# MLOps Framework for Fault Diagnosis in Air Conditioners Using Field Noise

Sang Uk Son[1], Yoojeong Noh[2], Sunhwa Park[3], Jangwoo Lee[3]

[1]*Pusan National University, Busan, 46241, Republic of Korea*
*gogandinhand@pusan.ac.kr*

[2]*Pusan National University, EPIC, Busan, 46241, Republic of Korea*
*yoonoh@pusan.ac.kr*

[3]*LG Electronics, Changwon, 51554, Republic of Korea*
*sunhwa1124.park@lge.com*
*jonathan.lee@lge.com*

## ABSTRACT

Fault diagnosis of heating, ventilation and air-conditioning (HVAC) equipment relies increasingly on data-driven models. However, real-world after-service recordings captured by technicians are noisy, imbalanced and often contain meaningless segments. These are labeled by domain experts but sometimes mislabeled. This paper proposes an initial noise-aware machine learning operations (MLOps) framework that enables robust classification, calibration as a prerequisite to uncertainty estimation and continuous improvement of air-conditioner sound diagnostics. The framework performs data preprocessing, uncertainty-based identification of label noise, systematic relabeling through gradient-based class activation maps (Grad-CAM++, hereafter referred to as CAMs), and clustering. A comprehensive metrics tracking facilitates reproducible experiments. Experiments on field recordings demonstrate that removing label noise leads to better generalization, as the learned representation forms more distinct clusters in the logits space, reducing the presence of mislabeled samples within each cluster. The proposed approach yields better generalization and provides a scalable pathway toward automated labeling and open-set recognition.

## 1. INTRODUCTION

Fault diagnosis of consumer air conditioners (ACs) in real-world field conditions is inherently challenging due to both data quality issues and the need for reliable classification performance. Various approaches have been explored for AC fault detection, including sensor-based monitoring of temperature, pressure, and current signals, vibration-based analysis, and visual inspections such as thermal imaging (Taheri et al., 2021; Chen et al., 2022; Matetić et al., 2022; Zhang et al., 2023). Among these, acoustic diagnostics have gained increasing attention for their non-invasive nature, low cost, and suitability for real-time application in operational environments. By capturing characteristic abnormal sounds from components such as compressors, fans, and valves, sound-based methods can identify a wide range of mechanical and structural faults without interfering with normal operation (Tang et al., 2023).

Recent advances in machine learning have further enhanced the capabilities of acoustic fault diagnosis. Convolutional neural networks (CNNs) trained on well-curated acoustic datasets have achieved high accuracy in controlled laboratory settings (Zhang et al., 2025; Liu et al., 2025). However, after-service (A/S) scenarios present a markedly different challenge. In practical field conditions, service technicians often record AC operational sounds using smartphones in noisy and uncontrolled environments, leading to class imbalance, severe background noise, and inaccurate label intervals that degrade model performance and hinder continuous improvement. Furthermore, over the product lifecycle, entirely new fault types that were absent during model training may appear, requiring rapid detection and integration into the diagnostic pipeline.

These conditions present two critical challenges. The first is label noise, caused by human annotation errors or ambiguous acoustic patterns, which can lead to persistent mislabeling and hinder performance improvements even as the dataset grows (Zhang et al., 2024). The second is the emergence of novel acoustic events, which requires distinguishing between closed-set noise—misclassified samples from known classes—and open-set noise, where the sample belongs to a previously unseen fault category (Lundgren & Jung, 2022).

To address both challenges, it is necessary to have a way of quantifying how confident the model is in each prediction. Predictive uncertainty allows us to determine whether a misclassification stems from label noise or from a genuinely novel fault type absent in the training data. From the perspective of uncertainty quantification, aleatoric uncertainty arises from intrinsic variability in the data—such as background noise or overlapping acoustic events—whereas epistemic uncertainty originates from the model's limited knowledge and can typically be reduced by adding high-quality labeled data (Abdar et al., 2021; Kendall & Gal, 2017). However, errors caused by label noise are not mitigated simply by enlarging the dataset.

