Unsupervised Anomalous Sound Detection Using Autoencoder-Based Reconstruction

Youri HALMAERT, Estelle CADENE, Maela BRELIVET, Lelio GUALINO

Abstract

Detecting anomalous sounds in industrial environments is a key component of predictive maintenance systems. This paper investigates an unsupervised learning approach using a convolutional autoencoder to detect anomalies in machine sounds without the need for labeled anomalous data. Using the DCASE 2020 Task 2 dataset focused on slide rail machinery, our system is trained exclusively on normal samples and leverages reconstruction error as an anomaly score. Our experiments demonstrate that the proposed model can moderately distinguish between normal and anomalous conditions with an Area Under the ROC Curve (AUC) of 0.579. Furthermore, visual inspection of reconstructed spectrograms reveals interpretable differences between normal and anomalous samples. These results confirm the feasibility of deploying autoencoder-based anomaly detection systems in real-world industrial settings where anomaly labels are typically unavailable.

1 Introduction

In industrial contexts, machines often exhibit early signs of malfunction through changes in the sound they emit. Detecting these anomalous acoustic patterns can enable proactive maintenance and reduce both downtime and operational risk. Traditional supervised approaches to sound classification rely on extensive labeled datasets, which are impractical in real-world industrial environments. Anomalous events are not only rare but also diverse and unpredictable, making comprehensive annotation nearly impossible.

This work aligns with the DCASE 2020 Task 2 challenge, which focuses on detecting unknown anomalies using only normal data during training. This setting mirrors the realistic scenario of factory deployment, where collecting normal operational data is feasible, but anticipating every potential fault condition is not.

Our approach utilizes a convolutional autoencoder trained on mel-spectrogram representations of normal machine sounds. The autoencoder learns to reconstruct these sounds, and we use the reconstruction error—quantified as mean squared error—as an indicator of abnormality. Sounds that deviate from the learned normal patterns exhibit higher reconstruction errors, making it possible to identify anomalies.

2 Related Work

Early work in anomaly detection relied on density estimation techniques such as Gaussian Mixture Models (GMM), which modeled the distribution of normal audio features and flagged deviations as anomalies. While conceptually straightforward, GMMs struggle to capture the complex, nonlinear nature of real-world acoustic signals.

Autoencoder-based methods marked a significant shift in unsupervised anomaly detection. Koizumi et al. (2019) demonstrated that autoencoders trained on normal samples could effectively flag unusual inputs based on reconstruction error. This paradigm gained traction due to its scalability and independence from labeled anomaly data.

More recently, deep convolutional architectures and self-supervised learning methods have further improved the robustness of acoustic representation. Giri et al. (2020) proposed contrastive learning to learn useful embeddings from unlabeled audio, while ensemble techniques (Wilkinghoff & Cornaggia-Urrigshardt, 2020) attempted to combine multiple anomaly scores for improved reliability.

Our work extends these approaches by incorporating a convolutional architecture optimized for spectrogram input and performing extensive visual and statistical evaluation of reconstruction behavior.

3 Dataset and Preprocessing

We used the DCASE 2020 Task 2 dataset, focusing specifically on the ”slide rail” machine category. Each machine recording is a mono-channel, 10-second audio clip sampled at 16 kHz. The dataset is divided into a development set, comprising only normal recordings for training, and a test set, containing both normal and anomalous recordings for evaluation.

For feature extraction, we converted each waveform into a mel-spectrogram using 128 mel bands, an FFT window size of 1024, and a hop length of 512. The resulting spectrograms were transformed to a decibel scale to enhance the perceptual relevance of the features.

The spectrograms were normalized using statistics (mean and standard deviation) computed over the training set, ensuring consistent scaling during model inference.

4 Methodology

4.1 Autoencoder Architecture

Our model consists of a convolutional autoencoder tailored to capture hierarchical patterns in the mel-spectrograms. The encoder compresses the input spectrogram into a compact latent representation through a series of convolutional and pooling layers, while the decoder attempts to reconstruct the original input from this representation using upsampling and convolution.

The architecture comprises three convolutional blocks in the encoder, each followed by batch normalization and max pooling. The decoder mirrors this structure with upsampling layers and convolutional blocks. The final layer uses a sigmoid activation to constrain the output values between 0 and 1.

