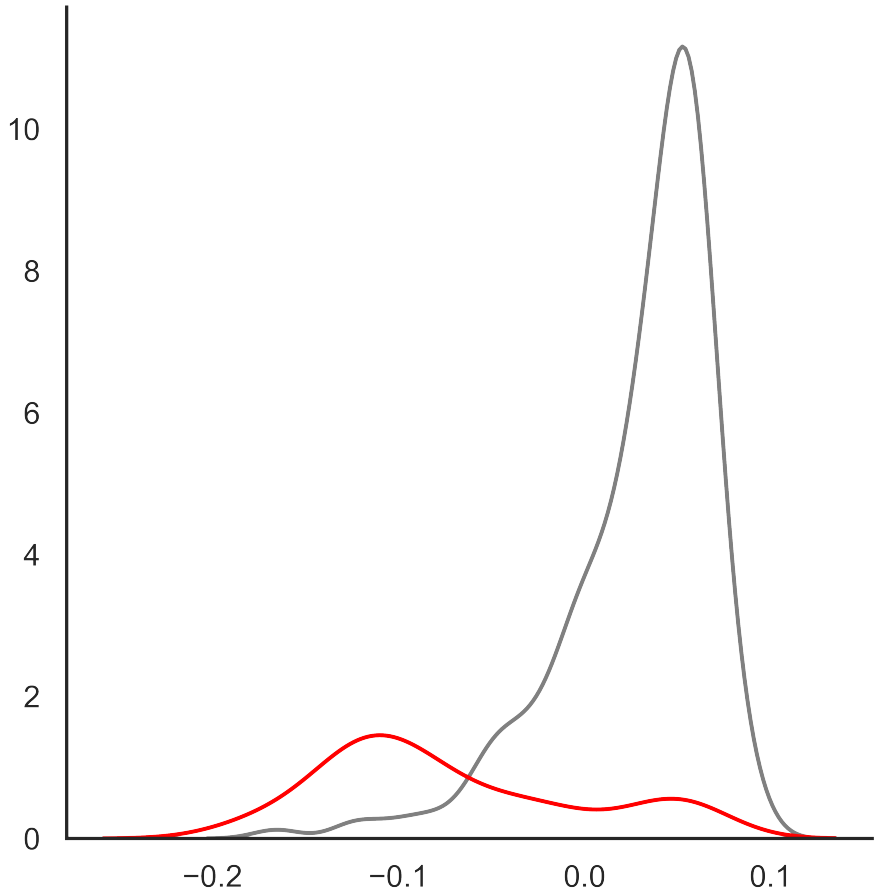
CHALMERS
UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG



Audio Anomaly Detection in Cars

An Evaluation of Unsupervised Anomaly Detection in Cars using Mel Frequency and Chroma Features

Master's thesis in Mathematical Sciences

ASMA HUSSEIN

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2023

Audio Anomaly Detection in Cars
Asma Hussein
Department Of Mathematical Sciences
University Of Gothenburg

Abstract

Audio anomaly detection in the context of car driving is a crucial task for ensuring vehicle safety and identifying potential faults. This paper aims to investigate and compare different methods for unsupervised audio anomaly detection using a data set consisting of recorded audio data from fault injections and normal "no fault" driving. The feature space used in the final modelling consisted of: CENS (Chroma energy normalized Statistic), LMFE (Log Mel Frequency Energy), and MFCC (Mel-frequency cepstral coefficients) features. These features exhibit promising capabilities in distinguishing between normal and abnormal classes. Notably, the CENS features which revealed specific pitch classes contribute to the distinguishing characteristics of abnormal sounds. Four Machine learning methods were tested to evaluate the performance of different models for audio anomaly detection: Isolation Forest , One-Class Support Vector Machines, Local Outlier Factor, and Long Short-Term Memory Autoencoder. These models are applied to the extracted feature space, and their respective performance was assessed using metrics such as ROC curves, AUC scores, PR curves, and AP scores. The final results demonstrate that all four models perform well in detecting audio anomalies in cars, where LOF and LSTM-AE achieve the highest AUC scores of 0.98, while OCSVM and IF exhibit AUC scores of 0.97. However, LSTM-AE displays a lower average precision score due to a significant drop in precision beyond a certain reconstruction error threshold, particularly for the normal class. This study demonstrates the effectiveness of Mel frequency and chroma features in modelling for audio anomaly detection in car and shows great potential for further research and development of effective anomaly detection systems in automotive applications.

Keywords: Audio Anomaly detection, Outlier detection, Machine learning, Mel Frequency, Chroma.

Acknowledgements

I would like to extend my sincere gratitude and appreciation to my supervisor, Michael Wallgren Fjellander, whose guidance, support and encouragements has been essential to finishing my thesis.

Asma Hussein, Gothenburg, June 2023

Contents

List of Figures

List of Tables

1

Introduction

The development of autonomous vehicles has led to an increased interest in the use of sensor data for various applications, including anomaly detection. Microphone(s) inside a vehicle is one key sensor which can provide valuable information about the environment and the vehicle's internal systems. For instance, Previous research has used smartphone audio to detect when the air particle filter in a car needs to be replace, using a machine learning approach with Mel-Cepstrum, Fourier and Wavelet features as input into a classification model [Siegel et al., 2017]. However, the use of audio data for anomaly detection is still a relatively new field, and there is a lack of understanding of suitable methods for profiling and processing as well as classifying sound.

Further, it is particularly interesting for vehicle manufacturers, such as Volvo Cars, to develop these methods. As audio anomaly detection (AAD) can be a useful tool for detecting mechanical faults in cars that give rise to noise. The microphone of a car can capture a wide range of sounds: produced by for instance the engine, transmission, and suspension. For example, unlubricated brakes can produce a high-pitched sound, while a stuttering engine can produce a knocking sound. These sounds can provide valuable information about the health of the car's mechanical systems. By using appropriate feature spaces generated from audio data captured in cars; models can be trained to recognize patterns in the audio signal that correspond to specific mechanical faults which can further be used to classify the audio signals as "normal" or "anomalous".

The advantage of using AAD for detecting mechanical faults is that it can be done in real-time and does not require any additional sensors or equipment. This makes it a cost-effective and non-invasive method for monitoring the health of a car's mechanical systems. Additionally, audio anomaly detection can be used to detect early warning signs of potential mechanical problems, allowing for preventative maintenance to be carried out before the problem becomes critical. Research has shown that AAD can be effective for detecting machine failures even in noisy factory environments[Tagawa et al., 2021].

The aim of this project has been to find the optimal feature space and classification algorithm for AAD in cars thru reviewing previous research and applying relevant methods to data collected in our test cars. This paper will first outline previous use of ADD, particularly in the context of fault detection, and describe relevant theory. We then describe the data set used in this study, which consists of various audio recorded in different environments as well as the prepossessing of the audio data by extraction of Mel frequency and chroma features. We then used the features to model Unsupervised machine learning algorithms for AAD, including the Local Outlier Factor, One Class

SVM, Isolation Forest and a Long short-term memory Autoencoder, and evaluate their performance using AUC, ROC curve and Precision-Recall metrics.

Our experimental results demonstrate the effectiveness of the proposed approach in detecting audio anomalies. Overall, this research provides a thorough evaluation of audio anomaly detection and highlights the potential of this approach in real-world in car application.

2

Theory

2.1 Audio Anomaly Detection and Classification

Audio anomaly detection (*AAD*) plays a crucial role in various domains, including car and machine fault detection. *AAD* methods work by identifying unusual patterns and deviations from normal behaviour in audio signals, enabling proactive maintenance, safety improvements, and enhanced system performance.

In the context of cars, previous research has been mostly focused on identifying engine faults, exploring the application of *AAD* for faults in engines such as Air filters faults and engine misfires. Researchers have done this by recording audio from engines via microphone and analysing the resulting signals with time and frequency domain features and classifying with different Machine learning algorithms. Choices of audio features have included Discrete Wavelet Transform (DWT), Mel-Cepstrum, and Fourier Transform features [Siegel et al., 2016][Kabiri and Makinejad, 2011][Siegel et al., 2017]. While classification methods have included tree based ensemble classification [Siegel et al., 2017] [Siegel et al., 2016] and Artificial Neural Network [Dandare and Dudul, 2013].

Additionally, researchers have studied other machine fault detection in the context of factory machines. Where the most popular audio features in primarily based on the Fast Fourier transform, and include Short term Fourier transform (STFT), Mel-frequency cepstral coefficients (MFCCs), log-Mel energy (LME) and other spectrum-based 'spectral' features [Nunes, 2021].

Furthermore, studies in audio classification, such as speech recognition and music genre classification, provide valuable insights and methodologies that can be adapted for audio anomaly detection in cars [Bonet-Solà and Alsina-Pagès, 2021].

A general summary of the typical AAD approach is illustrated in Figure 2.1 and will be reviewed in the following sections.

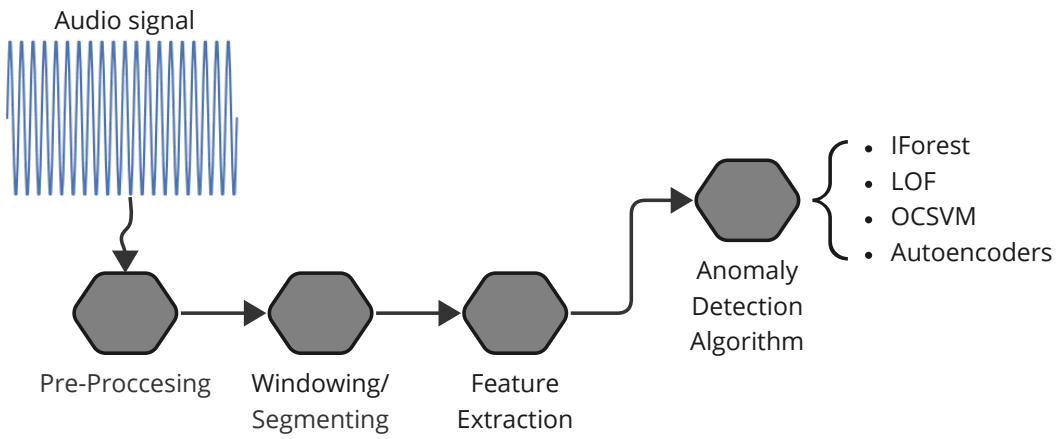


Figure 2.1: Typical approach in Audio anomaly detection

2.2 Feature extraction

Feature extraction is a fundamental step for using audio signals in AAD, where the raw audio signal needs to be transformed to features that characterise the sound. Different audio features have been used in the fields of audio classification and anomaly detection, each characterising specific aspects of the underlying audio data. This section will describe relevant features and their use in audio anomaly and event detection. While useful features can be obtained from the time-frequency domain. The features relevant for this project include spectrum and filter-spectrum features.