For this reason, the present study employs the Shannon entropy of the predicted class probability distribution as a compact measure of predictive uncertainty (Shannon, 1948). Low-entropy misclassifications usually correspond to overconfident predictions for the wrong class—a hallmark of closed-set noise—and are reviewed by experts with the aid of Class Activation Maps (CAMs) for visual interpretability (Zhou et al., 2016). Conversely, high-entropy predictions are treated as cases of elevated epistemic uncertainty, flagged as out-of-distribution (OOD) candidates, and subsequently analyzed via clustering to identify potential novel fault categories. This process enables a unified treatment of both data-quality issues and the closed-set/open-set noise problem, while addressing key weaknesses in existing field-deployed HVAC diagnostic systems.

The primary contribution of this work is the development of an MLOps-integrated diagnostic pipeline tailored for field-deployed HVAC systems, which combines entropy-based uncertainty analysis with CAM-assisted expert verification to detect and eliminate closed-set noise. While open-set detection and automated class discovery are recognized as critical for long-term adaptability, they are left as future research directions. By embedding this framework into operational workflows, this study aims to establish a foundation for sustainable, data-driven improvement and enhanced model reliability in the acoustic diagnosis of HVAC systems.

## 2. RELATED WORKS

Recent studies have focused on developing robust, computationally efficient, and practically deployable techniques for handling label noise. The One-step Anti-Noise (OSA) approach, for example, is a model-agnostic strategy that efficiently separates clean and noisy samples within a single inference step, thereby significantly reducing computational overhead in large-scale pre-training and fine-tuning tasks (Li et al., 2024). Likewise, the Tripartite method addresses realistic, large-scale noisy-label scenarios by precisely partitioning samples into clean, noisy, and uncertain subsets, achieving state-of-the-art performance across multiple benchmark datasets (Tang et al., 2024).

In the domain of audio and acoustic analysis, open-set recognition (OSR) remains a critical and unsolved challenge. Recent audio-based OSR research for sound event detection (SED) has explored methods capable of identifying known classes while simultaneously detecting unseen acoustic events by leveraging both spectral and temporal representations of audio signals (Zhang et al., 2022). Another line of work on unseen class discovery in the audio domain proposes a framework that applies novelty detection to learned embeddings in order to cluster and characterize unknown sounds, demonstrating high adaptability in real-world, dynamically evolving soundscapes (Xu et al., 2025).

Beyond classification accuracy, practical deployment requirements have driven the emergence of MLOps frameworks explicitly designed for resilience against noise, operational drift, and adversarial conditions. For instance, a resilience-aware MLOps methodology for medical diagnostic systems incorporates predictive-uncertainty calibration and post-hoc resilience optimization as dedicated stages in the MLOps lifecycle, thereby enhancing robustness to operational disturbances (Kisel et al., 2024). Furthermore, a recent survey on Secure MLOps highlights the vulnerabilities of ML pipelines to adversarial threats such as data poisoning, emphasizing the need for integrated security measures and continuous monitoring as essential components of a reliable MLOps ecosystem (Vasisht et al., 2025).

Unlike prior approaches that address label noise, OSR, and MLOps independently, our framework uniquely integrates these components into a single, cohesive MLOps pipeline for HVAC fault diagnosis. To the best of our knowledge, this is the first framework to combine classifier calibration with entropy-based noise detection in this domain. Our primary contribution is a workflow that first performs classifier calibration before entropy-based noise detection to ensure the reliability of our uncertainty estimations. This allows for the systematic identification of both closed-set and open-set noise, which is then integrated with expert review via CAMs and t-SNE clustering to enable a sustainable, continuously improving diagnostic loop in real-world industrial settings.

## 3. METHODOLOGY

### 3.1 MLOps Pipeline for Continuous Improvement

Figure 1 presents the core contribution of this study—a noise-aware MLOps pipeline specifically designed for sound-based fault diagnosis in HVAC systems. The pipeline operates in a continuous learning loop, starting with data preprocessing to ensure consistent input formats and extract features suitable for downstream modeling. Once the raw recordings are standardized, the data undergo label noise cleaning, a crucial step for reducing aleatoric uncertainty and preventing error propagation in subsequent training cycles.

