Exploring the Cross-Turbine Generalization Capacity of Anomaly Reinforced Autoencoder-based Normal Behaviour Models

Jai Parakh

25165933

Faculty of Science and Engineering

Department of Computer Science and Information Systems

University of Limerick

Submitted to the University of Limerick for the degree of

MSc in Artificial Intelligence and Machine Learning

August 25, 2026

1. Supervisor: Dr. Juan Albarracin
Department of Computer Science and Information Systems
University of Limerick
Ireland

2. Supervisor: Prof. Conor Ryan
Department of Computer Science and Information Systems
University of Limerick
Ireland

Abstract

Normal Behaviour Model (NBM) applied in the wind industry are machine learning algorithms that learn how wind turbines behave during normal operation, so that deviations from this behaviour can be flagged as anomalies. While Autoencoder-based architectures, and in particular the Anomaly Reinforced Autoencoder (ARAE) proposed by Bui et al. (2026) have proven highly effective for this task, existing paradigms require training and maintaining a separate model per turbine, resulting in substantial computational and maintenance overhead across large wind farms.

This dissertation investigates the cross-turbine generalisation capacity of the ARAE, evaluating whether a model trained on a single turbine can detect anomalies in other, previously unseen turbines from the same fleet. A systematic fifteen-experiment ablation study is conducted on real industrial Supervisory Control and Data Acquisition (SCADA) data from a ten-turbine wind farm, isolating the individual and combined contributions of three interventions: Condition-Binned Standardisation (CBS), a power-regime-conditioned re-normalisation scheme adapted from Pillay et al. (2025), decoder fine-tuning, a lightweight per-turbine calibration step adapted from Roelofs et al. (2024), and a Mixture of Experts (MoE) extension that routes between two specialised decoder heads.

The unmodified ARAE is shown to generalise no better than random guessing across the fleet (mean Area Under the Curve (AUC) ≈ 0.50), while combining CBS with decoder fine-tuning, applied to a single shared decoder, recovers the majority of usable discriminative performance (mean AUC = 0.735, mean recall = 0.835) and is adopted as this dissertation's proposed model. The MoE extension was investigated fully and showed that a power-bin-conditioned router is shown to collapse to near-uniform expert usage regardless of load-balancing regularisation, and a statistically rigorous, paired Wilcoxon comparison across all nine held-out turbines demonstrates that neither the learned router nor a fixed-weight ensemble

simplification of the same provides any significant improvement over the single-decoder baseline at any stage of simplification.

This dissertation adopts the single-decoder, CBS-conditioned, fine-tuned ARAE rather than either of the MoE variants. This flat fleet mean, however, is not evenly distributed. Stratifying by class balance shows the model recovers substantial generalisation capacity on turbines with a realistic, moderate rate of anomalous operation, performs best on turbines where the detection task is structurally easier due to an abnormal-majority operating history, and fails severely on at least one held-out turbine.

The findings demonstrate that architectural complexity is not what enables cross-turbine transfer in this setting, but power-regime-conditional data normalisation and lightweight per-turbine adaptation are. The conclusion documents a rigorous negative result for MoE routing in this domain, and offers a practical, low-maintenance path toward fleet-wide predictive maintenance deployment, while identifying the class-balance conditions under which it still fails.

Declaration

I herewith declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other Irish or foreign examination board.

The thesis work was conducted from 2025 to 2026 under the supervision of *Dr. Juan Albarracin* at University of Limerick.

Limerick, 2026

Declaration on Use of Generative AI

I hereby declare that I have followed all the guidelines and policies set forth by the University of Limerick regarding the use of generative AI in academic work ("Generative Artificial Intelligence, University of Limerick 2024"). Throughout the process of preparing the dissertation, I have used AI tools solely to rephrase, and improve the clarity of my own original ideas and writing so that it looks more: dissertation like. I ensured that the core content remained my own original contribution. AI was not used to generate ideas, or replace critical thinking and research efforts. AI was also used for analyzing the Wilcoxon values of the performed experiments. All AI usage was conducted in an ethical manner consistent with the university's standards for academic integrity and was discussed and agreed upon by my supervisor. I affirm that this declaration accurately reflects my adherence to the university's policies and that the final dissertation is a genuine and honest representation of my academic work

Limerick, 2026

Acknowledgements

I would like first and foremost, to express my sincere gratitude to my direct supervisor, Juan Albarracin. I am deeply thankful for his unwavering responsiveness in answering my questions and addressing my concerns. His willingness to provide essential knowledge support and meticulous proof-reading were fundamental to the progress of this work.

My special thanks go to Prof. Conor Ryan, Head Director of the BDS research group. His remarkable skills and charismatic leadership were truly inspiring and were instrumental in fostering my passion for academic research.

I am also indebted to the entire BDS group, a community of bright and talented individuals. Though our direct interactions may have been limited, their collective intellect and enthusiasm have been a significant source of influence and encouragement, solidifying my commitment to pursuing a path in artificial intelligence.

Contents

CONTENTS

List of Tables

List of Figures

Acronyms

AE Autoencoder. 1, 2, 8, 9, 11, 12, 13, 14, 17, 18, 19

ARAE Anomaly Reinforced Autoencoder. iv, v, vii, 1, 2, 3, 9, 11, 12, 13, 14, 18, 19, 20, 22, 23, 25, 26, 27, 28, 32, 33, 42, 49, 51, 52, 53, 55

AUC Area Under the Curve. iv, v, vii, 1, 30, 31, 33, 34, 38, 39, 40, 41, 42, 46, 47, 48, 49, 50, 51, 52, 53, 54

CBS Condition-Binned Standardisation. iv, v, vii, 3, 15, 16, 19, 23, 24, 25, 26, 27, 32, 33, 34, 35, 38, 39, 49, 50, 51, 52, 53, 54, 55

CNN Convolutional Neural Network. 8

FN False Negative. 30, 33

FP False Positive. 30, 33

FPR False Positive Rate. 30

FT Fine-Tuning. 32

GMM Gaussian Mixture Model. 23

LSTM Long Short-Term Memory. 8

MMD Maximum Mean Discrepancy. 12, 28, 29

MoE Mixture of Experts. iv, v, 3, 15, 17, 19, 25, 26, 28, 29, 32, 39, 41, 46, 47, 49, 52, 53

MSE Mean Squared Error. 8, 12, 28, 29

Acronyms

NBM Normal Behaviour Model. iv, 1, 2, 3, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18

SCADA Supervisory Control and Data Acquisition. iv, 1, 2, 3, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 20, 23, 27

TL Transfer Learning. 9, 10

TN True Negative. 30, 33

TP True Positive. 30, 33

TPR True Positive Rate. 30

WT Wind Turbine. 6, 7, 8, 11, 13, 14, 17, 18

1

Introduction

Wind energy continues to expand as a central pillar of global decarbonisation efforts, entailing an increasing operational burden to maintain the reliability of ever-larger turbine fleets while keeping maintenance costs proportionate to energy output. Unplanned downtime carries substantial cost across a wind turbine fleet, with failure frequency and repair duration varying significantly by subassembly (Faulstich et al., 2011). This cost is further compounded in offshore deployments, where the logistics of accessing installations for corrective maintenance are inherently more constrained than onshore. Early and reliable fault detection is therefore a critical capability for wind farm operators. Supervisory Control and Data Acquisition (SCADA) systems are standard equipment on virtually all commercial turbines, and provide the data substrate for this task. Normal Behaviour Models (NBMs) are semi-supervised anomaly detection models that learn to characterise healthy operation from historical SCADA data, and flag deviations from this learned normality as potential anomalies. Among the multiple approaches to NBM construction, deep reconstructive architectures, like Autoencoders (AEs), have become a dominant approach within this space, as reviewed in Chapter 2.

1.1 Motivation

Bui et al. (2026) proposed the Anomaly Reinforced Autoencoder (ARAE), a reconstructive autoencoder trained to actively reinforce reconstruction error on known-abnormal windows rather than simply tolerating it. He also demonstrated that this reinforcement mechanism produces strong anomaly detection performance with an AUC of 0.8607, when evaluated on a dataset pooled across multiple turbines from a single fleet, with training, validation, and test splits stratified only by class label rather than held out by

turbine. This is a meaningful result, but it leaves open a question of direct practical consequence for fleet-wide deployment: Does the ARAE's learned representation transfer to a turbine that contributed no data to training at all? The standard deployment pattern for NBM-based condition monitoring is to train and maintain one model per turbine, an approach that scales poorly as fleets grow and that is particularly ill-suited to newly commissioned turbines or units recently returned from major maintenance which, by definition, lack the historical operating data required to train a model of their own.

Within this NBM paradigm, the underlying assumption is that a model exposed only to healthy operating data will fail to capture patterns it has never encountered. An Autoencoder (AE) operationalises this assumption through a compress-and-rebuild architecture as an encoder maps each input window to a compact latent code, and a decoder attempts to recover the original window from that code alone (Jonas and Meyer, 2025). Because training uses exclusively normal operating windows, the network's internal representation becomes tuned specifically to the statistical regularities of healthy behaviour. A window that departs from those regularities is correspondingly reconstructed with higher error, and this elevated reconstruction error at inference time constitutes the anomaly score (Jonas and Meyer, 2025). The ARAE of Bui et al. (2026) builds on this same reconstruction-error mechanism, but modifies the training objective so that the decoder is not merely permitted to reconstruct abnormal windows poorly, but actively reinforced toward doing so, as detailed in Chapter 2.

A small number of recent studies have investigated cross-turbine transfer for autoencoder-based NBMs specifically. Roelofs et al. (2024) show that fine-tuning only the decoder of a pre-trained autoencoder, with the encoder frozen, recovers most of the performance of a full-year, target-turbine-specific training run using only one to three months of target data, but their study evaluates a conventional, non-reinforced autoencoder, not the ARAE's reinforcement objective. Jonas and Meyer (2025) propose an alternative generative domain-mapping strategy and show it can outperform fine-tuning under severe data scarcity, again for a conventional reconstruction-based NBM. Separately, Pillay et al. (2025) show that conditioning SCADA feature standardisation on active power, the turbine's dominant operating-condition signal substantially reduces the model complexity required for effective anomaly detection with a Gaussian Mixture Model, though their evaluation is confined to a single fleet with no turbine held out from training. No existing work, to the best of our knowledge, combines an anomaly-reinforced training objective, operating-state-conditioned standardisation, and a genuine turbine-level holdout evaluation protocol within a single study, which is the specific combination this dissertation investigates.

1.2 Problem Statement

Despite growing evidence that cross-turbine transfer is achievable for conventional autoencoder-based NBM, it remains untested whether the same holds for an anomaly-*reinforced* objective specifically, whose defining mechanism is actively degrading reconstruction on known-abnormal inputs and has only ever been trained and evaluated using data from turbines also present, in some proportion, in that same training run. Whether this mechanism generalises to a turbine excluded from training entirely, and what combination of preprocessing and adaptation techniques might be required to make it do so, has not been established. Furthermore, while several individually promising techniques exist for related sub-problems, operating-condition based standardisation, decoder-only fine-tuning, and conditional-computation architectures such as Mixture-of-Experts, which have shown promise in other domains but have not, to the best of our knowledge, been applied to wind turbine condition monitoring at all. There is no existing study that evaluates these techniques together, individually and in combination, under a statistically rigorous, genuinely turbine-level holdout protocol. This gap is addressed directly in this dissertation.

1.3 Aims

This dissertation is motivated by a single question: *Does the ARAE's learned representation transfer to a turbine that contributed no data to training at all?* We investigate this cross-turbine generalisation capacity using real industrial SCADA data from a ten-turbine wind farm, designating one turbine as the sole source of training data and holding the remaining nine out entirely, evaluating generalisation under a protocol substantially stricter than the one used in the ARAE's original proposal.

First, we establish a baseline by evaluating the unmodified ARAE, trained on a single turbine, directly on the nine held-out turbines with no adaptation of any kind, testing whether the reinforcement mechanism transfers on its own.

Second, we adapt and evaluate two techniques for closing any observed generalisation gap: CBS, adapted from a Gaussian Mixture Model setting to this dissertation's deep reconstructive architecture, and decoder-only fine-tuning, adapted from a conventional-autoencoder setting to the ARAE's reinforced objective, evaluated both individually and in combination.

Third, we investigate whether a MoE extension that routes between specialised decoder heads conditioned on operating power regime provides any additional generalisa-

tion benefit beyond standardisation and fine-tuning alone, an application of conditional computation not previously explored in this domain. We conduct this investigation to a statistically rigorous standard throughout, using matched-pairs significance testing across the held-out turbines rather than raw point-estimate comparison.

Finally, we propose a final model on the basis of this evidence, characterise its performance honestly by stratifying results according to each turbine's class balance rather than reporting a single aggregate figure, and identify the specific conditions under which the proposed approach still fails.

1.4 Contribution

This dissertation makes four contributions:

1. It introduces an operating-point-conditioned standardisation mechanism to align data across turbines, applied here for the first time to a deep reconstructive architecture.

2. It extends a lightweight, decoder-only calibration strategy, previously validated only for conventional autoencoders, to an anomaly-reinforced architecture for the first time.

3. It provides a statistically grounded, fleet-scale evaluation of cross-turbine generalisation across nine held-out turbines, at a scale not attempted by prior comparable studies.

4. It introduces a conditional-computation extension to the decoder, rigorously assessed throughout using matched-pairs statistical testing rather than point-estimate comparison, a standard of rigour not found in comparable prior work.

Together, these contributions constitute the first systematic investigation of how domain-aligning standardisation, lightweight decoder calibration, and conditional-computation specialisation interact to enable or constrain cross-turbine generalisation of an anomaly-reinforced autoencoder.

1.5 Thesis Structure

The remainder of this dissertation is organised as follows: Chapter 2 introduces the background of wind turbine condition monitoring, Normal Behaviour Monitoring, and

deep reconstructive models, establishing the theoretical basis for this work. Chapter 3 reviews related studies spanning cross-turbine transfer learning, operating-condition-aware standardisation, and Mixture-of-Experts architectures, positioning this dissertation's contribution against the landscape. Chapter 4 describes the dataset, preprocessing pipeline, model architectures, training procedures, and evaluation protocol. Chapter 5 presents and discusses the results of the fifteen-experiment ablation study, comparing model configurations and analysing the statistical significance of each architectural and preprocessing decision. Finally, Chapter 6 summarises the key findings and contributions of this dissertation and outlines directions for future work.

2

Background

2.1 Wind Turbine Condition Monitoring

Wind energy continues to expand as a central pillar of global decarbonisation efforts. The global installed wind capacity reached 1,299 GW by the end of 2025 (Global Wind Energy Council, 2026), entailing an increasing operational burden to maintain the reliability of ever-larger turbine fleets while keeping maintenance costs proportionate to energy output. Faulstich et al. (2011) analysed failure frequency and downtime duration across wind turbine subassemblies using onshore operational data from Germany's "250 MW Wind" programme, and used this onshore experience to project the likely reliability outcomes for turbines subsequently deployed offshore. Condition monitoring systems that can detect the early onset of degradation therefore occupy a critical role in modern Wind Turbine (WT) operations, particularly as deployment shifts toward offshore sites where the reliability challenges identified onshore are expected to be compounded.

The primary data source for industrial-scale WT condition monitoring is the SCADA system, which is embedded as standard equipment in virtually all commercial turbines. A SCADA system records a broad array of operational signals, including generator speed, rotor speed, blade pitch angles, nacelle temperature, transformer temperature, active and reactive power output, and wind speed that are usually averaged over ten-minute intervals. This averaging is widely reported to smooth out the high-frequency transients that are often most diagnostic of incipient faults, a persistent limitation of SCADA-based (as opposed to dedicated vibration-based) condition monitoring. Tautz-Weinert and Watson (2017) provide a comprehensive review of SCADA-based condition monitoring approaches, categorising the literature into trending, clustering, normal behaviour modelling, damage modelling, and alarm/expert-system approaches.

The number of SCADA channels available varies substantially across turbine fleets: this dissertation's own dataset comprises 33 variables (Bui et al., 2026), while other published studies report as few as 28 or as many as several hundred sensor channels per turbine before feature selection (Jonas and Meyer, 2025; Roelofs et al., 2024). Fleets of tens or hundreds of turbines generate commensurately large volumes of multivariate time-series data regardless of the exact channel count.

The taxonomy of anomaly types relevant to WT SCADA analysis was formalised in the survey by Yan et al. (2024), who distinguished three fundamental categories. A *point anomaly* occurs when a single observation deviates markedly from the expected range of values, for example a sudden spike in gearbox temperature. A *contextual anomaly* is one in which an observation is anomalous only with respect to its local temporal or operational context. A generator temperature reading that would be unremarkable at high wind speeds may be anomalous during periods of near-zero generation. A *collective anomaly* manifests when a sequence of individually plausible observations is anomalous in aggregate, a pattern of subtle, correlated drift across multiple signals over several hours or days that no single reading would betray. WT health degradation is reasoned here to predominantly produce collective and contextual anomalies rather than point anomalies, since early-stage faults typically express themselves as gradual, multivariate departures from expected operating relationships rather than abrupt single-channel excursions.

2.2 Normal Behaviour Monitoring

A Normal Behaviour Model (NBM) is a data-driven algorithm that characterises the healthy, fault-free operating regime of a WT from historical SCADA data, learning normal operational patterns so that deviations in posterior predictions can be flagged as potential faults (Chesterman et al., 2023). The underlying paradigm is semi-supervised: the model is trained exclusively on data drawn from confirmed normal operating periods, and anomalies are subsequently identified at inference time by measuring how far new observations depart from the learned normal behaviour. This semi-supervised approach is well suited to WT condition monitoring because fault events are, by design, rare relative to the volume of healthy operation, making it impractical to obtain labelled fault examples in sufficient quantity to train fully supervised classifiers (Yan et al., 2024).