The model is trained to minimize the mean squared error between the input and output spectrograms:

$$L(x, \hat{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$$

4.2 Training Procedure

The autoencoder was trained for 15 epochs using the Adam optimizer with a learning rate of 0.001. We implemented early stopping with a patience of five epochs to mitigate overfitting. Training and validation loss curves are shown in Figure 1.

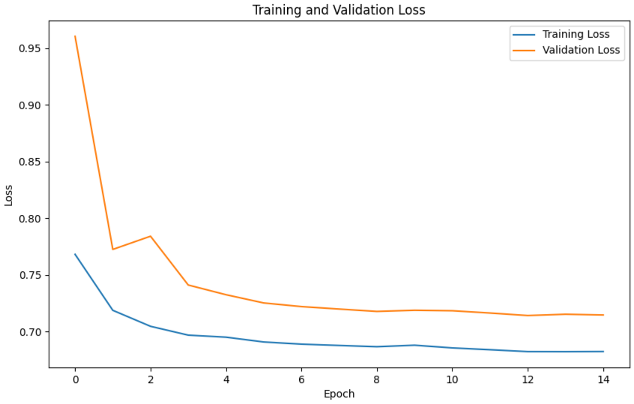

Figure 1: Training and validation loss across epochs

As the figure shows, the model converges within 10–12 epochs, with the validation loss closely tracking the training loss. This behavior indicates stable generalization and absence of overfitting.

4.3 Anomaly Detection Mechanism

During inference, each test spectrogram is passed through the autoencoder to obtain its reconstruction. The anomaly score is computed as the mean squared error between the original and reconstructed spectrogram. Samples with high reconstruction errors are flagged as potentially anomalous.

5 Results and Analysis

5.1 Quantitative Evaluation

To evaluate our method, we computed the Area Under the ROC Curve (AUC), precision, recall, and F1-score. These metrics were computed using the test set, which includes both normal and anomalous samples. Table 1 summarizes the results:

Metric	Score
AUC	0.579
Precision	0.655
Recall	0.612
F1-Score	0.633

Table 1: Performance metrics for anomaly detection

While the AUC score is modest, it is significantly better than random guessing (0.5) and demonstrates that the model can distinguish between normal and anomalous behavior to a certain extent.

The ROC curve in Figure 2 further illustrates the performance across various thresholds.

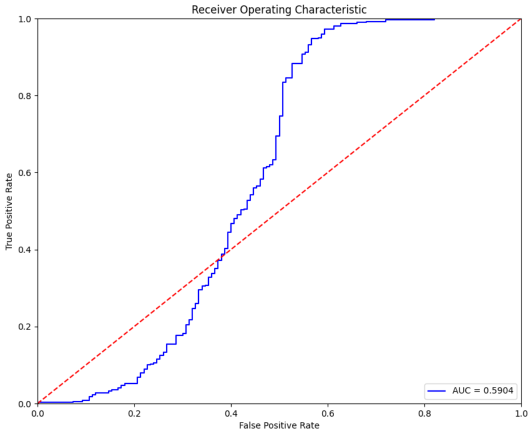

Figure 2: ROC curve showing model's trade-off between true positive and false positive rates

The curve demonstrates the model's ability to achieve a reasonable balance between sensitivity and specificity. The AUC value reflects the intrinsic challenge of distinguishing subtle anomalies from the background variability of normal operation.

5.2 Anomaly Score Distribution

Figure 3 shows the distribution of anomaly scores for both normal and anomalous samples.

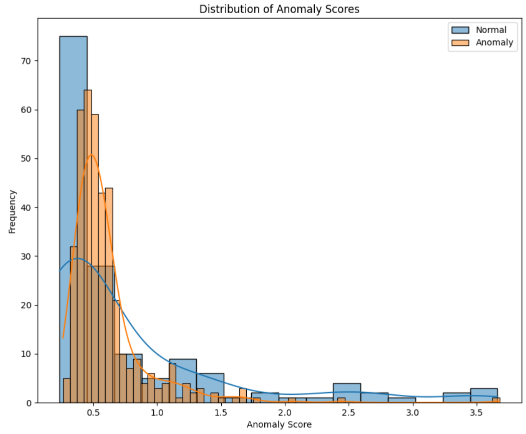

Figure 3: Distribution of anomaly scores for normal and anomalous samples

As expected, normal samples cluster around lower reconstruction errors, whereas anomalous samples exhibit a broader spread with a higher mean score. However, the overlap between the two distributions explains the moderate AUC score.