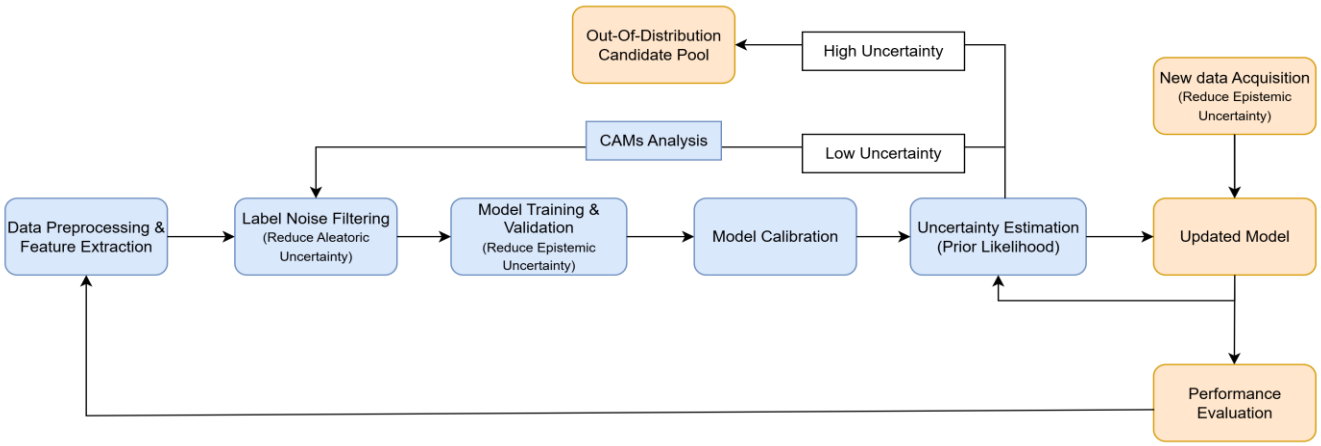


Figure 1. Proposed noise-aware MLOps pipeline for HVAC fault diagnosis

Following noise cleaning, the refined dataset is used for model training. A calibration step is then applied to adjust predictive probabilities, providing a well-calibrated output distribution that serves as a prerequisite for reliable uncertainty estimation. After calibration, epistemic uncertainty is quantified separately and reported alongside conventional performance metrics. This process not only produces a trained and calibrated model but also generates a comprehensive set of artefacts—including loss curves, confusion matrices, t-SNE visualizations of the logit space, and CAMs reports—all of which are systematically archived in the MLOps environment. The model registry stores each trained version, enabling a rollback mechanism that restores a previous model if the performance of a newly trained model fails to meet predefined thresholds.

An additional uncertainty-based filtering process is integrated after each training cycle. Here, softmax entropy is employed to identify potential out-of-distribution samples and closed-set label noise in newly ingested recordings. These anomalous samples are subsequently clustered in the logit space, allowing for clear visualization of their separability from in-distribution data. The clustered results, along with corresponding CAM visualizations, are then presented to domain experts for joint review. By combining cluster-based anomaly grouping with visual evidence from CAMs, experts can more accurately determine whether each sample is an out-of-distribution case or mislabeled instance. This expert-in-the-loop stage not only reduces the review workload by prioritizing the most informative and visually verifiable samples but also accelerates the model evolution cycle by feeding high-quality, expert-verified data back into the pipeline for the next iteration.

Through these integrated components—noise cleaning, calibration, uncertainty estimation, expert-guided refinement, and automated version control—the proposed pipeline in Figure 1 provides a robust, repeatable, and scalable approach to maintaining high generalization while adapting to evolving acoustic fault patterns in real-world environments.

### 3.2 Data Description

The dataset used in this study was collected from in-field recordings of residential HVAC systems, including various indoor unit types such as 1-Way Cassette, 4-Way Cassette, Ducted, and Floor Standing models. Recordings were captured using smartphones with a 44.1 kHz sampling rate and stereo channels during maintenance or inspection performed by technicians. The dataset was labeled into nine classes, including normal operation sounds and fault types such as fan motor noise by LG Electronics.