The canonical detection mechanism within an NBM framework is the reconstruction error, also referred to as the residual. At inference time, an input observation $\mathbf{x}$ is

passed through the model to obtain a reconstruction $\hat{\mathbf{x}}$, and the scalar anomaly score s is computed as the Mean Squared Error (MSE) between $\mathbf{x}$ and $\hat{\mathbf{x}}$:

$$s = \frac{1}{D}\sum_{d=1}^{D}(x_d - \hat{x}_d)^2, \tag{2.1}$$

where D denotes the dimensionality of the input. An observation is flagged as anomalous when s exceeds a chosen threshold. A common practical approach, adopted by Bui et al. (2026) among others, is to select this threshold by maximising a chosen classification metric (such as the F_1-score) on a held-out validation set. Deep learning methods, especially neural network-based architectures have rapidly become the dominant approach in recent research (Chesterman et al., 2023). These models offer superior capability to capture non-linear dependencies in high-dimensional data but come at the cost of increased computational demand and reduced interpretability. Within this category, AEs have emerged as a popular choice for anomaly detection, particularly in unsupervised or semi-supervised settings.

2.3 Autoencoder-based Anomaly Detection

Among the deep-learning architectures surveyed for industrial anomaly detection by Yan et al. (2024), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) networks, and AE are reported as the dominant paradigms. The AE has become particularly prominent in WT condition monitoring because it naturally implements the NBM paradigm: an AE trained on healthy data learns a compressed latent representation of normal behaviour, and anomalous inputs that lie outside the training manifold are, in principle, reconstructed poorly, yielding elevated residuals that serve as fault indicators. The general reconstruction-based detection mechanism was surveyed comprehensively by Ruff et al. (2021) across both deep and shallow anomaly detection methods.

For multivariate time-series inputs such as SCADA windows, one-dimensional convolutional layers are a natural architectural choice because they can capture local temporal patterns efficiently without the sequential dependencies of recurrent networks. Renström et al. (2020) demonstrated that a single deep autoencoder is capable of learning system-wide normal behaviour across the full set of SCADA signals in a WT, rather than the one-model-per-component approach typical of contemporary systems. Jonas et al. (2023) subsequently applied the AE NBM paradigm in the vibration-signal domain, showing that convolutional autoencoders can extract fault-relevant features directly from raw vibration spectra without hand-engineered features.

2.3.1 The Anomaly-Reinforced Autoencoder

A limitation of standard AE-based NBMs is that minimising reconstruction error on normal data does not explicitly discourage the model from also reconstructing anomalous data well. If the decoder is sufficiently expressive, it may generalise beyond the training manifold and reconstruct anomalies with low residual, thereby suppressing fault scores. The ARAE addresses this limitation through a modified training objective that simultaneously maximises normal reconstruction fidelity and deliberately degrades the reconstruction of anomalous inputs (Bui et al., 2026). The auxiliary penalty on anomalous samples creates a gradient signal that pushes the decoder away from the complement of the training manifold, sharpening the effective decision boundary between healthy and anomalous operating states. The specific loss formulation and the empirical evaluation of the ARAE against alternative AE variants are reviewed in Chapter 3. The precise batch-level mechanics of this loss, which differ subtly from its conceptual equation as stated by Bui et al. (2026), are established directly against both the original and this dissertation's own implementation in the chapter 4.

2.4 Transfer Learning for Industrial Time Series

Transfer Learning (TL) refers to the family of machine learning techniques that leverage knowledge acquired from a *source domain* or task to improve learning efficiency or performance on a related *target domain* or task. The formal framework adopted in this dissertation follows Yan et al. (2024), who provides a unified taxonomy of deep TL for anomaly detection in industrial time-series settings.

Following the transfer-learning formalism of Pan and Yang (2010), as adopted by Yan et al. (2024), let a domain $\mathcal{D}$ be defined as a pair $\{\mathcal{X}, P(\mathbf{X})\}$ where $\mathcal{X}$ is the feature space and $P(\mathbf{X})$ is the marginal distribution over inputs, and let a task $\mathcal{T}$ be defined as a pair $\{\mathcal{Y}, P(Y|\mathbf{X})\}$ where $\mathcal{Y}$ is the label space and $P(Y|\mathbf{X})$ is the conditional predictive distribution. Given a source domain $\mathcal{D}_S$ with associated task $\mathcal{T}_S$ and a target domain $\mathcal{D}_T$ with task $\mathcal{T}_T$, TL aims to improve the learning of $P_T(Y|\mathbf{X})$ by exploiting knowledge from $\mathcal{D}_S$ and $\mathcal{T}_S$. When $\mathcal{T}_S = \mathcal{T}_T$ but $\mathcal{D}_S \neq \mathcal{D}_T$, the setting is *transductive* (sometimes termed domain adaptation). When $\mathcal{T}_S \neq \mathcal{T}_T$, the setting is *inductive*. Cross-turbine condition monitoring constitutes a transductive problem: the task is to learn a NBM from SCADA signals and is identical across turbines, but the input distributions differ due to hardware heterogeneity, site-specific meteorological conditions, and operational strategies (Yan et al., 2024).

2.4.1 Categories of Deep Transfer Learning

Building on the taxonomy above, Yan et al. (2024) further categorise deep TL implementations into four families:

1. *Instance Transfer*: Re-using source-domain samples alongside a small target-domain sample.
2. *Parameter Transfer*: Adapting some or all of a pretrained model's weights on target-domain data, of which fine-tuning is the most common instance.
3. *Mapping Transfer*: Learning a transformation of the feature space that aligns source and target distributions.
4. *Domain-adversarial Transfer*: Training a feature extractor to be simultaneously discriminative for the task and indistinguishable across domains, typically via an adversarial domain classifier.

Across the industrial applications surveyed by Yan et al. (2024), parameter transfer is reported to be by far the most commonly deployed approach, owing to its comparative implementation simplicity, while instance transfer and domain-adversarial transfer are reported as rarely used effectively in industrial time-series anomaly detection specifically. This taxonomy is used throughout Chapter 3 to position the specific transfer strategies reviewed there.

3

Related Works

This chapter reviews the literature relevant to the cross-turbine generalisation problem investigated in this dissertation, structured to trace a path from established foundations to the specific gap this work addresses. Section 3.1 reviews autoencoder-based normal behaviour models for wind turbine condition monitoring, including the Anomaly Reinforced Autoencoder (ARAE) this dissertation extends. Section 3.2 reviews cross-turbine transfer learning specifically, covering both parameter-transfer approaches such as decoder-only fine-tuning and generative domain-mapping alternatives. Section 3.3 reviews operating-condition-aware feature standardisation, situating Condition-Binned Standardisation against predictive alternatives. Section 3.4 reviews Mixture-of-Experts and conditional computation more broadly, given the absence of prior work applying it within this domain specifically. Section 3.5 then draws these strands together to position this dissertation's contributions against the reviewed literature.

3.1 Autoencoder-based Fault Detection in Wind Turbines

The application of AE-based NBMs to WT condition monitoring has developed steadily over the last half-decade, with successive works establishing the architectural and methodological foundations on which the present research builds.

Renström et al. (2020) presented a systematic investigation into system-wide anomaly detection in WT using deep autoencoders, explicitly motivated by the observation that most contemporary systems monitor only one component at a time, requiring a separate model per component to cover the full turbine. Trained on multi-channel SCADA

time series from healthy operating periods, their approach instead learns a joint representation of the full sensor suite within a single model. The work demonstrates that the semi-supervised AE paradigm is viable at industrial scale for this broader, multi-component framing.

Subsequent work by Roelofs et al. (2021) extended the AE NBM framework to fault localisation through a method they name ARCANA, which frames anomaly explanation as an optimisation problem: rather than reading feature importance directly of the raw reconstruction residual, ARCANA solves for a modified reconstruction that removes anomalous properties from the input while staying close to the original sample, yielding a small, sparse set of highly explanatory features responsible for the anomaly. Roelofs et al. explicitly demonstrate that the raw, non-optimised reconstruction residual fails to correctly attribute a known injected fault (a wind-speed sensor error) to the correct feature, while ARCANA's optimisation-based attribution succeeds.

Jonas et al. (2023) demonstrated that the AE NBM framework extends beyond ten-minute averaged SCADA signals to high-frequency vibration measurements, applying convolutional autoencoders to extract fault-relevant features directly from the vibration half-spectrum without any hand-engineered spectral features, achieving reliable fault detection on both commercial turbine and test-rig data.

The closest predecessor to the present dissertation is the work of Bui et al. (2026), who conducted a systematic capacity study of deep AE-based NBMs on a real-world fleet dataset and proposed an Anomaly Reinforced Autoencoder (ARAE). The ARAE modifies the standard reconstruction loss through an additional penalty term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{normal}} + \frac{1}{1 + \mathcal{L}_{\text{abnormal}}}, \tag{3.1}$$

where $\mathcal{L}_{\text{normal}}$ is the reconstruction MSE on healthy training samples and $\mathcal{L}_{\text{abnormal}}$ is the reconstruction MSE on anomalous samples. The second term is a monotonically decreasing function of $\mathcal{L}_{\text{abnormal}}$. As the model reconstructs anomalies more accurately, this term increases, creating a gradient signal that pushes the decoder to reconstruct anomalies poorly. Equation 3.1 is Bui's own stated formulation and is reproduced here as such, establishing his model as a strong baseline within this fleet. Bui et al. further found that Wasserstein distance was the best-performing auxiliary loss variant across Maximum Mean Discrepancy (MMD), Fourier-domain, and Mahalanobis alternatives, and that min-max feature normalisation was preferred over z-score standardisation on the grounds that the relative spacing of SCADA signals carries diagnostic significance that z-scoring destroys. The ARAE loss formulation, its auxiliary loss variants, and the min-max normalisation preference documented in Bui et al. (2026) collectively

constitute the direct methodological starting point for the present work. However, the evaluation protocol of Bui et al. (2026) pooled windows from multiple turbines and applied a random split stratified only by class label (normal/abnormal), with no turbine-level holdout i.e the same turbines contributing to training also appeared in validation and test, differing only in which time-windows were assigned to each split. The study therefore did not address whether an ARAE generalises to a turbine entirely absent from training, the specific question this dissertation investigates.

3.2 Cross-Turbine Transfer Learning

A growing body of literature has addressed the challenge of adapting WT condition monitoring models across turbines, motivated by the practical constraint that newly commissioned assets lack the months or years of healthy operational history required to train a reliable NBM from scratch.

3.2.1 Parameter Transfer and Fine-Tuning

Roelofs et al. (2024) conducted the most directly relevant study of parameter transfer for AE-based anomaly detection in WTs. Their experimental setting involved two offshore wind farms in the German North Sea, and a fully connected feed-forward AE architecture with PReLU activations and a small bottleneck. Three transfer strategies were evaluated: transferring only the detection threshold from source to target, fine-tuning only the decoder layers of a source-trained AE on target data, and fine-tuning the full AE on target data.

The central finding of Roelofs et al. (2024) is that decoder-only fine-tuning outperforms both threshold-only transfer and full-AE fine-tuning in most of their experimental conditions. The explanation offered is that the encoder, having been trained on the source turbine, has learned low-level temporal feature extractors that are broadly transferable, while the decoder must adapt its reconstruction to the specific signal statistics of the target turbine. Full-AE fine-tuning risks overfitting on the small target sample when the entire network is updated, despite their use of a reduced learning rate (10^{-4} versus 10^{-3} for decoder-only) intended to mitigate this. Furthermore, Roelofs et al. (2024) found that asset-to-asset transfer using a single source turbine as the pre-training base consistently outperformed multi-asset pooling. This directly motivates the present dissertation's own choice to train its base model on a single source turbine rather than a pooled fleet subset (Section 4.1).

The present dissertation extends the decoder-only fine-tuning strategy of Roelofs et al. (2024) from a fully connected AE architecture to the 1D convolutional ARAE of Bui et al. (2026), and evaluates it across a larger set of nine held-out target turbines within the same fleet, thereby providing a more comprehensive picture of how fine-tuning performance varies with turbine-to-turbine similarity, as explained in detail in Section 4.4.2.

3.2.2 Generative Domain Mapping

Jonas and Meyer (2025) proposed an alternative approach to the data-scarcity problem through generative domain mapping. Rather than adapting a pre-trained model to the target turbine, they train a CycleGAN to translate SCADA data from a data-scarce target turbine into the distributional space of a data-rich source turbine, so that an NBM trained entirely on source data can be applied to the translated target data without modification.

The experimental setting of Jonas and Meyer (2025) comprised seven WTs from distinct wind farms sharing the same manufacturer but differing substantially in model type, rated power, and site conditions. Twenty-eight source–target turbine pairs were evaluated under scarcity scenarios ranging from one week to two months of target data. Domain mapping performed better than the no-transfer baseline at one week of scarcity, compared to direct fine-tuning, with the advantage of domain mapping being most pronounced under the most severe scarcity conditions of one to two weeks.

To prevent the CycleGAN from mapping anomalous target states to healthy-looking source states which would suppress anomaly scores and defeat the purpose of the NBM, Jonas and Meyer (2025) introduced three content-preservation losses:

1. A zero-state loss requiring that idle operating states map to idle states.
2. A rated-power loss requiring that full-load states map to full-load states.
3. The standard cycle-consistency loss of the CycleGAN framework.

The authors acknowledge, however, that these constraints only partially address the risk of anomaly suppression, and that the method may remain unreliable for anomaly types that overlap with the operational states used to define the preservation constraints. A direct empirical comparison between this dissertation's fine-tuning approach and Jonas and Meyer's domain mapping, within the same fleet and evaluation protocol, was not conducted here. Jonas and Meyer's underlying SCADA data was obtained under a data-sharing agreement with a named industrial partner and is not publicly available.

A further practical obstacle is specific to this dissertation's fleet: Jonas and Meyer's content-preservation losses require identifying idle and rated-power operating states as distinct subgroups within the training data, and several turbines in this fleet already exhibit severe class imbalance (Table 4.2). A further subdivision into operating-state subgroups would compound this imbalance rather than resolve it. A direct comparison is accordingly identified as a concrete direction for future work by a dedicated reimplementation effort on this dissertation's own fleet.

3.3 SCADA Feature Standardisation

The preprocessing of SCADA signals prior to model training is a consequential methodological choice that interacts with both model capacity and cross-turbine transferability. Conventional standardisation applies a global z-score transformation to each SCADA channel independently, mapping each feature to zero mean and unit variance computed over the full training set. This approach is straightforward but conflates the multimodal distribution of SCADA signals across different operating regimes: a generator temperature signal, for example, has a qualitatively different distribution at low load than at rated power, and a global z-score that averages over these regimes may obscure the conditional structure that a NBM must learn.

Prior work addressing this dependence falls broadly into two families, distinguished by whether the operating-condition relationship is explicitly *predicted* or instead used to *condition* a subsequent standardisation step. Bilendo et al. (2022) take the predictive approach. They fit a heterogenous stacked regressor algorithm to model a turbine's expected power output as a function of environmental and operational variables (wind speed, air density, rotor speed), then monitor the residual between predicted and observed power over time as the health indicator.

Pillay et al. (2025) introduced *Condition-Binned Standardisation* (CBS) as a conditioning based alternative that requires no such auxiliary model, it computes per-bin summary statistics directly from the training data, adding negligible complexity to the existing preprocessing pipeline. This dissertation adopts CBS over the predictive alternatives above for two reasons specific to its own setting: first, it integrates directly with the Min-Max-based feature scaling this dissertation inherits from Bui et al. (2026), requiring no separate model-fitting stage and second, specific to this dissertation's own extensions, the same power-bin assignment used by CBS is repurposed as the conditioning signal for the MoE router investigated in Section 4.3.1. The core idea of CBS is to partition the training data into bins defined by the active power output of the turbine,

used as a proxy for the overall operating condition and to standardise each dependent SCADA feature within each bin separately, using the bin-specific mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$.

Following the general procedure of Pillay et al. (2025), this dissertation's implementation standardises each dependent SCADA feature within its assigned bin using the bin-specific mean and standard deviation:

$$y' = \frac{y - \hat{\mu}_{\text{bin}}}{\hat{\sigma}_{\text{bin}}}. \tag{3.2}$$

The conditioning variable (active power) is excluded from the transformed feature set so that the model trains only on the within-bin residuals of the dependent channels, not on the power level itself.

The motivation of Pillay et al. (2025) was primarily model simplification for edge deployment. By removing the operating-point-driven multimodality from the feature distributions before training, a simpler model with fewer components suffices to describe the standardised data. In the context of the present dissertation, CBS assumes an additional and distinct role beyond parameter reduction. The active-power-conditioned standardisation aligns each turbine's SCADA signals to a common within-bin reference frame, reducing the covariate shift between turbines operating at comparable power levels but in different absolute thermal or mechanical states. This cross-turbine domain alignment function of CBS was not the primary intent of Pillay et al. (2025), whose own evaluation is confined to a single fleet of four turbines with no turbine-level holdout protocol, and has not been previously evaluated in an explicit transfer-learning context.

The preference of Bui et al. (2026) for min-max normalisation over z-score standardisation is worth examining against the CBS framework. Bui et al. (2026) argued that min-max normalisation preserves the relative spacing of SCADA signal values, which they considered diagnostically significant, and applied a single min-max scaler fit globally across their pooled, multi-turbine dataset. Because their evaluation protocol never holds out an entire turbine (Section 3.1), this design choice is never actually stress-tested against a turbine whose true operating range or calibration might differ from the fitted scaling bounds. CBS, by anchoring each channel to its own within-bin distribution, is more robust to these inter-turbine scale differences and is therefore better suited to the cross-turbine generalisation setting of this dissertation.

3.4 Mixture-of-Experts and Conditional Computation

The Mixture-of-Experts (MoE) paradigm routes different inputs to different specialised sub-networks ("experts") via a learned gating mechanism, rather than processing every input through an identical computation path. It was popularised in its modern deep-learning form by Shazeer et al. (2017), who introduced a sparsely-gated MoE layer for large-scale language and translation models, demonstrating that conditional computation could increase model capacity substantially without a proportional increase in computational cost, by activating only a small subset of experts per input.