2.2.1 Transforms

2.2.1.1 Fast Fourier Transform

The majority of commonly used audio features are in the frequency or time-frequency domain. To obtain these features the raw time domain audio signal is transformed to the frequency domain using either the FFT or the Wavelet Transform (WT). Using the FFT frequencies can be analysed without taking time into account. The FFT computes the discrete Fourier transform of an audio signal formulated as in Equation 2.1, where $X(k)$ is the complex frequency-domain representation of the signal, $x(n)$ is the input signal in the time domain, N is the number of samples in the input signal and and k is the frequency bin index respectively.

Alternatively, frequencies and their magnitude can be analysed over time using STFT, where the time dimension is reduced using different windowing and hop lengths. The Discrete STFT is defined as in Equation 2.2, where $w(n - m)$ is a window function. Given a windowing method, window size and hop length the audio signal is transformed into frames for which the FFT is performed. The resulting frames provides information

of the frequency and amplitude content of the time points which can be illustrated in a spectrogram like in Figure 2.2[Sharma et al., 2020].

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j \cdot 2\pi \cdot \frac{kn}{N}} \tag{2.1}$$

$$X(m, \omega) = \sum_{n=0}^{N-1} x(n) \cdot w(n-m) \cdot e^{-j \cdot \omega n} \tag{2.2}$$

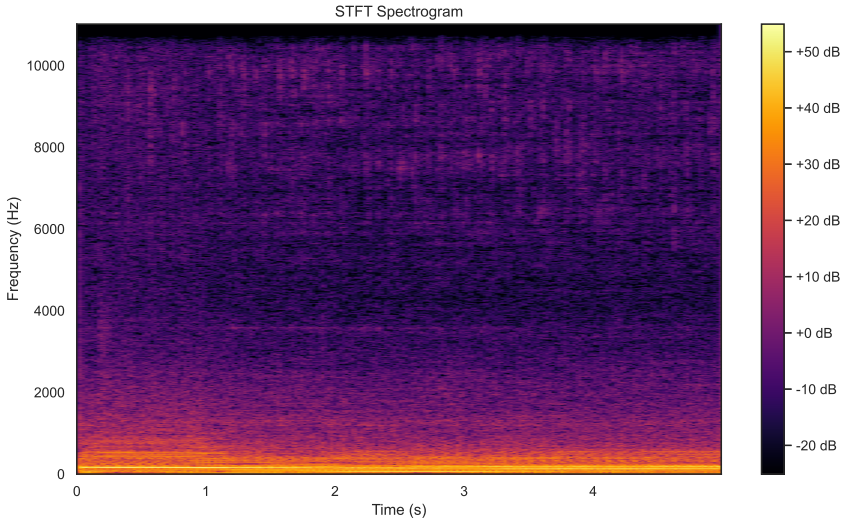


Figure 2.2: STFT Spectrogram Example

2.2.1.2 Constant-Q transform

The Constant-Q transform[Brown, 1991] is similar to the Fast Fourier Transform but uses a logarithm spaced frequency axis to better distinguish pitch classes. It is therefore often used to extract chroma features[Ewert, 2011].

2.2.2 Spectrum based features

A Spectrograms visualizes frequency contents of an audio signal over time, allowing for the identification of abnormal spectral patterns [Khan et al., 2021]. While the spectrogram alone can be used for audio characterisation. Spectral features can be extracted from the spectrogram for lower dimensional representation of the audio signal[Sharma et al., 2020]

2.2.2.1 Mel filterbank

Filter based frequency features apply filter banks to the spectrum of an audio signal. Mel filter banks are often used to produce a Mel spectrum which mimics the human perception of sound in that the distance between units of pitch are equally spaced [O'Shaughnessy, 2000].

The Mel spectrogram can then be used to compute Mel-frequency cepstral coefficients (MFCCs), derived from the power spectrum of audio signals, and Log-Mel Frequency Energy features.

Filter based features, and MFCC's in particular have been shown to be significant features for audio anomaly detection as well as general audio classification [Nunes, 2021][Bonet-Solà and Alsina-Pagè Pereira et al [Pereira et al., 2021] built an in-vehicle anomaly detection model with simulated anomalies by mixing noise of e.g. coughing and breaking glass to normal in-vehicle using exclusively filter-spectrum based features: including MFCCs, Gammatone Frequency Cepstral Coefficients (GFCCs), Mel Frequency Energy Coefficient (MFEC) achieving a AUC value of 78% using a Artificial Neural network.

2.2.2.2 Chroma

Chroma based features represents audio by mapping its spectrum into the 12 traditional pitch classes or 'chromas'. This audio feature is often, unsurprisingly, used in music classification [Muller and Kurth, 2006] but has otherwise been shown effective in classification of Environmental Sounds including traffic sound such as car horns and engine idling as well as other indoor and outdoor sounds [Mushtaq and Su, 2020]. It would thus be interesting to explore if chroma features could be useful in audio anomaly detection. Particularly Chroma energy normalized statistics (CENS) which has been used for audio similarity matching [Muller et al., 2005]

2.3 Machine learning algorithms

Machine learning algorithms are commonly used in audio anomaly detection to learn complex patterns and identify anomalies within the data. Unsupervised methods such as One Class Support Vector Machines and Autoencoders are particularly popular as they allow for anomaly detection by deviation from the norm without needing to train on anomalous data. This section will give a brief overview of One Class Support Vector Machines, Isolation forest, the Local outlier Factor and Autoencoders.

One Class Support Vector Machines (OCSVM) works by learning a decision boundary that separates normal data points from anomalies in a high-dimensional feature spaces by finding the optimal hyperplane that maximizes the margin around the normal data points while minimizing the number of support vectors; capture the inherent structure of the distribution of normal data. This method has been used by researchers in audio anomaly detection for machine failures[Tagawa et al., 2021] and abnormal event detection[Lecomte et al., 2011].

The *Isolation Forest* (IF)[Liu et al., 2012] algorithm is another unsupervised method based on ensemble learning and makes the assumption that anomalous data points are few and different. the algorithm isolates anomalies by partitioning the data into sub spaces, and provides an anomaly score based on number of splits needed to isolate a data point. The *Local Outlier Factor* (LoF) algorithm detects anomalies by measuring the local density deviation of a data point compared to its neighbors and, while not commonly used for audio

data, has shown good performance on high dimensional spectral data[Yu et al., 2020]

Lastly, *Autoencoders*, a type of neural network, have been successfully used to handle high-dimensional audio features by reconstructing the input data while capturing abnormal patterns[Dandare and Dudul, 2013][Nunes, 2021]. Making it particularly suited to anomaly detection, as the reconstruction error can be used as a measure of deviation from the norm.

3

Methods

This section will describe the data collected, features extracted and Machine Learning methods chosen for ADD modeling

3.1 Data

3.1.1 Data Collection

The data used in this research consisted of audio recorded from two vehicle of model: Volvo XC90. Audio was recorded of sound during normal driving in different environments and of sound after fault injections. The data was recorded via three smartphones: an iPhone 12, Samsung s10E and a Samsung s10 connected to the in-car microphones. These will be referred to as the iPhone, Samsung and In-car microphones respectively throughout this paper.

Mobile phones were used in order to assess the potential for audio anomaly detection that is accessible and reproducible. Since smartphones are widely used a smartphone audio recordings-based model would be easy to replicate. Multiple microphones were used in order to test if a difference in microphone and location makes a difference for our AAD modelling.

The iPhone was positioned on the centre of the windshield and the Samsung was placed between the driver and passenger seat. Both phones were set up using car mounts. We chose these locations in order to capture sounds like that which is heard by the driver and passengers.

The in-car microphones are also located along the center of the car and are the same microphones used for in-car phone calls.

The final data set largely consisted of normal recording form driving on normal roads: on country roads, regular city and suburb asphalt roads and highways. A second set of recordings was collected at the Volvo Hällered Proving Grounds, where there are 15 different tracks that simulate different type of roads. Three Hällered tracks were used for data collection: the *High speed track*, simulating high speed high way driving, the

Country road track, simulating bumps and turn of country roads, and the *Skid pad*, a large circular track allowing for turning and manoeuvring of a car. Normal 'no-fault' driving data was collocated at the High speed and Country road tracks and data from the fault injections were collocated at one or two or all three tracks depending on the nature of the fault injection.

The location (Hällered or Normal road), duration (in seconds), and weather conditions of the normal driving recordings are described in Table 3.1.

Table 3.1: Description of Normal Recordings

Location	Road type	Microphone	Weather	Total duration (sec)
Normal Roads	Country road	in-car	Clear	866
			Rain	1078
		iPhone	Clear	910
			Rain	1075
	Highway	in-car	Clear	1027
			Cloudy	1651
			Rain	666
			Snow	541
		iPhone	Clear	1495
			Cloudy	1674
			Rain	675
			Snow	541
		Samsung	Cloudy	602
	Regular asphalt road	in-car	Cloudy	1411
			windy/snow	1110
		iPhone	Clear	490
			Cloudy	1411
			Snow	1110
		Samsung	Cloudy	1411
Hällered Tracks	Country road track	in-car	Clear	397
			Rain	215
		iPhone	Clear	404
			Rain	219
	High speed track	in-car	Clear	423
			Rain	425
		iPhone	Clear	409
			Rain	446

3.1.1.1 Fault injections

The faults injection chosen for this research were:

- A 5g weight put on the front drive shaft to simulate a U-joint fault.
- A loosened left front wheel
- Two exhaust pipe leaks of size 0.5mm & 1mm to simulate different-sized exhaust pipe holes.

The fault injections made noise at different points of driving. The exhaust pipe leak would make a rough louder-than-normal noise at acceleration while the axle/joint fault would make an unusual singing noise at acceleration, and deceleration and a higher pitched singing noise when turning, the later of which was most prominent when recording at the *Skid pad*. The total amount of abnormal noise recorded for each fault injection varies based on the fault and track and are describe in Table 3.2 along with information on weather condition during recording. The microphone type is not described as all phones were used for data collation of fault injections.

