning effectively replaces slow, complex thermodynamic models with fast data-driven approximations that deliver real-time insights. For individual equipment, monitoring typically relies on vibration or process data. Supervised algorithms perform well when labeled fault data is available, but unsupervised approaches are often preferred since normal baseline data is abundant while actual failure data is rare. Highlighting this challenge in wind energy, Stetco et al. (2019) noted that classification methods dominate roughly two-thirds of published research, with the scarcity of labeled fault samples remaining a primary bottleneck across the discipline.

1.2.3 The Generalization Problem

Standard machine learning relies on the assumption that training and test data share the exact same distribution—a premise that quickly falls apart in industrial power plants. Differences in sensor placement, operational age, fuel quality, and seasonal conditions mean that no two machines behave identically, nor does a single machine remain constant from summer to winter. Zhao et al. (2024) identify this as domain shift, demonstrating across ten benchmark datasets that conventional models suffer severe performance drops when deployed on unseen machinery, and that current domain-adaptation techniques only address the issue partially..

1.2.4 Why the Measuring Method Matters

Evaluation metrics in this field also present a serious reliability issue. A common practice in anomaly detection research is applying 'point adjustment' before calculating F1 scores, where flagging a single data point counts as successfully detecting an entire fault event. As Kim et al. (2022) demonstrated, this technique can artificially inflate performance, making even random guessing look like a state-of-the-art method. Consequently, high scores achieved through point adjustment offer little assurance for actual control room operations, where false positives cause unnecessary costs and delayed alerts lead to critical failures.

1.3 Problem Statement

Machine learning has found broad application in power plant operations, ranging from output and emissions forecasting (Tufekci, 2014; Tan et al., 2019) to early fault detection (Khalid et al., 2021; Fahmi et al., 2024). However, a key limitation in current literature is that most models are validated using data from a single facility over limited timeframes. For instance, Tufekci (2014) evaluated a single combined-cycle plant running only at full capacity, Tan et al. (2019) relied on a seven-day window from one boiler, and Fahmi et al. (2024) analyzed just one month of data from a gas turbine. As a result, it remains unclear whether these models can maintain their performance on different equipment or across unobserved seasonal conditions.

This limitation goes beyond a simple lack of empirical testing. As Zhao et al. (2024) demonstrated across ten benchmark datasets, accuracy drops caused by domain shift are systematic, and existing generalization techniques only recover a fraction of the lost performance. Ultimately, current research still fails to offer a reliable answer to the very first question a plant engineer would ask before deployment.

This situation is further complicated by three critical challenges. First, labeled fault data remains scarce and severely imbalanced (Stetco et al., 2019), while few-shot learning techniques aimed at solving this have largely been benchmarked on lab-scale bearings rather than operational power plant machinery (Liang et al., 2023). Second, data accessibility poses a major hurdle; most studies rely on proprietary datasets that prevent external validation or direct comparison (Carvalho et al., 2019), largely because strict company policies restrict public data sharing (Ahn et al., 2023). Third, evaluation protocols themselves are inconsistent—varying metrics across papers and widespread reliance on point adjustment often mask model deficiencies by inflating random predictions into seemingly top-performing results (Kim et al., 2022). Compounding these issues, purely data-driven models frequently generate physically impossible predictions when deployed beyond their original training range—a scenario inevitable when transitioning to an unseen machine (Zhu et al., 2025).

Addressing these limitations is critical because the target failure modes—such as boiler tube leaks and major turbine breakdowns—represent primary causes of unscheduled plant outages, carrying steep financial penalties in lost generation and emergency repairs (Khalid et al., 2021). Yet, current methods remain constrained by narrow operational bounds: the Bagging model functions exclusively under full-load conditions (Tufekci, 2014), the sensor selection approach was validated on a single plant with only two fault classes (Khalid et al., 2021), the LSTM implementation provides forecasting without diagnostic capability (Tan et al., 2019), and the TCN-Autoencoder omits vital performance indicators like precision, recall, and false alarm rates entirely (Fahmi et al., 2024).

Consequently, there is a clear need for a framework that is validated across multiple machines and operating regimes, functions effectively with scarce labeled fault data, utilizes accessible datasets, and is evaluated using rigorous, transparent metrics. This research is designed specifically to address this gap.

