

Article

Robust Unsupervised Acoustic Anomaly Detection for Turbo Molecular Pumps Using ResNet–Convolutional Block Attention Module and Structural Similarity Loss

Chu-Hui Lee ^{1,*} , Po-Jui Chiang ¹ , Chien-Ming Wu ², Chih-Chyau Yang ² and Chun-Ming Huang ²

¹ Information Management, Chaoyang University of Technology, Taichung 413310, Taiwan; s11314622@gm.cyut.edu.tw

² Taiwan Semiconductor Research Institute, National Institutes of Applied Research, Hsinchu 300091, Taiwan; wucm@nlar.org.tw (C.-M.W.); ccyang@nlar.org.tw (C.-C.Y.); cmhuang@nlar.org.tw (C.-M.H.)

* Correspondence: chlee@cyut.edu.tw

Abstract

In semiconductor and optoelectronics manufacturing, the reliability of turbo molecular pumps (TMPs) is vital for maintaining vacuum integrity and ensuring product yield. However, acoustic monitoring in cleanrooms faces severe challenges due to ambient noise levels routinely exceeding 80 dBA and the scarcity of labeled anomaly data. This study proposes an unsupervised acoustic anomaly detection and localization system to address these issues. The performance of the proposed framework is evaluated using an acoustic dataset collected from operational turbopumps in an industrial semiconductor cleanroom environment, encompassing both normal operations and naturally occurring failure conditions. We introduce a Convolutional Autoencoder (CAE) based on ResNet-18, integrated with a Convolutional Block Attention Module (CBAM) to adaptively suppress high-decibel environmental noise. To enhance sensitivity to structural spectral defects, a hybrid loss function combining Mean Squared Error (MSE) and structural similarity index measure (SSIM) is implemented. Experimental results, supported by rigorous hyperparameter sensitivity analysis, demonstrate that the proposed model achieves an outstanding AUC of 0.9535 and a fault recall of 99.06% under a validation-calibrated threshold, significantly outperforming standard U-Net architectures. Furthermore, the system generates anomaly heatmaps for precise time–frequency localization, enabling explainable diagnostics. With a model inference throughput of 330.28 FPS and an end-to-end processing rate of $\approx 73\times$ in real time, the proposed framework provides an efficient and robust solution for real-time predictive maintenance in noisy industrial settings.

Keywords: turbo molecular pump (TMP); acoustic anomaly detection; unsupervised learning; CBAM; SSIM loss; predictive maintenance (PdM)



Received: 29 June 2026

Revised: 27 July 2026

Accepted: 28 July 2026

Published: 30 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the advent of Industry 4.0, smart manufacturing has transitioned from traditional automation to AI-driven self-diagnostic systems. In high-tech industries, predictive maintenance (PdM) has become a cornerstone strategy for enhancing operational efficiency. By leveraging artificial intelligence, manufacturers can detect early signs of equipment degradation, thereby preventing catastrophic failures and minimizing unplanned downtime.

In the semiconductor and optoelectronics sectors, the stability of the manufacturing environment is paramount. Turbo molecular pumps (TMPs) play a critical role in

maintaining the ultra-high-vacuum conditions required for sensitive processes such as etching and thin-film deposition [1]. Any mechanical failure in these high-speed rotating components—such as bearing wear, rotor imbalance, or foreign object ingestion—can lead to vacuum loss, resulting in massive wafer scrapping and significant economic losses.

Despite its importance, implementing effective monitoring for TMPs in actual cleanroom settings remains a formidable challenge. First, these environments are saturated with high-intensity background noise from cooling towers, air circulation fans, and other heavy machinery, routinely exceeding 80 dBA and approaching statutory occupational exposure limits. This noise masks subtle high-frequency anomaly signatures, rendering manual inspection and traditional acoustic sensors ineffective. Second, the scarcity of fault samples in production lines leads to severe data imbalance, which restricts the application of conventional supervised learning algorithms that rely on large labeled datasets [2]. Similar compounding challenges regarding severe background interference and diagnostic data scarcity have also been investigated in other extreme acoustic environments, such as condition monitoring for rotating machinery in underwater settings [3].

While deep learning reconstruction models like U-Net have shown promise in anomaly detection, their skip connections let anomalous content bypass the compression bottleneck and reappear in the reconstruction—the “identity mapping” problem, analyzed in detail in Section 2.2—which shrinks the reconstruction error gap between normal and abnormal samples. Furthermore, existing statistical methods often lack the interpretability required for engineers to perform rapid root cause analysis.

To overcome these barriers, this study proposes an unsupervised acoustic anomaly detection framework tailored to noisy cleanroom environments. By integrating a ResNet-18 backbone with attention mechanisms and a structural loss function, we aim to provide a robust solution for vacuum pump diagnostics. Our primary contributions are as follows:

- **Robust Feature Extraction:** We developed a ResNet-18-based Convolutional Autoencoder integrated with a Convolutional Block Attention Module (CBAM). The channel and spatial attention mechanisms adaptively filter the intense background noise while focusing on critical rotational harmonics.
- **Structural Anomaly Detection:** By incorporating structural similarity (SSIM) loss into the objective function, the model captures texture-based fault patterns on spectrograms more effectively than traditional pixel-wise MSE loss.
- **Explainable Localization:** The system generates anomaly heatmaps based on reconstruction errors, localizing anomalous energy in the time–frequency domain to support maintenance diagnostics.
- **Comprehensive Empirical Validation:** Through extensive ablation studies and hyperparameter sensitivity analysis, we systematically validate the necessity of removing skip connections to prevent “anomaly leakage.” The proposed framework demonstrates superior robustness against the high-decibel background noise while maintaining an inference speed exceeding 330 FPS on a desktop-class GPU ($\approx 73\times$ real time end-to-end), supporting real-time monitoring; validation on embedded platforms is left to future work.

2. Related Work

2.1. Traditional vs. Deep Learning Methods

Early anomaly detection relied on statistical methods such as Principal Component Analysis (PCA) and Isolation Forest (iForest) [4,5]. While these algorithms are computationally efficient, they typically require flattening 2D spectrograms into 1D vectors, destroying crucial temporal–spectral structural information. Furthermore, they lack interpretability; iForest outputs a single anomaly score but cannot localize the fault frequency or time, making it prone to false alarms in high-noise cleanroom environments (>80 dBA). To ad-

dress these limitations, Convolutional Autoencoders (CAEs) have become the mainstream unsupervised approach [6,7]. By learning to reconstruct normal samples, CAEs identify anomalies based on high reconstruction errors.

2.2. Reconstruction-Based Architectures and the U-Net Framework

The U-Net architecture, originally proposed for biomedical image segmentation [8,9], has become a cornerstone in various computer vision tasks due to its symmetric encoder–decoder structure. The “U” shape is formed by a contracting path (encoder) that captures contextual information and a symmetric expanding path (decoder) that enables precise localization. A defining feature of U-Net is the integration of “skip connections,” which concatenate high-resolution feature maps from the encoder directly to the corresponding upsampling layers in the decoder [8,10].