Table 3.2: Description of fault injection recordings

Track	Fault	Weather	Duration (sec) (total recording)
Country road track	Drive shaft weight	Clear	40 (2200)
	Exhaust pipe leak [0.5 mm]	Cloudy	60 (1916)
	Exhaust pipe leak [1 mm]	Rain	130 (2244)
High speed track	Drive shaft weight	Clear	80 (4108)
	Exhaust pipe leak [1 mm]	Rain	40 (1460)
Skid pad	Drive shaft weight	Clear	40 (480)
	Wheel loose nut	Clear	90 (300)

3.1.1.2 Semi-controlled environment

Since this research is of exploratory nature, we wanted to record data in the 'best case scenario'. Therefore, sound was recorded without any talking and all unexpected sound events, such as the dropping of an object, were excluded during data cleaning.

Further, only the segments where abnormal noises occurred from the fault injections were used during modeling. While it would be interesting to use the full recordings for 'sequential' anomaly detection that it will not be explored within the scope of this thesis. The total amount of recorded time is presented together with the abnormal sound duration in Table 3.2

3.1.2 Data Pre-processing

The data was recorded in *.m4a* format, at a sampling rate of 48khz, and converted to *.wav*.

3.1.2.1 Data Segmentation

The recorded audio data was segmented into smaller units of non-overlapping audio clips, based on fixed time intervals to maintain consistency and ensure comparability among the different segment lengths in anomaly detection to analyze the impact of segment length on modeling performance. In this study, we chose to test three different segment lengths: 3 seconds, 5 seconds, and 10 seconds, 5 and 10 second segments are commonly segmentation lengths in audio anomaly detection research [Nunes, 2021] and 3 second clips was included to test if shorter clips preform well in modelling. However, only the results from the best preforming, 10 second, segment length are presented in the results section.

3.1.3 Test - Development - Train split

The data was split into three sets. The first, training set, was used for training the models while the second, development, set was used for validation and evaluation of the model parameters. The last, testing set, was used as a final evaluation of the models on unseen data.

The focus of this project has been to develop a unsupervised learning model, which has only normal data in the training set, the normal/abnormal recordings are shown in Table 3.3. The train, development and test set consisted of 1072, 613 and 526 clips respectively.

In order to have a representative training set, it was sampled from a normal subset of all types of roads recorded during both rainy and clear days. The development set was then sampled with the inclusion of normal Hällered track and exhaust fault recordings, while the reaming normal driving recordings, the drive shaft weight and loose wheel nut fault injection recordings where left as a test set. These samples and their proportions can be seen in 3.4

Table 3.3: Train - Dev - Test split proportions

set	Normal/Abnormal	
Train	Normal	1.000000
Dev	Normal	0.859316
	Abnormal	0.140684
Test	Normal	0.768293
	Abnormal	0.231707

Table 3.4: Fault and Normal recording Road types

Set	Fault injection / road	Proportion
Train	Regular asphalt road	0.5179
	Highway	0.396
	Country road	0.0871
Dev	Regular asphalt road	0.468
	Highway	0.269962
	Exhaust pipe leak [1 mm]	0.099
	High speed track	0.084
	Exhaust pipe leak [0.5 mm]	0.042
	Country road track	0.023
	Country road	0.015
Test	Highway	0.405
	Regular asphalt road	0.226
	Drive shaft weight	0.149
	Country road track	0.0762
	High speed track	0.06
	Wheel loose nut	0.046
	Exhaust pipe leak [1 mm]	0.0244
	Exhaust pipe leak [0.5 mm]	0.0122
	Country road	0.003

3.2 Feature extraction and Anomaly Detection algorithms

Previous research in profiling of sound has focused on identifying patterns and features in the audio signal that can be used for various applications such as speech recognition, music classification, and anomaly detection [Darji, 2017]. This thesis will focus on features based on previous research described in 2.

The features used in the models presented in this paper were: Log Mel Frequency Energy, MFCC and Chroma features. which were extracted using python packages *torchaudio*[Yang et al., 2021], *nnAudio*, speechpy[Torfi, 2017] and librosa. MFCC and LMFE where extracted from the STFT while the chroma features where extracted from the Constant Q- transform of the clips.

Window size 2048 and hop lengths 512 are commonly chosen for audio analysis, but this study used longer windows to reduce the amount of features stored. The MFCC and chroma extraction used a window size of 8192 FFTs with a hop length of 1024, for 12 coefficients and pitch classes, corresponding to a window step of 0.17 and hop length of 0.04 resulting in 108x12, i.e. 1416, frames each. The LMFE was extracted using 0.5 second hop and window lengths and 79 filter banks and resulted in a total of 1404 LMFE frames.

While different windowing sizes where tried the best performing combination was kept for the final comparison of modelling, which included a total of *4236* features per 10 second clip.

3.2.1 Unsupervised Models

As mentioned in previous sections, the focus of this study was be to evaluate commonly used methods for unsupervised anomaly detection. All of which have a variety of parameters and settings. In order to get the optimal performance of these models, a grid search was preformed using the development set as a validation set. This was done for the Local outlier detectors k number of neighbours parameter and One Class Support Vector Machine' (OCSVM's) *nu* parameter. which resulted in $k = 86$ and *nu=0.005*. Further, since the training set contains no abnormal cases the '*contamination*' parameter for LOF and IF were set to *0.0001*.

The Autoencoder (AE) approach was implemented using a Long short-term memory (LSTM) neural network using *Tensorflow* with two hidden layers, a 'relu' activation and latent dimensions 64, with a batch size of 800 and 100 epochs. The LSTM-AE was trained using a Mean absolute error (MAE) loss function.

All the models where evaluated using the metrics described in the following section. The LOF, OCSVM and IF were evaluated using their respective anomaly score which consider data point with lower values as anomalous. Conversely higher reconstruction error indicate anomalous data for the LSTM-AE model.

3.3 Comparing methods - Evaluation metrics

The models were compared via relevant evaluation metrics. In the context of in car audio anomaly detection, it would be desirable to not falsely flag noise as abnormal, that is to have a high true positive rate (TPR) rate for the normal class, that is a high *Precision*. Yet we still want a model that correctly identifies the abnormal class, i.e. the false positive rate (FPR) needs be as low as possible.

Precision-recall curves and Area Under the Curve (AUC) of a ROC curve will therefore be used to assess the performance of the anomaly detection models. The ROC curve gives a visual representation of the TP-FP trade-offs made at different thresholds of a model. Similarly, the Precision-recall curve gives a visual representation of precision-recall trade-offs made by a model. Additionally, an advantage of these metrics in the context of unsupervised models that provide anomaly score is that a threshold can be chosen such that a desired Precision-recall score and/or TP-FP rate is reached in the development set. Lastly, AUC, TPR and FPR score have been the standard for evaluating Anomaly detection models in previous research [Nunes, 2021].

Further, given the diversity of conditions, fault injections and road types in our data set: various diagrams and plots will be used to illustrate different models' performance on different sub-categories.

4

Results

4.1 Feature Extractions and Illustrations

The extracted features, including CENS (Chroma energy normalized), LMFE (Log Mel Frequency Energy), and MFCC (Mel-frequency cepstral coefficients), demonstrate promising results in distinguishing between normal and abnormal classes. Particularly the CENS features which showed a notable difference when plotting the Chroma spectrum where the biggest difference can be seen in the Highway Track recordings.

A sample of Normal and abnormal clips are illustrated in Figures 4.1 and 4.1 where the pitch classes with the biggest difference are colored in red. For the exhaust pipe leak these are pitch classes B, G and F sharp, and for the Drive shaft weight they are pitch classes B, G and D. In the highway recordings the normal clips have a more even distribution of energy across frames while the abnormal clips seem to have a more energy concentrated in specific pitch classes. The normal country road recordings 4.2 do not have as even an energy distribution between classes yet they are still visually distinguishable from the fault injection recordings. This is not too surprising as country road driving has more variation in noise then highway driving.

The distribution of the means by class labels for our extracted features can be seen in Figure 4.3, 4.4 and 4.5. The CENS and LMFE show a clear difference in distributions of means, where the central point of the abnormal clips is different from the normal clips. While the MFCCs means do not show as clear of a difference between labels, these features were kept as they improved modelling. The results from modelling with CENS, LMFE and and MFCCs are presented in the following sections.

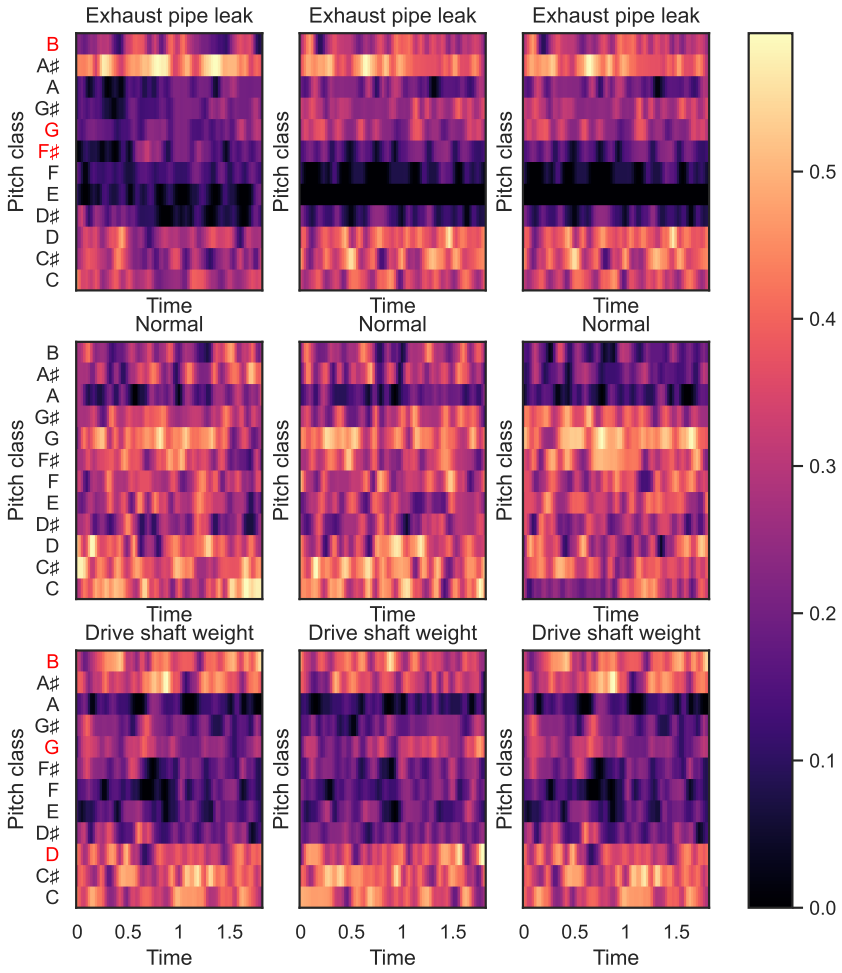


Figure 4.1: Chroma energy normalized (CENS) frames of Highway - normal and fault recordings

