Anomalous Sound Detection for Industrial Machines with Supervised and Unsupervised Learning Algorithms

Min-Gyu Kim (mk43469) Adhika Retnanto (apr928) Ryan Roby (rr53477) Sophia Shi (sbs3348)

Abstract

Anomalous sound detection (ASD) techniques based on machine acoustics have been proposed as alternative and more generalizable methods to prevent mechanical failure for industrial machines. In this study, we extended previous work on anomaly detection methods by exploring four feature extraction methods. We developed ASD methods relying on supervised and unsupervised approaches for four industrial machines from the Malfunctioning Industrial Machine Investigation and Inspection (MIMII) dataset with realistic background noise. By comparing the AUC with a baseline model, we show that a supervised approach performs better, indicating that analyzing the sound of a machine could be a cheaper alternative tool to supplement other sensors in a factory environment once the labels are present. However, we also show that our unsupervised approach performs poorer than the baseline MIMII model, indicating that further feature extraction and engineering techniques are required.

1. Introduction and Background

Industrial machinery can break down for various reasons, ranging from external strains to internal damage due to long-term use. Detecting malfunctioning factory machinery is essential to prevent economic and environmental losses caused by operation downtime, maintenance costs, resource waste, etc. For industrial machinery, anomaly detection techniques commonly rely on data obtained through internal sensors that detect structural vibrations, temperature, and pressure, which can be represented in the visual space. However, these sensors are specialized and can be sensitive when exposed to extreme conditions, potentially requiring high maintenance and repair costs.

Anomalous sound detection (ASD) techniques based on machine acoustics have been proposed as alternative and more generalizable methods to prevent mechanical failure for industrial machines as sound data are generally cheaper to collect. Mechanical failure can be associated with a distinguishable sound depending on the type of machinery. Despite the advantages of ASD, the main challenge with analyzing acoustic data is exacerbated by the noisy environments that can drown failure sounds, among many other issues. In this project, we produce a supervised and unsupervised machine-learning model that detects anomalous sounds that of industrial machines such as fans, pumps, valves, and slide rails from the Malfunctioning Industrial Machine Investigation and Inspection (MIMII) made available by Hitachi, Ltd ((Purohit et al., 2019)).

Previous work has been performed with anomalous sound detection on the MIMII dataset using machine learning techniques by focusing on 1) feature extraction and engineering to enhance anomaly detection. The authors of the dataset recommend representing the audio recordings as a visual representation as a Mel-spectogram ((Purohit et al., 2019)). Further feature extraction and engineering techniques have been explored. For example, Liu et al. use a spectral-temporal fusion-based self-supervised method to model the feature of the normal sound (Liu et al., 2022). Muller et al. use feature extraction methods using neural networks (NNs) that were pre-trained on image classification tasks (Müller et al., 2020). Tagawa et al. deploy a generative adversarial network (GAN) for sound signal reconstruction (Tagawa et al., 2021). Inspired by the previous work, we aim to extend the previous work on feature extraction techniques by representing the audio recordings through different transformations. By representing the audio data to visual data using commonly used transformations, we aim to improve the anomaly detection methods while retaining the explainability that is commonly encountered with feature extraction methods. More specifically, the purposes of our project are as follows:

- Perform feature extraction methods that convert audio to visual data such as Short-Time Fourier Transform (STFT), Mel-frequency cepstral coefficients (MFCCs), gammatone frequency cepstral coefficients (GFCCs), and Mel-spectrogram)
- Evaluate the performance of a supervised anomaly detection method based on the CNN-based classification model with the feature extraction methods
- Identify the underlying patterns of the dataset through

Table 1. MIMMI dataset content details (Purohit et al., 2019)

Machine type	Model ID	Normal samples	Abnormal samples
Valve	00	991	119
	02	708	120
	04	1,000	120
	06	992	120
Pump	00	1,006	143
	02	1005	111
	04	702	100
	06	1,036	102
Fan	00	1,011	407
	02	1,016	359
	04	1,033	348
	06	1,015	361
Slide Rail	00	1,068	356
	02	1,068	267
	04	534	178
	06	534	89
Total		14,719	3,300

an unsupervised anomaly detection method based on Gaussian Mixture Models (GMMs) with the feature extraction methods