1.4 Research Objectives

1. The overarching objective of this study is to develop and evaluate a robust machine learning framework for power plant performance prediction and fault detection—one that maintains high reliability across diverse machinery and varying operational environments.
2. **Research Objectives**
3. **Critique Existing Paradigms:** Systematically identify the key structural limitations of current machine learning models in power plant O&M, focusing on evaluation scale, data label dependence, and metric integrity.
4. **Construct Multi-Condition Benchmarks:** Curate, preprocess, and standardize public datasets spanning multiple machines and operational regimes, applying consistent time-windowing and class-imbalance techniques.
5. **Design a Hybrid Framework:** Develop a model integrating domain-guided feature selection, deep time-series architectures, unsupervised pre-training with few-shot adaptation, and physical constraints to ensure generalization and guard against non-physical predictions.
6. **Establish Unbiased Evaluation Standards:** Formulate a rigorous performance assessment protocol reporting MAE and RMSE for continuous predictions alongside precision, recall, and false alarm rates for fault detection, strictly avoiding point adjustment techniques.
7. **Cross-Machine Validation:** Benchmark the framework against established baselines (including bagged trees, LSTM, TCN-Autoencoders, and domain generalization models) on previously unseen machinery to assess true zero/few-shot transferability.
- 8.

1.5 Research Scope

- **Research Scope and Boundaries** To maintain a feasible and targeted research plan, the scope of this work is defined by the following boundaries:
- **Target Equipment:** Focuses exclusively on rotating and thermal assets within conventional power generation units—primarily boilers, turbines, and bearing systems. Nuclear and renewable energy installations (such as solar) fall outside this scope.
- **Functional Tasks:** Restricted to continuous performance forecasting (power output and NOx emissions) and early fault/anomaly detection. Remaining Useful Life (RUL) estimation and direct maintenance scheduling algorithms are deliberately excluded.
- **Data Modality:** Relies strictly on open-access, publicly available datasets. No proprietary industry data will be requested, nor will physical experimental rigs be constructed—a choice further detailed and justified in Section 3.2.

- **Methodological Framework:** Limited to domain-guided sensor selection, tree-based ensemble baselines, deep time-series networks (LSTM, TCN, autoencoders), few-shot adaptation, domain generalization, and basic physical consistency rules. Large AI foundation models and reinforcement learning architectures are not examined.
- **Validation Strategy:** Conducted purely offline using strictly held-out machines and time periods. While online deployment requirements are addressed in the discussion, live integration into an active plant control system is beyond the scope of this project.

Chapter 2: Literature Review

This chapter evaluates eleven key studies forming the foundation of this work. The initial six publications—carried over from the preliminary research—outline the current landscape of machine learning applications in power plant operations and maintenance. The five newly incorporated studies specifically target the systemic limitations identified in the earlier literature, including domain shift, severe label scarcity, ungrounded evaluation protocols, data privacy restrictions, and physics-agnostic modeling. The chapter concludes with a targeted gap analysis that maps these persistent challenges directly to the core research objectives of this proposal.

2.1 Machine Learning in Power Plants

2.1.1 Tufekci (2014)

Tufekci (2014) asked whether the full-load output of a combined cycle plant can be predicted from ambient conditions alone, without a physics model. Using 9,568 hourly records collected over six years and only four inputs, fifteen regression methods were compared. Bagging with a REP Tree gave the best result, with an MAE of 2.818 MW and an RMSE of 3.787 MW. The dataset was later published and became a standard benchmark, which is why it is used in this research. The limits are clear: the model covers full-load operation only, so it says nothing about part load or transients, and the data come from one plant, so transfer to another plant was never tested.

2.1.2 Stetco et al. (2019)

Stetco et al. (2019) conducted a comprehensive survey on machine learning applications in wind turbine condition monitoring, categorizing prior work by data sources, feature extraction techniques, model architectures, and validation schemes. Their analysis revealed a heavy reliance on SCADA and synthetic data, with decision trees, SVMs, and neural networks emerging as the dominant approaches. Importantly, they identified class imbalance and the rarity of labeled fault events as primary industry bottlenecks. Although their study is purely review-based and restricted to wind energy, the core operational challenges they highlight translate directly to thermal power plants.

2.1.3 Carvalho et al. (2019)

Carvalho et al. (2019) conducted a systematic literature review on machine learning in predictive maintenance using a structured selection protocol. Their findings identified Random Forest, neural networks, SVMs, and k-means as the most prevalent algorithms, with hybrid and ensemble configurations consistently yielding optimal performance. Crucially, the authors exposed a fundamental weakness in the field: a heavy dependence on proprietary datasets, which severely hinders reproducibility and direct model benchmarking. The review's main

limitations lie in its tight inclusion criteria, which restricted the number of evaluated primary studies, and its broad focus on general industrial maintenance rather than power generation specifically.