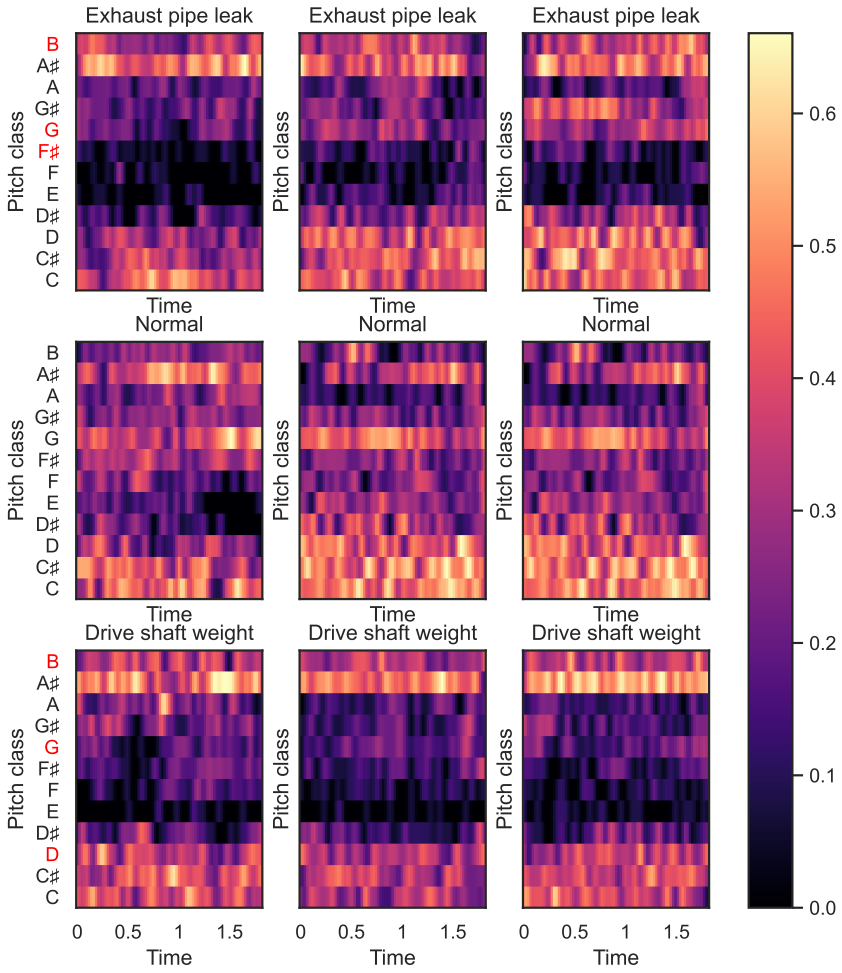


Figure 4.2: Chroma energy normalized (CENS) frames of Country road - normal and fualt recordnings

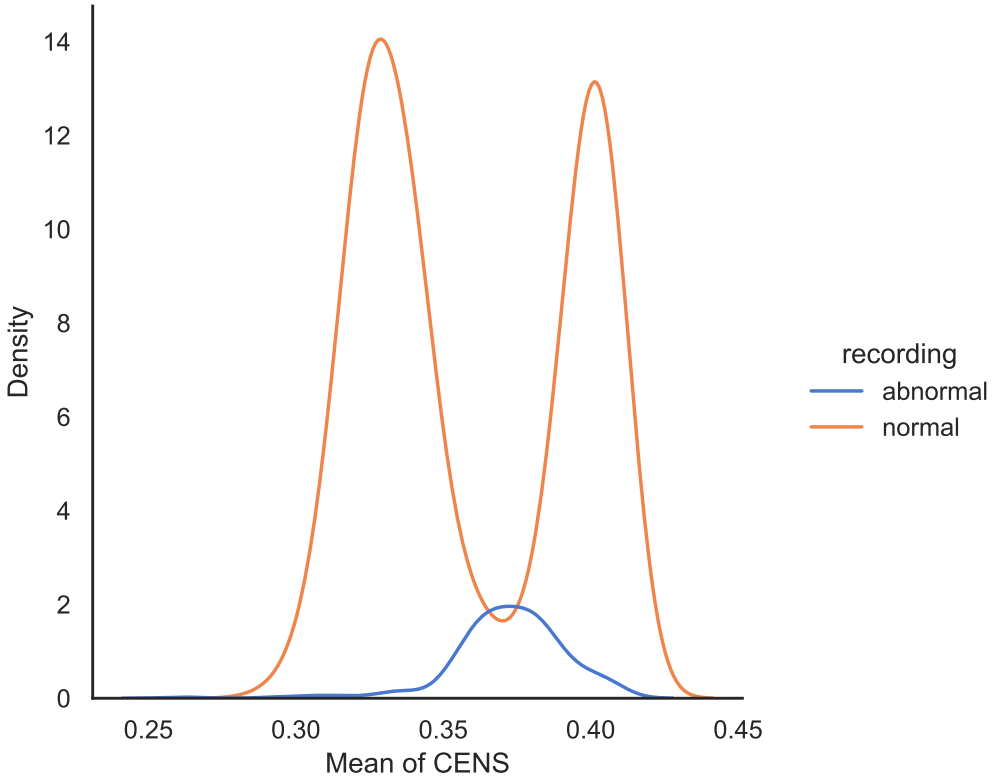


Figure 4.3: Chroma energy normalized (CENS) Means

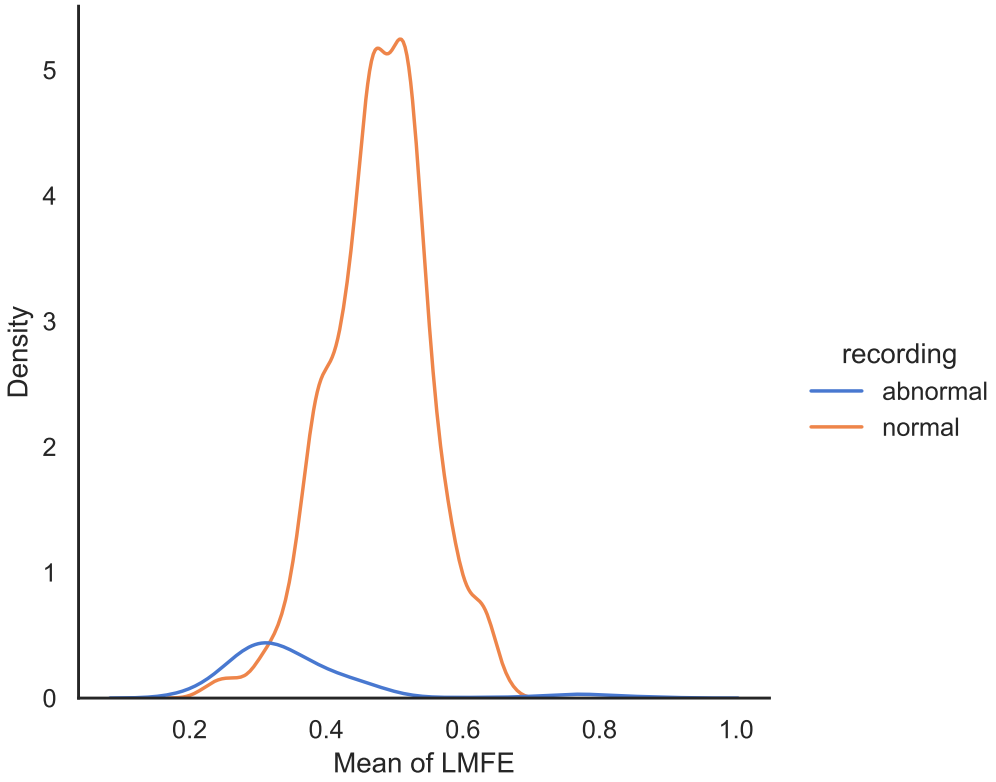


Figure 4.4: Log Mel Frequency Energy Means

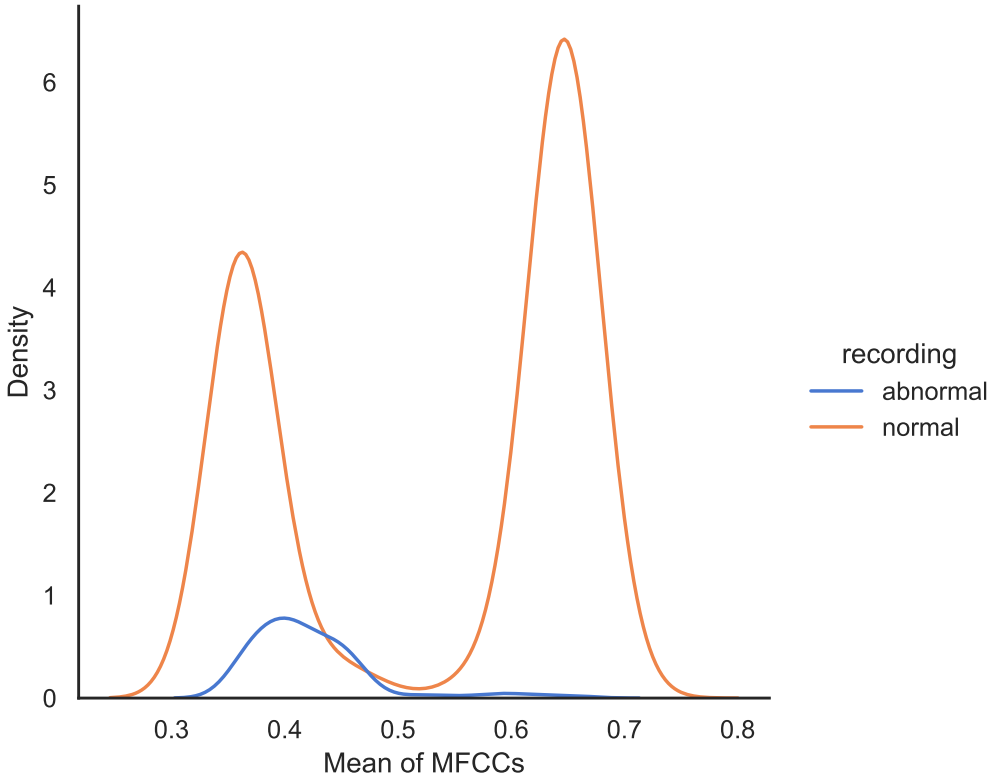


Figure 4.5: MFCC Means

4.2 Modelling results

When modeling with our feature space, all four models: LOF, IF, OCSVM and LSTM-AE showed good performance. Each methods ROC curve, AUC score, PR curve and AP score are displayed in 4.6 for the development set results and figure 4.7 for the test set results. LOF and LSTM-AE have the best AUC score at 0.98 while LSTM-AE has the lowest average precision score at 0.75. The LSTM-AE PR curve tells us that the models precision drastically drop past a certain reconstruction error threshold. Given that the positive class in this case is the normal class, its the precision of normal classes that drops.