Table 1. Result of classification on LG HVAC dataset

| Scratched | | | Fine tuning | | |
|---|---|---|---|---|---|
| Accuracy | TOP-2 Accuracy | F1-score (Macro) | Accuracy | TOP-2 Accuracy | F1-score (Macro) |
| 0.70 | 0.81 | 0.59 | 0.78 | 0.86 | 0.66 |

### 3.3 Data Preprocessing and Feature Extraction

Raw sound recordings were segmented to adjust the shape of the model input. Each segment was converted to a Mel-spectrogram with 224 frequency bins and 224 timeframes to leverage ImageNet finetuning. This approach is supported by prior research (Palanisamy et al., 2020), which demonstrates that CNN models pre-trained on ImageNet significantly improve performance in audio classification. For instance, on the ESC-50 dataset, a ResNet50 model's accuracy increased from 85.0% to 87.5%. On our proprietary dataset, applying the same pre-training also improved performance, as detailed in Table 1, with the model's F1 score increasing from 68% to 76% and accuracy from 70% to 78%. A global normalization consistent with ImageNet statistics was applied to stabilize training. Figure 2 illustrates representative Mel-spectrograms extracted from field recordings and their corresponding Class Activation Map (CAM) images. The figure includes spectrogram waveforms from multiple categories of input signals, such as drain pump AC motor noise, fan motor noise,

and voice recordings. As shown, the drain pump and fan motor noise appear highly similar, while the voice class often overlaps with fault-related noise. In such cases, relying solely on Top-1 accuracy may not fully reflect practical decision-making. Therefore, the performance evaluation of our model incorporates not only Top-1 but also Top-2 accuracy, which captures situations where multiple fault types coexist in a single recording. From an application perspective, technicians are advised to review the Top-3 classification results before repairing the air conditioner, ensuring more flexible and reliable after-sales service.

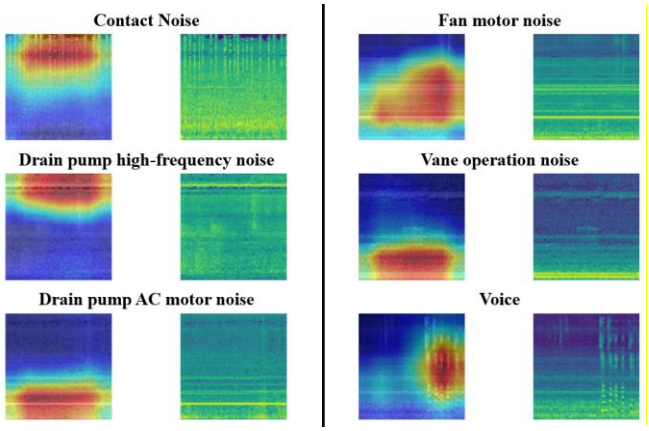


Figure 2. representative Mel-spectrograms & corresponding CAM images.

### 3.3.1 Calibration Prior to Noise Data Selection

Accurate uncertainty estimation requires that the model's predicted probabilities are well-calibrated. Without proper calibration, the softmax entropy values used for noise data detection may be misleading, resulting in incorrect identification of mislabeled or out-of-distribution samples. Therefore, calibration is performed before applying entropy-based filtering. Label smoothing is also applied during training to mitigate over-confidence and support better probability alignment. The calibrated probabilities obtained here form the basis for reliable entropy-based selection of noisy data.

### 3.3.2 Label Noise Detection and Removal

The quantity of information can be expressed following Hartley's definition as

$$I = n \log_2 s \tag{1}$$

Where $s$ denotes the number of possible symbols in the source alphabet and n represents the length of the transmitted sequence. For instance, when transmitting the symbol sequence "Hi" with an alphabet size of 52 (including both uppercase and lowercase letters), we have $n = 2$. To classify an individual symbol such as "H", the alphabet set may be conceptually partitioned in a binary fashion. Starting from the full 52 symbols ordered from uppercase to lowercase, repeated halving identifies "H" within the earlier subset, requiring approximately five iterations for unique determination. For the specific example of "H", the information quantity evaluates to $I \approx 5.7\ bits$.
Entropy can be expressed following Shannon's definition as