To the best of our knowledge, prior application of MoE architectures specifically to wind turbine condition monitoring, SCADA-based anomaly detection, or predictive maintenance did not surface any directly relevant prior work. The wind-turbine transfer-learning literature reviewed in Section 3.2, and the broader industrial time-series anomaly-detection survey of Yan et al. (2024) itself makes no mention of MoE or conditional-computation approaches despite covering the intersection of transfer learning, anomaly detection, and industrial time series comprehensively. MoE architectures are well established in large-scale sequence modelling, but their application to small-model, low-data, industrial time-series anomaly detection which is the setting of this dissertation, to the best of our knowledge, appears to be an unexplored territory rather than an area with an existing body of results this dissertation could be benchmarked against. This dissertation's MoE investigation should therefore be read as an exploratory application of a general-purpose technique to a new domain, rather than as extending or challenging a specific prior finding within that domain.

3.5 Summary

The preceding sections have surveyed a field in which deep learning-based approaches, and reconstruction-based architectures in particular, now dominate WT condition monitoring research. The majority of recent SCADA-based fault detection studies employ unsupervised or semi-supervised formulations, reflecting the structural scarcity of labelled fault data in industrial deployments (Yan et al. (2024)). Within this landscape, AE-based NBMs have emerged as the most widely adopted deep learning paradigm, owing to their natural alignment with the semi-supervised training protocol and their capacity for interpretable, per-channel residual analysis. Transfer learning methods for WT condition monitoring remain comparatively underexplored: the available literature is concentrated on parameter transfer for fully connected architectures (Roelofs

et al., 2024) and generative domain mapping (Jonas and Meyer, 2025), with fleet-scale evaluation across more than a handful of target turbines remaining absent.

The literature reviewed in the preceding sections can be organised into three largely independent research streams:

1. The first stream, exemplified by Renström et al. (2020), Roelofs et al. (2021), Jonas et al. (2023), and Bui et al. (2026) has established the AE-based NBM as an effective paradigm for WT condition monitoring, with the ARAE formulation of Bui et al. (2026) representing a strong baseline across their pooled multi-turbine dataset.

2. The second stream comprises Roelofs et al. (2024) and Jonas and Meyer (2025), and the further parameter-transfer studies discussed in Section 3.2.1 has demonstrated that cross-turbine transfer is feasible, with decoder-only fine-tuning and CycleGAN-based domain mapping each offering complementary advantages under different data-availability constraints.

3. The third stream, represented by Pillay et al. (2025), has shown that operating-point-conditioned standardisation can substantially reduce the complexity of the feature distributions that an NBM must model, within a single-fleet, non-transfer setting.

Critically, to the best of our knowledge, no prior work has jointly integrated all three streams. Existing cross-turbine transfer studies, including Roelofs et al. (2024) and Jonas and Meyer (2025), apply conventional global standardisation and do not consider operating-point-conditioned preprocessing as a domain alignment mechanism. Conversely, Pillay et al. (2025) evaluated their standardisation procedure exclusively in a single-fleet, within-farm model-simplification context and did not examine its effects on cross-turbine transferability, nor test a turbine-level holdout protocol. The ARAE architecture of Bui et al. (2026) has not been evaluated in any transfer learning setting. Bui et al.'s own study was confined to a pooled multi-turbine evaluation protocol using stratified split by label, not turbine-level, data splits. Furthermore, as established in Section 3.4, the Mixture-of-Experts extension has not been considered in any WT NBM context previously identified in this review, despite the theoretical motivation that operating-point-conditional expert routing may capture the multimodal dynamics of the turbine power curve more effectively than a monolithic decoder, which is a motivation this dissertation tests empirically and finds, via matched-pairs statistical testing, not to be supported in this setting (Chapter 5).

This dissertation addresses these gaps through five contributions:

1. It adapts CBS as a cross-turbine domain alignment mechanism, a novel use not considered by Pillay et al. (2025) and, unlike the equal-width binning scheme of that work, uses equal-frequency binning to ensure balanced statistical support across turbines with differently-shaped power distributions.

2. It extends the decoder-only fine-tuning strategy of Roelofs et al. (2024) from a fully connected AE to the 1D-CNN ARAE of Bui et al. (2026), providing the first transfer learning evaluation of this specific architecture.

3. It evaluates both zero-shot generalisation and decoder-only fine-tuning across a fleet of nine held-out target turbines, providing a statistically grounded characterisation of cross-turbine generalisation capacity at a fleet scale not attempted by any of the studies reviewed above.

4. It introduces and rigorously evaluates a MoE extension of the ARAE decoder, in which a power-bin-conditioned router selects between two specialised decoder heads.

5. It assesses each contribution via matched-pairs (Wilcoxon signed-rank) statistical tests across the fifteen experiment configurations of the ablation study (Appendix A), a level of statistical rigour not observed in any of the cross-turbine transfer-learning studies reviewed in Section 3.2, including Roelofs et al. (2024) and Jonas and Meyer (2025) themselves, both of which report point-estimate performance differences without significance testing.

Together, these contributions constitute the first systematic investigation of how CBS preprocessing, decoder-only fine-tuning, and MoE specialisation interact within the ARAE framework to enable or constrain cross-turbine generalisation in a real-world wind fleet.

4

Methodology

This chapter presents the methodology used to explore the cross-turbine generalization capacity of the Anomaly Reinforced Autoencoder (ARAE) and builds directly on the architecture and training procedure as proposed by Bui et al. (2026). It distinguishes explicitly between the components that are inherited and adapted from the original ARAE design, and those that are the novel contributions of this dissertation. To ensure experimental reproducibility, all procedures involving stochastic components like data splitting, model initialization, and training were executed using a fixed random seed. Unless stated otherwise, the seed value was set to 42 throughout this study.

4.1 Dataset

The proprietary dataset used in this research was provided by EnergyPro[1], an Irish industrial partner aiming to bridge academic research and industrial applications. The dataset consists of SCADA telemetry from a fleet of ten operational wind turbines operated by EnergyPro and preprocessed using the pipeline described by Bui et al. (2026). The missing values were recovered from time-aware linear interpolation, and the data was resampled to hourly resolution, with segmentation into overlapping windows of 128 consecutive hourly samples. A window was labelled abnormal if it contained fewer than 127 time steps labelled as normal, reusing the tolerance threshold used by Bui et al. (2026) to account for potential labeling noise.

Six operational and environmental features were retained as per Bui et al. feature selection rationale (average power output (`avgpower`), average rotor speed (`avgrotorspeed`),

[1] `https://www.energypro.ie/` *Retrieved August 1, 2026*

average wind speed (`avgwindspeed`), air density (`density`), ambient temperature (`ambienttemperature`), and average wind direction (`avgwinddirection`)), with average wind direction split into its sine and cosine components to preserve angular continuity, totalling to seven input channels. Feature scaling uses Min-Max normalization fit on typical operating conditions, with the operational ranges for each feature following the values as proposed by Bui et al. and shown in Table 4.1.

Table 4.1: Operational ranges for selected features. (Bui et al., 2026)

Feature	Code Name	Operational Range
Average Power Output	`avgpower`	$0\,\mathrm{kW}$ to $2050\,\mathrm{kW}$
Average Rotor Speed	`avgrotorspeed`	$0\,\mathrm{RPM}$ to $18\,\mathrm{RPM}$
Average Wind Speed	`avgwindspeed`	$3\,\mathrm{m\,s^{-1}}$ to $15\,\mathrm{m\,s^{-1}}$
Air Density	`density`	$1.0\,\mathrm{kg\,m^{-3}}$ to $1.3\,\mathrm{kg\,m^{-3}}$
Ambient Temperature	`ambienttemperature`	$-5\,^{\circ}\mathrm{C}$ to $30\,^{\circ}\mathrm{C}$
Average Wind Direction	`avgwinddirection`	0° to 360°

Of all the ten turbines, `T_09` was selected as the training turbine, while the remaining nine turbines were used exclusively to evaluate cross-turbine generalization, either directly or after lightweight fine-tuning, as described in below Section 4.4.2. This is a deliberate departure from Bui et al. (2026), in which the train/validation/test split was stratified by class within a pooled multi-turbine data rather than held out at the turbine level and was done to test the cross-turbine transfer rather than within-distribution generalization.

The class ratio varies substantially across the turbines, from near total normal samples in `T_04` to near majority abnormal samples in `T_08` and `T_10`

Table 4.2: Class Distribution (normal vs abnormal) per turbine

Turbine Id	Normal	Abnormal	Ratio (Normal:Abnormal)
T_01	972	196	4.96
T_02	941	267	3.53
T_03	847	23	36.83
T_04	844	1	844
T_05	580	428	1.35
T_06	688	347	1.98
T_07	1065	24	44.38
T_08	107	885	0.12
T_09	1501	707	2.12
T_10	231	1435	0.16

T_09 was chosen specifically, rather than any other turbine in the fleet, because it has the required normal-to-abnormal (closer to 2:1) class ratio of the ten turbines (2.12:1, Table 4.2) and the highest number of normal samples. This matters directly for the ARAE's reinforcement mechanism (Section 4.5): both the normal-window and abnormal-window loss terms are computed from the training turbine's own data, so a training turbine with too few abnormal windows would leave the reinforcement term with little signal to learn from. T_06, the next most balanced turbine (1.98:1), was used as a direct empirical comparison to validate this choice, the result of that comparison is reported in Section 5.5, and T_09 is used as the training turbine for every other experiment in this dissertation.

4.2 Preprocessing: Condition-Binned Standardisation (CBS)

The base pipeline's Min-Max scaling as explained in Section 4.1 is fit on operational ranges as observed across the fleet, and therefore does not account for the fact that different turbines have different typical operating conditions. A model trained on one turbine's operational conditions and evaluated on another is consequently exposed to a covariate shift that has nothing to do with genuine abnormal behaviour.

To address this shift, this dissertation adapts CBS, a preprocessing method proposed by Pillay et al. (2025) for Gaussian Mixture Model (GMM) based wind turbine anomaly detection. Pillay et al. (2025) demonstrated that binning SCADA features by active power output and standardising within each bin using local statistics rather than global mean and standard deviation, allows a GMM to achieve effective anomaly detection with reduced model complexity, since it no longer needs to represent the power-dependent structure that dominates the raw feature distributions.

This dissertation applies the same core idea to the ARAE model, instead of GMM and extends it to remove the inter-turbine operating-regime shift instead of the original purpose of model complexity reduction as proposed by Pillay et al. (2025). This distinction is important, since CBS as a normalization process is Pillay et al. (2025)'s contribution, and it's application to the ARAE model is a novel contribution of this dissertation.

Pillay et al. bin by *equal-width* intervals, whereas this dissertation's own implementation uses ten *equal-frequency* percentile-based bins. This is a genuine adaptation, not an oversight since equal-frequency binning guarantees comparable statistical support in every bin regardless of how a given turbine's power distribution is shaped, which is a more directly relevant property in a cross-turbine setting where different turbines power distributions are not assumed to share the same shape, whereas Pillay et al.'s equal-width scheme is well suited to their single-fleet, model-simplification objective but was not designed with cross-turbine distributional heterogeneity in mind.

CBS is called once during fitting, on the training turbine's normal windows only, and each window is assigned to one of 10 bins by it's mean `avgpower`, using equal frequency (percentile) bin edges as shown in Figure 4.1. A small epsilon is added and subtracted at the two extreme edges respectively, to guarantee that the boundary values fall inside either first or last bin. This means bins are equal in window count and not in power range. Within each bin, the mean and standard deviation of every feature is computed from the training turbine's windows and stored as bin's statistics.

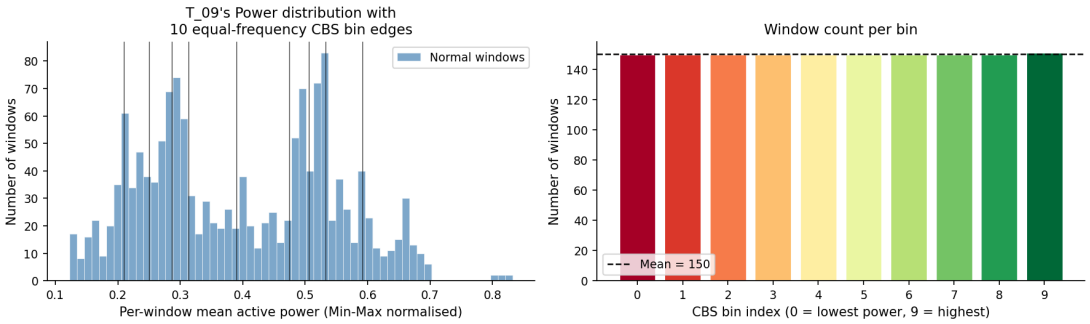

Figure 4.1: Equal-Frequency Power Binning (T_09 normal training windows)

This single fitted set of bin edges and per-bin statistics is then reused to transform every split and every turbine's windows, which includes the nine held out turbines for test and fine-tuning sets. No turbine specific refitting occurs anywhere in the pipeline. This was done to avoid data leakage and to ensure that the model is evaluated on unseen turbines without any prior knowledge of their operating conditions. In simplest terms, CBS's actual effect can be described as mapping every turbine's data into T_09's power regime as shown in Figure 4.2. A target turbine whose power distribution resembles T_09 will have its windows land roughly even across the bins, otherwise they will land unevenly, concentrated in whichever bin happens to overlap with the turbine's operating range. At transform time, every window is re-standardised using the statistics of the power bin it falls into. Following Pillay et al. (2025), the model is trained on all standardised features except avgpower, since it is the conditioning variable.

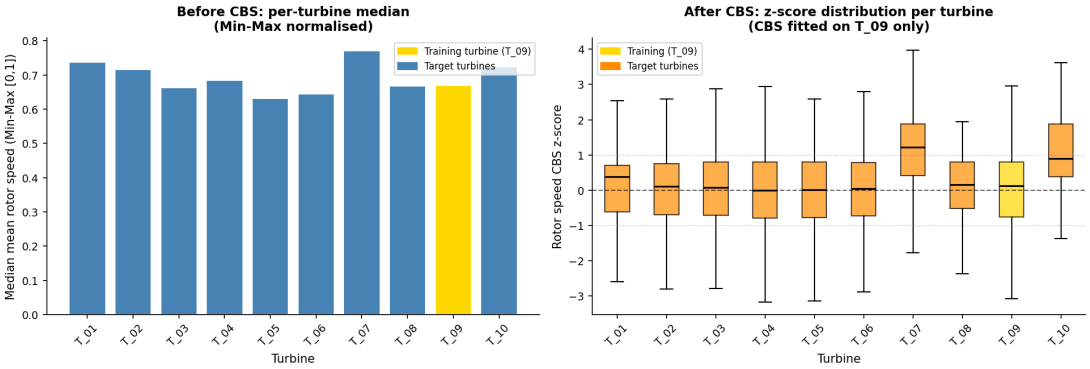

Figure 4.2: CBS cross-turbine generalisation on avgrotorspeed feature fitted on turbine T_09 only. Left: Per turbine spread before CBS; Right: CBS normalised distribution across the fleet.

It is important to note that this design choice sits in tension with Bui et al. (2026)'s justification for preferring min-max over z-score standardisation, stating that the relative spacing of signals may carry diagnostic significance. CBS reintroduces z-scoring but locally within a single power bin instead of globally over the entire dataset. The two design choices address different problems, while Bui concerns with pooled turbine diagnostics, this dissertation concerns with inter-turbine transferability.

4.3 Model Architectures

The base ARAE architecture is unchanged from Bui et al. (2026) as shown in Figure 4.3, and consists of a three-stage, one dimensional convolutional encoder-decoder pair. Each encoder stage applies an inverted-bottleneck transformation using a wide feature-extraction convolution of kernel size 3, followed by a ReLU activation and an average pooling downsampling step of factor 2, then a linear channel-compression convolution back to the stage's target width, followed by a channel-expanding convolution applied at each decoder stage. The resulting representation is then passed through a `Tanh` activation to produce the latent code, a 4 channel by 16 timesteps tensor.

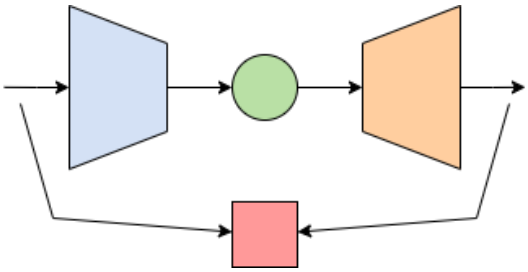

Figure 4.3: ARAE Architecture (Bui et al. (2026))

4.3.1 Mixture of Experts (MoE) Extension

To test whether conditional computation could further improve turbine-specific specialization, this dissertation extends the base ARAE architecture with a MoE mechanism, drawing on the general principle of input-conditioned expert routing introduced by Shazeer et al. (2017). Two structural changes are made to the base ARAE model:

1. Power is excluded from the encoder input because it is used to condition CBS binning and expert selection.

2. Two independent expert decoder heads replace the single decoder, both sharing the same encoder and latent representation but are structurally independent from the decompression stage onwards. Each expert reconstructs the full six channel, 128 timestep window from the shared latent space.

A `PowerBinRouter` module determines how the two experts' reconstruction outputs are combined. The router maintains a small learnable embedding table in which a window's raw `avgpower` mean is bucketed against the same bin edges used by CBS. The corresponding embedding row is looked up, and a softmax over that row produces a pair of gate weights. The final reconstruction is the gate weighted sum of the outputs from both the experts.

4.3.2 Fixed-Weight Ensemble Variant

A second MoE variant as shown in Figure 4.4 was developed after the router based design was found to converge towards nearly same gate weights, regardless of the tuning as explained in Section 5.6. It replaces the power router's output with a fixed weighting of 0.5 for each expert, effectively creating a fixed weight dual-decoder ensemble. The `PowerBinRouter` module remains present in the model definition but is bypassed in the forward pass and receives no gradient.