The other models, OCSVM and IF also perform well with AUC at 0.97 and AP score at 1. Figure 4.8 and 4.9 show the distribution of each models anomaly score on the development and test sets respectively, where the LSTM-AE's score is such that a higher score means more less normal while the other models have a negative anomaly score. By looking at the distribution we see that the models do indeed separate the classes very well. All model scores make a clear distinction between normal and abnormal audio data in the development set and test set, while the LSMT-AE model has the biggest overlap in scores between the two classes in the test set if compared to the other models.

As for the test set metrics in Figure 4.7, the models all preform less well with a drop around 0.10 in AUC score, which still gives a AUC score of over 0.8 for respective models. The models are thus all good at detecting anomalies. However, the performance of the models greatly depend on the choice of threshold for the anomaly scores, which if give varying results if chosen in the training and development of stages of the models. In this regard the LOF, IF and OCSVM preform the best. A comparison is made between the LOF and LSTM-AE as an illustrative example in the following section.

Additionally, table 4.1 shows the average model scores by class and microphone used. Here we can see that on average abnormal recordings lower and negative scores by the decision function models and a higher reconstruction error for the neural network model. There is also not a big difference in score by microphone. Given that the full recording were used for the normal recordings when modeling, and only a the subset of abnormal noise was selected from the full fault injection recordings, a model could be built to detect anomalies by on the full recordings of driving with a car fault by comparing the aggregated 10 second clip anomaly scores. However, this possibility was not explored in this thesis.