$$H(p) = -\sum_i p(x_i) \log_2 p(x_i) \tag{2}$$

In contrast to the Hartley measure, which assumes that all symbols are equally likely, real-world sources often exhibit non-uniform probability distributions. In such cases, the information content of an individual symbol $x_i$ is defined as

$$I(x_i) = \log_2 \frac{1}{p(x_i)} \tag{3}$$

and the overall entropy corresponds to the expected value of this information measure across the distribution of $X$. This formulation accounts for the variability in symbol likelihoods and provides a more general quantification of uncertainty.
Moreover, in the context of model training, the cross-entropy loss between the true label distribution $p$ and the predicted distribution $q$ can be expressed as

$$H(p, q) = H(p) + D_{KL}(p||q) \tag{4}$$

where $H(p)$ denotes the entropy of the label distribution and $D_{KL}(p||q)$ is the Kullback–Leibler divergence between $p$ and $q$. Since $H(p)$ is fixed by the labels, minimizing cross-entropy is effectively equivalent to reducing the KL divergence, thereby decreasing the epistemic error of the model although aleatoric uncertainty inherent in the data remains. From this perspective, softmax entropy itself can also be employed as an indicator of predictive uncertainty depending on the training data distribution. Mislabeled segments are known to degrade classifier performance, and we categorize such label noise into closed-set and open-set cases. For a discrete model-predicted class distribution $q = (q_1, q_2, \ldots, q_k)$, where $q_i$ denotes the predicted probability for class $i$ and $k$ is the total number of classes, the softmax entropy is defined as

$$H(q) = -\sum_i q(x_i) \log_2 q(x_i) \tag{5}$$

An initial classifier is trained, and the entropy of the softmax outputs is computed for each segment. Low-entropy misclassifications occur when the model is highly confident in an incorrect label, which typically corresponds to closed-set noise—cases where the true class exists in the training set but is mislabeled. In contrast, high-entropy misclassifications reflect low confidence and high output uncertainty, often indicating open-set noise, where the input does not belong to any known class. The latter are further examined using t-SNE visualization to detect such out-of-distribution instances. Segments flagged as potentially noisy are reviewed by

domain experts using both CAMs visualizations and t-SNE clustering of outliers. This expert-in-the-loop validation is justified since softmax entropy serves as a training data-dependent indicator of predictive uncertainty, thereby requiring human review to ensure reliable noise detection. This process improves label quality and enhances the model's generalization capability. To quantitatively validate the effectiveness of label noise removal on uncertainty estimation, the Expected Calibration Error (ECE) was measured before and after noise filtering. The ECE decreased from 0.055 to 0.048, indicating that the model became better calibrated and that entropy more accurately reflected predictive uncertainty after removing noisy labels.

### 3.4 Fault Classification Models

In this study, ResNet-34 was adopted as the classification backbone for the 224 × 224 Mel-spectrogram inputs prepared in Section 3.3. ResNet-34 leverages residual connections to mitigate the vanishing gradient problem in deep networks, while offering a balance between model size and computational efficiency, making it well-suited for periodic retraining and real-world deployment (He et al., 2016).

The model was initialized with ImageNet pretrained weights and fine-tuned on the target dataset. To improve training stability and accelerate convergence, the standard input resolution and global normalization were applied. During training, label smoothing was employed to reduce over-confidence, and a probability calibration procedure was performed to enhance the reliability of predicted probabilities. The calibrated outputs were then used for softmax-entropy-based label noise detection, enabling robust identification of mislabeled data and out-of-distribution samples.

By combining preprocessing, probability calibration, and noise-aware learning strategies, ResNet-34 operates reliably within the MLOps pipeline and maintains consistent generalization performance, even on noisy in-field HVAC recordings.