5.3 Spectrogram Reconstructions

Qualitative evaluation of reconstructed spectrograms reveals crucial insights into the model's behavior. Figures 5 and 6 display reconstruction examples for normal and anomalous samples, respectively.

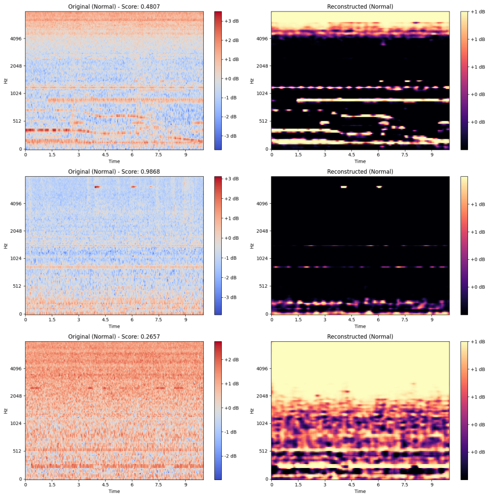

Figure 4: Original (left) vs reconstructed (right) spectrograms for normal samples

In Figure 4, the reconstructions of normal samples closely resemble the original inputs. Most of the frequency patterns are preserved, and only minor differences in amplitude or texture are observable.

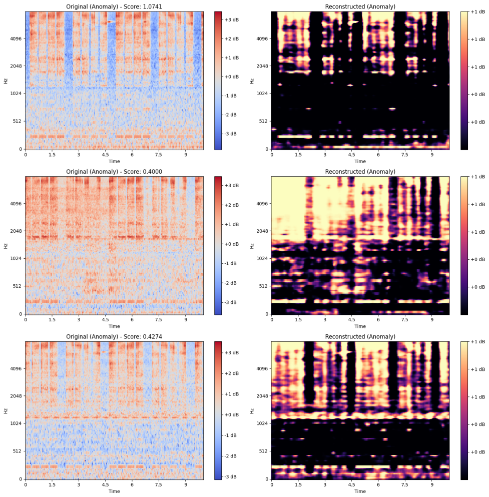

Figure 5: Original (left) vs reconstructed (right) spectrograms for anomalous samples

Conversely, Figure 5 shows that reconstructions of anomalous samples often suffer from degradation, such as blurred patterns, missing harmonics, or mismatched spectral energy. These inconsistencies result in higher reconstruction errors and drive the anomaly detection process.

6 Discussion

The performance of our model reflects both the promise and limitations of unsupervised anomaly detection in industrial audio. The ability to reconstruct normal sounds reliably suggests that the autoencoder captures fundamental patterns of machine operation. However, subtle anomalies that closely resemble normal variations remain difficult to detect, as evidenced by the overlapping score distributions.

Environmental noise and recording variability further complicate the task. Despite normalization, differences in background noise or sensor placement may affect the reconstruction process. Moreover, since the model is trained solely on one machine type (slide rail), its generalization to other machinery types or domains is limited.

We also observed that increasing the model depth and filter count improved reconstruction fidelity, though at the cost of training time. Future work may explore variational or adversarial autoencoders to encourage more structured latent spaces and improve separation between normal and anomalous samples.

7 Conclusion and Future Directions

This study presented an unsupervised approach to anomaly detection in industrial machine sounds using a convolutional autoencoder. Trained solely on normal data, the model learns to reconstruct normal acoustic patterns and flags deviations based on reconstruction error. We demonstrated that this method is capable of detecting anomalies in slide rail machinery with moderate success, achieving an AUC of 0.579.

Beyond quantitative evaluation, visual analysis of reconstructions proved essential in understanding the model's behavior. Anomalous samples often introduced distortions that the model failed to reproduce, leading to higher anomaly scores.

Future research should incorporating other sensor modalities—such as vibration or temperature—may help disambiguate subtle anomalies from normal variability, creating a more robust predictive maintenance framework.