# Enhancing Industrial Machine Sound Anomaly Detection Using STFT Integrated with DWT and Autoencoder-Based Neural Networks

Tawan Mattanaweerapong * and Kietikul Jearanaitanakij
*Department of Computer Engineering, School of Engineering*
*King Mongkut's Institute of Technology Ladkrabang, Lat krabang, Lat krabang, Bangkok, 10520, Thailand*
***Corresponding Author E-mail:** mtawan.mat@gmail.com
**Received:** Nov 14, 2025; **Revised:** Mar 09, 2026; **Accepted:** Mar 16, 2026

## Abstract

This study proposes a hybrid feature extraction approach that integrates the Discrete Wavelet Transform (DWT) with the Short-Time Fourier Transform (STFT) to improve the accuracy of anomalous sound detection in industrial machines. Conventional STFT-based methods, while effective in representing time–frequency characteristics, exhibit limitations in handling non-stationary noise and transient variations, which often lead to reduced anomaly detection performance in practical industrial environments. To address this problem, the proposed method incorporates multiresolution analysis through DWT, enhancing the system's capability to capture both spectral and temporal information with improved noise robustness. The MIMII dataset (valve, -6 dB, ID02) was used to evaluate the model, where the DWT–STFT feature, representation was applied to an autoencoder for unsupervised anomaly detection. Experimental results demonstrate that the integration of DWT effectively enhanced noise robustness and improved classification metrics, achieving higher AUC and F1-scores compared to the baseline STFT-based approach. In conclusion, the proposed DWT–STFT fusion provides a more resilient and discriminative feature representation, making it a promising technique for practical industrial anomaly detection systems.

**Keywords:** Discrete Wavelet Transform (DWT), Autoencoder, Unsupervised Learning, Machine Sound Anomaly Detection

## 1. Introduction

In modern industrial environments, early detection of anomalous machine sounds is highly beneficial, enhancing machine health monitoring capabilities and enabling efficient maintenance of equipment, while ensuring operational safety and reducing costs [1–2]. Traditional fault detection methods often rely on manual inspections or supervised learning models that require extensive labeled data. Unsupervised learning approaches, particularly AI-based models [3–5], effectively identify abnormal sound patterns without requiring labeled data, offering a scalable solution for data-driven machine condition monitoring.

Autoencoder-based anomalous sound detection offers several advantages for industrial monitoring. By analyzing acoustic signals, these systems can detect anomalies non-intrusively without integrating physical sensors into the machinery, minimizing interference with normal operations. They enable early detection, as subtle changes in machine sounds often precede mechanical faults such as bearing or motor failures. Furthermore, such systems can operate near real-time, continuously capturing and analyzing machine sounds, which allows timely interventions and helps reduce maintenance costs by preventing extensive damage and production downtime. Learning from normal operational data also improves diagnostic accuracy, often surpassing human inspectors in consistency by eliminating fatigue or subjective bias [2],[4],[6].

Among unsupervised methods, autoencoder-based neural networks have demonstrated strong potential for anomaly detection in acoustic data. An autoencoder compresses input data into a lower-dimensional latent space and reconstructs it back to its original form. By training the network exclusively on normal machine sounds, the model becomes sensitive to deviations from typical patterns. Consequently, anomalous sounds yield higher reconstruction errors, providing a robust metric for identifying faults.

In this study, we propose an autoencoder-based framework for machine sound anomaly detection with an enhanced preprocessing pipeline. The primary contribution of this work is the integration of the Discrete Wavelet Transform (DWT) [7–10] as a preprocessing step to improve multi-resolution and transient feature representation prior to time–frequency analysis. Specifically, the DWT-transformed signals are further processed using the Short-Time Fourier Transform (STFT) [11–12] to generate Mel-spectrograms, which are then used as input to a deep autoencoder trained exclusively on healthy machine sound data. While autoencoder-based anomaly detection is a well-established approach, this study systematically investigates the impact of different mother wavelets within the DWT stage under a controlled experimental setting. The results demonstrate that appropriate wavelet selection can enhance anomaly detection performance compared to the baseline configuration without DWT. For fair comparison, the baseline model follows the same architecture, training procedure, hyperparameters, and evaluation protocol, differing only in that the DWT preprocessing step is omitted.

The framework is evaluated using the MIMII Dataset [1]. To ensure a controlled experimental setting and to reduce environmental variability and computational complexity, this study focuses on a single machine configuration and on evaluating the performance of various wavelet types. The scope of the evaluation was also influenced by resource constraints.

Specifically, experiments are conducted on the Valve (machine ID02) subset. Among the four machine types, the valve category has the lowest average baseline AUC, as reported in the original MIMII benchmark study (Valve = 0.53, Pump = 0.68, Fan = 0.70, Slider = 0.70), indicating that anomaly detection for valves is comparatively more challenging and offers greater room for improvement. Furthermore, within the valve category, ID02 contains the smallest dataset among the available configurations, making it a particularly demanding and realistic test case.

Performance is assessed using multiple evaluation metrics, including AUC, precision, recall, and F1-score.

The primary contribution of this study lies in the integration of the Discrete Wavelet Transform (DWT) as a preprocessing step within an autoencoder-based anomalous sound detection framework. In addition, since DWT requires the selection of an appropriate mother wavelet, while the Fourier transform employs a fixed sinusoidal basis, this study systematically evaluates multiple wavelet types to examine their impact on anomaly detection performance.

The remainder of this paper is organized as follows: Section 2 reviews background theory and related work in anomalous sound detection and autoencoder-based methods. Section 3 describes materials and methods, including data preprocessing, feature extraction, and model architecture. Section 4 presents experimental results, and Section 5 discusses the findings and concludes with potential directions for future research.

## 2. Background and Related Work

### 2.1 Autoencoder Neural Networks

An autoencoder is a type of artificial neural network designed to learn compact and efficient representations of input data through an unsupervised learning process [2–5]. It comprises two main components: an encoder and a decoder. The encoder compresses high-dimensional input data into a lower-dimensional latent space, capturing its most salient features and underlying structures. The decoder then reconstructs the original input from this latent representation. The network is trained to minimize the reconstruction error, typically quantified by the Mean Squared Error (MSE) between the original input $x$ and the reconstructed output $\hat{x}$.

Mathematically, given an input vector $x$, the encoder function $f_\theta$ maps the input to a latent representation $z$, as defined in Eq. (1).

$$z = f_\theta(x) \tag{1}$$

The decoder function $g_\phi$ reconstructs the input from $z$, as defined in Eq. (2).

$$\hat{x} = g_\phi(z) \tag{2}$$

The model parameters $\theta$ and $\phi$ are optimized to minimize the loss function, as defined in Eq. (3).

$$L(x, \hat{x}) = |x - \hat{x}|^2 \tag{3}$$

Autoencoders are particularly effective for anomaly detection because they learn to reconstruct data that conforms to the distribution of normal training samples. When exposed to anomalous inputs that deviate from the learned normal patterns, the reconstruction error increases noticeably, providing a natural and reliable indicator for detecting anomalies.

### 2.2 Discrete Wavelet Transform (DWT)

The Discrete Wavelet Transform (DWT) is a time–frequency analysis technique that decomposes a signal into approximation and detail coefficients, effectively capturing both spectral and temporal information across multiple resolutions. It operates based on the concept of a mother wavelet, a prototype function that serves as the foundation for generating a family of wavelets through scaling and translation operations. By compressing or stretching the mother wavelet in time (scaling) and shifting it along the time axis (translation), the DWT enables localized signal analysis at different resolutions [7–10]. High-scale (stretched) wavelets correspond to low-frequency components that represent slowly varying trends, while low-scale (compressed) wavelets correspond to high-frequency components, emphasizing rapid transitions or transient phenomena.

During transformation, the signal is recursively passed through a pair of filters derived from the mother wavelet: a low-pass filter, which extracts the approximation coefficients (A) representing the global trend, and a high-pass filter, which yields the detail coefficients (D) that capture fine-scale variations. This hierarchical, multi-resolution decomposition allows the DWT to efficiently represent both global and local characteristics of a signal, making it particularly effective for analyzing non-stationary data such as acoustic or vibration signals.

The DWT decomposes a discrete-time signal $x[n]$ into approximation coefficients $A_j$ and detail coefficients $D_j$ across different resolution levels, as defined in Eqs. (4)–(5). This is achieved through recursive convolution with low-pass ($h[n]$) and high-pass ($g[n]$) filters, followed by downsampling:

$$A_j[k] = \sum_n x[n]\ h[2k - n] \tag{4}$$
$$D_j[k] = \sum_n x[n]\ g[2k - n] \tag{5}$$

Where $A_j[k]$ represents the low-frequency (approximation) component containing the overall trend or slowly varying information of the signal, while $D_j[k]$

captures the high-frequency (detail) components that emphasize rapid changes, transients, or noise at level $j$.