While these skip connections are instrumental in preserving fine spatial details and mitigating information loss during downsampling, they introduce a critical technical challenge when applied to unsupervised anomaly detection. In a reconstruction-based paradigm, the model is expected to learn a bottleneck representation that effectively “repairs” anomalies by projecting them back into the normal data distribution. However, directly applying U-Net to this task often triggers the “identity mapping” problem. Our preliminary research indicates that skip connections allow anomalous features to bypass the latent bottleneck, enabling the model to directly “copy” input noise and defects to the output—a phenomenon known as “anomaly leakage”. This perfect reconstruction of abnormal features significantly reduces the residual difference between normal and abnormal inputs, thereby degrading the sensitivity of the anomaly scoring mechanism. Consequently, for high-noise cleanroom environments, standard U-Net architectures may fail to distinguish subtle mechanical faults from background interference.

2.3. Attention Mechanisms and SSIM Loss

To mitigate noise interference and the identity mapping problem, attention mechanisms like the Convolutional Block Attention Module (CBAM) have been introduced [11,12]. The CBAM applies channel and spatial attention to adaptively weigh critical fault frequencies while suppressing background noise. Additionally, traditional Mean Squared Error (MSE) loss focuses solely on pixel-wise differences, ignoring local dependencies. In contrast, the structural similarity index (SSIM) [13] captures texture and structural integrity. Integrating SSIM helps the model focus on the structural patterns of pump failures (e.g., harmonic lines) rather than random noise fluctuations. Distinct from feature embedding methods such as PatchCore, which score anomalies by distance to a memory bank of pretrained patch features, and one-class objectives such as Deep SVDD, our approach couples a skip-free reconstruction pathway with an attention-refined bottleneck and a structure-aware loss, which outperforms both paradigms on high-noise TMP audio. Table 1 summarizes the characteristics of these method families.

Table 1. Comparison of anomaly detection methods for TMPs.

Category	Methods	Key Advantages	Critical Limitations
Traditional ML	iForest, PCA	Lightweight; fast.	No spatial features; unexplainable.
Basic DL	CNN-AE	Auto feature extraction.	Blurry outputs; noise-sensitive.
U-Net Based	U-Net AE	High detail retention.	Identity mapping; leakage issues.
Proposed	ResNet-CBAM	Noise robust; heatmap localization.	Slightly higher complexity.

3. Materials and Methods

3.1. Data Preprocessing

To capture high-frequency fault characteristics, raw acoustic signals are resampled to 44.1 kHz. Each recording is truncated or zero-padded to a fixed length of 10 s. We convert the 1D time-domain signals into 2D time–frequency features using Short-Time Fourier Transform (STFT) with an FFT size of 2048 and a hop length of 512. The resulting spectrograms are mapped to 128 Mel bands and converted to a logarithmic scale (dB) to mimic human auditory perception. Finally, the Log-Mel spectrograms are normalized and resized to 256×256 pixels to serve as the model input x . A Hann analysis window is used, with the 128 Mel bands spanning 0–22.05 kHz. Each Log-Mel spectrogram is min–max-normalized to $[0, 1]$ per spectrogram, bilinearly resized to 256×256 , and replicated to three channels to match the ResNet-18 input stem.

As shown in Figure 1, the proposed framework consists of three main stages: (1) the preprocessing of raw audio into Log-Mel spectrograms, (2) feature extraction and reconstruction using a ResNet-CBAM autoencoder, and (3) anomaly localization via a hybrid loss-based heatmap.

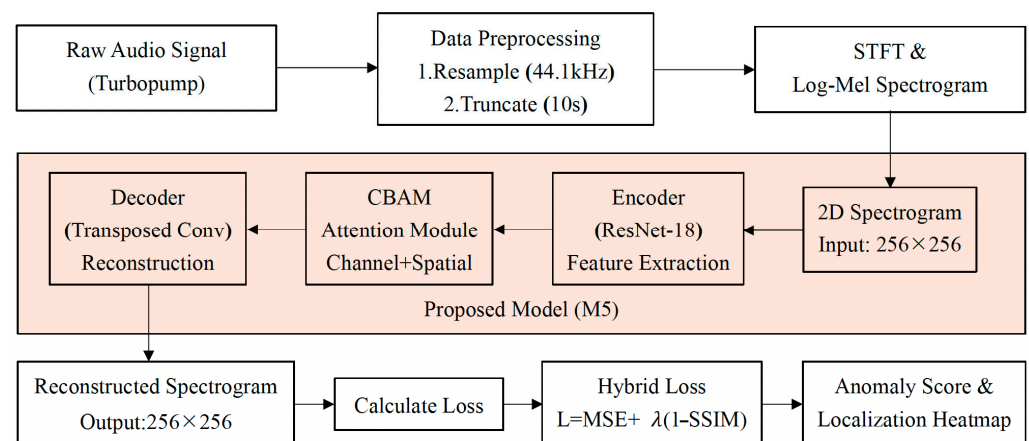


Figure 1. An overview of the proposed unsupervised anomaly detection framework. Training (normal data only): raw acoustic signals are converted to 256×256 Log-Mel spectrograms and fed to the ResNet-18 encoder; the CBAM module refines the bottleneck features, and the decoder reconstructs the input under the hybrid loss $L = L_{MSE} + \lambda \cdot L_{SSIM}$ ($\lambda = 2$). Testing: the anomaly score of an input is computed as $MSE + \lambda \cdot (1 - SSIM)$ between the input and its reconstruction, and compared against a threshold calibrated on the validation set; the pixel-wise squared reconstruction error yields the anomaly localization heatmap. Arrows indicate the direction of information flow; no skip connections exist between the encoder and decoder.

3.2. ResNet-CBAM Autoencoder Architecture

The proposed architecture, illustrated in Figure 2, is a Convolutional Autoencoder designed to solve the identity mapping problem found in standard U-Nets.

- **Encoder (Feature Extraction):** We utilize ResNet-18 [14,15] as the backbone, removing the fully connected layers to preserve spatial structures. This deep architecture effectively extracts complex texture features from the spectrograms.
- **Bottleneck with CBAM:** To function in such high-noise environments, a Convolutional Block Attention Module (CBAM) [11] is embedded between the encoder and decoder. Channel attention emphasizes specific fault frequency bands, while the apatial attention localizes the defect in the time domain.
- **Decoder (Reconstruction):** The decoder consists of transposed convolutional layers that upsample the latent features back to the original dimension.

- **Removal of Skip Connections:** A critical design choice is the complete removal of U-Net-style skip connections, eliminating the anomaly leakage pathway described in Section 2.2 (empirically analyzed in Section 4.4). By forcing information through the CBAM bottleneck, the model learns to reconstruct only the dominant normal features, resulting in higher reconstruction errors for anomalous inputs.