2.1.4 Khalid et al. (2021)

Khalid et al. (2021) focused on a practical problem: power plants monitor hundreds of sensors, but irrelevant data actually hurts model performance. Using real plant data, they used correlation analysis to select only the most relevant sensors for each fault, then trained models to detect boiler tube leaks and turbine issues. This reduced sensor set maintained detection accuracy while speeding up processing time. However, the study was tested on only one plant, covered just two fault types, relied on traditional algorithms, and lacked real-time validation.

2.1.5 Tan et al. (2019)

Tan et al. (2019) used an LSTM model to predict NO_x emissions in a 660 MW coal boiler, using seven days of operational data (10,000 samples). After tuning key parameters, the model achieved strong accuracy with errors under 3%, outperforming a standard SVM baseline in both steady and dynamic conditions. However, the study has clear limitations: seven days from a single boiler cannot account for seasonal changes, fuel variations, or long-term wear, and the work only covers prediction without providing control solutions.

2.1.6 Fahmi et al. (2024)

Fahmi et al. (2024) developed a TCN-Autoencoder with multi-head attention to detect anomalies in gas turbine vibration data. They evaluated their approach on a dataset of one million data points from four accelerometers at the Kirkuk power plant, successfully flagging real events like blade misalignment better than GRU and LSTM baselines. However, the study has two main limitations: the dataset covers only one month from a single turbine, and results are given only as reconstruction error without precision, recall, or false alarm rates, making its practical value hard to evaluate.

2.2 Studies That Address the Weak Points

2.2.1 Zhao, Zio and Shen (2024) - Domain Generalization

This study, published in *Reliability Engineering & System Safety*, directly addresses a major problem: most fault diagnosis models assume that training and test data share the same pattern, whereas domain shift is common in practice. The authors explore domain generalization, where a model learns from several machines to work on a new, unseen machine. Their main contribution is creating a unified benchmark using ten datasets (eight public and two private), allowing different methods to be compared fairly for the first time.

The findings show that adapting a model to an unseen machine is far harder than testing it on the same machine. For this research proposal, the paper provides three key elements: a clear problem definition, baseline methods for comparison, and proof that a unified testing framework is necessary. Its main limit is that most benchmark datasets focus on standalone bearings and gearboxes rather than full power plant systems.

2.2.2 Kim et al. (2022) - Honest Evaluation

Kim et al. (2022), presented at AAAI, focused on how models are evaluated rather than proposing a new model. They examined the common practice of using 'point adjustment' in anomaly detection, where flagging a single point counts as detecting an entire fault segment. The authors proved, both theoretically and experimentally, that this method severely inflates performance scores. As a result, even random predictions can appear to outperform top methods, rendering published rankings highly misleading.

This matters for the present research in two ways. It means some of the impressive numbers in the fault detection literature may not survive an honest re-check, and it gives a solid reason to report precision, recall and false alarm rate without adjustment. Its limit is that it is written for the general time-series community and does not discuss what a false alarm actually costs in a plant.

2.2.3 Liang et al. (2023) - Few-Shot Learning

Liang et al. (2023) reviewed few-shot learning for mechanical fault diagnosis using vibration data from 2018 to late 2023. They addressed a key issue highlighted by Stetco et al. (2019): critical machines rarely run until failure, and manual labeling is costly, leaving engineers with very few fault examples. The authors found that meta-learning dominates the field, making up over 90% of published approaches (mainly MAML variants and metric-based methods). However, a major limitation—and an opportunity for our study—is that most models were tested on laboratory bearing datasets rather than real power plants, likely overstating their practical accuracy.

2.2.4 Ahn et al. (2023) - Federated Learning

Ahn, Lee, Kim, Park and Jeong (2023), in *Sensors*, addressed the fact that industrial sites are spread out and that each site produces data with a different distribution, which makes central training difficult both technically and legally. They proposed a 1DCNN-BiLSTM model for time-series anomaly detection and combined it with federated learning, so each site trains locally and only model parameters are shared, not raw data. This is relevant because the same restriction is one reason why open power plant datasets barely exist. The limit is that the work is set in manufacturing rather than power generation, and the number of clients is small compared with a real fleet of units.