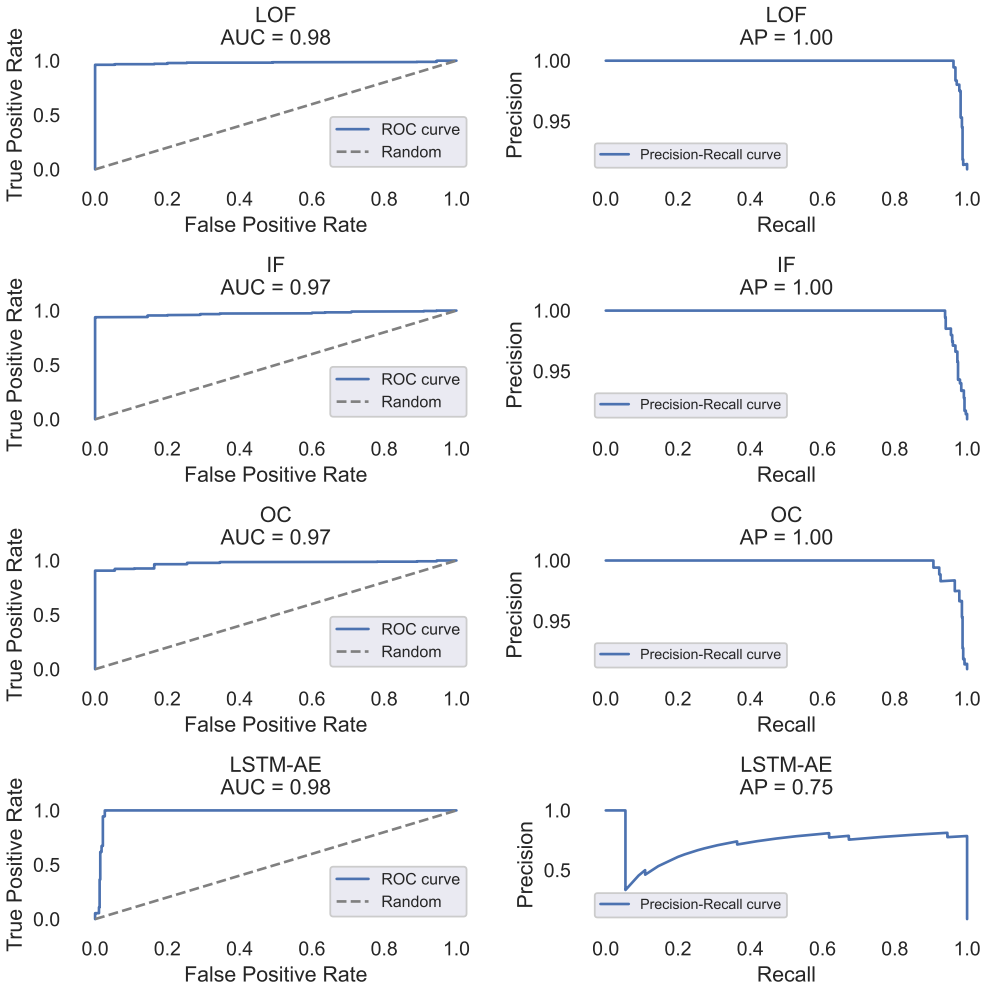


Figure 4.6: Development set Metrics for LOF, IF and OCSVM,and LSTM-AE

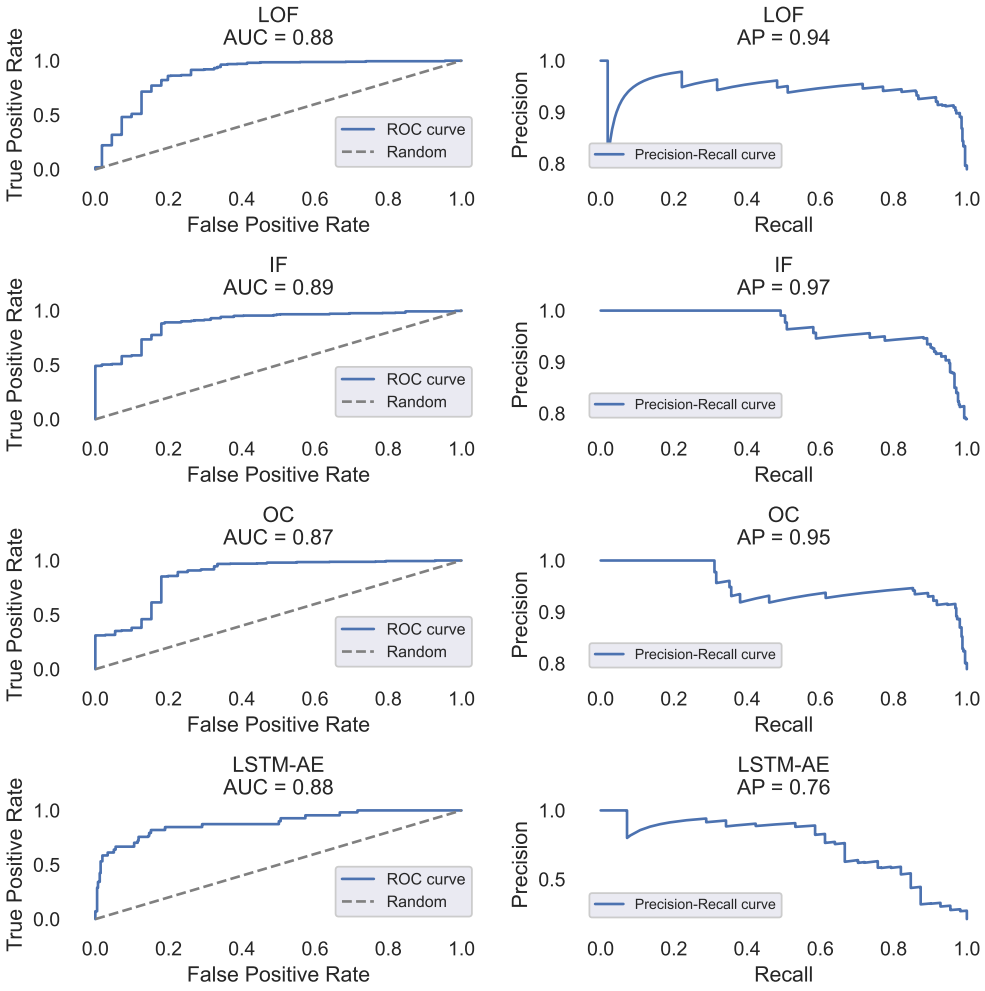


Figure 4.7: Test set Metrics for LOF, IF, OCSVM and LSTM-AE

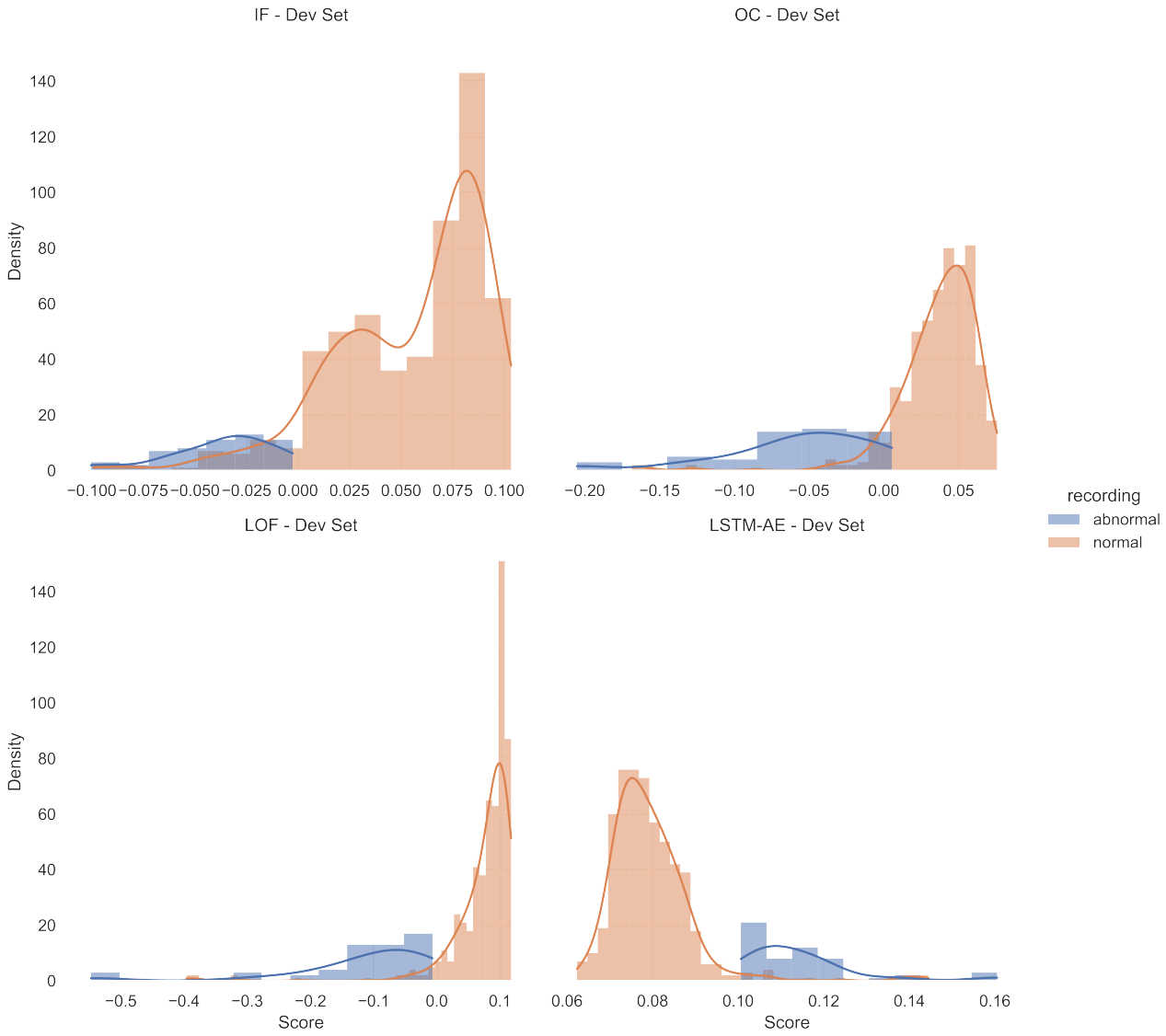


Figure 4.8: Distribution of scores on the Development set for all models

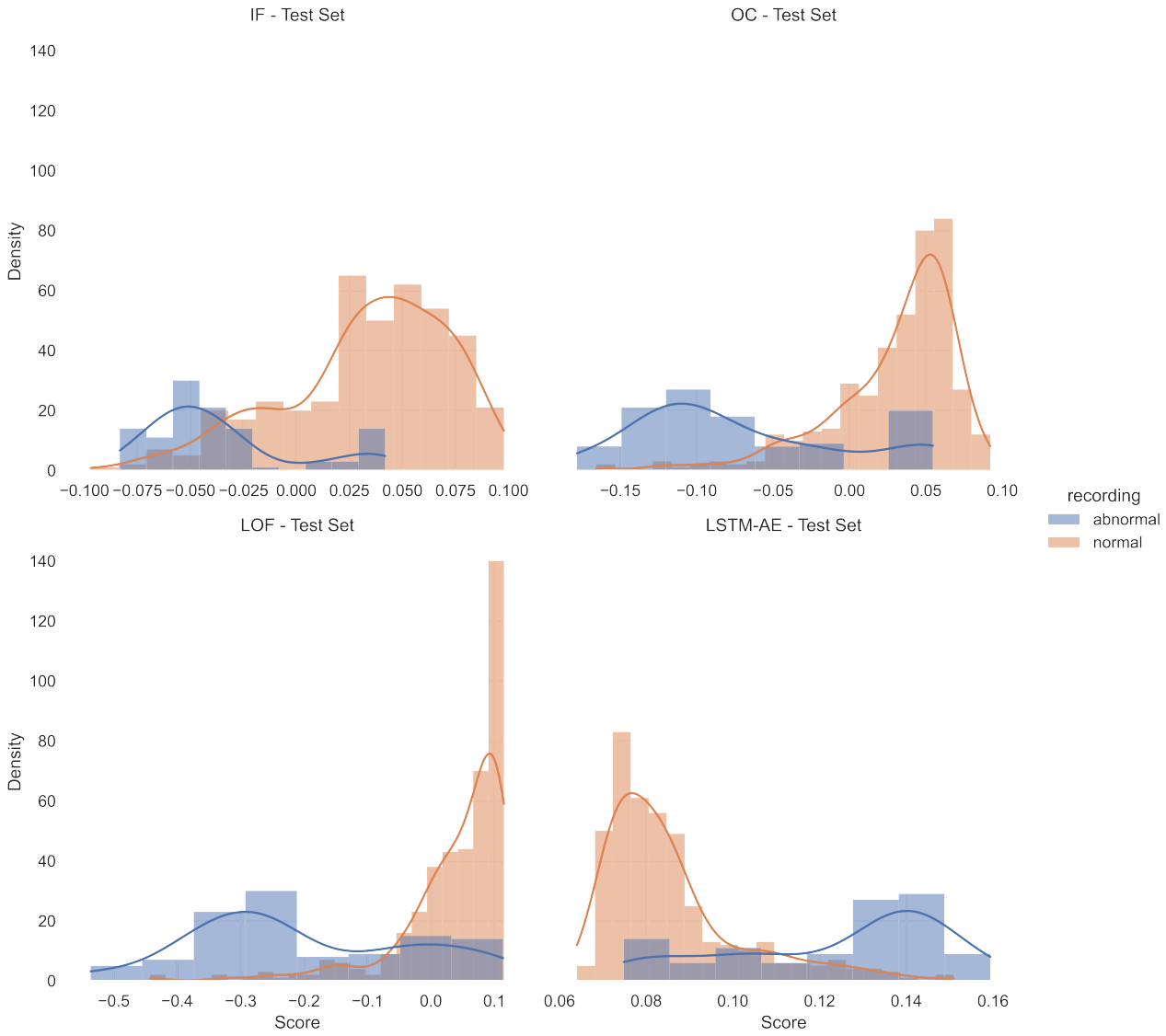


Figure 4.9: Distribution of scores on the Test set for all models

Table 4.1: Median prediction score by Class, Microphone and Model

recording	variable	brand	Anomaly score
abnormal	IF	in-car	-0.041937
		iphone	-0.042046
		samsung	-0.041991
	LOF	in-car	-0.163040
		iphone	-0.173543
		samsung	-0.149762
	LSTM-AE	in-car	0.119340
		iphone	0.120137
		samsung	0.119211
	OC	in-car	-0.075701
		iphone	-0.075990
		samsung	-0.073149
normal	IF	in-car	0.046265
		iphone	0.054393
		samsung	0.078799
	LOF	in-car	0.090711
		iphone	0.077525
		samsung	0.082236
	LSTM-AE	in-car	0.077853
		iphone	0.080064
		samsung	0.072077
	OC	in-car	0.041617
		iphone	0.041140
		samsung	0.041904

4.2.1 LSTM vs LOF

In this section a comparison is made between our LSTM-AE and LOF models. While the two models have similar AUC scores (see Figure 4.10) and anomaly score distributions as seen in Figure 4.9, a choice of anomaly detection in using LSTM-AE does not translate well to the test set. For instance, if we want 0.8 precision and recall in the development set using the LSTM-AE model we might chose the reconstruction error threshold of 0.10. Given a threshold of -0.05 for the LOF model chosen by the same criteria, the LOF preforms far better on the test set. This is due to the LOF being better at distinguishing between normal recordings and abnormal recording as represented by its AP score. Figure 4.12 shows the anomaly scores of the models by recording, in both the test and development set. Here we can see that it's the normal country road track driving that the LSTM-AE model has the most trouble reconstructing. The LOF model also gives lower scores to these recordings but classifies less normal recordings as anomalous at a -0.05 threshold. This shows that the Precision-Recall curve gives us a good representation of a models ability to detect anomalies.