- Assess the effects of varying signal-to-noise (SNR) ratios from realistic background noise on the anomaly detection models
- Analyze the anomaly detection model given an unbalanced dataset between normal and abnormal sound samples of four models of four industrial machines

2. Data Description

The dataset consists of sound samples that are generated from four types of industrial machines, including valves, pumps, fans, and slide rails. The data were collected using a circular microphone manufactured by System in Frontier Inc., TAMAGO-03. The microphone is a circular array that consists of eight distinct microphones with a sampling rate of 16 kHz and 16 bits per sample, while also enabling single- and multi-channel-based approaches. The microphone was placed 50 cm from the pumps, fans, and slide rails and 10 cm from the valves, collecting 10-second sound intervals.

Data for four models were obtained for each machine as shown in Table 1. The data in each model consists of up to 10,000 seconds of normal sound and 1,000 seconds of abnormal sound. Anomalous sounds include contamination, leakage, rotating unbalance, etc. In addition, the normal and abnormal sounds were mixed with background noise from real factory environments at several levels of signal-noise ratio (SNR): 6 dB, 0 dB, and 6 dB. Therefore, we use 14,719 normal and 3,300 abnormal sounds for model IDs 00, 02, 04, and 06 for our project.

3. Methodology

In this section, we present our methodology that was performed for each model within each machine as shown in Figure 1. First, we address the imbalance between normal to abnormal samples for the dataset (Sec. 3.1). Next, we trained and validated our model using the feature extraction methods (Sec. 3.2) with the described model architecture for the supervised and unsupervised methods (Sec. 3.3, 3.4). Finally, we analyzed our results through visualization (Sec. 3.5) and evaluation metrics 3.6).

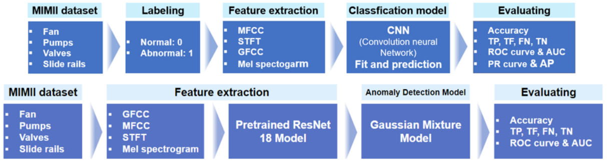

Figure 1. Top: supervised machine learning procedure, bottom: unsupervised machine learning procedure

3.1. Data preparation

The main purpose of the data preparation step was to address the imbalanced distribution between the normal and abnormal sound samples as shown in Table 1. For each model within each machine, we labeled the normal cases as 0 and the abnormal cases as 1. As shown in Figure 2, since there are fewer abnormal samples, we randomly selected the same numbers of normal samples, creating a more balanced dataset. Lastly, we split the selected samples into train and test datasets with a 1:1 ratio.

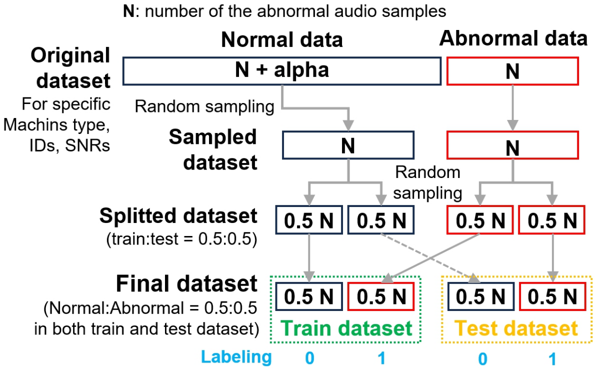

Figure 2. Data preparation for the supervised learning

3.2. Feature extraction methods

The main feature of our work is the feature extraction that were performed to transform the audio samples to visual datasets as the anomaly detection classifier is typically trained on an image format as input data. More specifically, we explored the following visual transformations: STFT (Short-term Fourier Transform), MFCCs (Mel-frequency cepstral coefficients), GFCCs (Gammatone frequency cepstral coefficients), and Mel-spectrogram where the hyperparameters applied for the transformations are shown in Table A1. As displayed in Figure 3 the results of the feature extraction techniques allowed for the abnormal samples to be distinguished more easily in the visual space.