At each level of decomposition, the approximation $A_j$ can be further decomposed into new sets of approximation and detail coefficients, forming a multi-resolution representation of the signal. This hierarchical structure allows DWT to zoom in on short-duration, high-frequency phenomena (via $D_j$) while also maintaining a broad view of long-term, low-frequency patterns (via $A_j$).

The approximation coefficients $A_j[k]$ are crucial because they preserve the coarse-scale or global structure of the signal — reflecting long-term behaviors, periodicity, or steady-state characteristics. These components describe how the signal evolves slowly over time, allowing the analysis of its general energy distribution or baseline pattern. In the context of mechanical or acoustic signals, the approximation part often corresponds to the normal operating conditions or the underlying vibration trend of the machine.

On the other hand, the detail coefficients $D_j[k]$ play an equally important role by emphasizing short-term variations, impulsive events, and localized discontinuities in the waveform. Such high-frequency fluctuations often indicate abrupt mechanical faults, such as impacts, friction, looseness, or bearing defects, which are not visible in the coarse-scale signal. Therefore, the detail components serve as sensitive indicators for transient and non-stationary phenomena that occur in localized time intervals.

Together, $A_j$ and $D_j$ form a complementary representation of the signal: $A_j$ provides the contextual background, while $D_j$ reveals fine-grained anomalies or dynamic changes. This dual perspective enables a more comprehensive analysis of real-world signals, allowing models to distinguish between normal and abnormal behaviors across multiple frequency scales.

While the Fourier transform provides only global frequency information, DWT applies localized basis functions (wavelets) that adapt to both high- and low-frequency components of the signal. This makes it particularly effective for analyzing non-stationary signals such as machine sounds, where anomalies may occur in localized time intervals. In anomaly detection, DWT enables the extraction of multi-scale features that emphasize transient changes and irregular patterns in acoustic data, improving the model's ability to distinguish between normal and abnormal machine conditions.

**Figure 1** demonstrates the hierarchical decomposition mechanism of the Discrete Wavelet Transform (DWT) [10], which analyzes a signal at multiple frequency scales. The process begins with the original signal being convolved with a pair of low-pass and high-pass filters to extract approximation and detail coefficients, respectively. The approximation coefficients, which retain the low-frequency content, are recursively decomposed into finer sub-bands to capture localized time-frequency information. This iterative filtering and down-sampling structure enables DWT to provide both temporal and spectral resolution, making it effective for analyzing non-stationary signals such as machine sounds or vibration data.

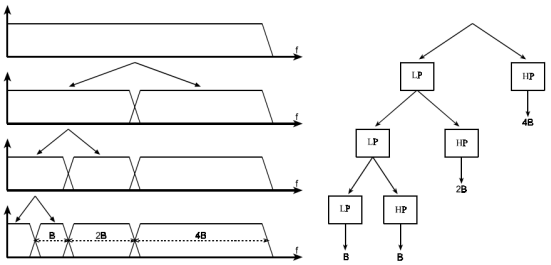


**Figure 1** Splitting the signal with iterated filter banks.

### 2.3 Short-Time Fourier Transform (STFT)

The Short-Time Fourier Transform (STFT) is a widely used method for analyzing non-stationary signals by applying the Fourier transform to short overlapping segments of the input signal [2],[11–12]. Unlike the conventional Fourier Transform, which assumes signal stationarity and provides only global frequency information, STFT divides the signal into shorter time intervals using a sliding window, allowing localized spectral analysis across time. By introducing this windowing process, STFT preserves temporal information while transforming the signal into the frequency domain. The choice of window size is crucial: larger windows provide higher frequency resolution but lower time resolution, whereas smaller windows provide finer time localization but coarser frequency resolution. This inherent trade-off allows STFT to balance between time and frequency representation, making it suitable for signals whose spectral characteristics change over time.

STFT has been extensively applied in sound and speech processing as it provides a compact yet informative representation of time–frequency features. In anomalous sound detection, STFT-derived spectrograms enable machine learning and deep learning models to capture both steady-state and transient acoustic behaviors, improving the detection of irregular events and machine faults.

### 2.4 Mel-Spectrogram

A spectrogram provides a time–frequency representation of an audio signal, illustrating how its spectral content evolves over time. It is typically computed by applying the Short-Time Fourier Transform (STFT) to overlapping frames of the input waveform, thereby decomposing the signal into sinusoidal components localized in both time and frequency domains. This linear-frequency spectrogram effectively visualizes harmonic and transient structures but does not reflect the nonlinear sensitivity of the human auditory system.

While STFT provides a linear frequency representation, the human auditory system perceives sound on a nonlinear scale, approximately logarithmic at higher frequencies. The Mel-spectrogram addresses this by remapping the frequency axis of the STFT output onto the Mel scale, which was designed to approximate the human perception of pitch [13–16]. This perceptually motivated transformation not only compresses the spectral dimensionality but also emphasizes frequency bands most relevant to auditory perception. Furthermore, higher-level acoustic descriptors such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral centroid, and spectral roll-off are frequently extracted from Mel-spectrograms to characterize the timbral and textural features of sound. These features have proven effective in distinguishing normal and abnormal machine operating conditions, as they capture subtle variations in spectral energy distributions that indicate potential mechanical faults.

### 2.5 Anomalous Sound Detection

Anomalous sound detection (ASD) aims to identify abnormal or unusual acoustic patterns that may indicate mechanical faults or system malfunctions. Machine sounds are typically characterized by time–frequency features, which capture both steady-state and transient behaviors associated with normal operation or emerging failures.

In this study, the raw audio signals are transformed into Mel-spectrograms, a perceptually motivated representation that approximates human auditory perception [13–16]. The Mel-spectrogram is obtained by first applying the Short-Time Fourier Transform (STFT) to the audio signal and then mapping the resulting frequency components onto the Mel scale. This transformation preserves the essential spectral characteristics of the sound, reduces data dimensionality, and enhances the model's ability to extract meaningful features for anomaly detection.

### 2.6 Unsupervised Learning for Anomaly Detection

Unlike supervised learning, which requires labeled datasets, unsupervised learning relies solely on the structure and distribution of input data to identify patterns. For anomaly detection, the unsupervised approach assumes that only normal data are available for training, and anomalies are treated as outliers during inference. Autoencoders naturally fit into this paradigm, as they model the normal data distribution during training. When the system encounters data that deviates from the learned patterns, the reconstruction error increases, providing a robust metric for anomaly detection without prior knowledge of fault types [5–6],[17].

This unsupervised approach offers significant advantages in industrial monitoring. It eliminates the need for extensive labeled datasets, which are often costly or impractical to obtain. Moreover, unsupervised autoencoders can continuously adapt to evolving machine behavior, allowing scalable and flexible monitoring of complex systems. By learning from normal operational data alone, the model can detect subtle anomalies that may precede mechanical faults, enable early intervention and reduce maintenance costs. In this work, we employ a Dense Autoencoder, also known as a Fully Connected Autoencoder (FCAE), in which each neuron in a layer is connected to all neurons in the subsequent layer, enabling the network to capture complex nonlinear relationships in the input data.

## 3. Material and Method

### 3.1 Tools and Software

The development and evaluation of the proposed unsupervised anomaly detection system were conducted using the following tools and software:

Python 3.8+: The primary programming language used for model development and data processing [18–19].

Scikit-learn: Used for evaluating performance metrics such as Area Under the Curve (AUC), Precision, Recall, and F1-score [18].

TensorFlow (Keras API): An open-source deep learning library employed to design, train, and evaluate the autoencoder neural network [19].

Librosa: A Python library for audio processing, used for feature extraction including Short-Time Fourier Transform (STFT) and Mel-spectrogram generation [20].

NumPy and SciPy: For numerical computation and signal processing operations, providing efficient array manipulation, linear algebra routines, and advanced mathematical functions [21].

Pandas and Matplotlib: For data manipulation, visualization, and performance analysis [21].

Google Colab: A cloud-based platform used for training and experimenting with deep learning models, leveraging GPU acceleration for faster computation [22–23].

### 3.2 Dataset

The experiments were performed using the MIMII dataset, which provides real-world recordings of industrial machine sounds under both normal and anomalous operating conditions [1]. The dataset was developed by the Research and Development Group, Hitachi, Ltd. in 2019. This dataset includes several machine types, including valves, pumps, fans, and slide rails. For this study, the valve category with -6 dB Signal-to-Noise Ratio (SNR) was selected to simulate a challenging detection scenario in noisy environments.

The MIMII dataset, which stands for Malfunctioning Industrial Machine Investigation and Inspection, was developed to serve as a benchmark for research on anomalous sound detection (ASD) in industrial settings. It contains high-quality recordings collected from actual factory environments, capturing the natural variability and complexity of machine operations. The recordings were made using multiple microphones placed at different positions to reflect real-world acoustic conditions, including the presence of environmental noise, reverberation, and machine-specific sound characteristics.