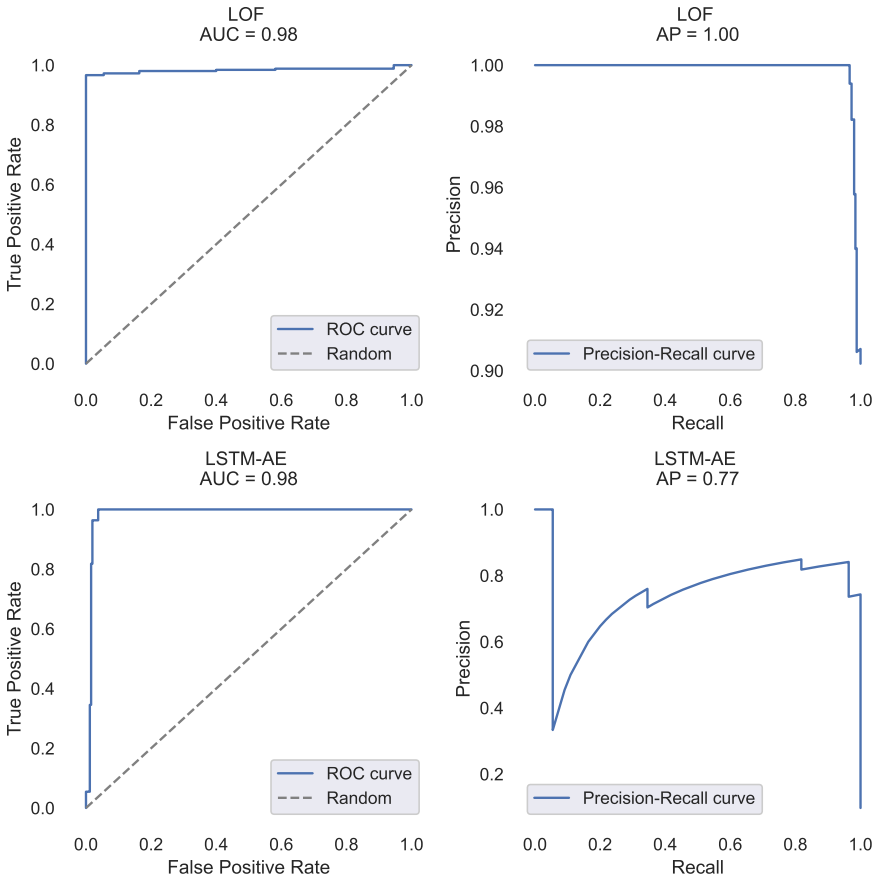


Figure 4.10: LOF vs LSTM-AE Model - Development set

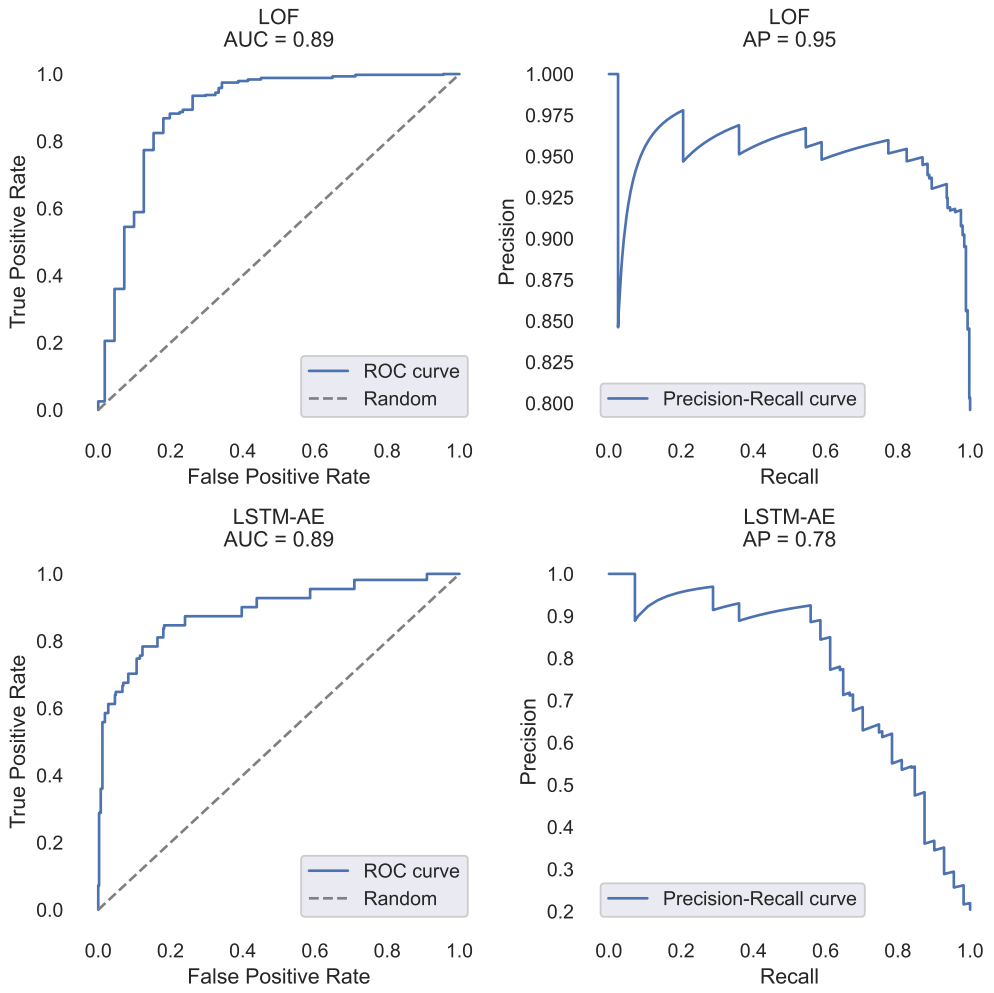


Figure 4.11: LOF vs LSTM-AE Model - Test set

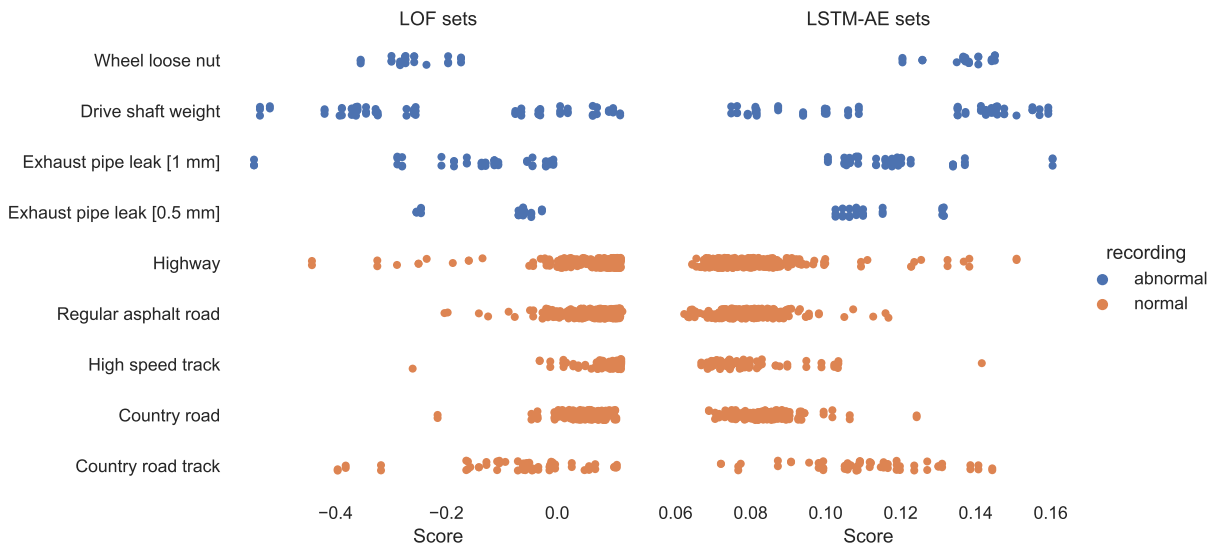


Figure 4.12: LOF vs LSTM-AE Model - Anomaly score by recording type

4.2.2 Effect of microphone

The presented results indicate that AAD modelling with Chroma, LMEF and MFCC features show promising results on audio recording from unseen environments and faults, When all microphone recordings are used during training and validation. To see how the performance is on recordings from an unseen microphone; the LSTM-AE and LOF models where remodeled, on only the Samsung and IPhone recordings using the same settings and parameters as in sections.

Anomaly detection thresholds were decided based on the loss distribution of the development set and a desired minimum of 0.80 recall/precision scores based on the Curves seen in table 4.13. Their performance was then finally evaluated on a test set with the test set recordings from the same microphones and a separate test set with the in-car subset recording previously excluded. It is clear from ROC and P-R curves of the LOF and LSTM-AE model that the LOF preformed much better on a unseen microphone, with an average precision of 0.99 vs the LSTM-AE models 0.63. Interestingly, the LOF increased in both AUC and AP score compared to previous results. The thresholds chosen for the LSTM-AE based on the development set is illustrated together with scores and recording type in figure 4.14, where we can see that a large chunk of the normal recordings are classified as abnormal in the in-car set. Conversely, figure 4.15, shows that the LOF model makes an even better distinction then when trained on all microphones.