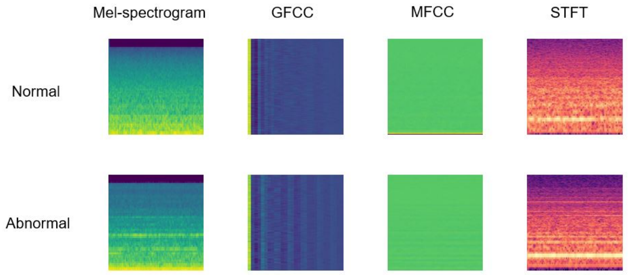

Figure 3. Visualization of feature extraction methods (machine: fan; from L to R: Mel-spectrogram, GFCC, MFCC, STFT)

3.3. Model Architecture for Supervised Learning

For anomalous sound detection using supervised learning, we implemented a Convolutional Neural Network (CNN) model as a classifier with inputs based on spectrograms extracted from the audio sound. CNNs were chosen for this study as they are particularly effective in tasks like image recognition, object detection, and classification, demonstrated by the several attempts to implement the CNN model for anomaly sound detection in previous works ((Morita et al., 2021; Zhao, 2020)).

As shown in Figure 4, the proposed CNN model architecture was constructed for each feature extraction method. 3x3 2D Convolutions and Max pooling techniques were used, followed by dense nonlinear activation functions. The detailed layer parameters are shown in Table C1.

Once constructed, model fitting was performed by applying the *binary cross entropy* loss function that was solved with the *Adam* optimizer with a 0.001 learning rate. We trained the model using 16 batch sizes, and 8 different epochs between 5-40.

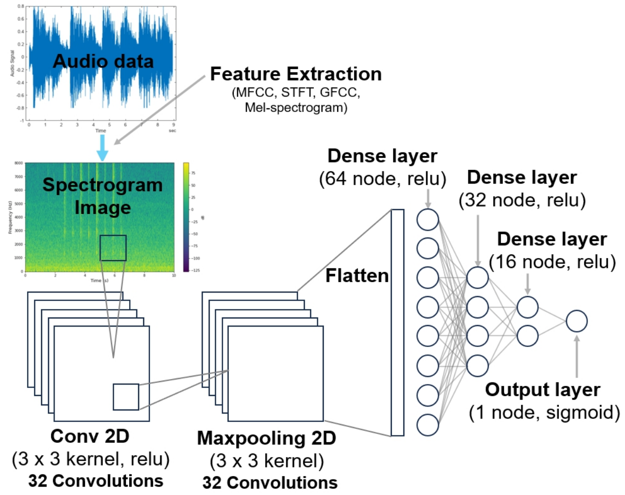

Figure 4. Conceptual image of the *spectrogram + CNN architecture* in this study for the supervised learning

3.4. Model Architecture for Unsupervised Learning

Similar to the supervised learning approach, the unsupervised learning for ASD was performed with the inputs based on spectrograms extracted from the audio sound. More specifically, the spectrogram was split into 8 pieces, where a pre-trained ResNet 18 model was implemented to enhance the extraction of spatial features from the spectrogram. Details on the model architecture can be found in Table **??**. Since the normal and abnormal labels were removed for the unsupervised learning approach, we used a two-component Gaussian Mixture Model to predict the underlying structure of the dataset for normal and abnormal sound samples.

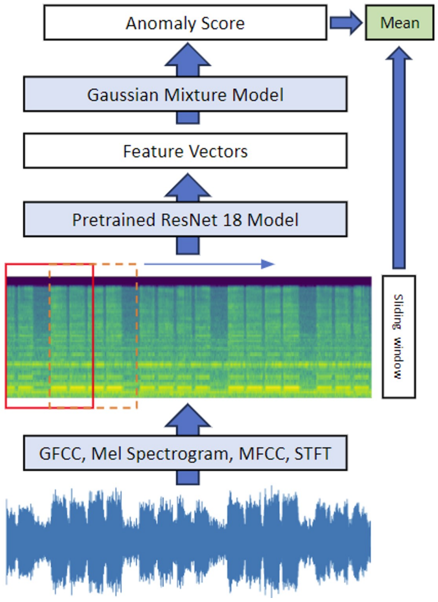

Figure 5. Conceptual image of the *spectrogram + ResNet18 + GMM* in this study for the unsupervised learning

3.5. Result Visualization

To evaluate the performance of individual datasets, we plot the predicted score [0,1] from the output of the ASD model with the labeled true status 0,1 for the train and test datasets.