Each machine type in the dataset exhibits distinct operational sound patterns, which are affected by both the machine's design and its operational status. The dataset includes both normal sound samples, representing correctly functioning machines, and anomalous samples, where specific faults or malfunctions were intentionally introduced. For example, in the valve category, anomalies may include faults such as loose parts, leakages, or irregular internal movements, all of which alter the machine's acoustic signature.

One of the key features of the MIMII Dataset is its inclusion of multiple Signal-to-Noise Ratio (SNR) levels: 6 dB, 0 dB, and −6 dB. Lower SNR levels simulate noisier environments where background noise can obscure machine sounds, making anomaly detection more challenging. According to the benchmark results reported in the original MIMII study, the average AUC decreases as SNR decreases (6 dB = 0.67, 0 dB = 0.61, and −6 dB = 0.53), indicating that −6 dB represents the most challenging condition. Therefore, this study focuses on the −6 dB subset to evaluate the robustness of the anomaly detection model under noisy industrial conditions.

The duration of each audio clip is typically 10 seconds, sampled at 16 kHz, and stored in 16-bit WAV format. The consistent recording length allows for standardized feature extraction and model training processes. Additionally, the dataset provides metadata annotations, including the machine type, operational condition (normal or anomaly), and SNR level, facilitating organized data management and supervised learning tasks.

In this study, the MIMII -6 dB Valve ID02 dataset was employed for experimental evaluation.

Although the experiments focus on the valve category (ID02, −6 dB) to maintain a controlled evaluation setting, the proposed DWT–STFT framework extracts general time–frequency characteristics of acoustic signals rather than machine-specific features. Therefore, the approach has the potential to generalize to other machine types exhibiting similar acoustic anomaly patterns, although the performance may vary depending on the acoustic characteristics of each machine type.

The dataset consists of 708 normal audio files and 120 anomalous audio files. The dataset was partitioned at the file level to prevent data leakage. The dataset consists of 708 normal audio files and 120 anomalous audio files. The dataset was partitioned at the file level to prevent data leakage. The 708 normal recordings were randomly divided into training, validation, and test subsets with a ratio of 80:10:10. Specifically, 80% of the normal recordings were used for training, 10% for validation, and the remaining 10% were included in the test set. All 120 anomalous recordings were strictly reserved for testing. Each audio file was entirely assigned to a single subset, and no frames extracted from the same recording appeared in multiple splits. The normal samples were used for training the autoencoder in an unsupervised manner, while the anomalous samples were reserved for performance evaluation.

By using the MIMII dataset, researchers can develop, train, and evaluate machine learning models under controlled yet realistic industrial conditions. Its diverse and comprehensive design makes it a valuable resource for advancing the state-of-the-art in industrial anomaly detection, contributing towards the development of more reliable predictive maintenance and fault diagnosis systems.

The raw waveform data were normalized prior to feature extraction. No additional waveform-level sliding-window segmentation was applied.

### 3.3 Feature Extraction

Feature extraction plays a crucial role in anomalous sound detection, as it transforms raw acoustic signals into informative representations that highlight meaningful patterns associated with machine health. In this study, a hybrid signal processing pipeline is employed to enhance the discriminative power of extracted features before feeding them into the neural network. The framework integrates wavelet-based preprocessing followed by time–frequency transformation and Mel-scale representation to capture both multi-resolution and perceptually relevant characteristics of machine sounds.

A 5-level Discrete Wavelet Transform (DWT) is first applied to the original 1-D time-domain signal, resulting in one approximation coefficient ($A_5$) and five detail coefficients ($D_1$–$D_5$). Wavelet-based denoising is performed by applying soft thresholding to the detail coefficients ($D_1$–$D_5$) using the universal threshold rule, while the approximation coefficient ($A_5$) is retained without modification. After thresholding, all coefficients ($A_5$ and the processed $D_1$–$D_5$) are reconstructed via inverse DWT to obtain a denoised 1-D time-domain signal. This reconstructed signal serves as the input to the subsequent time–frequency analysis. The decomposition depth was determined empirically by comparing Levels 3, 4, and 5 during preliminary experiments.

Level 3 retained more high-frequency noise, while Level 5 provided improved noise suppression and consistently higher validation AUC and F1 scores.

Therefore, a five-level decomposition was adopted in this study.

It should be emphasized that the proposed framework does not utilize individual wavelet sub-bands separately. Instead, the performance improvement is attributed to wavelet-based noise reduction prior to time–frequency transformation. This design ensures that downstream feature extraction is performed on a noise-reduced yet structurally preserved signal, thereby maintaining methodological consistency with the baseline pipeline.

The denoised signal is processed using the Short-Time Fourier Transform (STFT) with a window length of 4,096 samples and a hop length of 1,024 samples, resulting in a 75% overlap between adjacent frames. The spectrum is mapped onto 256 Mel filter banks to generate the Mel-spectrogram, followed by logarithmic scaling for dynamic range compression.

Additional sensitivity analysis was performed to evaluate the impact of STFT parameter choices. Alternative configurations (e.g., N_FFT values of 1,024,

2,048 and 4,096 with hop lengths including 256, 512, 1,024 and 2,048) were tested prior to finalizing the training pipeline. Although absolute detection performance varied slightly, the relative performance ordering of wavelet families remained stable, indicating that the primary conclusions regarding mother wavelet selection are robust to reasonable variations in time–frequency resolution.

Excessively small hop sizes (e.g., 256 samples) were observed to degrade performance, likely due to increased temporal redundancy and reduced discriminative spectral contrast. Therefore, the 4,096/1,024 configuration was selected as a balanced and stable setting. FFT sizes larger than 4,096 were not evaluated due to computational memory limitations.

The STFT inherently generates overlapping time frames according to the hop length configuration, producing a fixed-size time–frequency representation for each recording. No additional waveform-level sliding-window segmentation is applied; therefore, each 10-second recording is treated as a single training sample.

For reproducibility and clarity, the STFT configuration and derived frame characteristics are summarized below.

Summary of STFT and Segmentation Parameters

- Window length: 4,096 samples
- Hop length: 1,024 samples
- Overlap: 75%
- Clip duration: 10 seconds
- Sampling rate: 16 kHz
- Samples per clip: 160,000
- STFT padding mode: center=True (zero-padding of 2,048 samples at both boundaries)
- Partial segments: automatically handled via built-in zero-padding during STFT computation
- Time frames per clip: 157

Given the STFT configuration and segmentation parameters described above, the resulting time–frequency representations and their corresponding dimensional changes can be systematically determined. The complete dimensional flow of the proposed pipeline is summarized in **Table 1**.

**Table 1** Dimensional transformation across the proposed signal processing pipeline.

| Processing Stage | Representation | Matrix Dimension |
|---|---|---|
| Input (.wav) | Raw audio | 160,000 × 1 |
| DWT output | Reconstructed waveform | 160,000 × 1 |
| STFT output | Complex spectrogram | 2,049 × 157 |
| Mel-spectrogram output | Mel-scaled power spectrogram | 256 × 157 |
| Flattening | 1D feature vector | 40,192 × 1 |
| Autoencoder input | Training dataset matrix | 32 × 40,192 |
| Dense projection | First hidden layer | 32 × 2,048 |

The autoencoder was trained using mini-batch optimization with a batch size of 32. Each sample is represented as a 1D feature vector of length 40,192. Therefore, although the full training dataset consists of N=708 normal samples, the model processes 32 samples at a time during each forward pass. Consequently, the input to the network at each iteration has the dimensionality 32 × 40,192. This batch input is then linearly transformed by the weight matrix of the first dense layer 40,192 × 2,048, resulting in an output representation of size 32 × 2,048.

Since the baseline configuration already corresponds to the conventional STFT + Mel-spectrogram pipeline, the comparison with the proposed framework that incorporates DWT preprocessing prior to STFT and Mel-spectrogram extraction effectively represents an ablation experiment isolating the impact of wavelet-based preprocessing.

### 3.4 Autoencoder Model Architecture

The autoencoder model consists of an encoder and decoder implemented using fully connected (dense) layers:

The encoder network $E(\cdot)$ consists of four fully connected (FC) layers:

- FC(2048, 1024, LeakyReLU),
- FC(1024, 512, LeakyReLU),
- FC(512, 256, LeakyReLU), and
- FC(256, 256, LeakyReLU).

The output of the encoder is a 256-dimensional latent vector $z \in R^{256}$, representing the compressed feature space.

The decoder network $D(\cdot)$ mirrors the encoder structure and reconstructs the input from the latent representation through

- FC(256, 256, LeakyReLU),
- FC(256, 512, LeakyReLU),
- FC(512, 1024, LeakyReLU), and
- FC(1024, 2048, Linear).