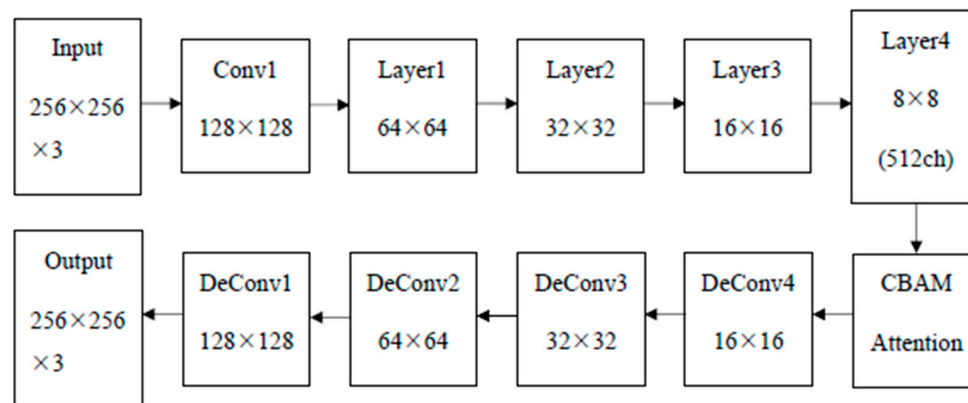


Figure 2. The architecture of the ResNet-CBAM autoencoder. Note the absence of skip connections to prevent identity mapping. The CBAM bottleneck filters background noise before the decoder reconstructs the signal.

3.3. Hybrid Loss Function

To detect both intensity differences and structural anomalies, we propose a hybrid loss function combining Mean Squared Error (MSE) and the structural similarity index (SSIM).

The MSE loss captures pixel-wise intensity differences:

$$L_{MSE} = \frac{1}{N} \sum_{i=1}^N (x_i - \hat{x}_i)^2 \quad (1)$$

where x is the input, $\hat{x}$ is the reconstruction, and N is the total number of pixels.

However, since pump faults often manifest as structural textures (e.g., harmonic lines) rather than simple intensity changes, we incorporate SSIM. SSIM measures similarity based on luminance (l), contrast (c), and structure (s). The SSIM loss is defined as follows:

$$L_{SSIM} = 1 - SSIM(x, \hat{x}) \quad (2)$$

The final objective function integrates both metrics to balance pixel-level accuracy and structural integrity:

$$L_{Total} = L_{MSE} + \lambda \cdot L_{SSIM} \quad (3)$$

where λ is a hyperparameter acting as a weighting coefficient that controls the relative importance of the structural similarity penalty. This hybrid approach ensures that the model preserves the global energy distribution while remaining highly sensitive to local structural defects caused by mechanical failures. The optimal configuration for λ , which is crucial for maximizing the model's robustness against the high-decibel background noise, will be empirically investigated and discussed in the subsequent hyperparameter sensitivity analysis.

4. Results and Discussion

4.1. Experimental Setup and Metrics

The proposed framework was implemented using Python 3.11.13 and the PyTorch 2.8.0 deep learning library on a Windows (64-bit) operating system. Audio signal processing,

including Mel spectrogram conversion and feature extraction, was conducted using the Librosa 0.11.0 library and Soundfile 0.13.1. The model was optimized using the Adam optimizer with a learning rate of 1×10^{-4} . All experiments and inference benchmarks were executed on a high-performance workstation equipped with an NVIDIA GeForce RTX 5070 GPU (NVIDIA, Santa Clara, CA, USA).

The experimental dataset consists of field-recorded acoustic signals from turbo molecular pumps (TMPs) operating in an industrial cleanroom environment with substantial ambient noise from facility systems and adjacent process equipment (routine facility measurements exceed 80 dBA, approaching the statutory occupational exposure limits of 90 dBA for 8 h and 95 dBA for 4 h of continuous exposure), provided by the Taiwan Semiconductor Research Institute (TSRI). Signals were captured from two pumps of the same model (one healthy, one faulty; Pfeiffer Vacuum/adixen ATH 500 M (Pfeiffer Vacuum GmbH, Aßlar, Germany), a five-axis magnetically levitated hybrid turbo molecular pump with a nominal rotation speed of 50,000 rpm, corresponding to a shaft frequency of 833 Hz) using an AVerMedia AM310 USB cardioid condenser microphone (AVerMedia Technologies, New Taipei City, Taiwan; frequency response 20 Hz–20 kHz) mounted on a fixed camera-mount bracket approximately 3–5 cm from the pump housing, aimed at the magnetic-bearing/motor section of the pump body, with no noise reduction, equalization, or gain automation applied, and digitized at 44.1 kHz with 16-bit resolution (single channel). The corpus comprises 2120 recordings of ≈ 65 s each: 1696 for training, 212 for validation, and a class-balanced test set of 106 normal and 106 abnormal recordings. Only normal recordings are used for training and validation; abnormal recordings appear exclusively in the test set. The abnormal class corresponds to a single naturally occurring fault condition on the faulty pump; the 106 abnormal samples are repeated recordings of this one fault instance, labelled as abnormal based on the facility's maintenance record.

To preclude data leakage, the dataset was split at the recording level: every recording, identified by its acquisition timestamp, belongs to exactly one subset, so no segment of a recording ever appears in more than one subset. We further verified that the normal subsets occupy non-overlapping, contiguous acquisition windows (test earliest, then validation, then training), as visualized in Figure 3. Consequently, the model cannot inflate its performance by memorizing recording-specific characteristics.

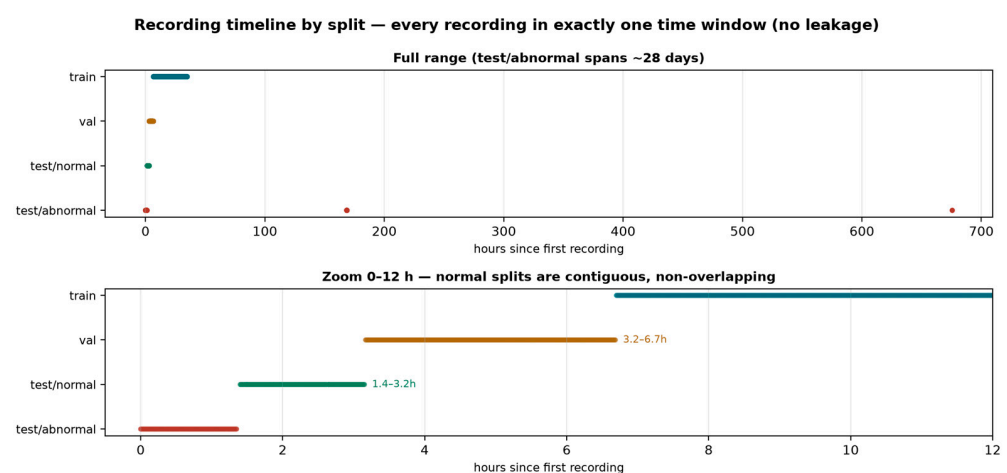


Figure 3. Recording timeline by subset. Every recording belongs to exactly one subset, and the normal subsets occupy non-overlapping acquisition windows, ruling out recording-level data leakage.