Table 2. Result of classification on LG HVAC dataset after label noise removal for Class 5

| ID | Noise type | F1-Score | ΔF1 (After-Before) |
|---|---|---|---|
| 0 | Refrigerant noise | 0.74 | -0.02 |
| 1 | Contact noise | 0.89 | +0.01 |
| 2 | Drain pump AC motor noise | 0.85 | +0.01 |
| 3 | Drain pump high-frequency noise | 0.71 | -0.01 |
| 4 | Drain pump mechanical noise | 0.32 | +0.03 |
| 5 | Voice | 0.81 | +0.02 |
| 6 | Vane operation noise | 0.96 | +0.14 |
| 7 | Normal operation | 0.46 | +0.23 |
| 8 | Fan motor noise | 0.61 | 0.00 |
| F1-score (Macro) | | 0.70 | +0.04 |
| Accuracy | | 0.79 | +0.01 |
| TOP-2 Accuracy | | 0.88 | +0.02 |

## 4. EXPERIMENTS AND RESULTS

We evaluated the proposed framework on a proprietary dataset of field recordings containing nine known classes plus normal operation, with a pronounced class imbalance. All experiments were conducted using five-fold cross-validation. Label noise removal was applied only to Class 5 (voice), which exhibited the highest proportion of mislabeled samples and whose acoustic characteristics enabled more reliable noise identification.

Removing noisy labels and retraining led to more compact and well-separated clusters in the logits space, thereby reducing mislabeled instances embedded within each class cluster. This effect, illustrated in the t-SNE plots of Fig. 3, reflects improved generalization rather than a mere boost in classification scores. In particular, Class 5 (highlighted in red) shows a distinct decision boundary after noise removal, forming a compact, isolated cluster with minimal intrusion from other classes. In contrast, the noisy-label setting produces a scattered distribution with greater class overlap.

As reported in Table 2, Class 5 achieved an F1 score of 0.81 ($\Delta = +0.02$) after cleaning, but the largest gains occurred in vane operation noise (Class 6: F1 = 0.96, $\Delta = +0.14$) and normal operation (Class 7: F1 = 0.46, $\Delta = +0.23$), indicating reduced confusion with voice samples. Minor decreases were observed for refrigerant noise (Class 0: $\Delta = -0.02$) and drain-pump high-frequency noise (Class 3: $\Delta = -0.01$), while fan motor noise (Class 8) remained unaffected. Overall, the system achieved Macro-F1 = 0.70 ($\Delta = +0.04$), Accuracy = 0.79 ($\Delta = +0.01$), and Top-2 Accuracy = 0.88 ($\Delta = +0.02$), aligning with the tighter Class 5 cluster in Fig. 3 and demonstrating system-wide benefits from targeted label cleaning.

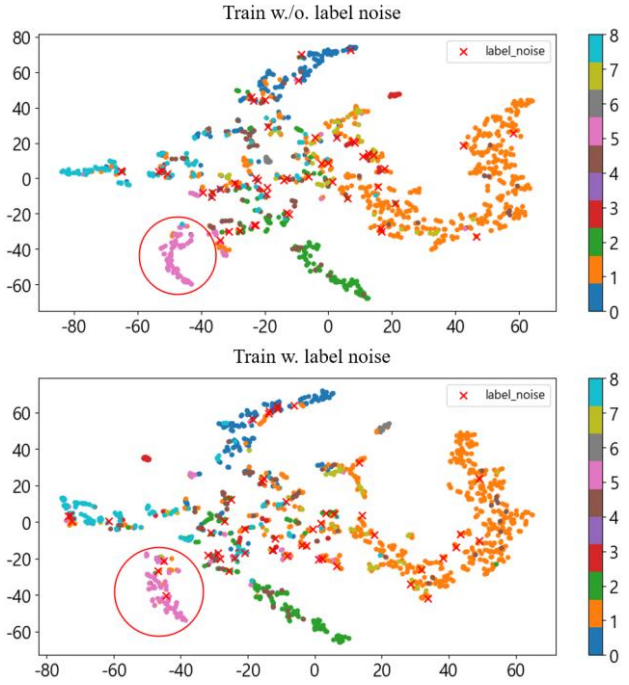


Figure 4. t-SNE visualization of logits space comparing training with and without label noise, highlighting improved separation for Class 5 after noise removal.

## 5. CONCLUSION

The experiments demonstrate that removing label noise yield significant gains in generalization. Entropy-based filtering distinguishes between closed-set and open-set noise, enabling targeted relabeling. The proposed MLOps pipeline automates data ingestion, training, evaluation and model management, providing a practical framework for industrial deployment. Future work will integrate active learning strategies to prioritize segments for expert review and extend the pipeline to other appliance categories.