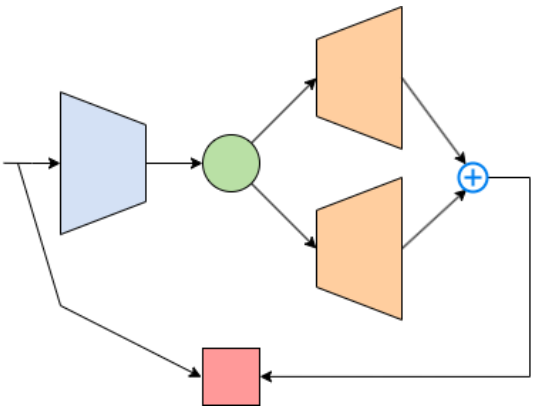

Figure 4.4: Fixed Weight Ensemble ARAE Architecture.

4.4 Training Procedure

4.4.1 Base Training

The encoder and decoder are trained on `T_09` only, using the Adam optimizer with an initital learning rate of 0.0001, decayed by a factor of 0.1 on validation-loss plateau, with a batch size of 96 and early stopping with a patience of 20 epochs on validation loss. Both normal and abnormal windows from the training turbine contribute to the base ARAE training loss as described in Section 4.5.

4.4.2 Decoder Fine-Tuning

The fine-tuning strategy used to calibrate the shared representation of each target turbine is adapted from Roelofs et al. (2024), who conducted a closely related cross-turbine transfer learning study for autoencoder based wind turbine anomaly detection and directly compared three fine-tuning strategies:

1. Fine-tuning the full autoencoder.
2. Fine-tuning only the decoder.
3. Fine-tuning only the detection threshold.

They reported that the threshold-only tuning underperformed the other two strategies, and the decoder-only one was the best out of the three. They also reported that full-autoencoder fine-tuning on a small target turbine sample carried a greater risk of overfitting and negative transfer despite their use of lower learning rate. This dissertation adopts their decoder-only variant, for two reasons:

1. Freezing the encoder addresses the dissertation's central research question, whether one shared cross turbine representation can generalize well across the entire fleet.
2. Keeping the encoder frozen avoids a possible confound between decoder calibration and the CBS normalization, since its bin edges are fit only on `T_09`, an unfrozen encoder could begin compensating for turbine specific CBS artifacts rather than adapting to its normal patterns.

Roelofs et al. (2024)'s autoencoder is a small, fully connected feed forward network with PReLU activations, trained on several hundred raw SCADA features, while the architecture used throughout this dissertation is the ARAE as proposed by Bui et al.

(2026) and explained in section 4.3. Adopting their fine-tuning strategy also functions as a test of whether their finding is sufficient to capture most of the cross-turbine variation.

A lightweight fine-tuning step is applied independently for every one of the nine held out test turbines, using only that turbine's own normal windows. The encoder and for MoE variants the router is frozen, with only the decoder parameters updated during fine-tuning. Since the latent bottleneck's only non-linearity is a `Tanh` activation, that carries no learnable parameters, freezing the encoder also freezes the latent representation. Fine-tuning, therefore, only adapts to how the shared turbine-agnostic representation is decoded back into reconstruction, not what that representation is.

The implemented fine-tuning uses a learning rate of 1×10^{-4}, lower than Roelofs et al. (2024)'s decoder-only rate of 1×10^{-3}. This conservative choice is motivated by the substantially smaller fine-tuning sets used here (N = 30 or 60 windows, versus the one-to-three months of target data used by Roelofs et al.), where a higher learning rate risked overfitting. A small sample of normal windows is drawn from each of the target turbines, and the decoder is trained via plain MSE against these windows only for a total of maximum 500 epochs with early stopping enabled. Two hyper parameters of this step were swept as part of the ablation study reported in Table 5.1, namely, fine-tune samples of 30 and 60, and early stopping patience of 30 and 60.

4.5 Loss Functions

The Anomaly Reinforced Autoencoder's (ARAE) core training objective is inherited and unchanged from Bui et al. (2026) and is defined in Equation 4.1,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{normal}} + \frac{1}{1 + \mathcal{L}_{\text{abnormal}}} \tag{4.1}$$

where $\mathcal{L}_{\text{normal}}$ and $\mathcal{L}_{\text{abnormal}}$ are computed identically but applied to disjoint subsets of a batch. Since minimizing $\frac{1}{1+\mathcal{L}_{\text{abnormal}}}$ is equivalent to maximizing $\mathcal{L}_{\text{abnormal}}$, this term actively reinforces poor reconstruction on windows already known to be abnormal. This is the defining mechanism of the ARAE model and is preserved without any modifications.

Both the normal-window loss and pre-inversion abnormal-window loss are a weighted composite of three terms, pointwise mean squared error, maximum mean discrepancy (MMD), and a mean-difference approximation to the Wasserstein distance as described below, unchanged from Bui et al. (2026)'s original design:

- **MSE**: Primary reconstruction loss, measuring the average squared difference between input and reconstruction.
- **MMD**: A kernel-based statistical measure used to quantify the difference between two probability distributions by comparing their mean embeddings in a reproducing kernel Hilbert space.
- **Wasserstein Distance**: Measures the minimum cost of transporting mass to transform one probability distribution into another, considering both the amount of mass and the distance it must be moved.

All three are combined using the uncertainty-based automatic weighting scheme of Yao et al. (2024), reused from Bui et al. (2026).

For the routed MoE variant, an additional entropy-based regularization term is applied to discourage the router from collapsing all windows into a single expert:

$$\mathcal{L}_{\text{balance}} = \sum_{k} \bar{p}_k \cdot \log(\bar{p}_k) \tag{4.2}$$

where $\bar{p}_k$ is the mean gate weight assigned to expert k across a batch. Since this quantity is the negative of the Shannon entropy of the mean gate distribution, adding $\lambda \cdot \mathcal{L}_{\text{balance}}$ to the total loss (with $\lambda \geq 0$) rewards configurations in which gate usage is spread more evenly across experts on average, discouraging the router from ignoring one expert entirely. Three values of λ $(0,\ 0.01,\ 0.05)$ were evaluated as part of the ablation study as reported in Table 5.1.

4.6 Evaluation

4.6.1 Anomaly Scoring

For every evaluated window, a per-channel reconstruction error vector is computed as the mean squared error between the output and the target. This vector is normalized per turbine using that turbine's own mean and standard deviation of error vectors, computed on its held-out evaluation windows before being reduced to a scalar anomaly score by taking the average across the channels. This per-turbine normalization is necessary because the raw scale of reconstruction error is not directly comparable across the turbines with different operating conditions. Without it, a fixed global flag would over or under flag a turbine whose baseline reconstruction error happens to sit at a different scale.

4.6.2 Thresholding and Metrics

A decision threshold is selected per turbine by maximizing Youden's J statistic over the turbine's validation windows (Youden, 1950), computed directly from the ROC curve (False Positive Rate (FPR) and True Positive Rate (TPR) at every candidate threshold). Youden's J implicitly treats a unit of recall gain as equally valuable to a unit of FPR increase. This is a simplifying choice rather than a domain-validated one: Roelofs et al. (2024), whose fine-tuning strategy this dissertation adopts, instead weights precision over recall (via an $F_{1/2}$-score) on the grounds that false alarms are operationally costlier than missed detections for wind farm operators. Youden's J is retained here as a standard, threshold-selection-agnostic default in the absence of a quantified cost model for this specific deployment context, and Section 5.10 shows explicitly where this assumption produces a degenerate result. This differs from Bui et al. (2026)'s threshold-selection method, which instead maximizes the F1-score directly. At this threshold, standard classification metrics are computed like precision, recall, accuracy, F1-score, and AUC.

The core metrics are derived from the counts of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which form the confusion matrix.

- **Precision** (Positive Predictive Value) measures how true positive predictions are. A high precision indicates that when the model flags an instance as positive, it is highly likely to be correct. This is crucial for minimizing false alarms, which can be costly or lead to alert fatigue in operational settings.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.3}$$

- **Recall** (Sensitivity or True Positive Rate) measures the model's ability to find all the anomalous samples in the dataset. A high recall indicates that the model misses very few true anomalies.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.4}$$

Fleet level metric is reported as the mean, minimum and maximum of each metric across the nine held-out test turbines. However, evaluating models solely on aggregate fleet means can be misleading, as small differences in mean values may be driven by extreme performance on a single turbine rather than a systematic improvement. Throughout this evaluation, the positive class is defined as *abnormal*, consistent with the anomaly-detection objective. A true positive is therefore a correctly-flagged abnormal

window, and a false positive is a normal window incorrectly flagged as abnormal. To rigorously evaluate the performance differences between model configurations, an exhaustive pairwise statistical testing framework is implemented. For every possible pair of the fifteen experiments evaluated in this study, a two-sided Wilcoxon Signed-Rank test is conducted. This non-parametric paired test is selected because:

1. The samples are naturally paired, with each held-out test turbine serving as its own matched control across different experimental configurations.
2. It doesn't assume normality of the data.
3. It does not require assuming that equal-sized differences in a performance metric represent equal amounts of genuine improvement across turbines with very different operating characteristics. For example, a given AUC gain on an abnormal-majority turbine, where the detection task is structurally easier (Section 5.11), may not reflect the same underlying improvement as an equal-sized gain on a more balanced turbine.

For each comparison between Experiment A and Experiment B, the test is evaluated over the paired metric vectors $\mathbf{x}_A, \mathbf{x}_B \in \mathbb{R}^9$, where each element corresponds to a test turbine's performance. The null hypothesis (H_0) posits that the distribution of paired differences is symmetric about zero, indicating no systematic performance difference between the two configurations. These exhaustive pairwise p-value matrices for all classification metrics are directly utilized in Chapter 5 to guide architectural ablation decisions and the full tables as shown in Appendix A.

4.7 Experimental Environment

All models were implemented using `PyTorch` (`v2.6.0+cu124`) and trained on a laptop equipped with an `NVIDIA GeForce RTX 3060` GPU and an `AMD Ryzen 9 5900HS` CPU. Training and evaluation were conducted under `Python` (`v3.13.9`) on a `Windows` system. Core dependencies included `NumPy` (`v2.3.4`), `Pandas` (`v2.3.3`), and `scikit-learn` (`v1.7.2`).

5

Results and Discussions

This chapter presents the results of fifteen experiment ablation study isolating the individual and combined contributions of CBS, Fine-Tuning (FT), and the MoE extension to cross-turbine generalization capacity of the ARAE model. Table 5.1 summarizes every configuration evaluated and subsequent sections walk through the findings in the order, the decisions were made.

Table 5.1: Summary of experimental configurations and fleet-level AUC performance

Exp	FT	FT-Epochs	FT-Records	CBS	MoE	λ	Mean AUC ± SD
1	-	-	-	-	-	-	0.4979 ± 0.2133
2	-	-	-	✓	-	-	0.4713 ± 0.1586
3	✓	500	60	-	-	-	0.6577 ± 0.1565
4	✓	500	60	✓	-	-	0.7354 ± 0.2258
5	✓	500	60	✓	-	-	0.5640 ± 0.2587
6	-	-	-	✓	✓	-	0.4932 ± 0.1807
7	-	-	-	✓	✓	-	0.5153 ± 0.1313
8	✓	500	30	✓	✓	-	0.6686 ± 0.2298
9	✓	500	60	✓	✓	-	0.7207 ± 0.1866
10	✓	500	60	✓	✓	-	0.6977 ± 0.2065
11	✓	500	60	✓	✓	0.05	0.7521 ± 0.2120
12	✓	500	60	✓	✓	0.01	0.6761 ± 0.2401
13	✓	500	30	✓	✓	0.05	0.7176 ± 0.2300
14	✓	500	60	✓	✓	0.00	0.7515 ± 0.2035
15	✓	500	30	✓	✓	0.00	0.7266 ± 0.2373

5.1 Baseline Cross-Turbine Generalization of the Unmodified ARAE

The raw ARAE trained on T_09, without any additional techniques (Experiment 1) and evaluated directly on the nine held out turbines with no adaption of any kind, achieves a fleet mean AUC of 0.4979, with a range of $0.1562 - 0.7321$, indistinguishable from random guessing. Despite achieving a strong within-turbine performance in the original ARAE, the model's learned representations do not generalize to unseen turbines, highlighting the need for additional techniques to improve cross-turbine generalization. This experiment is treated as baseline for the rest of the experiments, and all subsequent results are compared to it.

Table 5.2: Per-turbine metrics for the baseline experiment (Experiment 1)

Turbine	AUC	Precision	Recall	TP	FP	TN	FN
T_01	0.7321	0.276	0.76	149	391	581	47
T_02	0.7142	0.388	0.648	173	273	668	94
T_03	0.3254	0.036	1	23	608	239	0
T_04	0.5723	0.003	1	1	361	483	0
T_05	0.5653	0.891	0.266	114	14	566	314
T_06	0.2957	0.367	0.988	343	592	96	4
T_07	0.7291	0.065	1	24	346	719	0
T_08	0.1562	1	0.028	25	0	107	860
T_10	0.3905	0.876	0.523	750	106	125	685

5.2 Effect of CBS in Isolation

Applying CBS to the raw ARAE with no fine-tuning (Experiment 2) decreases the fleet average from 0.4979 to 0.4713 as can be seen in Table 5.1. This is a useful negative result, and the mechanism behind it follows directly from Section 4.2 as CBS's bin edges and per-bin statistics are fit on T_09's power regime and then reused for every target turbine. For a turbine whose power distribution differs from T_09, it concentrates its windows into a subset of the ten bins instead of distributing them evenly. CBS alone, therefore, only changes the statistical footing on which windows are presented to the

model, but does not teach the decoder anything about a target turbine's own dynamics. This result directly motivates the fine-tuning step introduced in Section 4.4.2.

5.3 Effect of Decoder Fine-tuning in Isolation

Applying decoder fine-tuning without CBS (Experiment 3: raw Min-Max scaled inputs, frozen encoder, decoder calibrated on 60 normal windows per target turbine) raises fleet mean AUC from 0.4979 to 0.6577, with a range of 0.4731 - 0.9056, as can be seen in Table 5.1. This is a significant improvement over the baseline, and demonstrates that even without CBS, the decoder can learn to adapt to the target turbine's dynamics through fine-tuning on a small number of normal windows. A small per-turbine calibration step recovers the majority of the generalization gap.

5.4 Combined Effect of CBS and Decoder Fine-tuning

Combining CBS with decoder fine-tuning (Experiment 4) raised fleet mean AUC to 0.7354, an improvement of 0.0775 over decoder fine-tuning alone, with a range of 0.2561 - 0.9809, as can be seen in Table 5.1. This shows that CBS and decoder fine-tuning are complementary rather than redundant, since CBS removes the between-turbine operating regime before the decoder is calibrated, giving the fine-tuning step a smaller gap to close on each target turbine.

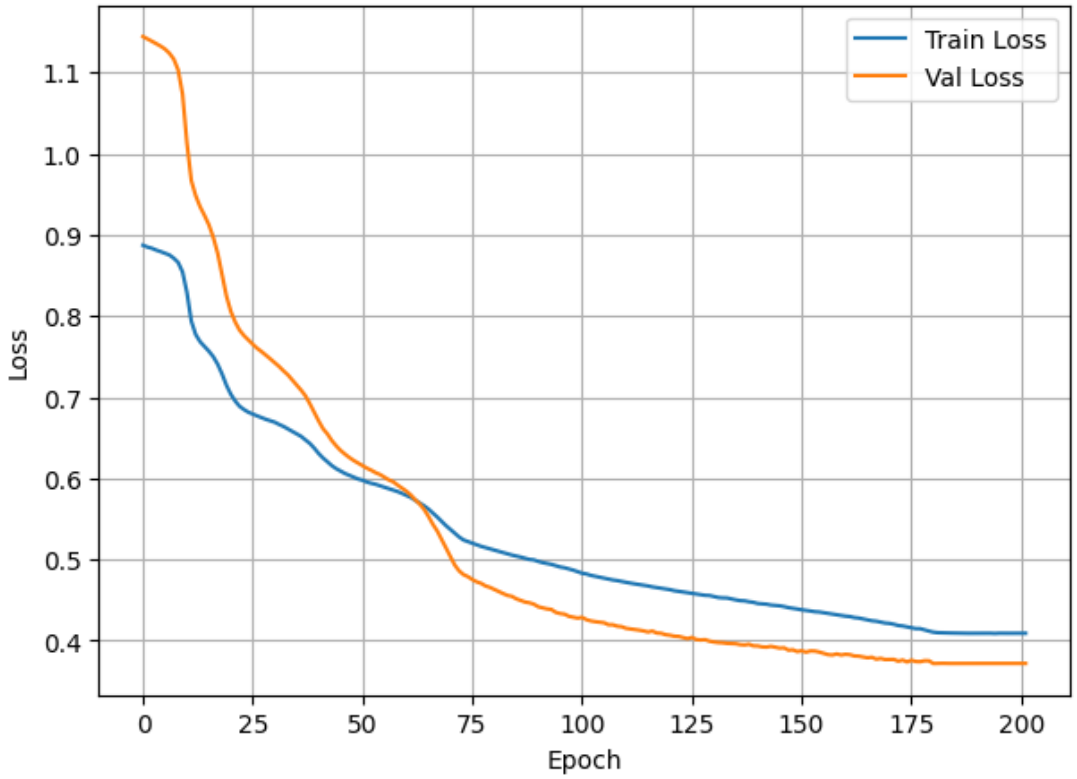

Figure 5.1: Training Curve of Experiment 4: CBS + Decoder Fine-tuning.