To ensure consistent input dimensions, each recording was processed by truncating the first 10 s of the signal to capture critical early-stage signatures; for audio files shorter than 10 s, zero-padding was applied to maintain a uniform segment length. All audio

segments were processed at a sampling rate of 44,100 Hz to ensure effective analysis of high-frequency responses up to 22 kHz. To evaluate the system’s effectiveness, we utilized the Area Under the Receiver Operating Characteristic Curve (AUC), Precision, Recall, and F1-Score. Furthermore, the inference speed, measured in Frames Per Second (FPS), was recorded to verify the system’s suitability for real-time monitoring. To comprehensively evaluate the robustness and effectiveness of the proposed framework, the subsequent sections present our findings across three analytical levels: hyperparameter sensitivity optimization, a quantitative ablation study, and qualitative visual analysis.

Unless otherwise stated, the binary operating threshold is calibrated on the validation set—as the 99th percentile of the validation (normal-only) anomaly scores—and then applied unchanged to the test set. The AUC is threshold-free. This policy is fully specified by the training data alone; no test set information is used to select the operating point. Under this policy, the proposed model attains a recall of 0.9906 (105/106 faults detected, one missed) at a precision of 0.7609.

4.2. Hyperparameter Sensitivity Analysis

To determine the optimal balance between pixel-wise intensity differences and localized structural integrity in the hybrid loss function ($L_{Total} = L_{MSE} + \lambda \cdot L_{SSIM}$), a rigorous hyperparameter sensitivity analysis was conducted. We evaluated the system’s performance across six different weighting coefficients: $\lambda \in \{0, 0.5, 1, 2, 3, 4\}$. Crucially, the configuration with $\lambda = 0$ relies entirely on the MSE loss function, which corresponds to the ResNet-CBAM baseline (designated as model M4 in the subsequent ablation study in Section 4.3). Note that the $\lambda = 0$ row of Table 2 stems from an independent training run of this architecture; its AUC therefore differs slightly from the M4 entry in Table 3 (cf. the seed-variance analysis below). Investigating these variants allows us to justify the mathematical necessity of the structural similarity penalty before presenting the comprehensive structural ablation study. The evaluation metrics for each λ configuration are summarized in Table 2.

Table 2. Performance comparison of different λ values in the hybrid loss function.

λ Value	Precision	Recall	F1-Score	AUC
0	0.8763	0.8019	0.8374	0.9151
0.5	0.7984	0.9717	0.8766	0.9365
1	0.5730	1.0000	0.7285	0.7750
2	0.7609	0.9906	0.8607	0.9535
3	0.6709	1.0000	0.8030	0.9171
4	0.7211	1.0000	0.8379	0.9324

Table 3. Ablation study of different model architectures.

Experiment	Precision	Recall	F1-Score	AUC
M1	0.9608	0.4623	0.6242	0.8855
M2	0.7114	1.0000	0.8314	0.9205
M3	0.7436	0.8208	0.7803	0.8769
M4	0.8721	0.7075	0.7812	0.9220
M5	0.7609	0.9906	0.8607	0.9535

Two additions strengthen this analysis. First, to verify that the $\lambda = 2$ optimum is not a single-run artifact, we retrained the $\lambda = 2$ and $\lambda = 3$ configurations with three random seeds each: $\lambda = 2$ attains an AUC of 0.9482 ± 0.0029 (mean $\pm$ std) versus 0.9364 ± 0.0025 for $\lambda = 3$, with $\lambda = 2$ superior in every individual seed (Figure 4a). Second, we motivate the search grid:

λ rescales the SSIM term to the same order of magnitude as the MSE term, so $\lambda \in [0.5, 4]$ spans the regime from pixel-dominated to structure-dominated training. Mechanistically, MSE enforces per-pixel energy fidelity but under-penalizes structural deviations such as smeared harmonic ridges, whereas SSIM explicitly scores local luminance, contrast, and structure; at $\lambda \approx 2$, the two gradients are comparable in magnitude, allowing both objectives to shape the reconstruction, while larger λ increasingly sacrifices the pixel fidelity on which the MSE component of the anomaly score relies. (The $\lambda = 1$ entry reflects an anomalously weak training run; across seeds, performance varies smoothly in λ . See the seed statistics above).

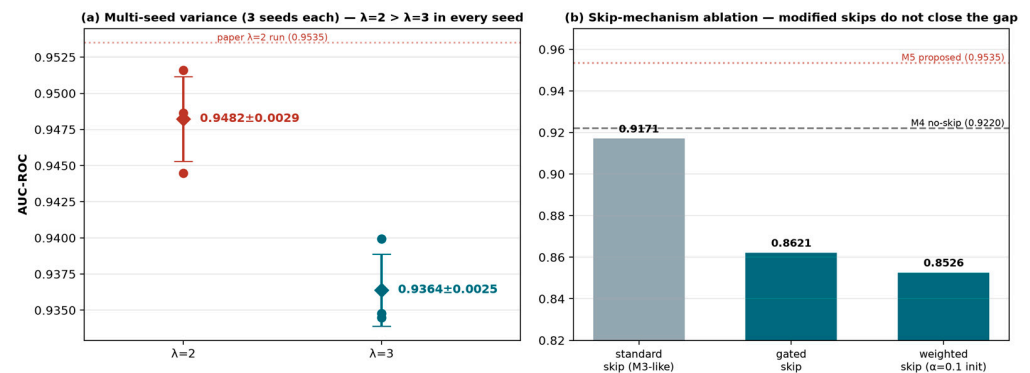


Figure 4. (a) Multi-seed variance of proposed model: $\lambda = 2$ outperforms $\lambda = 3$ in every seed (3 seeds each). (b) Skip mechanism ablation: gated and learnable-weighted skip variants fail to close gap to skip-free architectures.

4.3. Quantitative Analysis and Ablation Study

To systematically validate the effectiveness of the proposed modules, we compared five model variants (M1 to M5). Because the structural differences are critical to understanding the performance variations, the configurations of each variant are detailed below, and their comprehensive evaluation metrics are summarized in Table 3.

- **Baseline (M1—Simple CNN):** This variant uses a basic three-layer Convolutional Autoencoder. Under the validation-calibrated threshold, M1 exhibits the opposite failure mode: a conservative operating point with high precision (0.9608) but unacceptably low recall (0.4623), missing more than half of the faults: its score distribution separates the classes poorly, as its AUC (0.8855) confirms.
- **Encoder-Only Baseline (M2—ResNet-AE):** This variant utilizes a ResNet-18 backbone as the encoder but completely removes skip connections. It serves as a robust deep learning foundation, achieving a significantly better AUC of 0.9205.
- **Impact of Skip Connections (M3 vs. M2):** M3 (ResNet U-Net) adds standard U-Net skip connections to the M2 architecture, and its AUC drops from 0.9205 to 0.8769. This finding is scoped to reconstruction-based anomaly scoring on our TMP dataset (skip-based architectures remain effective in other formulations) and the exact magnitude of the deficit varies across training runs; what is consistent across runs and skip designs is the ordering skip < no-skip (see the extended ablation below).
- **Role of Attention Mechanism (M4 vs. M2):** M4 (ResNet-CBAM) builds upon the M2 architecture by integrating the CBAM into the bottleneck. The model's ability to filter out background noise improved, which is reflected in the AUC increasing from 0.9205 to 0.9254. This demonstrates that channel and spatial attention effectively focus on critical features. It is important to note that M4 relies purely on MSE loss, making it equivalent to the $\lambda = 0$ configuration discussed in Section 4.2.