LeakyReLU activation functions were used to improve the learning capacity of the model, while batch normalization and dropout (dropout rate $0.1$) were employed to prevent overfitting and stabilize training.

### 3.5 Training Procedure

The model was trained exclusively on normal sound samples using the following initial configurations:

Optimizer: Adam with a learning rate of 0.0005.

Batch Size: 32 samples per training batch.

Epochs: 100 full training cycles.

Early Stopping: Applied based on validation loss to prevent overfitting.

Loss Function: Mean Squared Error (MSE) between input and reconstruction.

The pipeline for training is illustrated in **Figure 2**.

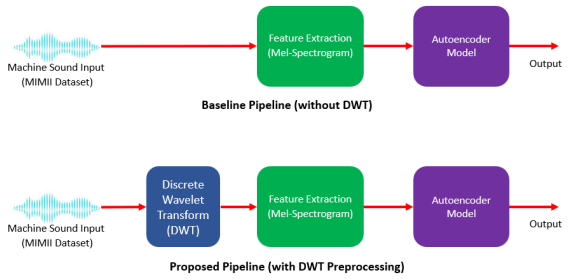


**Figure 2** Training pipeline

The diagram in **Figure 2** illustrates the processing pipeline for machine sound analysis using the MIMII dataset. The upper path represents the baseline approach without wavelet transformation, where the raw signal is directly converted into Mel-spectrogram features before being processed by the autoencoder. The model output is then used to evaluate the system's performance.

The lower path shows the integration of the Discrete Wavelet Transform (DWT) prior to feature extraction, where the wavelet-transformed signal is converted into Mel-spectrogram features and then fed into a dense autoencoder for anomaly detection.

The autoencoder was trained using mini-batch optimization with a batch size of 32. Each audio sample was represented as a 40,192-dimensional feature vector obtained from the flattened Mel-spectrogram. Although the full training dataset consisted of 708 normal samples, the model processed 32 samples at a time during each forward pass, resulting in an input matrix of size 32×40,192 per iteration.

The network parameters were optimized using the Adam optimizer with default momentum coefficients. The model was trained for a maximum of 100 epochs. To mitigate overfitting and improve generalization, early stopping was employed based on validation loss, with training terminated if no improvement was observed for a predefined patience period.

The dataset was strictly partitioned into training, validation, and test subsets prior to experimentation. The validation set was used for monitoring training progress, hyperparameter tuning, and wavelet performance evaluation. To prevent data leakage and inadvertent tuning on the test set, all model selection decisions were made exclusively based on validation performance. The test set remained completely unseen throughout model development and was used only once for final performance reporting.

### 3.6 Anomaly Detection

During inference, the reconstruction error for each input sample was computed using the same Mean Squared Error (MSE) formulation defined in Eq. (3), measuring the difference between the original input and the reconstructed output. This reconstruction error was directly used as the anomaly score, following the standard reconstruction-based autoencoder anomaly detection approach.

Since the model was trained exclusively on normal machine sound data, it learns the distribution of healthy operating conditions. Therefore, anomalous inputs that deviate from this learned distribution result in significantly higher reconstruction errors.

Anomaly detection was then performed by comparing the reconstruction error against a threshold value determined from the validation set. This threshold represents the maximum acceptable reconstruction error observed for normal samples, providing a statistical boundary for normal behavior. Samples with reconstruction errors exceeding this threshold are considered to deviate significantly from the normal pattern and are therefore classified as anomalous. By leveraging the reconstruction error in this way, the system can identify subtle deviations caused by machine faults or abnormal operating conditions without requiring labeled anomalous data, making it highly suitable for unsupervised anomaly detection in industrial environments.

### 3.7 Mother Wavelet Selection and Implementation Framework

Wavelets are mathematical functions that decompose signals into different frequency components while preserving temporal information, making them particularly suitable for analyzing non-stationary signals. A mother wavelet serves as a prototype function, and by scaling and translating it, a family of wavelets is generated, enabling multi-resolution analysis. Different wavelet families possess distinct properties such as symmetry, compact support, vanishing moments, and smoothness, which influence their ability to represent and extract meaningful features from signals. **Figure 3** presents examples of wavelet functions.

To ensure a comprehensive and unbiased evaluation, multiple wavelet families with diverse mathematical characteristics were considered, including compact support (Daubechies: db), near symmetry (Symlets: sym), enhanced vanishing moments (Coiflets: coif), smooth frequency-domain behavior (Discrete Meyer: dmey), and linear-phase reconstruction (Biorthogonal: bior; Reverse Biorthogonal: rbio) [7–10]. These properties are well established in classical wavelet theory and are particularly relevant to audio signal decomposition and reconstruction tasks.

In this study, six wavelet families (db, sym, coif, dmey, bior, rbio), comprising a total of 97 mother wavelets, were systematically applied as part of the preprocessing pipeline to investigate their influence on feature representation for anomaly detection. By exploring different wavelet characteristics, the study aims to analyze how wavelet selection affects the autoencoder's ability to capture salient patterns in machine sound signals.

All wavelet families were implemented within the discrete wavelet transform (DWT) framework using multilevel decomposition. The continuous wavelet transform (CWT) was not employed, as the objective of this study is compact signal representation and

reconstruction rather than continuous time–frequency analysis.

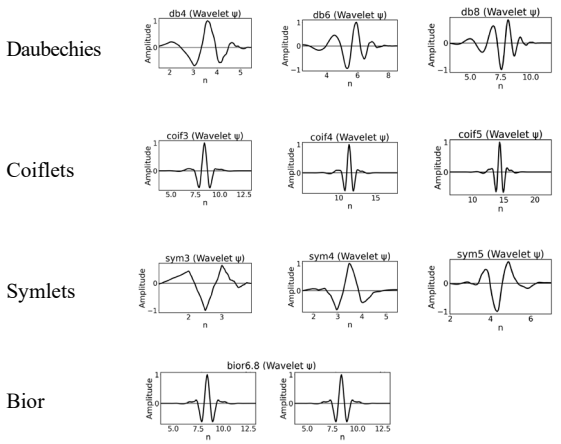


**Figure 3** Sample of wavelet signal

### 3.8 Threshold Selection Strategy

In unsupervised anomaly detection, the anomaly decision boundary plays a critical role, as it directly affects precision and recall. Therefore, the anomaly threshold in this study was defined using a fully reproducible and statistically grounded procedure.

After training the autoencoder using only normal training samples, reconstruction errors were computed for each validation sample using Mean Squared Error (MSE), as defined in Eq. (6).

$$e_i = \frac{1}{d}\sum_{j=1}^{d}\left(x_{ij} - \widehat{x_{ij}}\right)^2 \tag{6}$$

where:

$x_{ij}$ denotes the original feature value,

$\widehat{x_{ij}}$ is the reconstructed value, and

$d$ is the feature dimension.

To ensure numerical stability and comparability across configurations, the reconstruction errors were standardized using z-score normalization, as defined in Eq. (7).

$$z_i = \frac{e_i - \mu_e}{\sigma_e} \tag{7}$$

where $\mu_e$ and $\sigma_e$ represent the mean and standard deviation of reconstruction errors in the validation set.

The anomaly threshold was then determined using a Receiver Operating Characteristic (ROC)-based approach. Specifically, the threshold was selected by maximizing Youden's J statistic, as defined in Eq. (8)

$$J(\tau) = TPR(\tau) - FPR(\tau) \tag{8}$$

where TPR and FPR denote the true positive rate and false positive rate at threshold $\tau$. The optimal threshold is defined as Eq. (9)

$$\tau^* = \arg\max_{\tau} J(\tau) \tag{9}$$

This criterion identifies the operating point that provides the best trade-off between sensitivity and specificity.

Importantly, the threshold $\tau^*$ was determined exclusively using the validation set and was subsequently applied to the test set without further adjustment. This protocol prevents data leakage and ensures an unbiased evaluation of detection performance.

### 3.9 Evaluation Metrics

To quantitatively evaluate the model, multiple metrics were computed from the test set using the reconstruction error distribution. These include:

- AUC (Area Under ROC Curve): Measures the model's ability to distinguish between normal and anomalous data.
- Precision and Recall: Evaluate classification quality, especially for imbalanced datasets.
- F1 Score: Harmonic mean of Precision and Recall.

### 3.10 Experimental Setup

To ensure full reproducibility, the key experimental settings are summarized below.