This paper introduced a noise-aware MLOps framework for fault diagnosis in air-conditioning units. By leveraging entropy-based label noise detection and comprehensive pipeline automation, the framework achieves robust performance on noisy after-service recordings and lays the foundation for open-set recognition. The framework reduces expert labeling effort and accelerates model updates, facilitating sustainable deployment of PHM systems in consumer products.

## ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2020-NR049569 and RS-2024-00341872) and LG Electronics.

## REFERENCES

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... & Makarenkov, V. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76, 243–297.

Chen, J., Zhang, L., Li, Y., Shi, Y., Gao, X., & Hu, Y. (2022). A review of computing-based automated fault detection and diagnosis of heating, ventilation and air conditioning systems. Renewable and Sustainable Energy Reviews, 161, 112395.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

Kisel, M. et al. (2024). Resilience-aware MLOps for AI-based medical diagnostic system. ResearchGate.

Li, H. et al. (2024). One-step Noisy Label Mitigation. arXiv preprint arXiv:2410.01944.

Liu, Y., Xu, Z., He, Y., Guo, P., & Mu, K. (2025). Acoustic fault diagnosis method for rotating machinery based on improved spectral subtraction and CNN–TCN model. *Measurement*, 256(Part E), 11848.

Lundgren, J., & Jung, D. (2022). Data-driven fault diagnosis analysis and open-set classification of time-series data. *Knowledge-Based Systems*, 241, 108276.

Matetić, I., Štajduhar, I., Wolf, I., & Ljubić, S. (2022). A review of data-driven approaches and techniques for fault detection and diagnosis in HVAC systems. *International Journal of Energy Research*, 46(15), 21902–21928.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423.

Hartley, R. V. L. (1928). Transmission of information. *Bell System Technical Journal*, 7(3), 535–563.

Taheri, S., Ahmadi, A., Mohammadi-Ivatloo, B., & Asadi, S. (2021). Fault detection diagnostic for HVAC systems via deep learning algorithms. *Energy and Buildings*, 252, 111453.

Tang, L., Tian, H., Huang, H., Shi, S., & Ji, Q. (2023). A survey of mechanical fault diagnosis based on audio signal analysis. *Measurement*, 220, 113294.

Tang, Y. et al. (2024). Tripartite: Tackling Realistic Noisy Labels with More Precise Partitions. *Sensors*, 25(11), 3369.

Vasisht, A. et al. (2025). Towards Secure MLOps: Surveying Attacks, Mitigation Strategies, and Research Challenges. arXiv preprint arXiv:2506.02032.

Xu, J., Wang, Y., Xu, R., Wang, H., & Zhou, X. (2025). Research on Open-Set Recognition Methods for Rolling Bearing Fault Diagnosis. *Sensors*, 25(10), 3019.

Zhang, B., Zhou, C., Li, W., Ji, S., Li, H., Tong, Z., & Ng, S.-K. (2022). Intelligent Bearing Fault Diagnosis Based on Open Set Convolutional Neural Network. *Mathematics*, 10(21), 3953.

Zhang, B. et al. (2022). Intelligent Bearing Fault Diagnosis Based on Open Set Convolutional Neural Network. Mathematics, 10(21), 3953.

Zhang, F., Nausheen, S., & Sadeghian, P. (2023). Deep learning in fault detection and diagnosis of building HVAC systems: A systematic review with meta analysis. Building and Environment, 238, 110299. https://doi.org/10.1016/j.buildenv.2023.110299

Zhang, Y., Huang, D., & Togneri, R. (2024). Impact of noisy labels on sound event detection: Deletion errors are more detrimental than insertion errors. arXiv preprint, arXiv:2408.14771.

Zhang, Y., Yu, Y., Yang, Z., et al. (2025). Rolling bearing fault identification with acoustic emission signal based on variable-pooling multiscale convolutional neural networks. *Scientific Reports*, 15, 15644.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). Learning deep features for discriminative localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2921–2929.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778.