- **Effectiveness of Proposed Method (M5 vs. M4):** The final proposed model (M5) builds upon M4 by incorporating the optimal hybrid loss function ($\lambda = 2$) alongside the CBAM. This combination achieved the best overall performance, with the highest AUC (0.9535) and F1-score (0.8607) of all variants. Under the validation-calibrated threshold, M5 detects 105 of 106 faults (recall 0.9906), whereas M4, despite a higher precision (0.8721), misses nearly one-third of the faults (recall 0.7075). This contrast shows that while the CBAM effectively filters noise, the optimally weighted SSIM loss ($\lambda = 2$) is crucial for capturing the structural fault features (such as harmonic patterns) that MSE alone misses, converting them into a consistently separable anomaly score rather than an occasionally confident one.

Based on these configurations, we further analyzed their performance through ROC curves (Figure 5) and confusion matrices (Figure 6) to dissect their robustness and classification stability.

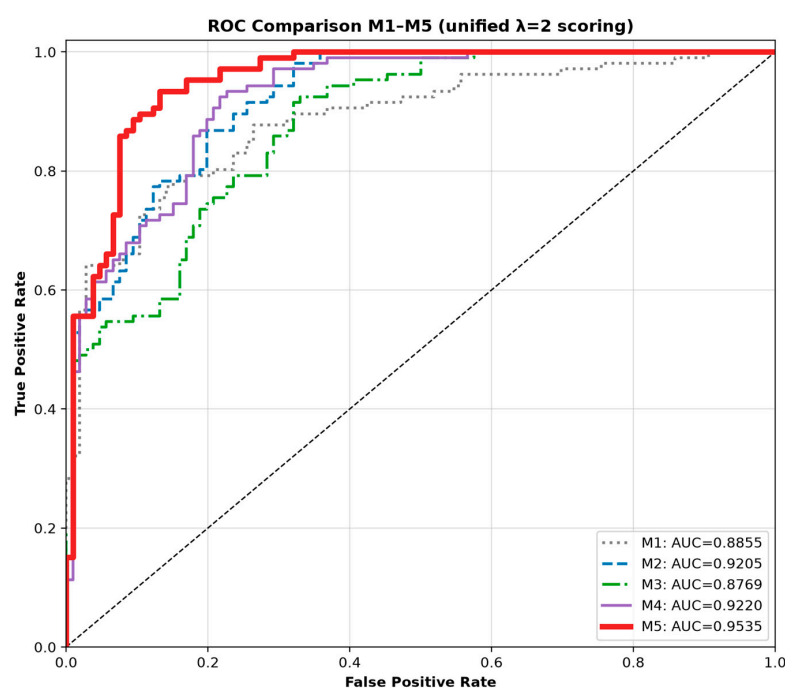


Figure 5. Receiver Operating Characteristic (ROC) curve comparison across the five model variants (M1 to M5) in the ablation study. The proposed M5 model (thick red line) consistently bounds the other variants and achieves the highest AUC of 0.9535, demonstrating superior diagnostic threshold stability under severe cleanroom noise interference.

Threshold Stability and Overall Performance (ROC Analysis): Figure 5 presents the ROC curves of all five variants. M1 is weakest (AUC = 0.8855), while the proposed M5 model envelopes the other curves across almost the entire false-positive-rate range, achieving the highest AUC of 0.9535 and demonstrating stable diagnostic behavior across operating thresholds.

Operating Point Analysis (Confusion Matrix Analysis): In predictive maintenance, a missed fault is far costlier than a false alarm. Under the validation-calibrated threshold (Figure 6), M1 raises only 2 false alarms but misses 57 of the 106 faults (recall 0.4623); M4 is precise but still fault-missing (11 FP, 31 FN); the proposed M5 model converts its superior score separability into near-zero missed faults—105 of 106 detected—at a moderate false alarm cost (33 FP, precision 0.7609). Because M5's score distributions are the best separated of all variants, its operating point can be shifted along the ROC curve to meet any site-specific false alarm budget without retraining.

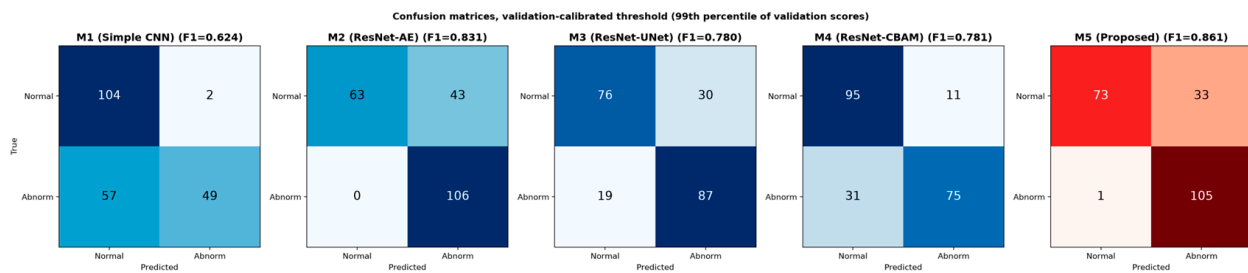


Figure 6. Confusion matrices for the evaluated model variants (M1 to M5) on the testing dataset, using a threshold calibrated on the validation set (99th percentile of validation anomaly scores). The proposed M5 model detects 105 of 106 faults (recall 0.9906); for predictive maintenance, where a missed fault is far costlier than a false alarm, this operating characteristic is preferred. In each matrix, darker cells indicate higher counts; the proposed M5 is shown with a red color scale to distinguish it from the baseline variants (blue).

4.4. Qualitative Analysis: Overcoming Identity Mapping

Figure 7 provides a qualitative reconstruction comparison that visualizes how the proposed framework behaves on an abnormal input, in contrast to a skip connection architecture.