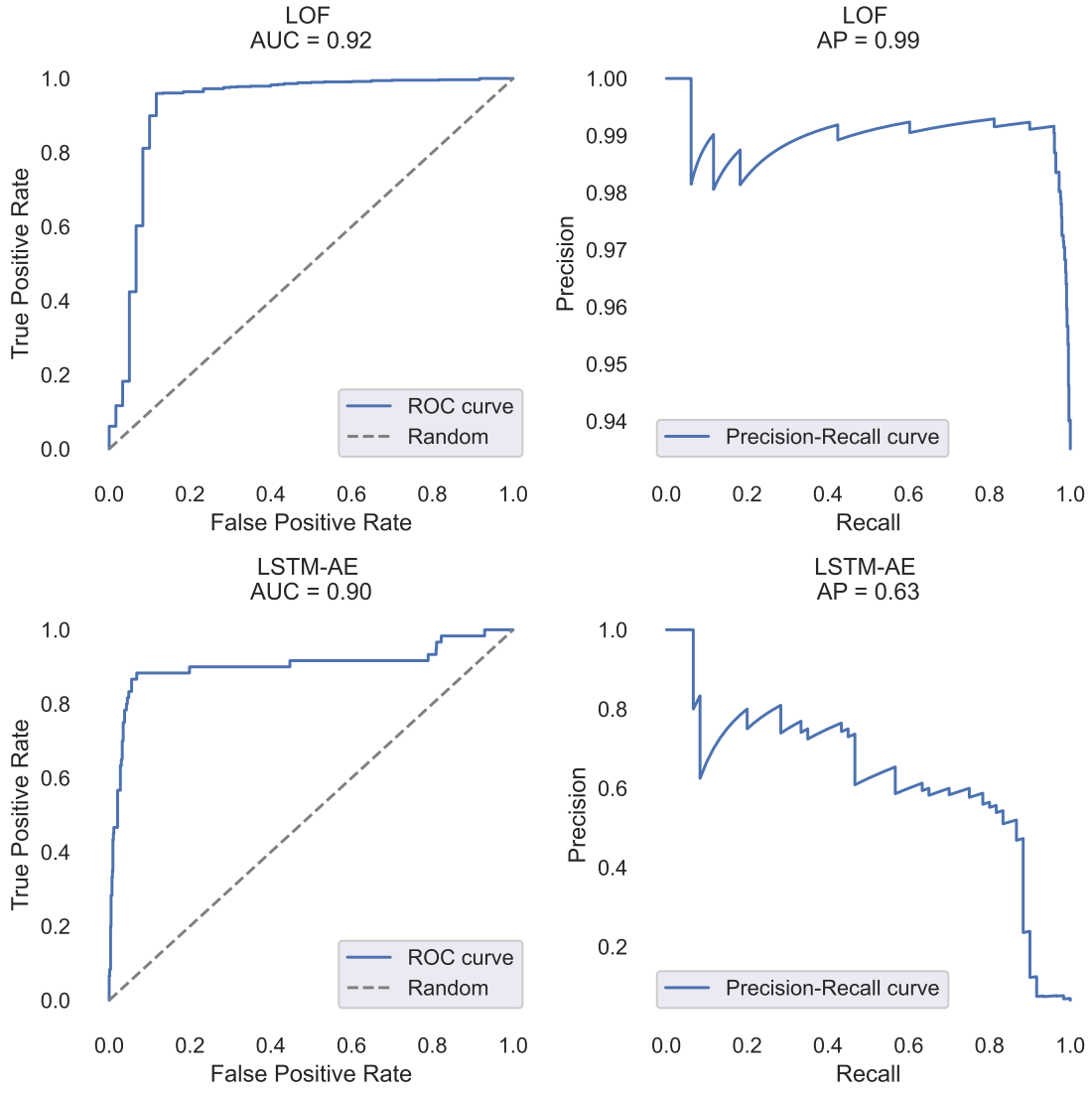


Figure 4.13: ROC and Precision-Recall curves on in-car set by Model

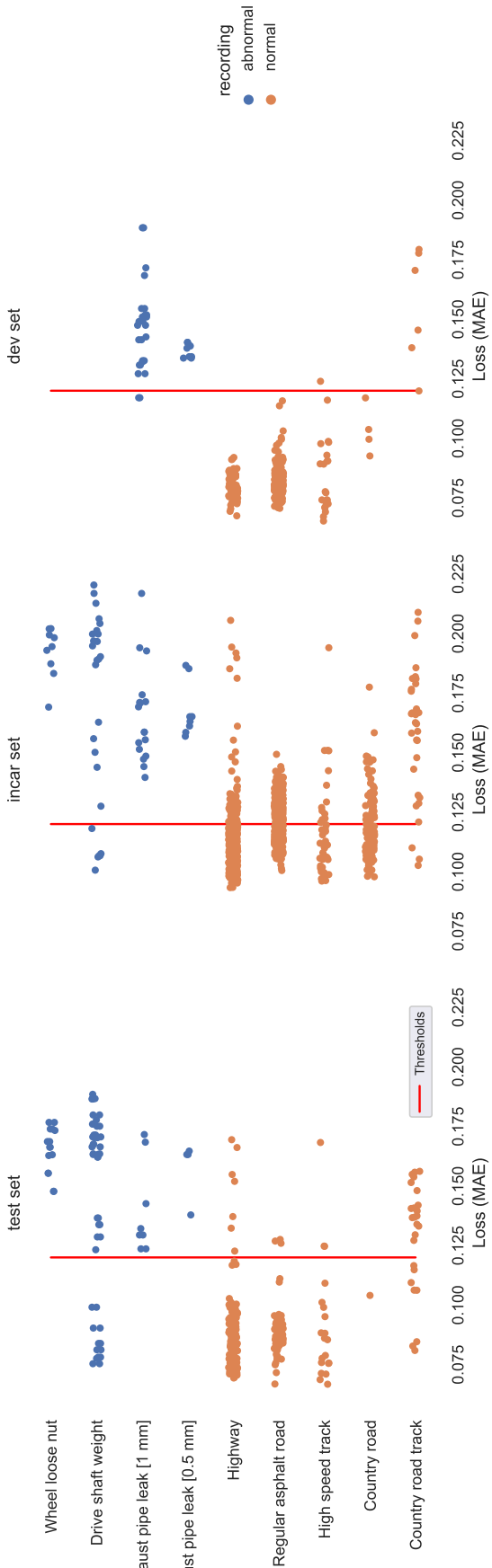


Figure 4.14: LSTM-AE Reconstruction errors by Fault and Road type - compared between dev, test and in-car set

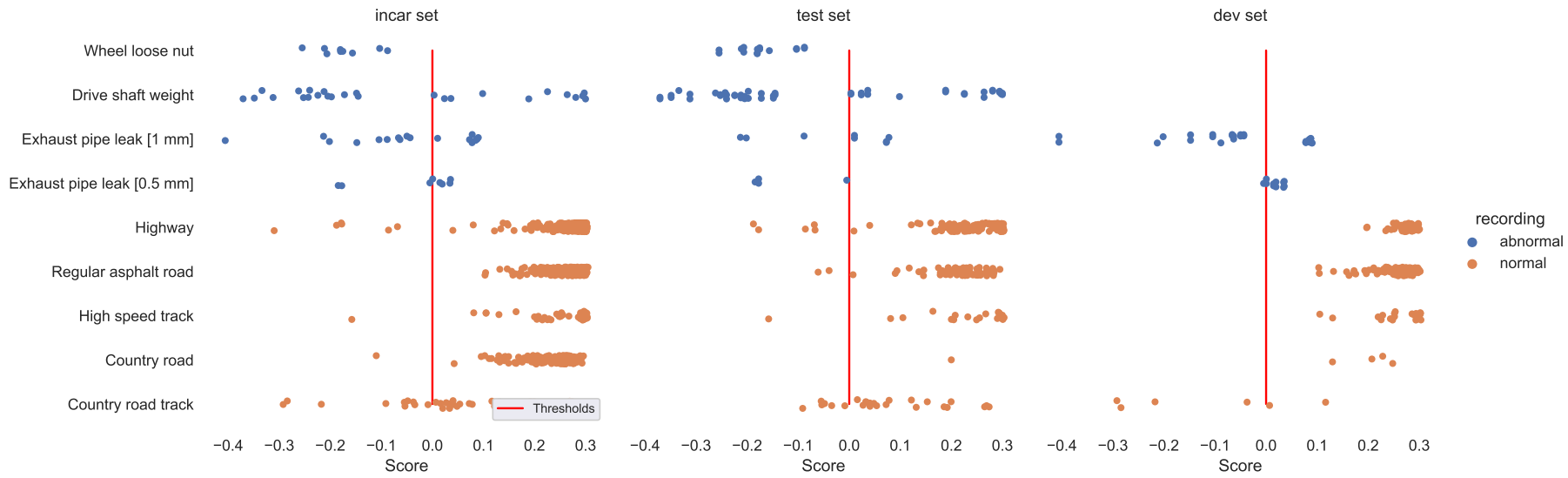


Figure 4.15: LOF Anomaly Scores by Fault and Road type - compared between dev, test and in-car set

5

Conclusion

Our results demonstrate the effectiveness Mel Frequency and Chroma features and unsupervised machine learning in detecting anomalies in car audio data. In particular Chroma features and the Local outlier factor, the first of which is more commonly used for music classification and the later not commonly used for audio data. While all the tested models preformed well, the Local Outlier factor excelled in both AUC score, Average Precision on our test set and on classifying audio data from an unseen microphone making it a great choice for audio anomaly detection in the context of cars. This project has been limited in that it has only explored a hand full of car faults, but shows great potential for further research, where the methodology presented can be further explored by testing on new car faults and normal driving data.

To conclude, these findings contribute to the understanding of different approaches and their performance in identifying anomalies in car audio data, which can provide insights for developing effective anomaly detection systems in automotive applications.

Bibliography

[Bonet-Solà and Alsina-Pagès, 2021] Bonet-Solà, D. and Alsina-Pagès, R. M. (2021). A comparative survey of feature extraction and machine learning methods in diverse acoustic environments. *Sensors*, 21(4).

[Brown, 1991] Brown, J. C. (1991). Calculation of a constant q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434.

[Dandare and Dudul, 2013] Dandare, S. and Dudul, S. (2013). Multiple fault detection in typical automobile engines: A soft computing approach. *WSEAS Trans. Signal Process*, 10:254–262.

[Darji, 2017] Darji, M. C. (2017). Audio signal processing: A review of audio signal classification features. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2(3):227–230.

[Ewert, 2011] Ewert, S. (2011). Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features. In *Proc. ISMIR*.

[Kabiri and Makinejad, 2011] Kabiri, P. and Makinejad, A. (2011). Using pca in acoustic emission condition monitoring to detect faults in an automobile engine. In *29th European Conference on Acoustic Emission Testing (EWGAE2010)*, pages 8–10.

[Khan et al., 2021] Khan, A. S., Ahmad, Z., Abdullah, J., and Ahmad, F. (2021). A spectrogram image-based network anomaly detection system using deep convolutional neural network. *IEEE Access*, 9:87079–87093.

[Lecomte et al., 2011] Lecomte, S., Lengellé, R., Richard, C., Capman, F., and Ravera, B. (2011). Abnormal events detection using unsupervised one-class svm-application to audio surveillance and evaluation. In *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 124–129. IEEE.

[Liu et al., 2012] Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2012). Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39.

[Muller and Kurth, 2006] Muller, M. and Kurth, F. (2006). Enhancing similarity matrices

for music audio analysis. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5, pages V–V. IEEE.

[Muller et al., 2005] Muller, M., Kurth, F., and Clausen, M. (2005). Chroma-based statistical audio features for audio matching. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, pages 275–278.

[Mushtaq and Su, 2020] Mushtaq, Z. and Su, S.-F. (2020). Efficient classification of environmental sounds through multiple features aggregation and data enhancement techniques for spectrogram images. *Symmetry*, 12(11).

[Nunes, 2021] Nunes, E. C. (2021). Anomalous sound detection with machine learning: A systematic review. *arXiv preprint arXiv:2102.07820*.

[O'Shaughnessy, 2000] O'Shaughnessy, D. (2000). *Speech communications : human and machine*. IEEE Press, New York, 2. ed. edition.

[Pereira et al., 2021] Pereira, P. J., Coelho, G., Ribeiro, A., Matos, L. M., Nunes, E. C., Ferreira, A., Pilastri, A., and Cortez, P. (2021). Using deep autoencoders for in-vehicle audio anomaly detection. *Procedia Computer Science*, 192:298–307.

[Sharma et al., 2020] Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, 158:107020.

[Siegel et al., 2016] Siegel, J., Kumar, S., Ehrenberg, I., and Sarma, S. (2016). Engine misfire detection with pervasive mobile audio. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III 16*, pages 226–241. Springer.

[Siegel et al., 2017] Siegel, J. E., Bhattacharyya, R., Kumar, S., and Sarma, S. E. (2017). Air filter particulate loading detection using smartphone audio and optimized ensemble classification. *Engineering Applications of Artificial Intelligence*, 66:104–112.

[Tagawa et al., 2021] Tagawa, Y., Maskeliūnas, R., and Damaševičius, R. (2021). Acoustic anomaly detection of mechanical failures in noisy real-life factory environments. *Electronics*, 10(19):2329.

[Torfi, 2017] Torfi, A. (2017). SpeechPy: Speech recognition and feature extraction.

[Yang et al., 2021] Yang, Y.-Y., Hira, M., Ni, Z., Chourdia, A., Astafurov, A., Chen, C., Yeh, C.-F., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E. Z., Lian, J., Mahadeokar, J., Hwang, J., Chen, J., Goldsborough, P., Roy, P., Narenthiran, S., Watanabe, S., Chintala, S., Quenneville-Bélair, V., and Shi, Y. (2021). Torchaudio: Building blocks for audio and speech processing. *arXiv preprint arXiv:2110.15018*.

[Yu et al., 2020] Yu, S., Li, X., Zhao, L., and Wang, J. (2020). Hyperspectral anomaly

detection based on low-rank representation using local outlier factor. *IEEE Geoscience and Remote Sensing Letters*, 18(7):1279–1283.