2.2.5 Zhu et al. (2025) - Physics-Informed Model

Z Zhu et al. (2025) addressed two key issues in data-driven thermal models: high data requirements and predictions that violate physical laws. They proposed a physics-informed variational autoencoder that adds physical rules—such as parameter ranges and relationships—directly into the loss function. In simulations, adding these constraints increased model accuracy, raising R-squared rising from 0.902 for a plain autoencoder to 0.976. When tested on a real power plant heater, the model remained accurate and physically logical under unseen operating conditions. However, the study focuses on a single subsystem and targets modeling rather than fault detection.

2.3 Gap Analysis

Reviewing these eleven studies reveals a clear pattern. The field's strength lies in its use of real operational data and advanced techniques, ranging from tree ensembles to attention-based autoencoders. However, key weaknesses persist: most results are limited to single sites, labeled fault data remains scarce, data sharing is restricted, evaluation metrics are inconsistent, and models can produce non-physical outputs outside their training ranges. While recent studies address these issues individually, none combine these solutions within power plant applications. The table below links each identified gap directly to the objectives of this research.

Gap in the literature	What this research does about it
Models are tested on one plant or one unit only (Tufekci, 2014; Tan et al., 2019; Fahmi et al., 2024).	Train on some machines and test on a machine the model has never seen (Objectives 2 and 5).
Accuracy drops when the machine or the operating condition changes (Zhao et al., 2024).	Use domain generalization ideas and compare against the baselines from that benchmark (Objective 3).
Very few labelled faults, and they are imbalanced (Stetco et al., 2019; Liang et al., 2023).	Train first on normal data only, then fine-tune with a handful of fault examples (Objective 3).
Metrics differ from paper to paper, and point adjustment inflates scores (Kim et al., 2022).	Fix the metrics before the experiments and never use point adjustment (Objective 4).
Most datasets are private, so nobody can check the results (Carvalho et al., 2019; Ahn et al., 2023).	Use public datasets only, and share the code and settings (Objective 2).
Data-driven models can break basic physics outside the training range (Zhu et al., 2025).	Add simple physical rules to the loss function (Objective 3).

Table 2.1: Gaps in the literature and how this research responds.

Two main debates in the literature guided the methodology for this research. First, Tufekci (2014) showed that tree-based ensembles outperform neural networks for static regression, whereas Tan et al. (2019) and Fahmi et al. (2024) found deep learning superior for time-series tasks. Consequently, both approach types are included as baselines rather than assuming deep

models are always better. Second, Khalid et al. (2021) recommend reducing sensor inputs, while Fahmi et al. (2024) use high-dimensional raw signals directly. Because these strategies conflict, our experiments will compare both directly instead of choosing one beforehand.

Chapter 3: Research Methodology

This research uses a quantitative, experimental design. The reason is simple: the main claim is that a model can keep its accuracy on a machine it has never seen, and the only way to settle that is to run controlled experiments and measure what happens. The work is split into five phases, each linked to the objectives in Section 1.4.

3.1 Phase 1: Literature Review

Objective 1 Execution The literature review in Chapter 2 is expanded through a systematic search across IEEE Xplore, ScienceDirect, Scopus, and the ACM Digital Library. The search targets key terms including *power plant*, *fault detection*, *predictive maintenance*, *domain generalization*, *few-shot learning*, and *anomaly detection evaluation*. Papers are selected if they are peer-reviewed, published from 2014 onwards, and evaluate real machinery data.

A detailed matrix organizes each study by its task, data source, dataset size, model type, metrics, use of point adjustment, and cross-machine validation. Recording point adjustment and cross-machine testing provides a fresh view missing from past reviews. The final output is a documented gap analysis and a clear list of baseline models to benchmark.

3.2 Phase 2: Data

- **Objective 2 Execution** This research relies exclusively on public datasets. Using open data avoids reliance on private plant agreements while solving a key issue highlighted by Carvalho et al. (2019) and Zhao et al. (2024): private datasets prevent reproducibility and fair model comparison. By using open data, all results in this study can be independently verified.
- **Selected Datasets**
- **UCI Combined Cycle Power Plant (Tufekci, 2014):** Contains 9,568 hourly records over six years for performance prediction.
- **Open Rotating Machinery Datasets (Zhao et al., 2024):** Includes CWRU and Paderborn bearing data, which cover multiple machines and operating conditions, allowing specific machines to be held out for testing.
- **Open Acoustic and Vibration Datasets (Liang et al., 2023):** Includes MIMII and MFPT datasets, featuring multiple machine units of the same type for few-shot learning evaluation.
- **Standardized Data Preparation**
- **Cleaning:** Remove duplicate timestamps, filter noise using a moving average, and interpolate short missing gaps (Fahmi et al., 2024).
- **Sensor Selection:** Apply correlation analysis to rank inputs (Khalid et al., 2021) while retaining a full-input baseline to evaluate high-dimensional modeling strategies.