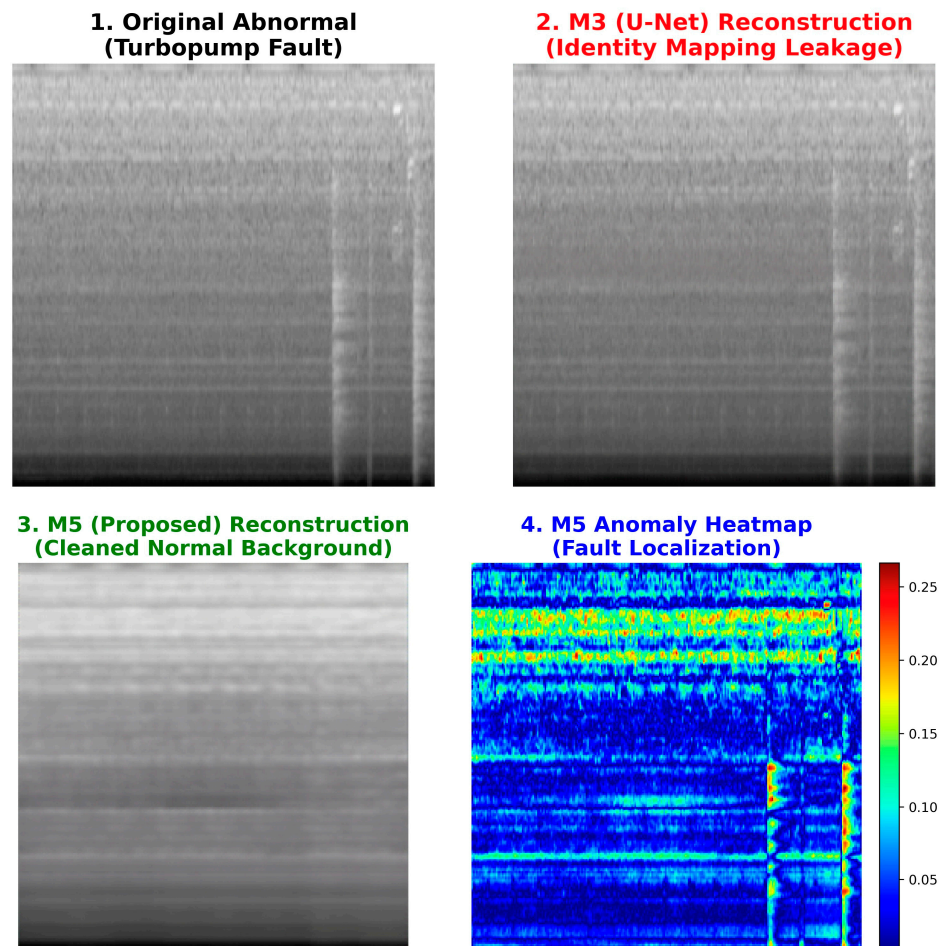


Figure 7. Qualitative comparison of time–frequency spectrogram reconstructions for turbopump fault localization. (Panel 1) shows the original abnormal spectrogram masked by background noise. (Panel 2) demonstrates the “identity mapping leakage” flaw in the M3 (U-Net) architecture, where fault signatures bypass the bottleneck. (Panel 3) illustrates the proposed M5 model’s ability to filter anomalies and reconstruct a cleaned normal background. (Panel 4) presents the final M5 anomaly heatmap, successfully isolating the structural harmonic lines for explainable diagnostics.

- **Analysis of Anomaly Leakage in U-Net (Panel 1 vs. Panel 2):** Panel 1 of Figure 7 displays the original abnormal Log-Mel spectrogram collected from a failing turbo molecular pump. The incipient fault manifests as a series of subtle, continuous horizontal harmonic lines situated in the high-frequency bands, heavily cloaked under the harsh factory background noise.

When the same signal is processed by the ResNet-UNet (M3), the fault signatures reappear almost perfectly in the reconstruction (Panel 2): the skip pathways copy both the background noise and the anomalous harmonic lines directly to the output. This visually confirms the anomaly leakage mechanism described in Section 2.2 and explains M3's reduced AUC (0.8769) in the quantitative ablation.

- **The Efficacy of the Proposed Framework (Panel 3 vs. Panel 4):** In contrast, the proposed M5 model (no skip connections, CBAM bottleneck, $\lambda = 2$) reconstructs a clean, normal-like background spectrogram in which the fault patterns are absent (Panel 3). Information is forced through the constrained bottleneck, the attention mechanism suppresses non-periodic ambient noise, and no shortcut pathway exists through which the anomalous lines could leak.

The absolute pixel-wise difference between the input (Panel 1) and the reconstruction (Panel 3) then yields the anomaly heatmap in Panel 4, in which the residual energy concentrates on the anomalous horizontal lines in the high-frequency region. This suggests that the framework does not merely react to pixel-level intensity variations but learns the structural regularities of normal pump operation, providing a useful visual aid for maintenance diagnostics.

Beyond the qualitative heatmaps, we quantify where the model's decision resides in frequency via an occlusion analysis [16]: each frequency band of a test spectrogram is replaced with the corresponding content of the average normal spectrogram, and the resulting drop in anomaly score measures that band's contribution. The decision is strongly localized (the top 20% of frequency rows account for 41% of the total occlusion importance (Gini coefficient 0.43 versus 0 for a uniform attribution)) and peaks at $\approx 471\text{--}580\text{ Hz}$, adjacent to the machine's dominant tonal component (Figure 8). Crucially, this model-intrinsic attribution correlates with the independent reconstruction error heatmap (Pearson $r = 0.60$), i.e., two different explanation mechanisms agree on the responsible bands, supporting the internal consistency of the localization. Direct mapping of this band onto the pump's characteristic mechanical frequencies (shaft rotation, blade-pass, and bearing defect frequencies) requires the pump's mechanical specifications and is left for future work.

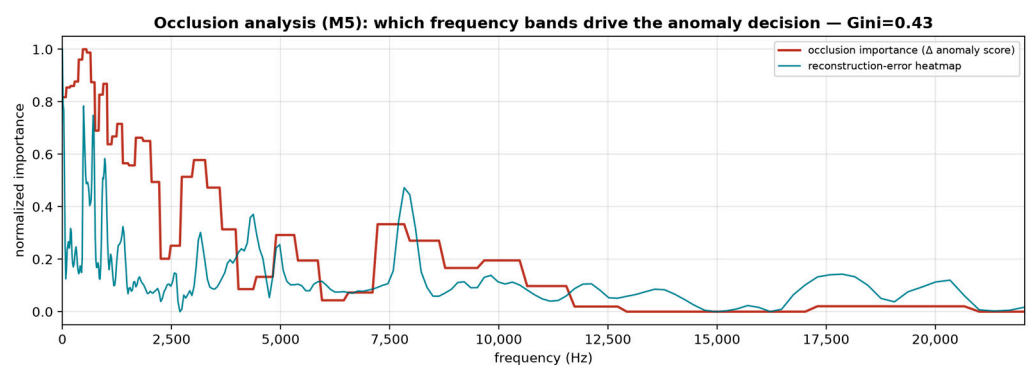


Figure 8. Occlusion-based frequency attribution of the anomaly decision (red) against the reconstruction error profile (blue). The two independent mechanisms agree ($r = 0.60$), with importance concentrated near 471–580 Hz.

4.5. Efficiency Analysis for Real-Time Monitoring

Table 4 presents the trade-off between model complexity, inference speed, and detection performance. M3 and M4 are omitted: M3's computational footprint is nearly identical to M2's, and M4 shares its architecture and inference cost with M5, differing only in the training loss.

Table 4. Performance analysis: parameters, latency, and FPS. The proposed M5 model achieves the highest AUC while maintaining real-time inference speed (330.28 FPS), striking a better balance than the lightweight but less accurate M1.