- Implementation Environment
  - Python 3.10
  - TensorFlow 2.13 (Keras API)
  - Librosa 0.10
  - NumPy 1.24
  - Scikit-learn 1.3
  - Matplotlib 3.7
  - Model training was conducted using Google Colab Pro with access to a CUDA-enabled NVIDIA GPU environment.
- Dataset Configuration
  - Experiments were performed using the MIMII Dataset (Valve ID02, −6 dB SNR condition).
  - Audio duration: 10 seconds
  - Sampling rate: 16 kHz (160,000 samples per clip)
  - Normal samples: 708
  - Anomalous samples: 120
  - The dataset was partitioned at the file level with an 80:10:10 split for normal recordings (training:validation:test), while all anomalous recordings were reserved exclusively for testing.
- Wavelet Processing
  - Discrete Wavelet Transform (DWT)
  - Decomposition level: 5
  - Wavelet families evaluated: db, sym, coif, dmey, bior, rbio
  - Soft thresholding applied to detail coefficients ($D_1$–$D_5$)
  - Reconstruction via inverse DWT using A5 and processed $D_1$–$D_5$ components
- Time–Frequency Representation
  - FFT size: 4,096
  - Hop length: 1,024 (75% overlap)
  - Mel filter banks: 256
  - STFT padding mode: center=True

- Each Mel-spectrogram had a dimension of 256 × 157 and was flattened into a 40,192-dimensional feature vector before being fed into the autoencoder.

- Model Configuration
  - A fully connected dense autoencoder was employed:
  - Latent dimension: 256
  - Batch size: 32
  - Optimizer: Adam
  - Learning rate: $5 \times 10^{-4}$
  - Maximum epochs: 100
  - Loss function: Mean Squared Error (MSE)
  - Early stopping based on validation loss
  - Threshold Selection and Evaluation
  - The anomaly detection threshold was determined from the validation set using ROC analysis. The optimal threshold was selected by maximizing Youden's J statistic (J = TPR − FPR).

## 4. Result

### 4.1 Discrete Wavelet Transform of MIMII file

When the original signal is processed using the Discrete Wavelet Transform (DWT), it can be decomposed into multiple sub-signals that represent different frequency components at various resolutions. This hierarchical decomposition enables the analysis of both time and frequency characteristics of the signal simultaneously.

As shown in the **Figure 4**, the original signal is successively decomposed into one approximation component ($A_5$) and several detail components ($D_1$–$D_5$). The approximation signal ($A_5$) captures the low-frequency content, representing the overall trend of the signal, while the detail signals ($D_1 - D_5$) capture high-frequency variations and transient features at progressively finer scales. This multi-level structure allows for effective signal denoising, feature extraction, and anomaly detection in time-series analysis.

**Figure 4** illustrates the five-level Discrete Wavelet Transform (DWT) decomposition of the original machine sound signal using the coif5 wavelet. The top panel shows the raw signal, which contains both low- and high-frequency components mixed together. After applying the DWT, the signal is successively divided into one approximation component ($A_5$) and five detail components ($D_1$–$D_5$). The approximation ($A_5$) represents the smoothed, low-frequency structure of the signal, where most of the overall energy is preserved.

In contrast, the detail components capture higher frequency variations at each corresponding level — with $D_1$ containing the finest details and higher frequency transients, and $D_5$ reflecting lower-frequency fluctuations. This hierarchical separation enables a clearer observation of subtle or localized changes that may be masked in the original signal, making DWT a powerful tool for analyzing non-stationary or noisy time-domain data.

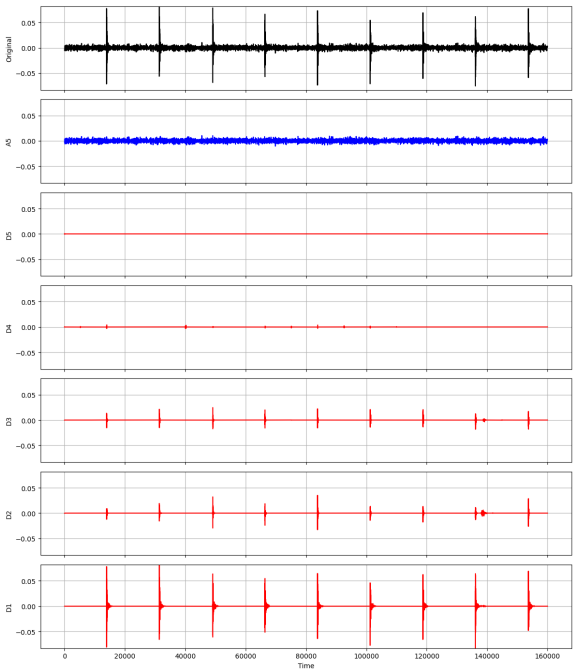


**Figure 4** Wavelet decomposition of the -6dB Valve ID02 MIMII sound signal using the coif5 wavelet.

This hierarchical decomposition demonstrates how wavelet transform separates the signal into multi-resolution components, enabling simultaneous analysis of both long-term patterns and localized high-frequency events. Such representations are valuable in machine condition monitoring and anomaly detection, as they enhance the visibility of abnormal events that may not be clearly observable in the raw waveform.

### 4.2 Mel Spectrogram

**Figure 5** illustrates the comparison of Mel-spectrograms of a MIMII signal without (top) and with coif5 wavelet preprocessing (bottom). The top spectrogram, generated without wavelet preprocessing, exhibits a relatively dense distribution of spectral components across the entire frequency range. While the periodic structure of the machine sound is visible, the background energy appears more widespread, resulting in a less distinct separation between dominant frequency bands and noise components.

In contrast, the bottom spectrogram, obtained using the Coiflet-5 (coif5) wavelet transform, demonstrates a clearer representation of the signal. The major harmonics and transient events are more sharply emphasized, particularly in the mid-to-high frequency regions, while irrelevant background energy is significantly suppressed. This enhancement highlights the periodic structure of the machine operation more distinctly, which facilitates more effective feature extraction for anomaly detection.

Overall, the comparison suggests that wavelet preprocessing (coif5) improves the time–frequency representation of the spectrogram and produces cleaner feature patterns. It should be noted that **Figure 5** provides a qualitative visualization of the preprocessing effect.

The quantitative impact of this improvement is evaluated using anomaly detection performance metrics (AUC and F1-score) reported in **Table 2** section 4.

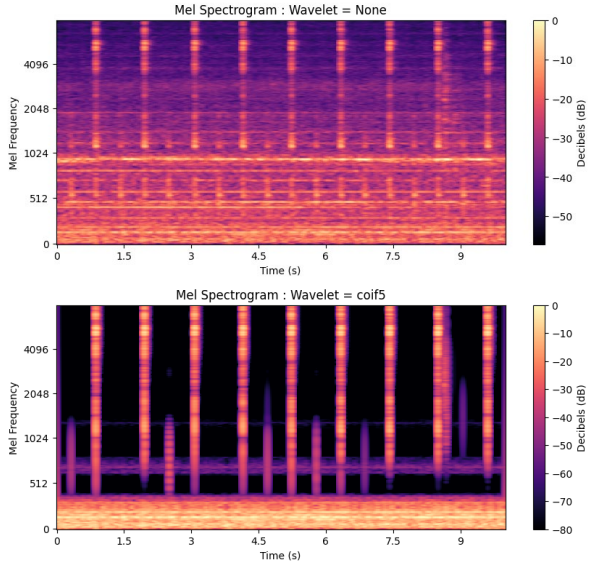


**Figure 5** Comparison of Mel-spectrograms of MIMII signal without (top) and with coif5 wavelet (bottom) preprocessing

### 4.3 Comparison Between No-Wavelet and DWT-Based Approaches

After obtaining the reconstructed signals from both cases, namely without applying the Discrete Wavelet Transform (DWT) and with DWT using various mother wavelets, the model performance was evaluated using several metrics, including AUC, Accuracy, F1-score, Precision, and Recall. The results were organized by placing the baseline case (No wavelet) at the top, followed by the DWT-based results sorted in descending order of AUC. To improve experimental reliability, each configuration was executed three times with different random initializations, and the averaged performance values are reported in **Table 2**.

From the experimental results summarized in **Table 2**, it is evident that applying the Discrete Wavelet Transform (DWT) as a preprocessing step generally enhances the model's performance across most metrics, including AUC, Accuracy, and F1-score, compared with the baseline case without wavelet transformation (AUC = 0.6105).

To provide a comprehensive evaluation of the proposed anomaly detection framework, both Receiver Operating Characteristic (ROC) curves and Precision–Recall (PR) curves are reported. While single-point metrics such as Accuracy and F1-score offer a summary of performance at a selected threshold, they do not fully characterize the trade-off between detection sensitivity and false alarm rate.

The ROC and PR curves shown in **Figures 6–7** are generated from a representative run, whereas the reported performance metrics correspond to the average results obtained over three independent experiments.