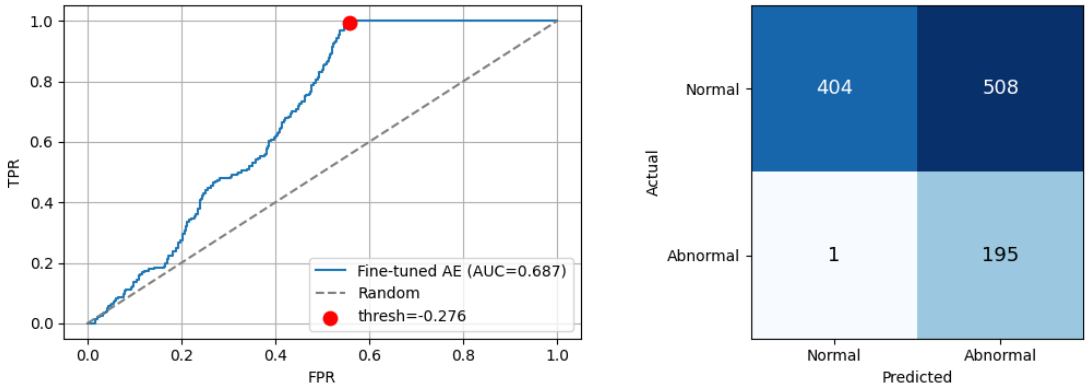

Figure 5.2: Left: ROC curve for Turbine 1; Right: Confusion Matrix for Turbine 1.

5. Results and Discussions

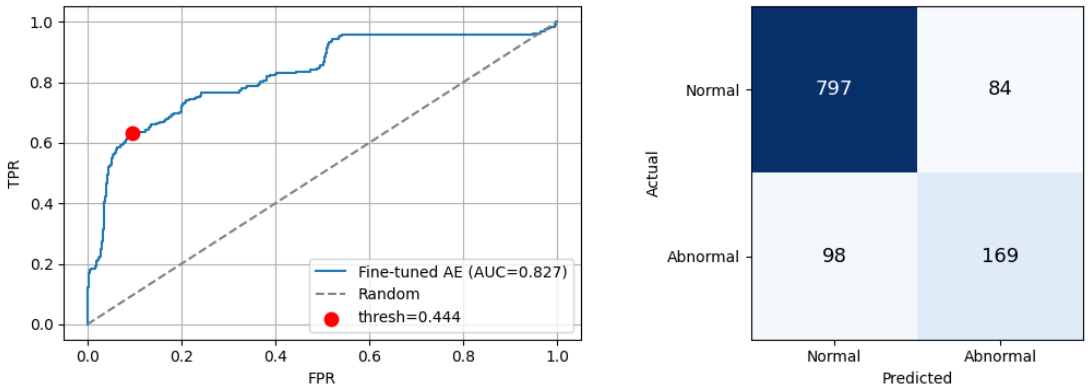

Figure 5.3: Left: ROC curve for Turbine 2; Right: Confusion Matrix for Turbine 2.

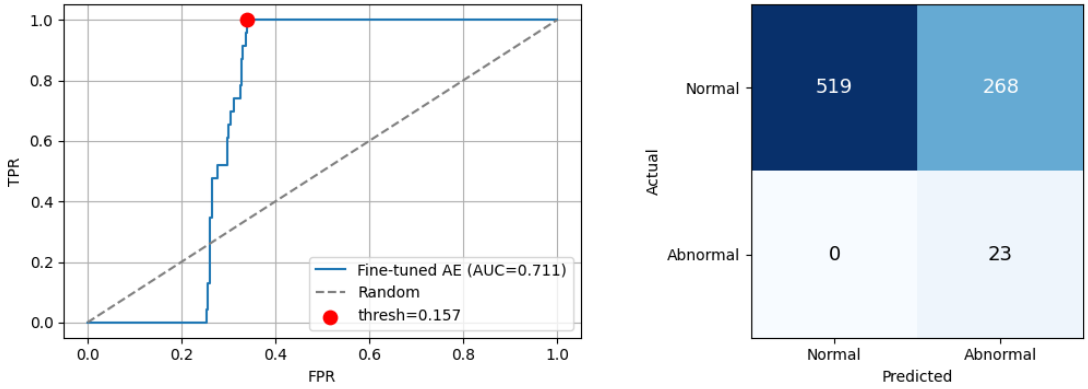

Figure 5.4: Left: ROC curve for Turbine 3; Right: Confusion Matrix for Turbine 3.

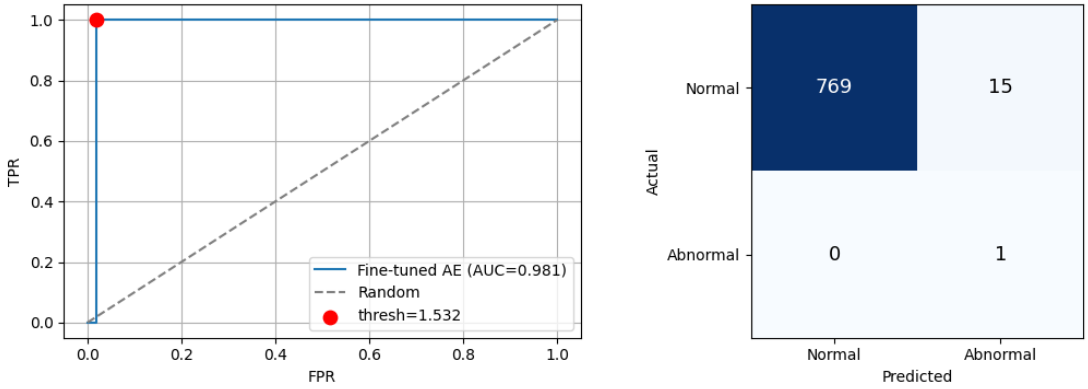

Figure 5.5: Left: ROC curve for Turbine 4; Right: Confusion Matrix for Turbine 4.

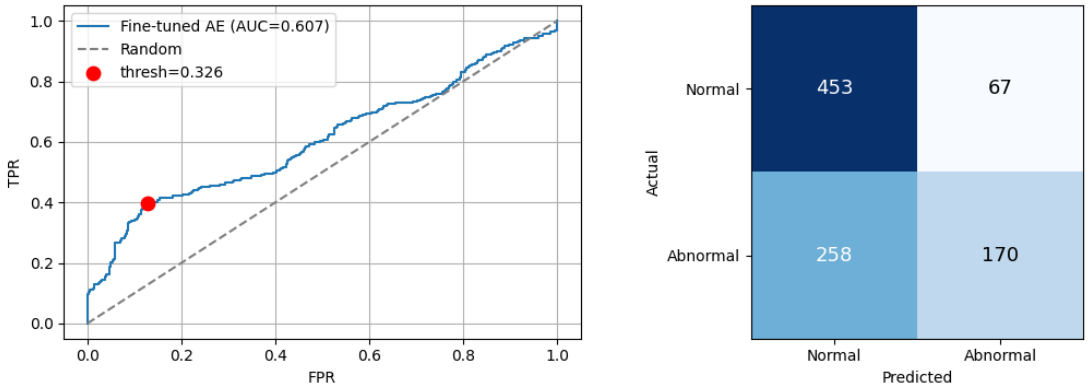

Figure 5.6: Left: ROC curve for Turbine 5; Right: Confusion Matrix for Turbine 5.

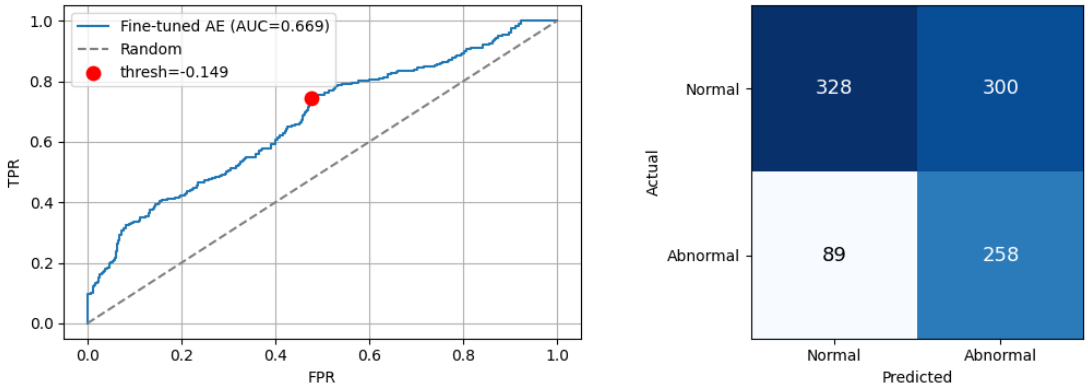

Figure 5.7: Left: ROC curve for Turbine 6; Right: Confusion Matrix for Turbine 6.

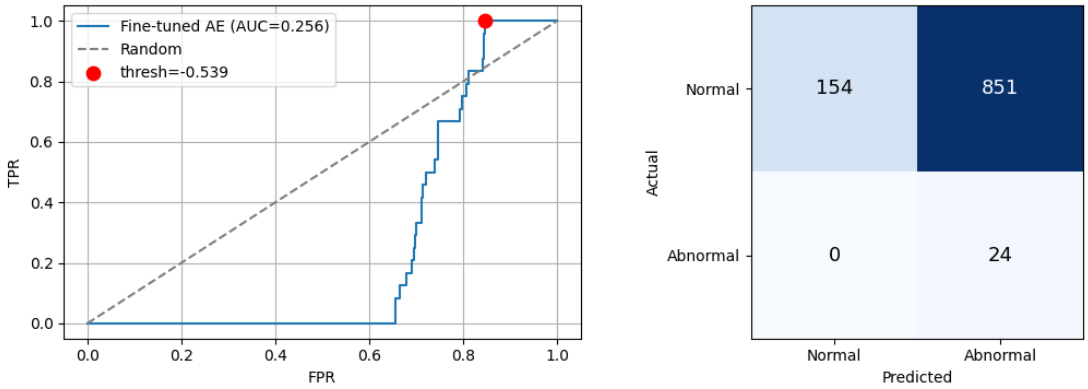

Figure 5.8: Left: ROC curve for Turbine 7; Right: Confusion Matrix for Turbine 7.

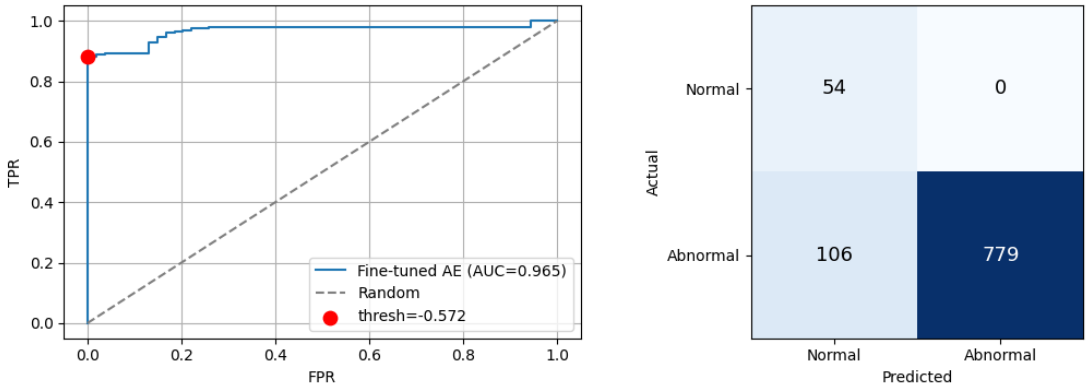

Figure 5.9: Left: ROC curve for Turbine 8; Right: Confusion Matrix for Turbine 8.

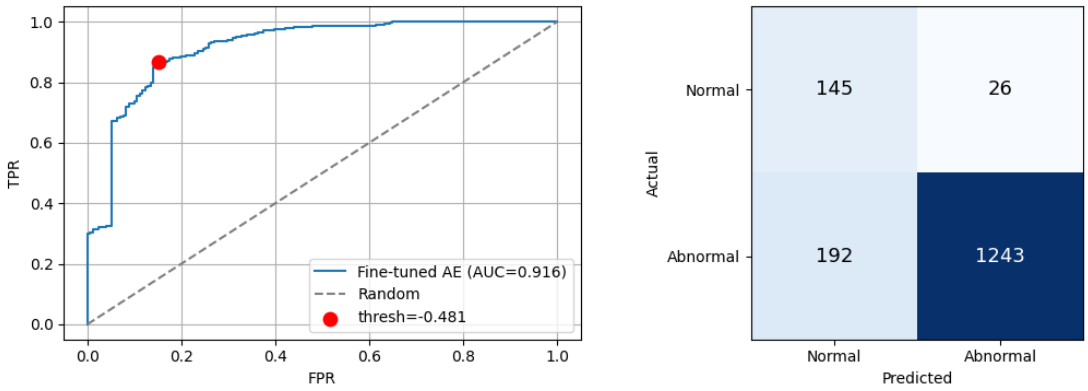

Figure 5.10: Left: ROC curve for Turbine 10; Right: Confusion Matrix for Turbine 10.

5.5 Choice of Training Turbine

Section 4.1 establishes T_09 as the training turbine on the grounds that it has the normal-to-abnormal class ratio in the fleet (2.12:1) closest to (2:1) and more samples than T_06. To validate this choice empirically rather than by ratio alone, Experiment 5 repeats Experiment 4's configuration (CBS plus decoder fine-tuning) exactly, but with T_06 as the training turbine, the fleet's second most suitable turbine (1.98:1) with the nine remaining turbines (now including T_09 itself) held out for evaluation.

Training on T_06 instead of T_09 drops fleet mean AUC from 0.7354 to 0.5640, with a range of 0.0303 - 0.8332, as shown in Table 5.1, a substantially larger and more variable drop than the difference in class ratio between the two turbines alone would

predict. This result supports T_09 as the training turbine for the remainder of this dissertation, and all further experiments use T_09 exclusively.

This finding is a useful limitation already identified in this dissertation. The choice of training turbine is not interchangeable within the fleet, and generalization capacity is sensitive to which turbine's power regime the shared encoder, and CBS's bin statistics, are fit to. A systematic comparison across all ten turbines as training-turbine candidates, rather than the two most class-balanced only, is identified as a direction for future work (Section 6.4).

5.6 Mixture of Experts Extension

The first attempt at introducing power-conditioned expert routing (Experiment 6), with both experts initialized identically and no load balance load regularization, produced router gate weights of exactly (0.51, 0.49) for every single target turbine regardless of the power bin. This means that the two experts receive identical gradients throughout training and never differentiate. The router exists but the model is a single decoder with it's output split in half. Fleet mean AUC was 0.4932, with a range of 0.3058 - 0.8636, as can be seen in Table 5.1, which is consistent with this interpretation.

Breaking the symmetry by initializing the two experts with different random noise (Experiment 7) does produce differentiated gate weights per turbine, ranging from 55/45 to 59/41. This resulted in a fleet mean AUC of 0.5153, with a range of 0.4929 - 0.6953, far below Experiment 4's 0.7354, as can be seen in Table 5.1.

Keeping the same configuration as Experiment 7, but enabling decoder fine-tuning on 30 normal windows per turbine (Experiment 8) raises fleet mean AUC to 0.6686, with a range of 0.3716 - 0.9727. Increasing the number of fine-tuning windows to 60 (Experiment 9) further raises the fleet mean AUC to 0.7207, with a range of 0.4622 - 0.9864, as can be seen in Table 5.1. This shows that just like Experiment 3, decoder fine-tuning is a critical component of cross-turbine generalization, and that the MoE extension is not sufficient to close the generalization gap on its own. For fine-tuning, in experiments 8 and 9, the epochs were set to 500, with a patience of 30, which simply means if the loss does not decrease for 30 epochs, the fine-tuning for that turbine is stopped. Changing it to 60 in Experiment 10, with all other parameters the same, results in a fleet mean AUC of 0.6977, with a range of 0.4182 - 0.9853, this minor drop shows that in some target turbines overfitting starts to occur after 30 epochs, and that the patience parameter is effective at preventing this. Figure 5.11 shows the per-turbine AUC for Experiments 6-10.

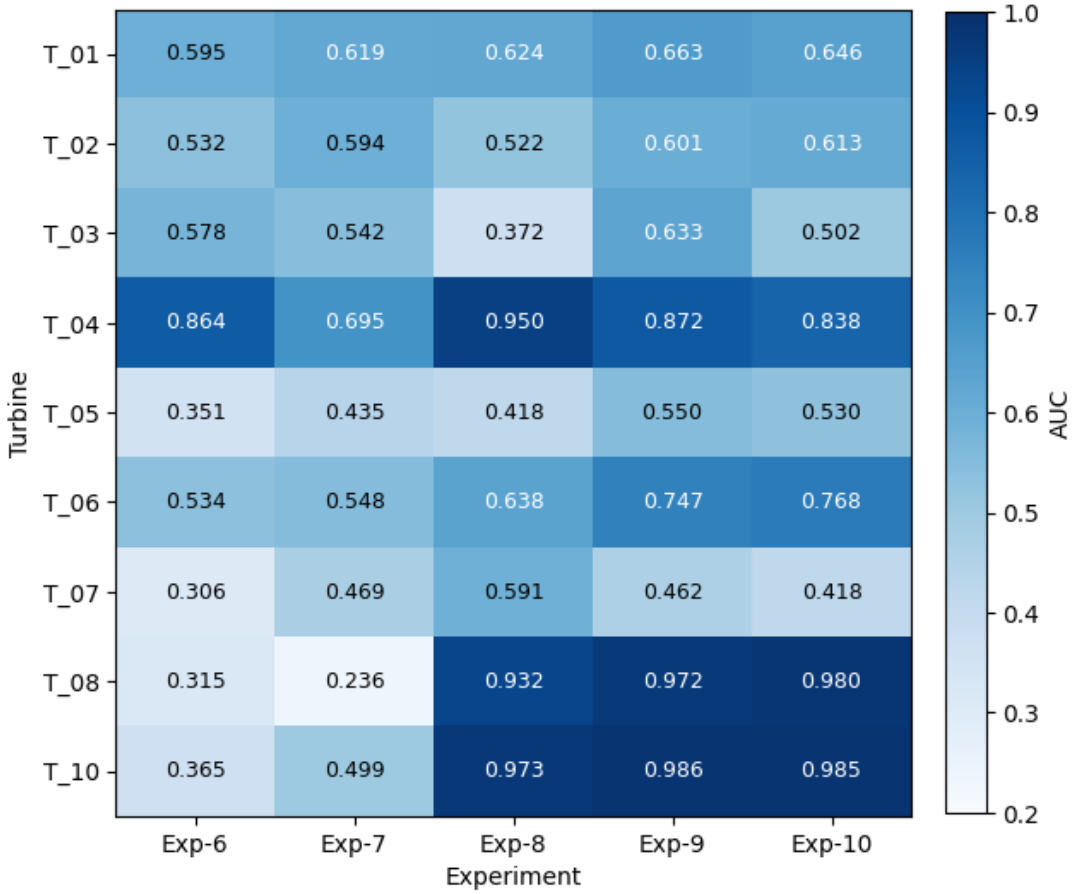

Figure 5.11: Heatmap of per-turbine AUC for Experiments 6-10.