Model	Parameters	Latency (ms)	Inference Speed (FPS)	AUC
M1	0.19 M	0.5994	1668.47 frames/s	0.8855
M2	13.96 M	2.5725	388.73 frames/s	0.9205
M5	14.00 M	3.0278	330.28 frames/s	0.9535

While M1 offers the highest throughput (1668.47 FPS), its detection performance (AUC = 0.8855) is insufficient for industrial use. The proposed M5 model, despite its 14.00 M parameters, still runs at 330.28 FPS (3.03 ms latency on an NVIDIA RTX 5070), far above the >30 FPS requirement for real-time acoustic monitoring, providing the best balance between detection performance (AUC = 0.9535) and speed.

For deployment-oriented assessment, we additionally measured the complete processing chain per 10 s clip: PCM loading is 13.7 ms, Mel-spectrogram computation is 64.1 ms, image preparation is 42.2 ms, and model inference is 17.3 ms—137 ms in total, i.e., preprocessing dominates (87%) and the end-to-end pipeline runs at approximately $73\times$ in real time. We therefore report the model-only figure of 330.28 FPS (RTX 5070) as an upper bound and the $\approx 73\times$ real-time factor as the deployment-relevant metric. Benchmarking on embedded platforms (e.g., NVIDIA Jetson) is left as future work; the desktop-class results reported here do not by themselves establish edge readiness.

4.6. Comparison with External Baselines

To position the proposed method beyond the internal ablation, we evaluated two representative unsupervised anomaly detection paradigms on identical data: PatchCore, a feature embedding method using ImageNet-pretrained ResNet-18 features (layers 2–3) with a nearest-neighbor memory bank subsampled from the normal training spectrograms, and Deep SVDD, a one-class method trained to map normal spectrograms into a minimal hypersphere. These two methods were selected because they represent the two dominant paradigms of unsupervised anomaly detection—feature embedding with a nearest-neighbor memory bank, and one-class objective learning—and because both operate directly on the same spectrogram inputs without any task-specific supervision, providing strong, architecture-agnostic reference points; recent acoustic anomaly detection systems largely build on these same two principles [7]. Both use only normal data for training and are evaluated on the same test split. PatchCore attains an AUC of 0.9010 and Deep SVDD 0.8742, versus 0.9535 for the proposed M5 model (Figure 9). We attribute the margin to the fact that generic ImageNet patch features and global one-class embeddings are less sensitive to the faint, structured spectral deviations that the hybrid MSE–SSIM reconstruction objective explicitly targets. We note that these baselines use standard rather than maximal configurations (e.g., WideResNet-50 features and full core-set selection for PatchCore); given the ≈ 0.05 AUC margin, the ranking is unlikely to be configuration-dependent.

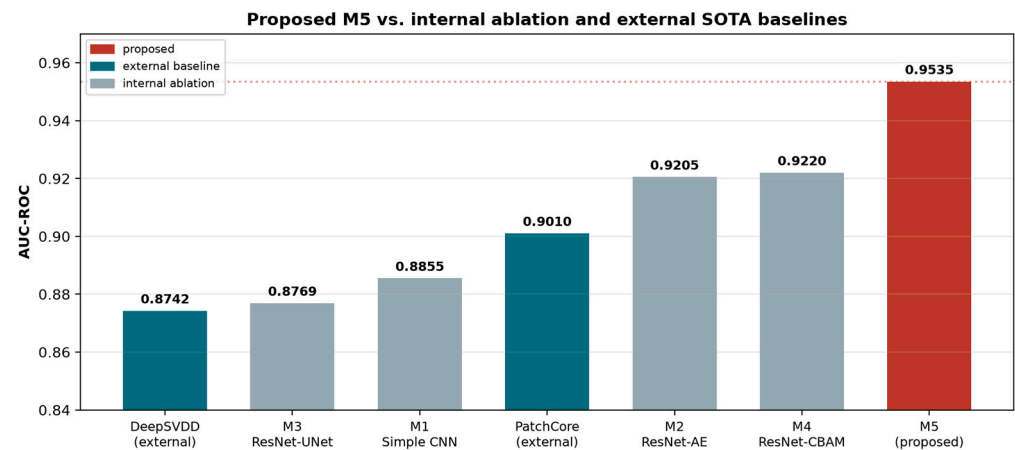


Figure 9. AUC comparison of the proposed M5 model against the internal ablation variants and two external unsupervised baselines (PatchCore, Deep SVDD) trained on the identical normal-only spectrograms. The horizontal dashed line marks the AUC of the proposed M5 (0.9535).

4.7. Robustness Under Controlled Noise Injection

The dataset's inherent high-noise (>80 dBA) background reflects a single acoustic environment; to characterize robustness systematically, we injected additive white noise into the raw test audio (before spectrogram generation) at signal-to-noise ratios from 20 dB down to −5 dB and re-evaluated M2 (no attention), M4 (CBAM), and M5 (proposed) without retraining. The proposed M5 model degrades most gracefully, from 0.9535 (clean) to 0.827 at 20 dB and 0.701 at −5 dB, whereas M2 collapses to chance level (0.488) at −5 dB (Figure 10). The sweep separates the two design contributions: at 0 dB, the CBAM alone lifts AUC from 0.565 (M2) to 0.708 (M4), and the hybrid loss adds a further margin to 0.746 (M5), with the ordering M5 > M4 > M2 stable across all noise levels. This corroborates that the attention bottleneck suppresses noise-dominated features while the structural loss preserves the fault-relevant spectral structure.

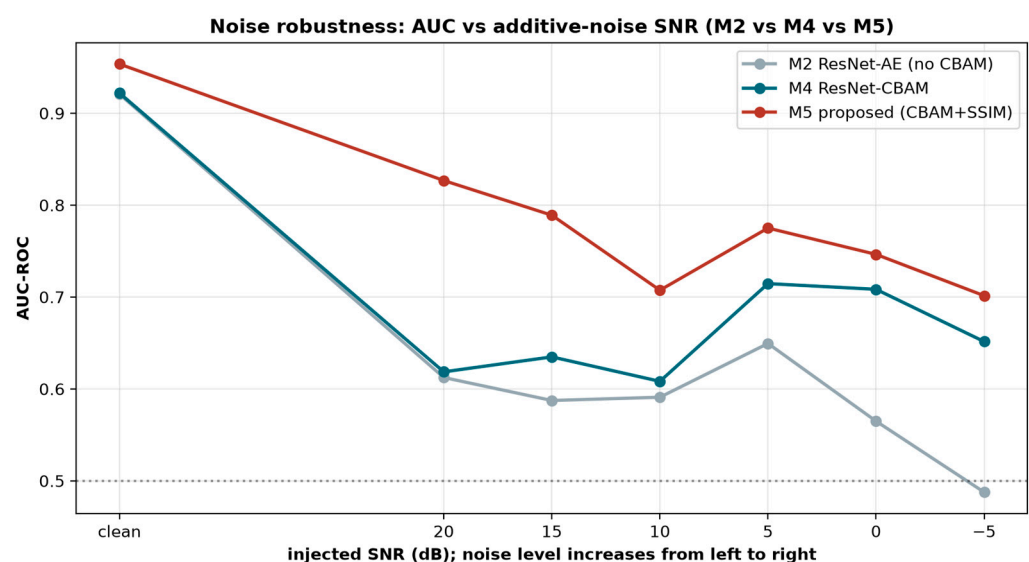


Figure 10. Detection performance versus injected noise level. M5 is the most robust at every SNR; the M4-over-M2 margin at a low SNR isolates the contribution of the CBAM.