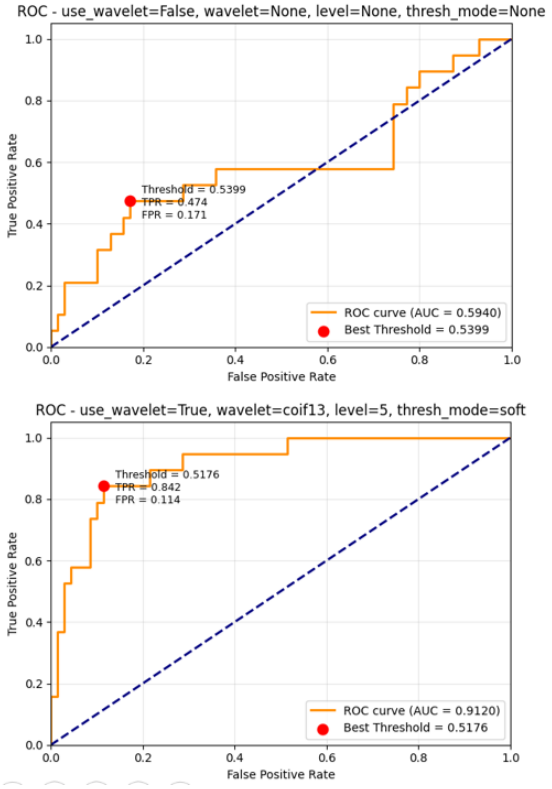


**Figure 6** ROC curves comparing baseline and DWT-based models. The wavelet-enhanced configuration achieves higher AUC, indicating improved anomaly discrimination.

**Table 2** Performance comparison on the MIMII dataset between the baseline model (without DWT preprocessing) and DWT-based models using various mother wavelets, sorted by AUC.

| Rank | Baseline | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| - | No wavelet | 0.6105 | 0.7191 | 0.4267 | 0.3806 | 0.4912 |
| **Rank** | **With Wavelet** | **AUC** | **Accuracy** | **F1** | **Precision** | **Recall** |
| 1 | coif13 | 0.9135 | 0.8727 | 0.7259 | 0.6719 | 0.7895 |
| 2 | coif5 | 0.9133 | 0.9101 | 0.7778 | 0.8235 | 0.7368 |
| 3 | sym13 | 0.9058 | 0.8951 | 0.7628 | 0.7381 | 0.7895 |
| 4 | sym15 | 0.9028 | 0.8577 | 0.7032 | 0.6341 | 0.7895 |
| 5 | coif11 | 0.8985 | 0.8539 | 0.6977 | 0.6250 | 0.7895 |
| 6 | sym16 | 0.8967 | 0.8764 | 0.7317 | 0.6818 | 0.7895 |

**Table 2** Performance comparison on the MIMII dataset between the baseline model (without DWT preprocessing) and DWT-based models using various mother wavelets, sorted by AUC. (cont.)

| Rank | With Wavelet | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| 7 | coif14 | 0.8925 | 0.8427 | 0.6818 | 0.6000 | 0.7895 |
| 8 | sym19 | 0.8900 | 0.8652 | 0.7143 | 0.6522 | 0.7895 |
| 9 | coif3 | 0.8897 | 0.8539 | 0.6977 | 0.6250 | 0.7895 |
| 10 | bior5.5 | 0.8882 | 0.8839 | 0.7439 | 0.7035 | 0.7895 |
| 11 | bior3.7 | 0.8872 | 0.8652 | 0.7143 | 0.6522 | 0.7895 |
| 12 | bior3.9 | 0.8835 | 0.8502 | 0.6879 | 0.6222 | 0.7719 |
| 13 | bior3.5 | 0.8754 | 0.8539 | 0.6977 | 0.6250 | 0.7895 |
| 14 | coif17 | 0.8752 | 0.8502 | 0.6692 | 0.6278 | 0.7193 |
| 15 | sym20 | 0.8717 | 0.8502 | 0.6783 | 0.6284 | 0.7369 |
| 16 | rbio3.9 | 0.8712 | 0.7790 | 0.6266 | 0.5182 | 0.8246 |
| 17 | DB38 | 0.8707 | 0.8652 | 0.7143 | 0.6522 | 0.7895 |
| 18 | rbio5.5 | 0.8699 | 0.8689 | 0.7201 | 0.6621 | 0.7895 |
| 19 | sym9 | 0.8663 | 0.8884 | 0.6867 | 0.7222 | 0.6692 |
| 20 | DB12 | 0.8634 | 0.7715 | 0.6442 | 0.5538 | 0.8421 |
| 21 | coif15 | 0.8629 | 0.7903 | 0.6269 | 0.5282 | 0.7895 |
| 22 | DB9 | 0.8626 | 0.8453 | 0.6447 | 0.5879 | 0.7168 |
| 23 | rbio6.8 | 0.8611 | 0.7790 | 0.6289 | 0.5232 | 0.8246 |
| 24 | coif10 | 0.8601 | 0.8090 | 0.6412 | 0.5539 | 0.7719 |
| 25 | sym17 | 0.8566 | 0.8239 | 0.6429 | 0.5705 | 0.7369 |
| 26 | DB16 | 0.8544 | 0.8127 | 0.6656 | 0.5846 | 0.8070 |
| 27 | coif8 | 0.8540 | 0.8528 | 0.6569 | 0.6095 | 0.7168 |
| 28 | coif4 | 0.8529 | 0.8240 | 0.6446 | 0.5739 | 0.7369 |
| 29 | DB15 | 0.8503 | 0.8188 | 0.6069 | 0.5270 | 0.7168 |
| 30 | DB17 | 0.8497 | 0.8338 | 0.6286 | 0.5618 | 0.7168 |
| 31 | DB23 | 0.8490 | 0.8258 | 0.6203 | 0.5507 | 0.7168 |
| 32 | DB10 | 0.8481 | 0.8413 | 0.6402 | 0.5815 | 0.7168 |
| 33 | sym6 | 0.8479 | 0.8127 | 0.6602 | 0.5827 | 0.7895 |
| 34 | demy | 0.8466 | 0.8240 | 0.6464 | 0.5777 | 0.7369 |
| 35 | bior2.8 | 0.8461 | 0.8389 | 0.6572 | 0.6119 | 0.7193 |
| 36 | rbio4.4 | 0.8426 | 0.8277 | 0.6522 | 0.5875 | 0.7369 |
| 37 | bior2.6 | 0.8409 | 0.8427 | 0.6578 | 0.6311 | 0.7017 |
| 38 | DB27 | 0.8405 | 0.7589 | 0.5828 | 0.4817 | 0.7882 |
| 39 | coif16 | 0.8401 | 0.8053 | 0.6300 | 0.5560 | 0.7369 |
| 40 | rbio3.5 | 0.8391 | 0.7940 | 0.6245 | 0.5303 | 0.7719 |
| 41 | rbio3.7 | 0.8364 | 0.7790 | 0.6241 | 0.5310 | 0.7895 |
| 42 | DB6 | 0.8361 | 0.7753 | 0.6157 | 0.5252 | 0.7719 |
| 43 | DB28 | 0.8354 | 0.8292 | 0.6282 | 0.5656 | 0.7168 |
| 44 | DB37 | 0.8344 | 0.8183 | 0.6093 | 0.5326 | 0.7168 |
| 45 | DB30 | 0.8338 | 0.8128 | 0.6111 | 0.5433 | 0.7018 |
| 46 | DB35 | 0.8333 | 0.8053 | 0.6367 | 0.5586 | 0.7544 |
| 47 | sym10 | 0.8326 | 0.7228 | 0.5851 | 0.4694 | 0.8421 |
| 48 | DB11 | 0.8325 | 0.8212 | 0.6307 | 0.5617 | 0.7406 |
| 49 | sym11 | 0.8323 | 0.7266 | 0.5963 | 0.4827 | 0.8597 |
| 50 | DB25 | 0.8306 | 0.8314 | 0.6345 | 0.5932 | 0.6842 |
| 51 | bior6.8 | 0.8293 | 0.7228 | 0.5840 | 0.4575 | 0.8596 |
| 52 | coif6 | 0.8239 | 0.7474 | 0.5499 | 0.4476 | 0.7519 |
| 53 | sym8 | 0.8208 | 0.7603 | 0.6191 | 0.5328 | 0.7895 |
| 54 | DB4 | 0.8203 | 0.7678 | 0.6111 | 0.5031 | 0.8070 |
| 55 | sym12 | 0.8138 | 0.7228 | 0.5748 | 0.4667 | 0.8070 |
| 56 | sym5 | 0.8111 | 0.7905 | 0.5830 | 0.5209 | 0.7105 |
| 57 | sym14 | 0.8075 | 0.7153 | 0.5643 | 0.4638 | 0.7895 |

**Table 2** Performance comparison on the MIMII dataset between the baseline model (without DWT preprocessing) and DWT-based models using various mother wavelets, sorted by AUC. (cont.)