5.7 Load Balance Regularization for MoE

Building on Experiment 9's configuration, Experiment 11 additionally enables the entropy-based load balance term with $\lambda = 0.05$, keeping fine-tune records to 60 and patience of 30 unchanged. This raises fleet mean AUC to 0.7521, with a range of 0.3758 - 0.9964, the best-performing router-based configuration in the study, as can be seen in Table 5.1. Reducing the load balance weight to $\lambda = 0.01$ while holding all other parameters fixed (Experiment 12) lowers fleet mean AUC to 0.6761, with a range of 0.3038 - 0.9549. This is a substantial drop for a five-fold reduction in λ, and shows that at $\lambda = 0.01$ the regularization pressure is too weak to produce the same effect achieved at $\lambda = 0.05$. Returning to $\lambda = 0.05$ but reducing the fine-tune records from 60 to 30 (Experiment 13) lowers fleet mean AUC to 0.7176, with a range of 0.4267 -

0.9982, as can be seen in Table 5.1. This mirrors the pattern already observed between Experiments 8 and 9, and confirms that a larger fine-tuning sample improves cross-turbine generalization regardless of whether load balance regularization is enabled. Table 5.3 shows the per-turbine AUC and gate weights for Experiments 11-13.

Table 5.3: Per-turbine AUC and expert gate weights for MoE experiments 11 - 13

Turbine	Exp-11		Exp-12		Exp-13	
	AUC	Gate (A, B)	AUC	Gate (A, B)	AUC	Gate (A, B)
T_01	0.6250	0.50, 0.50	0.4074	0.50, 0.50	0.5824	0.51, 0.49
T_02	0.7968	0.50, 0.50	0.6174	0.50, 0.50	0.7001	0.51, 0.49
T_03	0.7800	0.49, 0.51	0.6900	0.51, 0.49	0.4267	0.49, 0.51
T_04	0.9120	0.50, 0.50	0.9184	0.51, 0.49	0.9877	0.50, 0.50
T_05	0.5128	0.50, 0.50	0.5003	0.51, 0.49	0.5122	0.50, 0.50
T_06	0.7826	0.50, 0.50	0.7434	0.51, 0.49	0.7887	0.50, 0.50
T_07	0.3758	0.49, 0.51	0.3038	0.51, 0.49	0.4856	0.49, 0.51
T_08	0.9964	0.49, 0.51	0.9498	0.51, 0.49	0.9982	0.49, 0.51
T_10	0.9881	0.49, 0.51	0.9549	0.51, 0.49	0.9772	0.50, 0.50

An interesting finding here is that Experiment 11 scored the highest AUC but the gate weights across the target turbines were either (0.51, 0.49) or (0.50, 0.50) as seen in Table 5.3, which is the main motive behind running Experiments 14 and 15, to see if the router is necessary at all, or if a fixed gate of (0.5, 0.5) is sufficient to achieve the same performance.

5.8 Ensemble ARAE

In experiments 14 and 15 the router was discarded for a fixed ensemble between the two decoder heads. The results are denoted to have $\lambda = 0$ in tables and figures, in order to differentiate from the ones that have router. Both expert decoders continue to run on every window, with their outputs simply averaged. Holding fine-tune records at 60 and patience at 30, matching Experiment 11's fine-tuning configuration exactly (Experiment 14) achieves a fleet mean AUC of 0.7515, with a range of 0.4271 - 1.0000, as can be seen in Table 5.1. This is statistically indistinguishable from both Experiment

11's router-based result (0.7521) and Experiment 4's single-decoder result (0.7354), as established in Section 5.9.

Reducing the fine-tune records from 60 to 30, with all other parameters unchanged (Experiment 15), lowers fleet mean AUC to 0.7266, with a range of 0.3629 - 0.9872. This is the same direction and a broadly similar magnitude of effect as the corresponding comparisons already established between Experiments 8 and 9, and between Experiments 11 and 13, and constitutes a third independent confirmation, that a larger fine-tuning sample improves cross-turbine generalization regardless of the routing mechanism.

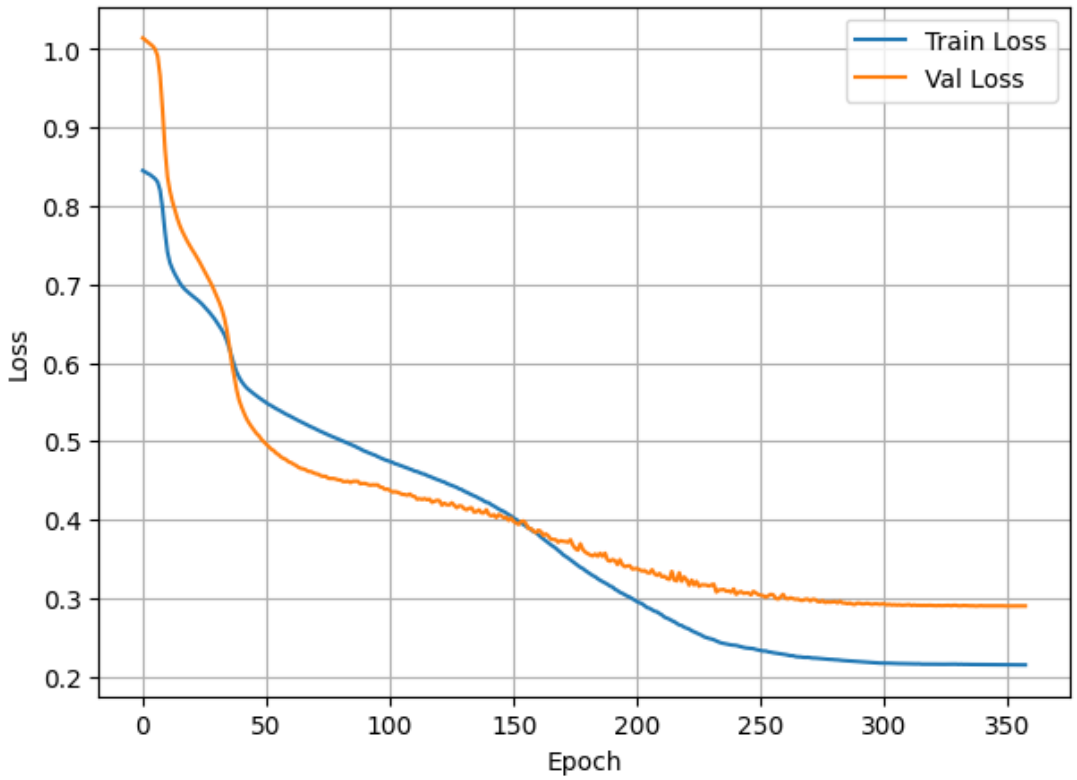

Figure 5.12: Training Curve of Experiment 14: ARAE Ensemble + Decoder Fine-tuning.

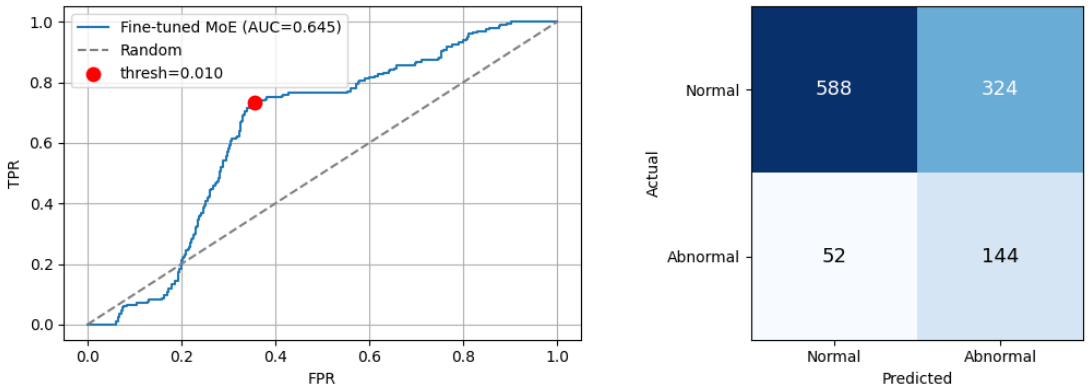

Figure 5.13: Left: ROC curve for Turbine 1; Right: Confusion Matrix for Turbine 1.

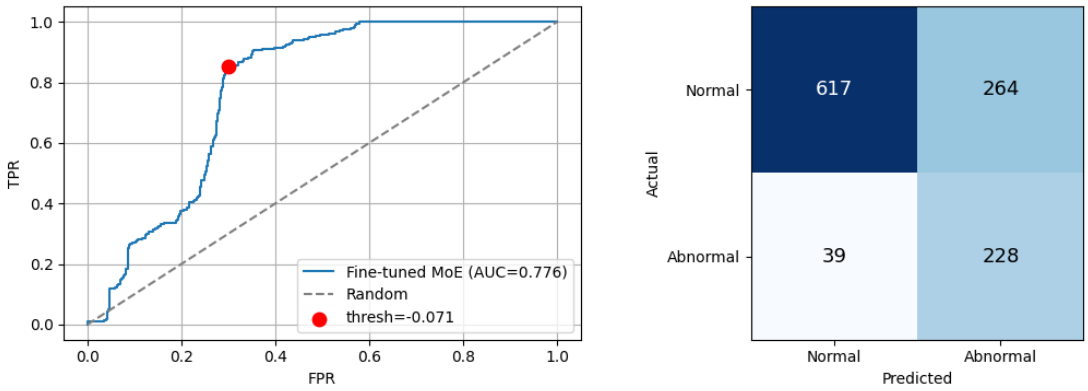

Figure 5.14: Left: ROC curve for Turbine 2; Right: Confusion Matrix for Turbine 2.

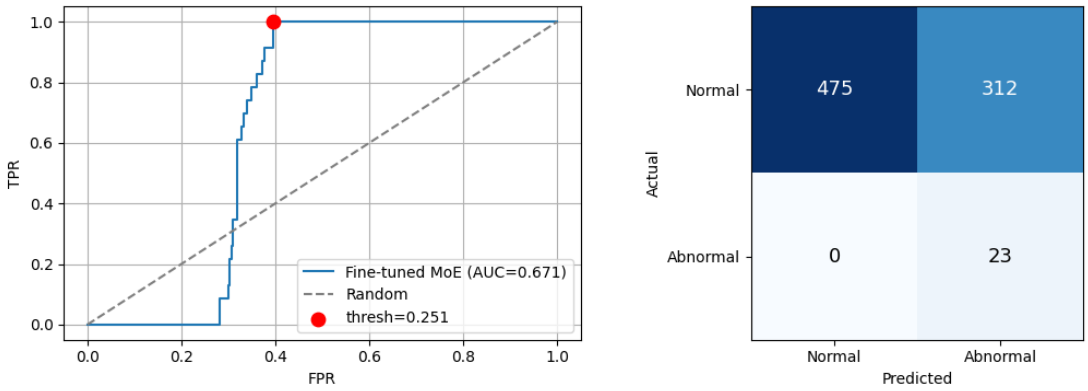

Figure 5.15: Left: ROC curve for Turbine 3; Right: Confusion Matrix for Turbine 3.

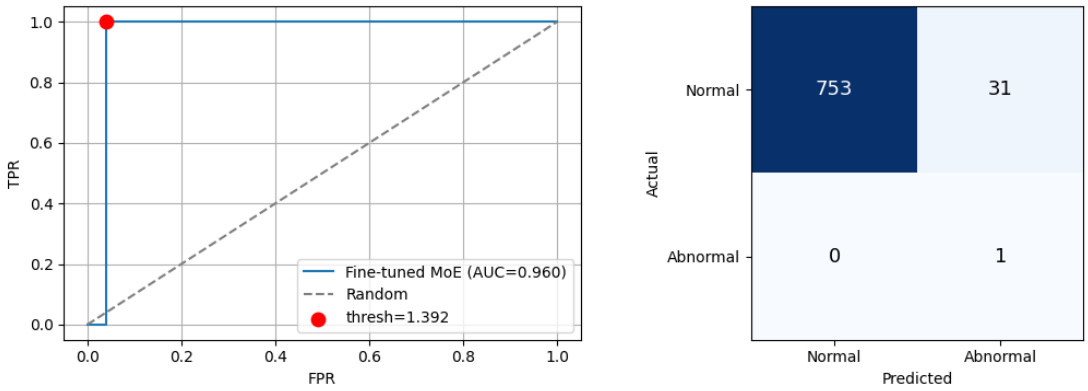

Figure 5.16: Left: ROC curve for Turbine 4; Right: Confusion Matrix for Turbine 4.

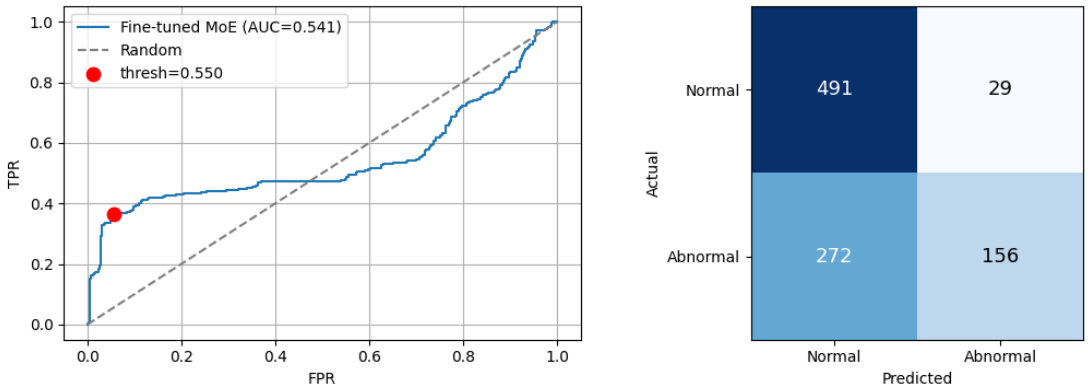

Figure 5.17: Left: ROC curve for Turbine 5; Right: Confusion Matrix for Turbine 5.

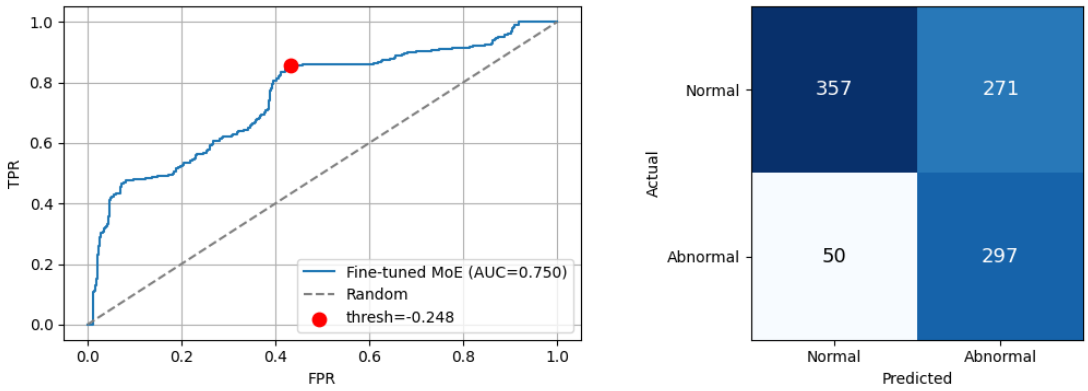

Figure 5.18: Left: ROC curve for Turbine 6; Right: Confusion Matrix for Turbine 6.

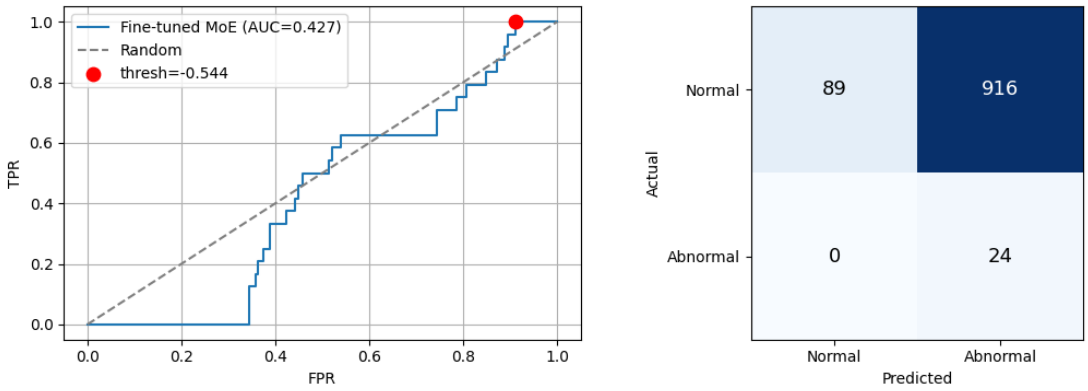

Figure 5.19: Left: ROC curve for Turbine 7; Right: Confusion Matrix for Turbine 7.