5. Conclusions

This study proposes a robust unsupervised anomaly detection system for turbo molecular pumps (TMPs) in high-noise cleanroom environments. By integrating a ResNet-18

autoencoder with the CBAM attention mechanism and an optimally weighted structural similarity (SSIM) hybrid loss function, we successfully addressed the limitations of traditional methods and the “identity mapping” problem found in standard U-Net architectures.

Extensive experimental results demonstrate that our proposed method achieves a superior AUC of 0.9535 and near-perfect fault recall (0.9906) under a validation-calibrated threshold. Averaged over three random seeds, the proposed configuration attains an AUC of 0.9482 ± 0.0029 , confirming that the single-run figure is representative. The removal of skip connections, combined with attention mechanisms, allows the model to effectively filter the intense (>80 dBA) background noise while preventing anomaly leakage. Furthermore, the generated anomaly heatmaps provide interpretable diagnostics, localizing anomalous spectral patterns in the time–frequency domain. With a highly efficient inference speed of 330.28 FPS, the proposed system is well suited for real-time predictive maintenance in semiconductor manufacturing.

All experiments derive from two pumps of the same model (one healthy, one faulty) in a single cleanroom; although the recording-level split precludes leakage and the noise injection study probes changing acoustic conditions, generalization across pump models, rotating speeds, and factory noise profiles remains to be validated. The reported operating points correspond to a validation-calibrated threshold (99th percentile of validation scores), stated explicitly in Section 4.1, and the quantitative localization analysis, while internally consistent across two attribution mechanisms, awaits confirmation against the pump’s mechanical characteristic frequencies. Future work will focus on validating the framework on embedded platforms and across additional pump models, rotating speeds, and factory sites, extending the dataset accordingly.

Author Contributions: Conceptualization, C.-H.L. and C.-M.H.; methodology, C.-H.L. and C.-M.H.; validation, C.-H.L.; investigation, P.-J.C.; resources, C.-M.H., C.-M.W. and C.-C.Y.; data curation, C.-M.H., C.-M.W. and C.-C.Y.; writing—original draft preparation, P.-J.C.; writing—review and editing, C.-H.L., C.-M.H., C.-M.W. and C.-C.Y.; supervision, C.-H.L.; project administration, C.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Science and Technology Council (NSTC), Taiwan, grant number NSTC-115-2221-E-324-016-.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions imposed by the Taiwan Semiconductor Research Institute (TSRI).

Acknowledgments: The authors would like to thank the Taiwan Semiconductor Research Institute (TSRI), National Institutes of Applied Research, for providing the acoustic dataset of the turbo molecular pumps and for their technical support.

Conflicts of Interest: Authors Chien-Ming Wu, Chih-Chyau Yang and Chun-Ming Huang were employed by the Taiwan Semiconductor Research Institute, National Institutes of Applied Research. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Bangcheng, H.; Zan, H.; Xu, Z.; Xu, L.; Tong, W.; Shiqiang, Z. Loss Estimation, Thermal Analysis, and Measurement of a Large-Scale Turbomolecular Pump with Active Magnetic Bearings. *IET Electr. Power Appl.* **2020**, *14*, 1283–1290. [CrossRef]
2. Lee, J.; Yoon, D.; Jeong, G.M.; Kim, H. Acoustic Anomaly Detection for Propeller Crack Defects: A Drone-Based Surrogate Dataset for UAM. *IEEE Access* **2025**, *13*, 195564–195578. [CrossRef]

3. Kim, K.S.; Lee, J.H. Unsupervised Acoustic Anomaly Detection for Rotating Machinery Under Submarine-like Environments: Considering Data Scarcity and Background Noise via Proxy Data Generation. *Sensors* **2026**, *26*, 2659. [[CrossRef](#)] [[PubMed](#)]
4. Xu, H.; Pang, G.; Wang, Y.; Wang, Y. Deep Isolation Forest for Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 12591–12604. [[CrossRef](#)]
5. Gewers, F.L.; Ferreira, G.R.; De Arruda, H.F.; Silva, F.N.; Comin, C.H.; Amancio, D.R.; Costa, L.D.F. Principal component analysis: A natural approach to data exploration. *ACM Comput. Surv.* **2022**, *54*, 70. [[CrossRef](#)]
6. Serin, G.; Sener, B.; Ozbayoglu, A.M.; Unver, H.O. Review of tool condition monitoring in machining and opportunities for deep learning. *Int. J. Adv. Manuf. Technol.* **2020**, *109*, 953–974. [[CrossRef](#)]
7. Lv, Q.; Gong, Y.; Yang, B.; Zhou, Y.; Zhang, L.; Liu, J. A self-supervised acoustic anomaly detection model with a multi-branch attention mechanism. *Eng. Res. Express* **2026**, *8*, 035236. [[CrossRef](#)]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 9351, pp. 234–241. [[CrossRef](#)]
9. Anand, V.; Gupta, S.; Koundal, D.; Singh, K. Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images. *Expert Syst. Appl.* **2023**, *213*, 119230. [[CrossRef](#)]
10. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [[CrossRef](#)] [[PubMed](#)]
11. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 3–19. [[CrossRef](#)]
12. Alirezazadeh, P.; Schirrmann, M.; Stolzenburg, F. Improving Deep Learning-based Plant Disease Classification with Attention Mechanism. *Gesunde Pflanz.* **2023**, *75*, 49–59. [[CrossRef](#)]
13. Fan, J.; Chen, M.; Gu, Z.; Yang, J.; Wu, H.; Wu, J. SSIM over MSE: A new perspective for video anomaly detection. *Neural Netw.* **2025**, *185*, 107115. [[CrossRef](#)] [[PubMed](#)]
14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*; IEEE: New York, NY, USA, 2016; pp. 770–778. [[CrossRef](#)]
15. Sunnetci, K.M.; Kaba, E.; Çeliker, F.B.; Alkan, A. Comparative parotid gland segmentation by using ResNet-18 and MobileNetV2 based DeepLab v3+ architectures from magnetic resonance images. *Concurr. Comput. Pract. Exp.* **2023**, *35*, e7405. [[CrossRef](#)]
16. Elrashid, M.; Deschênes, A.; Subakan, C.; Ravanelli, M.; Georges, R.; Morin, M. Toward Faithful Explanations in Acoustic Anomaly Detection. In *Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; IEEE: New York, NY, USA, 2026; pp. 5276–5280. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.