An example of our result visualization is shown in Figure 6, where it includes the predicted values [0,1] by the ASD model (red circles) with the true values (blue small circles). In this figure, we can see the distribution of the results visually, and which threshold is appropriate for the

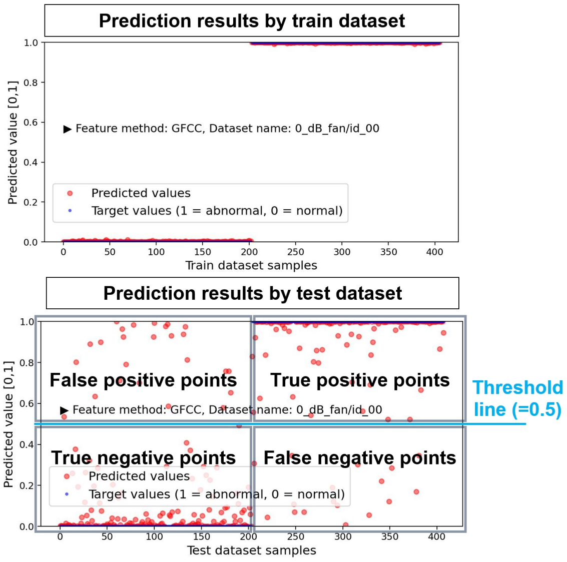

Figure 6. Visualization example of the model prediction results for each dataset

classification and clustering of supervised and unsupervised learning, respectively. However, because the threshold can have very different values according to the model IDs and SNRs (signal-to-noise ratios), to avoid over-fitted classification, we assume the threshold is 0.5 for all datasets. In this study, we made a total of 6,144 figures considering four feature extraction methods, four machine types, four model IDs, three SNRs, and 8 different epochs.

3.6. Evaluation Metrics

The evaluation metrics for the ASD models are based on a binary cross entropy loss function as the sound labels are either normal or abnormal. Therefore, we evaluated the models by determining an ROC curve with the AUC score and the PR curve with the AP score. In addition, we also assessed the performance of the ASD models through other methods like accuracy and confusion matrix.

Figure 7 shows the example of the ROC and PR curve for specific datasets in this study.

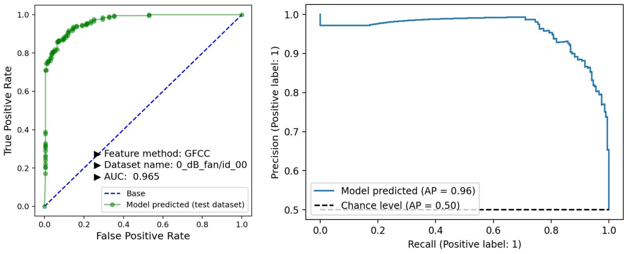

Figure 7. A representative visualization example of the model prediction results for each dataset

We also use the confusion matrix to show the result visually and compare the model result between different datasets.

4. Results and Discussions

4.1. Result for the Supervised Learning Model

4.1.1. AUC Score Comparison with the MIMII Baseline Model

After performing the ASD supervised learning described in Section 3.3, we compared our results with the baseline model based on the unsupervised auto-encoder model provided by Purohit et al. (Purohit et al., 2019). Comparing the AUC scores, all four feature extraction methods with the supervised CNN model outperform the baseline unsupervised model as shown in. Figure 8 shows the average AUC scores for the four feature extraction methods for each of the machines in the dataset. Among the feature extraction methods, the STFT performs best, further amplified by the apparent differences in the visualization in Figure 6. Additionally, the figure shows the results after training the model at epoch 5 and epoch 25. Generally, increasing epochs improves the model performance before over-fitting occurs. When the model was trained for 25 epochs, the AUC scores were improved by 0.09-0.15 for the valve, 0.01-0.09 for the pump, 0.01-0.06 for the Fan, and 0.03-0.08 for the slide rail compared to the model trained after 5 epochs.