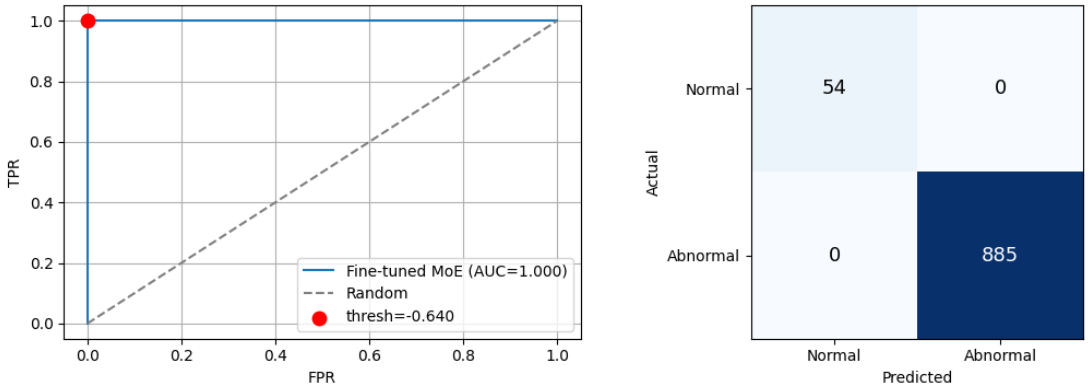

Figure 5.20: Left: ROC curve for Turbine 8; Right: Confusion Matrix for Turbine 8.

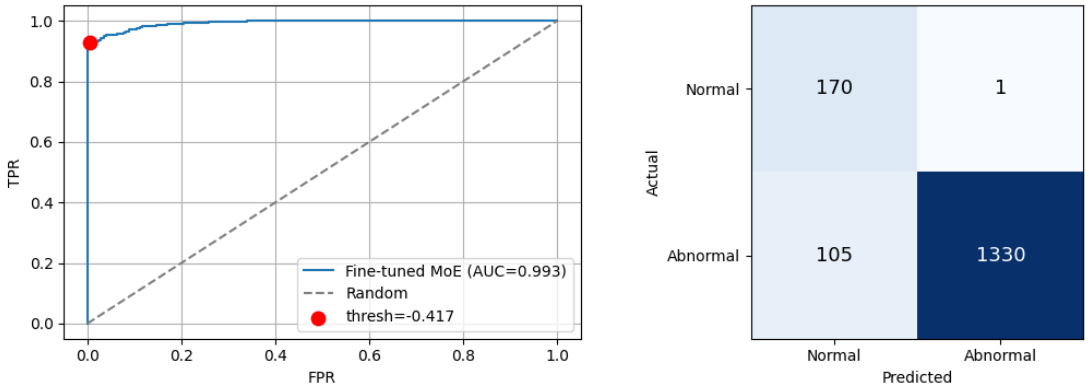

Figure 5.21: Left: ROC curve for Turbine 10; Right: Confusion Matrix for Turbine 10.

5.9 Is the Router or Ensemble Necessary?

Three configurations reach comparable fleet mean AUC: Experiment 4 (CBS + fine-tuning: 0.7354), Experiment 11 (the best-tuned router-based MoE, 0.7521), and Experiment 14 (the fixed-weight ensemble, 0.7515).

Table 5.4: Wilcoxon signed-rank test results for pairwise comparisons of Experiments 4, 11 and, 14.

Comparison	Mean AUC Difference	Wilcoxon p-Value
Exp-14 (Fixed Ensemble) vs Exp-4 (CBS + fine-tuning)	+0.016	0.7343
Exp-11 (Router-based MoE) vs Exp-4 (CBS + fine-tuning)	+0.017	0.4257
Exp-14 (Fixed Ensemble) vs Exp-11 (Router-based MoE)	-0.0006	0.7343

None of these experiments are statistically distinguishable from each other, as shown in Table 5.4. This result points to a two-step argument rather than a single one. First, Experiment 11 and Experiment 14 turn out to be statistically indistinguishable ($p = 0.7343$). The learned router performs no better than a fixed (0.5, 0.5) gate. This isn't surprising given Table 5.3, which already shows the router never settles on meaningfully differentiated gate weights. On that basis, the router can be dropped in favour of the simpler fixed-weight ensemble, since it isn't worth the added complexity. The same table then shows that this fixed-weight ensemble (Experiment 14) is also statistically indistinguishable from Experiment 4, a plain, single-decoder, CBS-conditioned, fine-tuned ARAE with no dual-decoder structure at all ($p = 0.7344$). The same logic applies a second time: when two configurations perform equivalently, there's no reason to keep the more elaborate one. Here, that means the dual-decoder ensemble itself fails to justify its existence over the single-decoder model.

This conclusion is reinforced by the per-turbine AUC values underlying Table 5.4: Experiment 11 has a higher AUC than Experiment 4 on 5 of the 9 test turbines, and Experiment 14 on only 4 of the 9. Neither MoE variant beats the single-decoder baseline on a majority of turbines, despite both showing a positive mean AUC difference. The positive mean is driven by a small number of turbines with large swings in the MoE

variants' favour (T_06, T_07, T_08, T_10), while Experiment 4 has the numerical edge on the remaining turbines. A positive mean built from a minority of turbines showing large, inconsistent swings is characteristic of sampling noise rather than a genuine, generalizable effect.

Experiment 4 is therefore adopted as the final proposed model for the remainder of this dissertation. The full MoE investigation, router design, the expert collapse failure mode, load balance regularization, and the fixed-weight ensemble simplification is retained in full as this dissertation's contribution to understanding why conditional computation does not improve cross-turbine generalization in this setting, rather than as the architecture ultimately recommended.

5.10 Precision - Recall Trade-off

Before the conclusion of Section 5.9, it is worth checking whether recall alone might suffice for arguing for the adoption of the ensemble after all. Experiment 14 shows a higher recall than Experiment 4 on six of the nine test turbines. Comparing the ROC curves and confusion matrices for both experiments turbine-by-turbine (Figures 5.2 - 5.10 and 5.13 - 5.21) shows this is not, on its own, a reason to reverse that conclusion.

On T_06, T_08, and T_10 (Figures 5.7/5.18, 5.9/5.20, 5.10/5.21), the ensemble's higher recall is accompanied by improved precision and a lower false-positive rate, alongside a genuine AUC improvement, the closest thing to real evidence of better discrimination anywhere in the MoE comparison, though still not statistically significant at the fleet level. On T_02 (Figures 5.3/5.14), higher recall is achieved at the cost of a disproportionately larger increase in false-positive rate, and AUC decreases from 0.827 to 0.776, indicating the Youden's-J threshold has simply moved to a different point on a marginally worse ROC curve, not that detection improved. On T_05 (Figures 5.6/5.17), the pattern runs the other way entirely. Experiment 4 has both higher recall (0.397 vs. 0.364) and higher AUC (0.607 vs. 0.541) than the ensemble (Experiment 14), which instead trades away both recall and ranking quality for a modest precision and false-positive-rate improvement. A reminder that the single-decoder baseline is not uniformly the weaker configuration on a per-turbine basis, even though it is the configuration this dissertation ultimately adopts on the strength of the fleet-level statistical comparison. On T_07 (Figures 5.8/5.19), the ensemble reaches a recall of 1.000 by flagging 91% of all normal windows as anomalous, on a model whose AUC (0.427) is below chance level, and Experiment 4's own result on this turbine is worse still (AUC 0.256). This is a degenerate operating point, not a meaningful detection result, for either configuration.

The main reason for the poor performance on T_07 is shown in Figure 4.2, since it differs greatly from the power distribution of the reference turbine T_09.

5.11 Stratifying the Fleet Mean by Class Balance

The single fleet-mean AUC reported for the adopted final model (0.7354, Experiment 4) aggregates across turbines with very different class balance. Table 4.2 gives each turbine's full normal/abnormal window count. The abnormal-window percentages derived from it are used to group turbines below.

Table 5.5: Experiment 4 AUC, stratified by turbine class balance

Group	Turbines	Abnormal Ratio	Group Mean AUC
Moderate Imbalance	T_01, T_02, T_05, T_06	16.8% - 42.5%	0.6975
Extreme Normal-majority	T_03, T_04, T_07	0.1% - 2.6%	0.6493
Extreme Abnormal-majority	T_08, T_10	86.1% - 89.2%	0.9406

Two points follow directly from this breakdown. First, the turbines that most closely resemble a realistic deployment scenario, a meaningful but non-dominant rate of anomalous operation, sit at a mean AUC of **0.6975**, below the flat fleet mean of 0.7354. T_01's class ratio, established in Table 4.2 at 16.8% abnormal, places it within this group. Second, the extreme abnormal-majority group (T_08, T_10) contributes the two highest AUC scores in the fleet (mean 0.9406), but this is a substantially easier discrimination task by construction: when 85%+ of a turbine's operating history is already labelled abnormal, correctly ranking anomalous windows above the small remaining normal minority is a lower bar than the reverse, and this group's contribution to the flat fleet mean should not be read as evidence of strong detection capability under realistic operating conditions. The extreme normal-majority group (T_03, T_04, T_07) is presented separately because its members do not behave as a coherent group at all, the AUC ranges from a severe failure (T_07, 0.2561, well below random discrimination) to a near-best-in-fleet result (T_04, 0.9809) within turbines that share very similar (near-zero) abnormal ratios. Two of the three (T_04 at a single abnormal window,

T_07 at 24) have sample sizes small enough that their individual AUC estimates carry substantial variance regardless of model quality, though T_07's failure is severe enough (well below chance, not merely weak) that sample-size variance alone is unlikely to be the full explanation, and this turbine's failure is treated as a genuine limitation of the proposed model rather than dismissed as noise. The honest statement is that the model's behaviour on severely normal-majority turbines is inconsistent, ranging from near-best to worse-than-random within the same stratum. Taken together, this stratification supports a more precise version of this dissertation's central claim than the flat fleet mean permits, the proposed model recovers substantial cross-turbine generalization capacity on turbines with a realistic operating profile, achieves its highest apparent scores on turbines where the classification task is structurally easier, and fails severely on at least one turbine that a deployed system would need to handle correctly.

5.12 Final Model and Comparison to the Original ARAE

The dissertation's proposed model is **Experiment 4**: CBS conditioned inputs and decoder fine-tuning (60 fine-tune records, patience 30), applied to a single, unmodified decoder, no MoE router, no dual-decoder ensemble. This is not the architecture originally anticipated at the outset of this dissertation, which set out specifically to investigate whether a MoE extension would improve cross-turbine transfer. It is the outcome of following that investigation through to a statistically consistent conclusion (Section 5.9). The MoE extension router design, the expert collapse failure mode and its mitigation, load balance regularization, and the fixed-weight ensemble simplification is retained in full (Sections 5.6 - 5.10), not because it is the recommended architecture, but because a rigorous, fully-documented negative result is itself a legitimate contribution of this dissertation.

Experiment 4 achieves a fleet mean AUC of 0.7354 (min 0.2561, max 0.9809) and a fleet mean recall of 0.8350 across the nine held-out test turbines, with zero exposure of the encoder, and only a 60-window per-turbine slice of decoder-level exposure, to any target turbine's data. As Section 5.11 establishes, this flat mean should not be quoted in isolation, a more representative figure for a realistic deployment scenario is the moderate-imbalance group mean of 0.6975, alongside the explicit acknowledgement that the model fails severely (AUC 0.2561, well below random discrimination) on at least one held-out turbine. This is compared against Bui et al. (2026)'s original, pooled multi-turbine ARAE, which reports AUC 0.8607 and recall 0.9651 evaluated within the same turbine's own data distribution. The two figures answer different questions

under different evaluation protocols (within-distribution vs. genuinely held-out turbines never seen by the encoder), and are not directly comparable on a like-for-like basis. The finding is that a single trained representation, fine-tuned with a small, cheap, per-turbine calibration step, recovers approximately 85% of within-turbine AUC ($0.7354/0.8607$) and 87% of within-turbine recall ($0.8350/0.9651$) when transferred to entirely new turbines without requiring a full model to be trained and maintained per unit, and without requiring any additional architectural complexity beyond CBS and fine-tuning. Given the stated motivation of this dissertation, reducing the computational and maintenance overhead of per-turbine modelling across large fleets, it supports the practical viability of the proposed approach even though the raw numbers are numerically lower than the single-turbine figure.

6

Conclusions

6.1 Key Findings

This dissertation set out to determine whether the ARAE, proposed by Bui et al. (2026), could be adapted to generalise across a fleet of turbines without requiring a separately trained model per unit. Nine findings emerge from the fifteen-experiment ablation study in Chapter 5:

1. The unmodified ARAE does not generalise across the turbines. When evaluated directly on nine held-out turbines with no adaptation, it achieves a fleet mean AUC of 0.4979, indistinguishable from random guessing (Experiment 1). This confirms the motivating premise of this dissertation.

2. Normalisation and adaptation are complementary but individually insufficient. CBS alone, without decoder fine-tuning, does not improve but marginally worsens the cross-turbine performance (Experiment 2, 0.4713), a result traced directly to the mechanism by which CBS's bin statistics are fit exclusively on the training turbine and applied unchanged elsewhere (Section 5.2). Decoder fine-tuning alone recovers substantially more of the gap (Experiment 3, 0.6577). Combined, the two interventions reach a fleet mean AUC of 0.7354 (Experiment 4), confirming they address distinct sources of the generalisation gap rather than a shared one.

3. The choice of training turbine is not interchangeable within the fleet. Repeating Experiment 4's configuration with `T_06` in place of `T_09` as the training turbine (Experiment 5) drops fleet mean AUC from 0.7354 to 0.5640 (Section 5.5),

despite T_06 being the fleet's second best turbine. This empirically validates T_09 as the training turbine used throughout the remainder of this dissertation.

4. MoE routing collapses under symmetric expert initialisation: with both experts initialised identically, the router converges to an input-independent gate split of exactly (0.51, 0.49) on every turbine (Experiment 6), functionally equivalent to a single decoder with its output arbitrarily halved.

5. A statistically rigorous, paired Wilcoxon comparison across the nine held-out turbines (Section 5.9) shows that neither the learned router ($p = 0.4257$), nor the fixed-weight ensemble simplification of it ($p = 0.7343$), provides a significant improvement in AUC over the single-decoder CBS-fine-tuning baseline.

6. Applying the same argument consistently, rather than stopping at the first favourable simplification, leads this dissertation to adopt the single-decoder configuration (Experiment 4) as its final proposed model. The router-versus-ensemble comparison justifies simplifying the router away. The same statistical standard, applied one step further, shows the ensemble itself is indistinguishable from the plain single-decoder baseline. The full MoE investigation is retained as a documented account of why conditional computation does not help in this setting, not as the recommended architecture.

7. The final model's flat fleet mean AUC (0.7354) is not evenly distributed as explained in Section 5.11. Turbines with a realistic, moderate rate of anomalous operation achieve a mean AUC of 0.6975, below the flat fleet figure. On the other hand, turbines with an overwhelming abnormal-majority operating history achieve the highest scores in the fleet (mean AUC 0.9406) on what is structurally an easier discrimination task. The model fails severely with AUC 0.2561, well below random discrimination on T_07, a turbine with a realistic anomaly rate.

8. T_07's power distribution differs substantially from the training turbine's, and because CBS's bin statistics are fit exclusively on the training turbine, this mismatch degrades the standardisation that T_07 receives specifically.

9. The comparison of this dissertation's cross-turbine results to Bui et al. (2026)'s own reported ARAE performance (AUC 0.8607, recall 0.9651) must be read precisely. Bui et al.'s evaluation protocol pools windows from multiple turbines and applies a random split stratified only by class label, with no turbine held out entirely from training, an easier setting than a genuine turbine-level holdout. This

dissertation's model recovers approximately 85% of within-distribution AUC and 87% of within-distribution recall under a strictly harder evaluation protocol, using only a single shared encoder and a 60-window per-turbine calibration step, which is not a weaker result than the original comparison implied, but a more precisely characterised one.

6.2 Contributions

This dissertation makes the following contributions: It provides the first evaluation of the ARAE architecture specifically under a genuine turbine-level holdout protocol, in contrast to the pooled, class-stratified evaluation used in its original proposal. It adapts CBS, originally proposed by Pillay et al. (2025) for Gaussian Mixture Model-based anomaly detection, using equal-width power bins, to a deep reconstructive autoencoder for the first time, using equal-frequency binning to ensure balanced statistical support across turbines with differently-shaped power distributions, and explicitly positions this choice against predictive, physically-grounded alternatives (Section 3.3) rather than adopting it by default. It combines CBS with decoder-only fine-tuning, adapted from Roelofs et al. (2024)'s conventional-autoencoder setting to the ARAE's reinforced objective for the first time, and demonstrates the two techniques are complementary rather than redundant. It empirically validates the choice of training turbine within the fleet, rather than assuming it, showing that a plausible alternative candidate (selected on the same class-balance criterion) produces a substantially worse and more variable outcome. It provides a documented, reproducible case study of MoE expert collapse under symmetric initialisation in a small-model, low-data regime, together with the partial mitigation achieved through initialisation noise and entropy-based load-balancing, an application of conditional computation not previously explored in wind turbine condition monitoring, to the best of our knowledge. Most substantively, it applies matched-pairs statistical testing, rather than raw point-estimate comparison, throughout its ablation study, and demonstrates that this materially changes the conclusion drawn about the MoE extension's value, and that applying the same standard consistently, not stopping at a single favourable simplification, changes it further still, a methodological standard not observed in the closest comparable prior transfer-learning studies in this domain (Jonas and Meyer, 2025; Roelofs et al., 2024). Finally, it identifies and mechanistically explains a specific, severe failure mode CBS's single-source-turbine bin fitting degrading standardisation for a distributionally dissimilar target turbine, rather than reporting only the fleet-aggregate figures that would obscure it.