| Rank | With Wavelet | AUC | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|---|---|
| 58 | DB21 | 0.8073 | 0.7996 | 0.5532 | 0.5000 | 0.6291 |
| 59 | rbio2.8 | 0.8070 | 0.7566 | 0.6188 | 0.5264 | 0.8070 |
| 60 | coif7 | 0.8065 | 0.7260 | 0.5516 | 0.4457 | 0.7882 |
| 61 | sym7 | 0.8057 | 0.8175 | 0.6153 | 0.5472 | 0.7168 |
| 62 | coif2 | 0.8042 | 0.7605 | 0.5650 | 0.4968 | 0.7281 |
| 63 | DB29 | 0.8031 | 0.7607 | 0.5540 | 0.4588 | 0.7231 |
| 64 | bior4.4 | 0.7990 | 0.8202 | 0.6191 | 0.5652 | 0.6842 |
| 65 | DB13 | 0.7987 | 0.8156 | 0.5722 | 0.5264 | 0.6291 |
| 66 | sym4 | 0.7952 | 0.7971 | 0.6036 | 0.5549 | 0.7055 |
| 67 | DB26 | 0.7908 | 0.7734 | 0.5489 | 0.4646 | 0.6817 |
| 68 | rbio2.6 | 0.7897 | 0.7528 | 0.5614 | 0.4906 | 0.7193 |
| 69 | coif12 | 0.7862 | 0.7116 | 0.5701 | 0.4942 | 0.7719 |
| 70 | sym3 | 0.7781 | 0.8467 | 0.5659 | 0.6135 | 0.5514 |
| 71 | DB2 | 0.7763 | 0.7816 | 0.5387 | 0.4657 | 0.6466 |
| 72 | DB22 | 0.7741 | 0.7730 | 0.5989 | 0.5295 | 0.7406 |
| 73 | DB20 | 0.7729 | 0.7153 | 0.5363 | 0.4385 | 0.7368 |
| 74 | coif1 | 0.7689 | 0.7552 | 0.5040 | 0.4208 | 0.6291 |
| 75 | DB36 | 0.7683 | 0.6639 | 0.5131 | 0.3980 | 0.8120 |
| 76 | sym2 | 0.7669 | 0.7814 | 0.5253 | 0.4627 | 0.6115 |
| 77 | DB3 | 0.7643 | 0.8046 | 0.5199 | 0.4958 | 0.5652 |
| 78 | DB19 | 0.7622 | 0.7584 | 0.5072 | 0.4368 | 0.6115 |
| 79 | bior3.3 | 0.7619 | 0.6854 | 0.5689 | 0.4967 | 0.7544 |
| 80 | coif9 | 0.7614 | 0.7091 | 0.4979 | 0.3797 | 0.7293 |
| 81 | DB31 | 0.7611 | 0.7184 | 0.5656 | 0.4705 | 0.7882 |
| 82 | DB24 | 0.7611 | 0.6780 | 0.5212 | 0.4142 | 0.7882 |
| 83 | DB18 | 0.7607 | 0.7570 | 0.4976 | 0.4270 | 0.6053 |
| 84 | DB7 | 0.7550 | 0.6216 | 0.4957 | 0.3750 | 0.8409 |
| 85 | DB5 | 0.7548 | 0.7195 | 0.5225 | 0.4281 | 0.7281 |
| 86 | DB33 | 0.7478 | 0.7634 | 0.4939 | 0.4472 | 0.6040 |
| 87 | sym18 | 0.7471 | 0.7004 | 0.4932 | 0.3905 | 0.6842 |
| 88 | DB34 | 0.7386 | 0.6310 | 0.4542 | 0.3287 | 0.7694 |
| 89 | DB14 | 0.7118 | 0.7088 | 0.4983 | 0.4293 | 0.6592 |
| 90 | DB1 | 0.7024 | 0.6305 | 0.4523 | 0.3374 | 0.7531 |
| 91 | rbio1.1 | 0.6960 | 0.5768 | 0.4420 | 0.3163 | 0.7719 |
| 92 | DB8 | 0.6816 | 0.7257 | 0.4179 | 0.3660 | 0.4887 |
| 93 | bior1.1 | 0.6677 | 0.5918 | 0.4306 | 0.3280 | 0.7193 |
| 94 | bior3.1 | 0.6647 | 0.7453 | 0.4521 | 0.4192 | 0.4912 |
| 95 | rbio3.3 | 0.6466 | 0.4944 | 0.4166 | 0.2819 | 0.8246 |
| 96 | DB32 | 0.6408 | 0.6508 | 0.3910 | 0.3006 | 0.5639 |
| 97 | rboi3.1 | 0.4592 | 0.7491 | 0.2256 | 0.3555 | 0.1930 |

The ROC curve illustrates the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across varying decision thresholds, and the Area Under the ROC Curve (AUC) reflects the model's overall discriminative ability independent of a specific threshold. A higher AUC indicates improved separability between normal and anomalous sound patterns. In our experiments, the wavelet-enhanced configurations demonstrate a consistent upward shift of the ROC curve compared to the baseline without wavelet preprocessing, indicating improved anomaly separability in the latent reconstruction space.

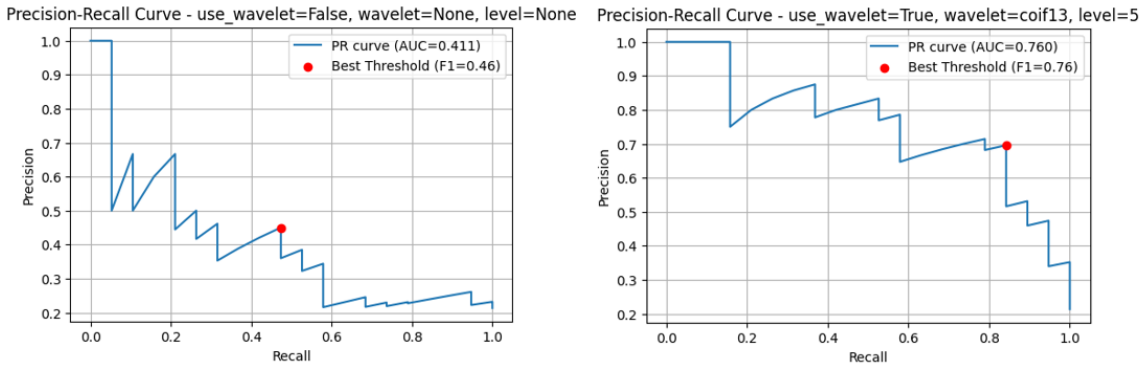


**Figure 7** Precision–Recall curves under imbalanced conditions, illustrating the improvement in Average Precision achieved by the DWT-based preprocessing method compared to the baseline model.

However, anomalous sound detection is inherently an imbalanced classification problem, where normal samples significantly outnumber anomalous ones. In such scenarios, the Precision–Recall (PR) curve provides a more informative evaluation than the ROC curve. The PR curve illustrates the trade-off between precision and recall and directly reflects the impact of false positives on detection reliability. In highly imbalanced datasets, ROC curves may present an overly optimistic view of performance, whereas PR curves more clearly reveal performance degradation when false alarms increase.

Therefore, the Average Precision (AP), corresponding to the area under the PR curve, is additionally reported. As illustrated in **Figure 7**, the proposed wavelet-based preprocessing method substantially improves the PR curve compared with the baseline model without wavelet processing. The improvement in Average Precision indicates that the proposed DWT-based feature enhancement increases the model's ability to detect rare anomalous events while maintaining higher precision.

This result is consistent with the findings in **Table 2**, where wavelets such as coif13, coif5, sym13, and sym15 achieve significantly higher ROC-AUC values (above 0.90). The improvement can be attributed to the ability of the Discrete Wavelet Transform (DWT) to capture both transient and stationary characteristics of acoustic signals at multiple resolutions, which helps suppress background noise while emphasizing local variations associated with mechanical anomalies.

Among the evaluated wavelets, coif13 achieved the highest AUC score of 0.9135, indicating strong discriminative capability for anomaly detection. coif5 obtained a nearly identical AUC value (0.9133) while providing higher Accuracy (0.9101) and a stronger F1-score (0.7778). This suggests that coif5 offers a more balanced trade-off between precision and recall, whereas coif13 demonstrates stronger anomaly detection sensitivity through higher recall. These results indicate that both wavelets provide strong detection performance, but with slightly different operational characteristics.

However, not all wavelet configurations contributed equally to the model's detection performance. While the majority of wavelets improved the AUC compared with the baseline STFT-only model, several configurations produced relatively lower AUC scores, indicating weaker discriminative capability. In particular, some low-order or poorly matched wavelet functions yielded performance close to or even below the baseline result (AUC = 0.6105), suggesting that the effectiveness of wavelet preprocessing strongly depends on the compatibility between the selected mother wavelet and the characteristics of the acoustic signal.

This reduction in performance may arise from mismatches between the wavelet characteristics and the statistical properties of the analyzed acoustic signals. Different signals often require different wavelet functions for effective representation rather than a universal wavelet choice.