- **Windowing:** Convert time series into fixed-length overlapping windows with tuned window sizes (Tan et al., 2019).
- **Imbalance Handling:** Apply class weighting and controlled resampling strictly to the training set, leaving the test set untouched.
- **Data Splitting:** Hold out at least one complete machine strictly for final testing. Random window shuffling across splits is avoided to prevent data leakage.
-

3.3 Phase 3: Building the Model

Objective 3 Execution The proposed architecture consists of four components, each addressing a specific requirement identified in the literature:

- **Input Stage:** Integrates optional correlation-based sensor selection (Khalid et al., 2021) to directly test the impact of reduced inputs.
- **Core Model:** Combines a Temporal Convolutional Network (TCN) Autoencoder trained on normal data (Fahmi et al., 2024) with a domain-invariant objective (Zhao et al., 2024) to handle cross-machine variations.
- **Few-Shot Module:** Uses a metric-based meta-learning head (Liang et al., 2023) to adapt to new faults using very few labeled examples.
- **Physics-Informed Loss:** Adds physical constraints—such as expected parameter ranges and load relationships—directly to the loss function (Zhu et al., 2025).

Implementation Details The framework is built using Python with PyTorch, alongside scikit-learn for traditional baselines. All code, hyperparameter settings, and random seeds are logged to ensure full reproducibility. Training will be conducted on a GPU workstation or cloud GPU instances.

3.4 Phase 4: How Results Are Measured

Objective 4 Execution

Evaluation rules are fixed prior to running experiments to ensure unbiased reporting. Prediction performance is evaluated using MAE and RMSE (Tufekci, 2014; Tan et al., 2019). Fault detection is measured using precision, recall, F1-score, and daily false alarm rates.

Crucially, all metrics are computed point by point without point adjustment, following the findings of Kim et al. (2022). Where direct comparison with published studies requires point adjustment, both raw and adjusted scores are reported transparently. Detection delay—the time elapsed between fault onset and the initial alarm—is also recorded. Every experiment is executed across five random seeds, reporting results as mean $\pm$ standard deviation with paired statistical significance tests applied before asserting model superiority.

3.5 Phase 5: Experiments and Comparison

Objective 5 Execution Three main experiments are designed to validate the framework:

- **Experiment A (Single-Machine Baseline):** Models are trained and tested on the same machine, establishing a benchmark that reflects standard literature conditions.
- **Experiment B (Unseen Machine Test):** Models are trained on a set of machines and evaluated on a completely unseen target machine to test domain generalization.
- **Experiment C (Label Scarcity Sensitivity):** Repeats Experiment B while restricting the available labeled fault instances for fine-tuning to 5, 10, and 20 samples to identify performance breakdown thresholds.

The comparison set includes bagged trees (Tufekci, 2014), an LSTM (Tan et al., 2019), a TCN-Autoencoder (Fahmi et al., 2024) and at least two domain generalization baselines from Zhao et al. (2024). Ablation studies remove one part of the proposed model at a time, so the contribution of the sensor selection, the domain-invariant objective, the few-shot head and the physical rules can each be measured on its own.

3.6 Ethical Considerations

This research utilizes machine datasets exclusively, with no personal data involved. All datasets are publicly available for research, with full adherence to their respective licensing terms and proper citation. The primary ethical requirement of this study is rigorous, unbiased reporting: negative results and failure cases of the proposed model will be disclosed fully, upholding the core argument of this proposal against performance inflation in the literature.

3.7 Limitations

Three limitations are stated in advance. First, there is no access to live plant data, so the cross-machine test is done on open machinery datasets rather than on two real generating units. These datasets are smaller and cleaner than a plant historian, so the results should be read as evidence that the approach works, not as proof that it is ready for a control room. Second, the number of labelled faults in open datasets is limited, which limits the statistical power of the detection results; this is reported rather than hidden. Third, the evaluation is offline, so it cannot capture things that matter in a live plant, such as integration with the control system, operator trust, and model drift over years. These are discussed as future work.

