Scrutinizing Sounds: A Study on Reconstruction Based Anomaly Detection in Vinyl Records' Audio Data

JORIS JONKERS, University of Twente, The Netherlands
LISANDRO JIMENEZ ROA, University of Twente, The Netherlands
MARIELLE STOELINGA, University of Twente, The Netherlands

Anomalous sound detection is an important task in many real-world applications such as surveillance, quality control, and healthcare. For this research, we propose a novel application for state-of-the-art anomaly detection techniques. The proposed application is the detection of anomalous sounds within the recordings of vinyl records, which poses significant challenges due to the context-sensitive nature of the music stored on these records. This application poses further challenges, as due to the laborious process of labelling no labels were available for the used datasets. Considering these limitations use was made of a reconstruction-based detector featuring a denoising autoencoder (**DAE**) which uses a Long Short-Term Memory (**BLSTM**) recurrent neural network (**RNN**). The study will investigate features which may serve to detect such anomalous sounds and the effectiveness of using a reconstruction-based anomaly detector for the stated problem.

Additional Key Words and Phrases: Denoising Auto-Encoders, Minimum Redundancy Maximum Relevance, Novelty Detection, Anomalous Sound Detection, Vinyl Records

1 INTRODUCTION

Despite the contemporary shift from tangible to digital media, vinyl records have seen a surprising resurgence in popularity [26]. However, the very nature of the vinyl medium poses several limitations, primarily the vulnerability of the grooves that encode the audio data. This structural fragility often leads to the generation of anomalous sounds during playback, such as pops and crackles. These anomalies compromise the listener's experience and lead to issues in determining the quality and therefore value of vinyl records.

1.1 Problem Definition

Current solutions for assessing the condition of used vinyl records rely on grading standards [1]. However, these methods are often subject to the seller's bias, potentially leading to inflated estimations of the record's quality and, consequently, its market value. The absence of an unbiased, objective grading system that operates without human intervention is a clear gap in the field.

Moreover, there is a lack of studies focusing specifically on detecting anomalies in vinyl records. Previous research has covered anomaly detection within data extensively [5, 6, 20], but few, if any, have applied these techniques to the detection of anomalies within musical recordings. Lu et al. [15] attempted to detect anomalies within musical datasets, but the study was limited to the detection of wholely anomalous samples instead of the localization of anomalies within musical compositions.

The application of anomaly detection to vinyl records introduces unique challenges due to the time-series nature of audio data [6]. Identifying the anomalous sounds in a recording without considering their temporal context is a significant challenge due to the complex nature of musical compositions. This study aims to address these gaps and provide a preliminary step towards developing an unbiased, automated grading system for vinyl records.

1.2 Research Goals and Questions

The study focuses on identifying the most effective features and anomaly detection techniques for detecting anomalies in recordings of vinyl records. The primary research questions guiding this study are as follows:

- What features are most effective at distinguishing between musical compositions and the types of anomalies that occur on vinyl records?
- To what extent can a reconstruction-based model be used to identify anomalous sounds within vinyl record recordings?

2 BACKGROUND AND RELATED WORK

Vinyl records store information through variations in the depth of their grooves. For mono sound, these variations occur along one axis, while for stereo sound, they occur along two axes. During playback, a needle moves linearly through these grooves, its sideways movements corresponding to the varying groove depths. This mechanism retrieves the stored information and translates it into sound. However, any surface damage or debris within these grooves can cause the needle to deflect. Such deflections produce anomalous sounds during playback. This study aims to find a method to detect these anomalies, the following sections will describe the chosen models and the data used during this study.

2.1 Anomaly Detection Approaches

Anomaly detection has been a topic of active research since 1999 [24]. Previous studies have compared different anomaly detection techniques, namely **Classification-Based Novelty Detection**, **Statistical Novelty Detection**, **Distance-Based Novelty Detection**, **Clustering-Based Novelty Detection**, and **Reconstruction-Based Novelty Detection** [5, 20].