These findings highlight that the selection of an appropriate mother wavelet is crucial. The wavelet should provide a suitable balance between time and frequency localization while matching the spectral characteristics of the target signal. Wavelets with moderate symmetry and compact support—such as Coiflets and mid-order Symlets—appear to be particularly compatible with the transient and noisy nature of industrial acoustic data, leading to improved reconstruction accuracy and better anomaly discrimination.

Overall, incorporating the Discrete Wavelet Transform (DWT) consistently enhances the model's performance in anomaly detection tasks. While a small number of wavelets exhibit inferior performance compared with the non-wavelet baseline, the majority provide substantial improvements. The results further suggest that different wavelets may be preferable under different operational requirements, such as maximizing anomaly detection sensitivity or reducing false alarms. Therefore, careful wavelet selection based on data characteristics and empirical evaluation is essential for achieving optimal anomaly detection performance.

### 4.4 Baseline and Reference Comparison

**Table 3** presents a comparative summary of the Average AUC scores under the −6 dB condition. The table includes (i) the benchmark results reported in the

original MIMII publication, (ii) our baseline implementation without wavelet preprocessing, and (iii) the proposed framework using the best-performing mother wavelet (coif13). This comparison is provided to contextualize the performance improvement achieved through wavelet-based signal enhancement within a controlled experimental setting.

**Table 3** Baseline and Reference Average AUC Comparison for Valve (−6 dB Condition)

| Item | Method | Average AUC |
|---|---|---|
| 1 | MIMII Reported Result | 0.53 |
| 2 | Proposed (No Wavelet) | 0.6105 |
| 3 | Proposed (coif13 : Best Wavelet) | 0.9135 |

It should be noted that the architecture and training protocol used in this study differ from those reported in the original MIMII paper. Therefore, the comparison is intended for contextual reference only and should not be interpreted as a direct state-of-the-art benchmark evaluation.

## 5. Discussion and Conclusion

This study proposed a feature enhancement framework for anomalous sound detection in industrial environments by integrating the Discrete Wavelet Transform (DWT) with Short-Time Fourier Transform (STFT) to generate Mel-spectrogram representations. The motivation behind this approach lies in combining the multi-resolution capability of wavelet analysis with the time–frequency localization of the STFT, resulting in a richer and more discriminative feature space for subsequent model training.

Experimental results demonstrated that the proposed method achieved AUC values above 0.91 on the MIMII dataset. These results indicate that incorporating wavelet-based preprocessing can enhance the model's ability to distinguish between normal and abnormal machine sounds compared to conventional spectral representations.

Among the evaluated wavelets, the Coiflet family demonstrated particularly strong performance. Specifically, coif13 achieved the highest AUC, while coif5 provided the most balanced performance across accuracy, precision, and F1-score, suggesting that different wavelets may be preferable depending on the operational requirements of anomaly detection systems.

This performance improvement can be attributed to the DWT preprocessing stage, which effectively reduces high-frequency noise and preserves transient signal structures before spectral decomposition. Such characteristics highlight the strength of wavelet transforms in representing non-stationary signals with localized temporal features.

The subsequent STFT analysis converts the denoised and smoothed signal into a time–frequency representation while maintaining local phase and magnitude information. When combined with the Mel-scale transformation, the resulting Mel-spectrogram captures perceptually meaningful frequency bands that correspond closely to human auditory perception. Moreover, Mel-based feature representations such as MFCCs and Mel-spectrograms have been widely shown to yield robust and consistent performance in audio classification tasks, supporting their suitability for machine condition monitoring.

The autoencoder-based unsupervised learning model, implemented using the Keras deep learning framework and trained on the Google Colaboratory environment, was capable of learning compact latent representations of normal machine sounds and detecting deviations that indicate anomalies. The achieved performance suggests that combining DWT with STFT serves as an effective preprocessing strategy for improving the separability of normal and abnormal features in the latent space.

A limitation of this study is that the evaluation was conducted on a single machine configuration from the MIMII dataset. Although the proposed DWT–STFT framework captures general acoustic time–frequency characteristics and is therefore expected to generalize to other machine types and SNR conditions, comprehensive cross-machine and cross-SNR evaluations remain an important direction for future work.

In conclusion, the integration of DWT and STFT prior to Mel-spectrogram generation enhances audio feature extraction for industrial anomalous sound detection, particularly in noisy environments. This hybrid approach effectively leverages the strengths of both time–frequency and multi-resolution analysis, leading to improved anomaly detection performance.

## 6. Reference

[1] H. Purohit, R. Tanabe, K. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. Detection and Classification of Acoustic Scenes and Events 2019*, Oct. 25–26, 2019, pp. 209–213, doi: 10.33682/m76f-d618.

[2] T. Ye, T. Peng and L. Yang, "Review on Sound-Based Industrial Predictive Maintenance: From Feature Engineering to Deep Learning," *Mathematics*, vol. 13, no. 11, 2025, Art. no. 1724, doi: 10.3390/math13111724.

[3] C. M. Bishop, "Introduction," in *Pattern Recognition and Machine Learning*, New York, NY, USA: Springer, 2006, ch. 1, sec. 1.2, pp. 21–24.

[4] S. Russell and P. Norvig, "Learning from Examples," in *Artificial Intelligence: A Modern Approach*, 3rd ed., Upper Saddle River, NJ, USA: Pearson, 2010, ch. 18, sec. 18.1, pp. 693–695.

[5] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015, doi: 10.1016/j.neunet.2014.09.003.

[6] J. Guan, Y. Liu, Q. Kong, F. Xiao, Q. Zhu, J. Tian and W. Wang, "Transformer-based autoencoder with ID constraint for unsupervised anomalous sound detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2023, no. 1, 2023, doi: 10.1186/s13636-023-00308-4.

[7] S. Mallat, "Introduction to a Transient World," in *A Wavelet Tour of Signal Processing*, 2nd ed., San Diego, CA, USA: Academic Press, 1999, ch. 1, sec. 1.3, pp. 28-34.

[8] I. Daubechies, "Discrete Wavelet Transform: Frames," in *Ten Lectures on Wavelets, Philadelphia*, PA, USA: SIAM, 1992, ch. 3, pp. 53–105.

[9] A. Graps, "An introduction to wavelets," *IEEE Computational Science and Engineering*, vol. 2, no. 2, pp. 50–61, 1995, doi: 10.1109/99.388960.

[10] C. Valens, "A Really Friendly Guide to Wavelets," The University of New Mexico, Albuquerque, NM, USA, 1999. Mar. 1, 2026. [Online]. Available: http://agl.cs.unm.edu/~williams/cs530/arfgtw.pdf.

[11] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, 1965, doi: 10.2307/2003354.

[12] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, no. 11, pp. 1558–1564, 1977, doi: 10.1109/proc.1977.10770.

[13] S. S. Stevens, J. Volkmann and E. B. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937, doi: 10.1121/1.1915893.

[14] B. Logan, "Mel frequency cepstral coefficients for music modeling," in *International Symposium on Music Information Retrieval*, Plymouth, MA, USA, Oct. 23–25, 2000, pp.1–11.

[15] T. Ganchev, N. Fakotakis and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in *Proc. 10th Int. Conf. Speech and Computer (SPECOM)*, Patras, Greece, Oct. 17–19, 2005, pp. 191–194.

[16] A. S. B. Saharom and F. Ehara, "Comparative Analysis of MFCC and Mel Spectrogram Features in Pump Fault Detection Using Autoencoder," in *2024 2nd International Conference on Computer Graphics and Image Processing (CGIP)*, Kyoto, Japan, Jan. 12–14, 2024, pp. 1–6, doi: 10.1109/CGIP62525.2024.00030.

[17] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi and N. Harada, "Unsupervised Detection of Anomalous Sound Based on Deep Learning and the Neyman–Pearson Lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp.212–224, 2019, doi: 10.1109/TASLP.2018.2877258.

[18] R. Garreta and G. Moncecchi, "Supervised Learning," in *Learning scikit-learn: Machine Learning in Python*, Birmingham, U.K.: Packt Publishing, 2013, ch. 2, pp. 25–60.

[19] Keras. "Keras: Developer guides." keras.io. https://keras.io/guides/ (retrieved Mar. 19, 2026).

[20] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg and O. Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. 14th Python in Science Conference*, Austin, TX, USA, Jul. 6–12, 2015, pp. 18–24, doi: 10.25080/majora-7b98e3ed-003.

[21] NumPy. "NumPy Documentation." numpy.org. https://numpy.org/doc/ (retrieved Mar. 19, 2026).

[22] A. Sharma, "A comprehensive guide to Google Colab: Features, usage, and best practices." analyticsvidhya.com. https://www.analyticsvidhya.com/blog/2020/03/google-colab-machine-learning-deep-learning/ (accessed Mar. 19, 2026).

[23] Google Research, "Welcome to Colaboratory." research.google.com. https://research.google.com/colaboratory/ (accessed Mar. 19, 2026).