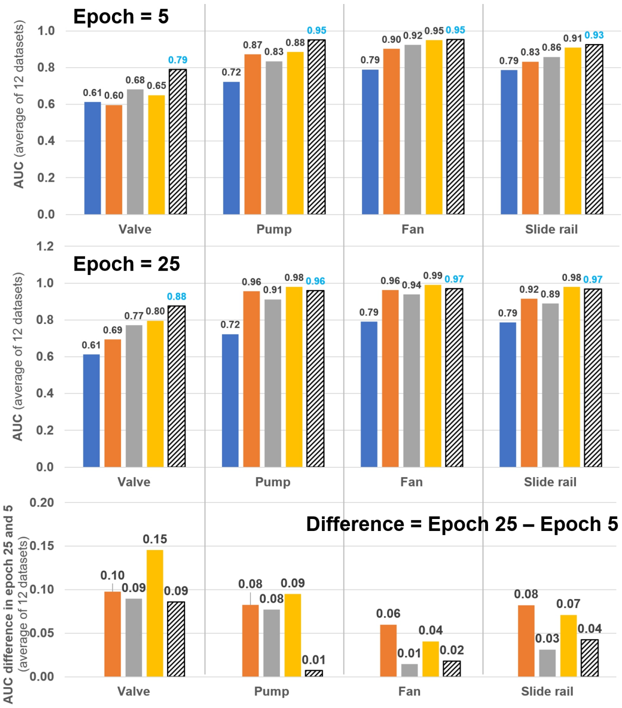

Figure 8. Comparison of the AUC scores between MIMII baseline model and proposed CNN model by different feature extraction methods (STFT, MFCCs, GFCCs, and Mel-spectrogram)

4.1.2. Impact of the SNRs on the Model Performance

The background noise of the environment can greatly affect the performance of the ASD models. Here, we explored the effects of the SNRs at values of -6 dB, 0 dB, and 6 dB. According to the confusion matrix, the SNRs significantly display a positive correlation with the model performance as shown in Figure 9. The confusion matrix displayed here is an example for the machines, where Figure 9a) shows the best result, and Figure 9b) shows the worst result. As a result, the model performances show higher accuracy at SNRs (i.e. 6 dB), while the model performes worse at lower SNR (i.e. -6 dB). In other words, the model performs better when the audio samples are dominated by the noise of the machines. Since our results indicate that improving model performance with the lower SNR is difficult, further research aiming to filter the background noise is required to overcome this limitation of our model, which was beyond the scope of our studies.

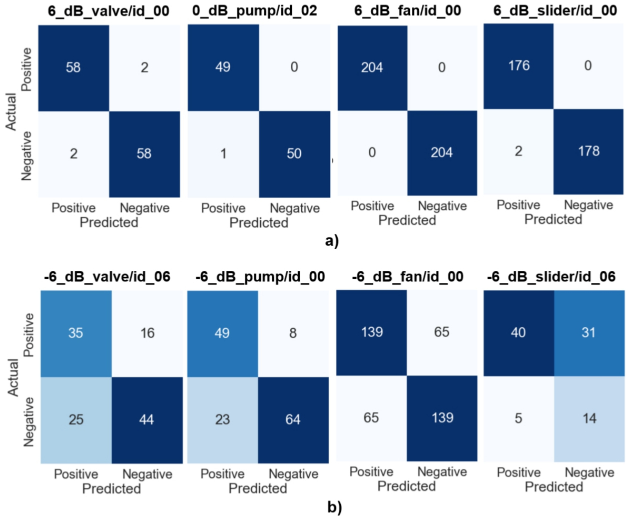

Figure 9. Confusion matrices for the best and worst cases for four machine type with the STFT method at epoch 25: a) best results, b) Worst results

4.1.3. Assessment for Over-fitting Training at Varying Epochs

We also explored the potential issues with respect to overfitting when training the models at higher epochs. Here, we observed that there are no significant over-fitting problems by increasing the epochs as the accuracy for each model tends to converge to certain values. The train and test accuracy of 192 cases representing different machine types, domains, SNR, and feature extraction methods are shown after 5, 15, 25, and 35 epochs in Figure 10a). These results indicate the general converging trend for all machines, models, and feature extraction methods. In addition, as shown in Figure 10b), the red lines show the over-fitting tendency in larger epochs. Based on these results, we conclude that training the model for 25-30 epochs may give the best results. With that said, training the models beyond 25 epochs is unnecessary as the accuracy of the training and testing datasets have already converged.