6.3 Limitations

Several limitations qualify these findings. All experiments were conducted on a single wind farm's fleet of ten turbines of a presumed homogeneous type. Whether the same approach transfers across manufacturers, capacities, or sites was not tested. The evaluation protocol, while a genuine turbine-level holdout for the encoder, still permits the decoder a small amount of target-turbine exposure via fine-tuning. This dissertation characterises efficient few-shot transfer, not strict zero-shot generalisation. Several test turbines have very small numbers of abnormal windows for example, `T_04` at a single abnormal window in particular, making individual AUC estimates for those turbines high-variance. Youden's J threshold selection assumes equal cost between false positives and false negatives, an assumption stated explicitly in Section 4.6.2. Roelofs et al. (2024), whose fine-tuning strategy this dissertation adopts, instead weight precision over recall for this exact application domain, and no quantified cost model specific to this dissertation's own deployment context was established. CBS's bin statistics are fit exclusively on the training turbine throughout, the mechanism by which this degrades standardisation for a distributionally dissimilar target turbine (Section 5.2) is identified but not corrected, and a per-target-turbine bin-refitting alternative was not tested. The training-turbine comparison in Section 5.5 is limited to the two most suitable candidates in the fleet, a systematic comparison across all ten turbines as training-turbine candidates was not conducted. Finally, this dissertation did not evaluate two credible alternative strategies identified in Chapter 3: generative domain mapping (Jonas and Meyer, 2025), shown in a directly comparable setting to outperform fine-tuning under severe data scarcity, and predictive, physically-grounded operating-condition modelling (Bilendo et al., 2022), an alternative to CBS's purely statistical conditioning approach.

6.4 Future Work

Building on the limitations above, several directions merit further investigation. The mechanism identified behind CBS's standalone underperformance and `T_07`'s specific failure, i.e, bin statistics fit, exclusively on the training turbine motivates a direct test of per-target-turbine bin refitting, using only each target turbine's own small fine-tuning sample to avoid leakage, to establish whether this recovers standalone CBS performance or whether fine-tuning already subsumes the correction. This dissertation adopted only the decoder-only fine-tuning variant of Roelofs et al. (2024). A direct comparison against their full-autoencoder variant, and separately against Jonas and Meyer (2025)'s

generative domain-mapping alternative, within the CBS-conditioned ARAE setting developed here, would clarify whether either represents a meaningfully stronger cross-turbine adaptation strategy. A direct comparison between CBS's conditioning-based approach and the predictive, physically-grounded alternative of Bilendo et al. (2022) was similarly not conducted and is a natural next step, particularly given the power curve's approximately physical grounding may transfer more robustly across turbines than CBS's empirically-fit bin statistics. Section 5.5 establishes that training-turbine choice is not interchangeable, but only compares two candidates. A systematic evaluation across all ten turbines as training-turbine candidates would establish whether class balance alone predicts a suitable training turbine, or whether other factors (e.g. representativeness of operating conditions relative to the rest of the fleet) are also required, and would generalise this dissertation's single-comparison finding into a proper selection criterion. The generalisation protocol established here should be extended to turbines from additional wind farms and manufacturers to test whether it holds beyond a single homogeneous fleet. As this study found no evidence that a power-conditioned router adds value, future work could explore whether a routing signal derived from the latent representation itself produces more genuinely differentiated expert usage than the power-bin signal tested here. Finally, combining the fine-tuned, CBS-conditioned representation developed here with Bui et al. (2026)'s supervised Bottleneck architecture, should sufficient labelled fault data become available across multiple turbines, represents a natural next step toward a model that is both label-efficient in the few-shot cross-turbine setting and capable of near-perfect within-distribution performance when labels are abundant.

Appendix A

Wilcoxon Signed-Rank Test Results

Reading the matrices below: Each entry is the two-sided exact Wilcoxon signed-rank p-value ($n = 9$ held-out turbines) comparing the row experiment against the column experiment for the named metric. Colour indicates significance: blue for $p < 0.01$, green for $0.01 \leq p < 0.05$, orange for $0.05 \leq p < 0.10$, and plain black text for $p \geq 0.10$ (not significant). The matrix is symmetric by construction, so only the upper triangle is populated. Experiment-5 is excluded throughout, because it has a different train-turbine (T_06) than the other fourteen experiments (T_09).

Table A.1: Pairwise exact Wilcoxon signed-rank p-values, AUC, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		0.65	0.20	0.13	0.91	1.00	0.36	0.16	0.25	0.13	0.25	0.25	0.10	0.13
2			0.04	0.02	1.00	0.25	0.10	0.02	0.02	0.03	0.13	0.04	0.01	0.01
3				0.13	0.20	0.10	0.82	0.20	0.43	0.16	0.82	0.25	0.20	0.30
4					0.01	0.04	0.25	0.73	0.50	0.43	0.30	0.82	0.73	0.82
6						0.57	0.07	0.00	0.03	0.00	0.07	0.03	0.00	0.04
7							0.20	0.01	0.07	0.02	0.16	0.07	0.01	0.07
8								0.20	0.36	0.16	0.82	0.16	0.07	0.25
9									0.20	0.50	0.36	0.82	0.30	0.82
10										0.43	0.43	0.43	0.05	0.50
11											0.01	0.73	0.73	1.00
12												0.13	0.01	0.43
13													0.36	0.65
14														1.00
15														

A. Wilcoxon Signed-Rank Test Results

Table A.2: Pairwise exact Wilcoxon signed-rank p-values, Precision, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		1.00	1.00	0.25	0.64	1.00	0.57	0.91	1.00	0.30	0.91	0.65	0.25	0.55
2			0.79	0.31	0.81	0.55	0.30	0.91	0.65	0.29	0.80	0.20	0.20	0.25
3				0.16	0.73	0.91	0.20	0.73	0.38	0.07	0.57	0.24	0.04	0.16
4					0.47	0.20	0.02	0.25	0.43	0.98	0.25	0.65	0.84	0.84
6						0.58	0.09	1.00	0.20	0.71	0.92	0.57	0.27	0.46
7							0.25	0.57	1.00	0.13	0.55	0.16	0.08	0.08
8								0.13	0.10	0.04	0.24	0.01	0.03	0.02
9									0.64	0.38	0.43	0.73	0.20	0.26
10										0.50	0.91	0.59	0.20	0.08
11											0.06	1.00	0.97	1.00
12												0.43	0.24	0.16
13													1.00	0.57
14														1.00
15														

Table A.3: Pairwise exact Wilcoxon signed-rank p-values, Recall, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		0.84	0.56	0.31	0.56	0.44	0.06	0.16	0.22	0.81	0.25	0.69	0.22	0.47
2			0.44	0.44	0.81	0.06	0.16	0.31	0.44	1.00	0.84	0.81	0.44	0.56
3				0.69	0.44	0.69	0.69	0.31	0.84	1.00	0.56	0.94	0.81	0.94
4					0.16	0.56	0.06	0.44	0.56	0.69	0.56	0.81	0.56	0.81
6						0.06	0.09	0.16	0.19	0.58	0.22	0.58	0.22	0.38
7							0.31	0.44	1.00	0.94	0.84	1.00	1.00	1.00
8								1.00	0.56	0.08	0.16	0.38	0.56	0.05
9									0.56	0.30	0.28	0.30	1.00	0.47
10										0.16	0.69	0.30	0.69	0.47
11											0.94	1.00	0.03	1.00
12												0.94	0.44	0.94
13													0.16	0.69
14														0.22
15														

Table A.4: Pairwise exact Wilcoxon signed-rank p-values, Youden's J, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		0.82	0.91	1.00	0.73	0.04	0.20	0.50	0.36	1.00	0.25	0.57	1.00	0.50
2			0.82	0.91	0.91	0.30	0.57	0.73	0.91	0.65	0.73	0.65	0.82	1.00
3				0.65	0.82	0.57	0.73	0.73	1.00	0.65	0.65	0.73	0.43	1.00
4					0.91	0.30	0.16	0.57	0.82	0.73	0.43	0.65	0.43	0.91
6						0.13	0.43	1.00	0.65	1.00	0.36	1.00	0.91	0.73
7							0.20	0.36	0.91	0.20	0.64	0.25	0.20	0.73
8								0.30	0.91	0.30	0.91	0.36	0.25	0.30
9									0.65	0.43	0.36	0.82	0.43	1.00
10										0.30	0.65	0.25	0.36	0.36
11											0.07	0.65	0.91	0.43
12												0.73	0.20	0.30
13													0.50	0.57
14														0.25
15														

Table A.5: Pairwise exact Wilcoxon signed-rank p-values, True Positives, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		1.00	0.44	0.31	0.44	0.56	0.06	0.22	0.22	0.69	0.31	0.58	0.22	0.30
2			0.31	0.44	1.00	0.06	0.09	0.31	0.44	0.97	0.84	0.69	0.31	0.44
3				0.69	0.31	0.31	0.56	0.56	0.69	0.81	0.56	0.81	0.62	0.72
4					0.16	0.56	0.06	0.19	0.31	0.58	0.44	0.92	0.22	1.00
6						0.06	0.09	0.16	0.19	0.22	0.16	0.22	0.16	0.16
7							0.31	0.44	1.00	0.94	0.84	0.94	1.00	0.94
8								0.69	0.69	0.16	0.16	0.61	0.69	0.03
9									0.69	0.58	0.22	0.58	0.69	0.47
10										0.16	0.44	0.47	0.69	0.58
11											0.62	1.00	0.16	1.00
12												0.81	0.25	0.98
13													0.22	0.56
14														0.22
15														

A. Wilcoxon Signed-Rank Test Results

Table A.6: Pairwise exact Wilcoxon signed-rank p-values, False Positives, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		1.00	0.50	0.46	0.95	1.00	0.73	0.73	0.82	0.30	1.00	0.30	0.31	0.46
2			0.43	0.30	0.48	0.84	0.82	0.38	0.50	0.07	0.65	0.02	0.15	0.05
3				0.43	1.00	0.73	0.30	1.00	1.00	0.57	0.73	0.65	0.82	0.57
4					0.46	0.38	0.10	0.36	0.73	0.43	0.30	0.50	0.95	0.64
6						0.46	0.50	0.65	1.00	0.07	0.57	0.25	0.15	0.20
7							0.71	0.74	0.46	0.03	0.65	0.13	0.15	0.08
8								0.30	0.43	0.03	0.50	0.00	0.10	0.01
9									0.95	0.10	0.65	0.43	0.50	0.38
10										0.10	0.91	0.21	0.25	0.25
11											0.04	0.84	0.57	1.00
12												0.65	0.43	0.30
13													1.00	0.43
14														0.87
15														

Table A.7: Pairwise exact Wilcoxon signed-rank p-values, True Negatives, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		1.00	1.00	0.73	0.91	0.82	0.43	0.76	0.73	0.82	0.73	0.36	0.91	0.82
2			0.65	0.91	0.89	0.57	0.50	1.00	0.91	0.20	0.73	0.13	0.52	0.36
3				0.43	1.00	0.73	0.43	1.00	1.00	0.57	0.73	0.43	0.82	0.36
4					1.00	0.73	0.36	0.36	0.73	0.43	0.30	0.36	0.95	0.43
6						0.46	0.50	0.91	0.91	0.50	0.91	0.25	0.57	0.20
7							0.71	0.89	1.00	0.05	0.91	0.13	0.20	0.08
8								0.80	0.82	0.16	0.82	0.00	0.43	0.01
9									0.95	0.10	0.65	0.36	0.50	0.20
10										0.10	0.91	0.07	0.25	0.04
11											0.04	0.57	0.57	0.50
12												0.43	0.43	0.20
13													0.43	0.43
14														0.43
15														

A. Wilcoxon Signed-Rank Test Results

Table A.8: Pairwise exact Wilcoxon signed-rank p-values, False Negatives, across the fourteen experiments sharing a common evaluation protocol. Colour legend as defined above.

	1	2	3	4	6	7	8	9	10	11	12	13	14	15
1		1.00	0.44	0.31	0.44	0.56	0.06	0.22	0.22	0.69	0.31	0.58	0.22	0.30
2			0.31	0.44	1.00	0.06	0.09	0.31	0.44	0.97	0.84	0.69	0.31	0.44
3				0.69	0.31	0.31	0.56	0.56	0.69	0.81	0.56	0.81	0.62	0.72
4					0.16	0.56	0.06	0.19	0.31	0.58	0.44	0.92	0.22	1.00
6						0.06	0.09	0.16	0.19	0.22	0.16	0.22	0.16	0.16
7							0.31	0.44	1.00	0.94	0.84	0.94	1.00	0.94
8								0.69	0.69	0.16	0.16	0.61	0.69	0.03
9									0.69	0.58	0.22	0.58	0.69	0.47
10										0.16	0.44	0.47	0.69	0.58
11											0.62	1.00	0.16	1.00
12												0.81	0.25	0.98
13													0.22	0.56
14														0.22
15														

References

Francisco Bilendo, Hamed Badihi, Ningyun Lu, Philippe Cambron, and Bin Jiang. A normal behavior model based on power curve and stacked regressions for condition monitoring of wind turbines. *IEEE Transactions on Instrumentation and Measurement*, 71, 2022. ISSN 15579662. doi:10.1109/TIM.2022.3196116. 15, 54, 55

Hoang Tu Bui, Juan Albarracín, Kyro Keown, and Conor Ryan. On the capacity of deep autoencoder-based normal behaviour models in wind turbine condition monitoring. *Proceedings of the 18th International Conference on Agents and Artificial Intelligence*, pages 2316–2323, 2026. doi:10.5220/0014229200004052. iv, v, vii, 1, 2, 7, 8, 9, 12, 13, 14, 15, 16, 18, 19, 20, 21, 25, 27, 28, 29, 30, 49, 51, 52, 55

Xavier Chesterman, Timothy Verstraeten, Pieter-Jan Daems, Ann Nowé, and Jan Helsen. Overview of normal behavior modeling approaches for SCADA-based wind turbine condition monitoring demonstrated on data from operational wind farms. *Wind Energy Science*, 8(6):893–924, June 2023. ISSN 2366-7451. doi:10.5194/wes-8-893-2023. 7, 8

S. Faulstich, B. Hahn, and P. J. Tavner. Wind turbine downtime and its importance for offshore deployment. *Wind Energy*, 14:327–337, 2011. ISSN 10991824. doi:10.1002/WE.421. 1, 6

Global Wind Energy Council. Global wind report 2026. Technical report, GWEC, 2026. URL `https://www.gwec.net/reports/globalwindreport`. 6

Stefan Jonas and Angela Meyer. Fault detection in new wind turbines with limited data by generative transfer learning. *Energy and AI*, 22:100626, 12 2025. ISSN 2666-5468. doi:10.1016/J.EGYAI.2025.100626. 2, 7, 14, 15, 18, 19, 53, 54

Stefan Jonas, Dimitrios Anagnostos, Bernhard Brodbeck, and Angela Meyer. Vibration fault detection in wind turbines based on normal behaviour models without feature engineering. *Energies*, 16:1760, 2 2023. ISSN 1996-1073. doi:10.3390/EN16041760. 8, 12, 18

REFERENCES

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010. ISSN 10414347. doi:10.1109/TKDE.2009.191. 9

Sahil Pillay, Pieter de Villiers, P. S. Heyns, and Allan De Freitas. Standardisation of active power dependent features in wind turbine scada data for gmm based anomaly detection. *Social Science Research Network Preprints*, 2025. doi:10.2139/SSRN.5615670. iv, 2, 15, 16, 18, 19, 23, 24, 53

Niklas Renström, Pramod Bangalore, and Edmund Highcock. System-wide anomaly detection in wind turbines using deep autoencoders. *Renewable Energy*, 157:647–659, 9 2020. ISSN 0960-1481. doi:10.1016/J.RENENE.2020.04.148. 8, 11, 18

Cyriana M.A. Roelofs, Marc Alexander Lutz, Stefan Faulstich, and Stephan Vogt. Autoencoder-based anomaly root cause analysis for wind turbines. *Energy and AI*, 4:100065, 6 2021. ISSN 2666-5468. doi:10.1016/J.EGYAI.2021.100065. 12, 18

Cyriana M.A. Roelofs, Christian Gück, and Stefan Faulstich. Transfer learning applications for autoencoder-based anomaly detection in wind turbines. *Energy and AI*, 17:100373, 9 2024. ISSN 26665468. doi:10.1016/j.egyai.2024.100373. iv, 2, 7, 13, 14, 17, 18, 19, 27, 28, 30, 53, 54

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus Robert Muller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109:756–795, 5 2021. ISSN 15582256. doi:10.1109/JPROC.2021.3052449. 8

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1 2017. URL `https://arxiv.org/pdf/1701.06538`. 17, 25

Jannis Tautz-Weinert and Simon J. Watson. Using scada data for wind turbine condition monitoring - a review. *IET Renewable Power Generation*, 11:382–394, 3 2017. ISSN 17521424. doi:10.1049/IET-RPG.2016.0248;PAGE:STRING:ARTICLE/CHAPTER. 6

Peng Yan, Ahmed Abdulkadir, Paul Philipp Luley, Matthias Rosenthal, Gerrit A. Schatte, Benjamin F. Grewe, and Thilo Stadelmann. A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. *IEEE*

Access, 12:3768–3789, 2024. ISSN 21693536. doi:10.1109/ACCESS.2023.3349132. 7, 8, 9, 10, 17

Yueyue Yao, Jianghong Ma, Shanshan Feng, and Yunming Ye. Svd-ae: An asymmetric autoencoder with svd regularization for multivariate time series anomaly detection. *Neural Networks*, 170:535–547, 2 2024. ISSN 0893-6080. doi:10.1016/J.NEUNET.2023.11.023. 29

W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. doi:10.1002/1097-0142(1950)3:1¡32::aid-cncr2820030106¿3.0.co;2-3. 30