3.8 Expected Outcomes

- **Documented Gap Analysis:** A systematic literature matrix recording cross-dataset validation and scoring methodologies for each study—addressing a key gap in existing reviews.

- **Standardized Multi-Machine Pipeline:** A fully prepared multi-machine dataset collection paired with a reproducible data-preprocessing pipeline to enable independent verification.
- **Integrated Domain-Generalization Architecture:** A novel framework combining sensor selection, domain-invariant temporal features, a few-shot head, and physics-informed constraints, designed to maintain high accuracy on unseen machinery.
- **Fault-Label Scarcity Thresholds:** Quantitative guidelines detailing the minimum number of labeled fault samples required before model performance degrades, serving as a practical guide for industrial deployment.
- **Unbiased Evaluation Protocol:** A standardized benchmark protocol reporting prediction errors, detection metrics, and detection delays without point adjustment, resolving the performance inflation issues identified by Kim et al. (2022).
- **Practical Recommendations & Future Roadmap:** Operational guidelines for engineers alongside an honest summary of open challenges, detailing the exact requirements for real-world plant deployment.

3.9 Gantt Chart

The research is planned over twelve months, from September 2026 to August 2027. Writing starts in month seven and runs alongside the experiments instead of being left to the end, so results are written down while they are still fresh.

Task / Month	S	O	N	D	J	F	M	A	M	J	J	A
1. Literature review and gap analysis												
2. Fix the evaluation rules												
3. Download and clean the datasets												
4. Build the baseline models												
5. Build the proposed model												
6. Run the experiments												
7. Analyse and compare results												
8. Write the thesis												
9. Revision and submission												
10. Prepare the presentation												

Figure 3.1: Gantt chart of the research plan (S = September 2026 to A = August 2027).

References

- Ahn, J., Lee, Y., Kim, N., Park, C., & Jeong, J. (2023). Federated learning for predictive maintenance and anomaly detection using time series data distribution shifts in manufacturing processes. *Sensors*, 23(17), Article 7331. <https://doi.org/10.3390/s23177331>
- Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcala, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, Article 106024. <https://doi.org/10.1016/j.cie.2019.106024>
- Fahmi, A.-T. W. K., Reza Kashyzadeh, K., & Ghorbani, S. (2024). Advancements in gas turbine fault detection: A machine learning approach based on the temporal convolutional network-autoencoder model. *Applied Sciences*, 14(11), Article 4551. <https://doi.org/10.3390/app14114551>
- Khalid, S., Hwang, H., & Kim, H. S. (2021). Real-world data-driven machine-learning-based optimal sensor selection approach for equipment fault detection in a thermal power plant. *Mathematics*, 9(21), Article 2814. <https://doi.org/10.3390/math9212814>
- Kim, S., Choi, K., Choi, H.-S., Lee, B., & Yoon, S. (2022). Towards a rigorous evaluation of time-series anomaly detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7), 7194–7201. <https://doi.org/10.1609/aaai.v36i7.20680>
- Liang, X., Zhang, M., Feng, G., Wang, D., Xu, Y., & Gu, F. (2023). Few-shot learning approaches for fault diagnosis using vibration data: A comprehensive review. *Sustainability*, 15(20), Article 14975. <https://doi.org/10.3390/su152014975>
- Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Keane, J., & Nenadic, G. (2019). Machine learning methods for wind turbine condition monitoring: A review. *Renewable Energy*, 133, 620–635. <https://doi.org/10.1016/j.renene.2018.10.047>
- Tan, P., Xia, J., Zhang, C., Fang, Q., & Chen, G. (2019). Dynamic modeling of NO_x emission in a 660 MW coal-fired boiler with long short-term memory. *Energy*, 176, 429–436. <https://doi.org/10.1016/j.energy.2019.04.020>
- Tufekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems*, 60, 126–140. <https://doi.org/10.1016/j.ijepes.2014.02.027>
- Zhao, C., Zio, E., & Shen, W. (2024). Domain generalization for cross-domain fault diagnosis: An application-oriented perspective and a benchmark study. *Reliability Engineering & System Safety*, 245, Article 109964. <https://doi.org/10.1016/j.ress.2024.109964>
- Zhu, B., Ren, S., Weng, Q., & Si, F. (2025). A physics-informed variational autoencoder for modeling power plant thermal systems. *Energies*, 18(17), Article 4742. <https://doi.org/10.3390/en18174742>