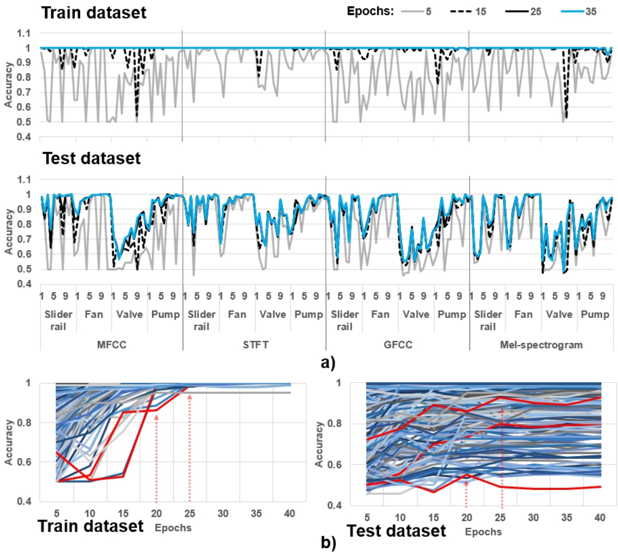

Figure 10. Accuracy changes of 192 individual models by increasing learning epochs

4.1.4. Limitations of the Supervised Learning Approach

The supervised learning approach has several limitations that are either limited by our design choice or by the available dataset:

- The models perform poorly at lower SNR. Further research is required to overcome this limitation.
- The models cannot adapt to the domain shift. In other words, the models might not be robust when exposed to other realistic industrial environments.
- The models struggle with generalization under the various condition changes since the model is trained for each machinery model.
- Obtaining and labeling the audio samples from all possible abnormal cases for supervised learning is difficult as the dataset is more distributed towards normally functioning machiens.

4.2. Unsupervised Learning Approach

As a preliminary measure to study the underlying patterns of the dataset, we removed the labels for the normal and abnormal samples and performed the unsupervised learning approach as described in Section 3.4. As a result, we focused solely on the performance of the unsupervised learning approach compared to the MIMII Baseline Model. Similar to the supervised learning approach, the performance were evaluated by taking the average of all models for each machine. As shown in Figure 11, the performance of our approach is lower than the baseline model. This can be explained by the relative simplicity of our models where our approach focuses on the feature extraction of the audio samples. The baseline model uses an auto-encoder that could represent the dataset at a higher dimensional space, which could help differentiate clusters between the normal and abnormal samples. Despite the poorer performance, we also see that the SNR have less of an effect on the unsupervised approach, which could be exploited further. Nonetheless, the Mel-spectrogram generally performs better out of the four feature extraction methods.

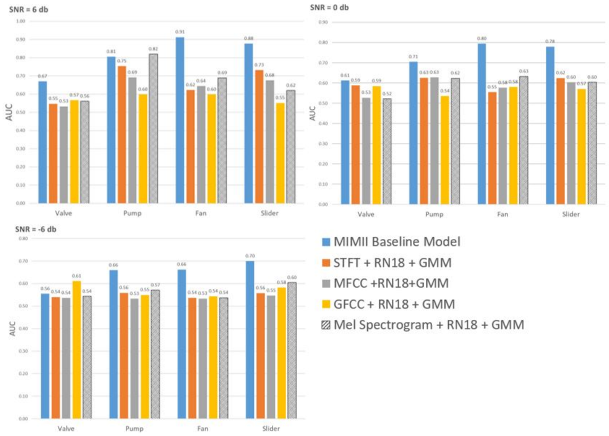

Figure 11. Comparison of the AUC scores between MIMII baseline model and unsupervised model by different feature extraction methods (STFT, MFCCs, GFCCs, and Mel-spectrogram)

5. Conclusion and Future Work

In this study, we extended previous work on anomaly detection methods by exploring various feature extraction methods that retain the explainability of the dataset. We developed ASD methods relying on supervised and unsupervised approaches. We show that a supervised approach performs better than the baseline MIMII model. This is justified as the broken machinery can typically be determined easily and ”labeled” through other sensory equipment. As a result, analyzing the sound of a machine could be a cheaper alternative tool to supplement other sensors in a factory environment. However, we also show that our unsupervised approach performs poorer than the baseline MIMII model, indicating that further feature extraction and engineering techniques are required to understand the underlying patterns of the dataset while sacrificing some of the explainability of our approach.

We propose future works to improve our results:

- Application of the transfer learning methods for generalization to various domain shifts.
- Incorporation of the ensemble algorithms as a classifier based on the CNN model’s flattened output data.
- Investigation of the unsupervised learning methods and incorporating the benefits of similar results across different SNRs.
- Exploration of other image classification algorithms using the same feature extraction methods

Accessibility

The Python code and relevant documents are provided in this project link: https://github.com/minsky97/ECE381K_AML_term_project.

References

Liu, Y., Guan, J., Zhu, Q., and Wang, W. Anomalous sound detection using spectral-temporal information fusion, 2022.

Morita, K., Yano, T., and Tran, K. Anomalous sound detection using cnn-based features by self supervised learning. *Tech. Rep., DCASE2021 Challenge*, 2021.

Müller, R., Ritz, F., Illium, S., and Linnhoff-Popien, C. Acoustic anomaly detection for machine sounds based on image transfer learning. *arXiv preprint arXiv:2006.03429*, 2020.

Purohit, H., Tanabe, R., Ichige, K., Endo, T., Nikaido, Y., Suefusa, K., and Kawaguchi, Y. MIMII Dataset: Sound Dataset for Malfunctioning Industrial Machine Investigation and Inspection. 2019. doi: 10.48550/ARXIV.1909.09347.

Tagawa, Y., Maskeliūnas, R., and Damaševičius, R. Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. *Electronics*, 10(19), 2021. ISSN 2079-9292. URL https://www.mdpi.com/2079-9292/10/19/2329.

Zhao, J. Anomalous sound detection based on convolutional neural network and mixed features. In *Journal of Physics: Conference Series*, volume 1621, pp. 012025. IOP Publishing, 2020.

A. Feature Extraction Methods

Each feature extraction method requires several hyperparameters and are applied as follows:

Table A1. Hyperparameters for feature extraction methods

Feature Extraction Methods	Hyperparameters
STFT	nperseg = 1024, noverlap = 512
MFCC	n_mfcc = 200
GFCC	nfilts = 250, num_ceps = 250
Mel-Spectogram	n_fft = 1024, hop_length = 512, n_mels = 128, power = 2

Where, nperseg: number of data points used in each STFT block, noverlap: number of points of overlap between blocks, n_mfcc: number of MFCCs to return, nfilts: number of filter banks, num_ceps: number of cepstral coefficients to return, n_fft: number of points for the FFT, hop_length: number of samples between successive frames, n_mels: number of Mel bands to generate, power: exponent for the magnitude spectrogram.

B. Supervised Learning: CNN Model Architecture

Table B1. CNN model architecture

	Output shape			
Layer	STFT	MFCCs	GFCCs	Mel-spectrogram
Feature extraction methods	(514 X 314)	(128 X 431)	(1000 X 250)	(128 X 431)
Convolution 2D (3X3)	(512 X 312, 32)	(126 X 429, 32)	(998 X 248, 32)	(126 X 429, 32)
Max pooling 2D (3X3)	(170 X 104, 32)	(42 X 143, 32)	(332 X 82, 32)	(42 X 143, 32)
Flatten	565,760	192,192	871,168	192,192
Dense ('relu' activation)	64			
Dense ('relu' activation)	32			
Dense ('relu' activation)	16			
Dense ('sigmoid' activation)	1			

C. Unsupervised Learning: ResNet and Gaussian Mixture Model

Table C1. GMM model architecture

	Output shape			
Layer	STFT	MFCCs	GFCCs	Mel-spectrogram
Feature extraction methods	(514 X 314)	(128 X 431)	(1000 X 250)	(128 X 431)
Convolution 2D (3X3)	(512 X 312, 32)	(126 X 429, 32)	(998 X 248, 32)	(126 X 429, 32)
Max pooling 2D (3X3)	(170 X 104, 32)	(42 X 143, 32)	(332 X 82, 32)	(42 X 143, 32)
Flatten	565,760	192,192	871,168	192,192
Dense ('relu' activation)	64			
Dense ('relu' activation)	32			
Dense ('relu' activation)	16			
Dense ('sigmoid' activation)	1			