Each of these techniques is better suited to some problems than others. This study considers the specifics of the vinyl records application and the available data to decide which technique is the most appropriate.

TScIT 38, July 7, 2023, Enschede, The Netherlands

Classification-Based Novelty Detection techniques are ruled out primarily due to the insufficient availability of anomalies. Classification-based approaches can over-fit on the specific representation of the available anomalies.

Statistical Novelty Detection techniques are not suitable due to the high dimensionality of the data used in this study. These techniques are less effective with complex distributions required for high-dimensional datasets like the musical data this study involves.

Distance-Based Novelty Detection techniques, which use distance measures, are also not suitable due to the high dimensionality of the target data and the variability between musical compositions.

Similarly, **Clustering-Based Novelty Detection** techniques were also excluded due to their tendency to over-fit on training data and the difficulty they have in dealing with high-dimensional spaces.

After considering all other techniques, we chose **Reconstruction-Based Novelty Detection**. These techniques aim to reconstruct the normal representation of data and are less likely to over-fit on the few examples of anomalies available. They also allow us to leverage the vast amount of available normal data during the training process.

2.2 Reconstruction-Based Detection Techniques

Reconstruction-Based Novelty Detection techniques have gained interest due to their efficacy in handling context-sensitive detection problems, particularly the use of autoencoders [17, 18].

In this study, we use Bidirectional Long Short-Term Memory (**BLSTM**) and Long Short-Term Memory (**LSTM**) Recurrent Neural Network (**RNN**) Denoising Autoencoders (**DAEs**). Which have shown markedly better performance than distance-based and probability-based novelty detection techniques, when applied to context-sensitive time-series data [16, 17, 22].

Reconstruction-based models learn how to remove noise from a signal instead of learning the representation of the noise itself. Thus, they learn the behaviour of the signal, not the behaviour of the noise, which allows these models to detect anomalies different from those they were originally trained on. This is particularly beneficial due to the limited number of available examples of anomalies.

2.3 Audio Features

Given the short duration of this study, a thorough custom feature derivation process for distinguishing between musical compositions was impractical. Therefore, at the offset of the study, a literature review was conducted to identify relevant features commonly used in such differentiation tasks, particularly focusing on anomaly detection and background noise differentiation.

We then analyzed different features that might detect specific audio signatures of the anomalies, concentrating on the Music Information Retrieval (**MIR**) field that employs unique features such as audio fingerprints. The **MIR** field, successfully applied in applications like Shazam and modern digital assistants, has invested significant effort in identifying features to discriminate between different musical compositions. Studies like [23, 25] emphasized the importance of 'texture' features and **Mel-frequency Cepstral Coefficients (MFCC)** in genre classification and speech/music discrimination problems respectively.

Preceding studies have explored anomaly detection within music datasets, identifying spectral, temporal, and rhythmic features for classification [13, 15]. In this study, we extend upon these features with additional features from the field of **MIR**.

From our literature review, we found numerous multi-dimensional features, such as **MFCC**, **Chroma**, and **Tonnetz**, which are frequently used for audio analysis. These features span both **MIR** and speech classification fields [4, 8–12, 25], an overview of these features and their related studies is given in Table 5.

The literature also emphasized the significance of certain derived features, originally proposed by [25] and subsequently validated by [13]. These features were found to be applicable in the domain of anomaly detection by [15], who explored the identification of anomalies within a dataset of musical compositions.

In this study, we examined these derived features due to their demonstrated relevance in related use cases. These features are derived from 'texture windows', as described by [25]. These texture features encompass the mean, variance, and low energy values measured over a predefined number of windows, which have been shown to significantly enhance accuracy in classification problems.

Specifically, the low-energy value, a term infrequently used outside of [25], is a binary statistical feature that indicates whether the energy of a window falls below 90% of the feature's mean energy, thereby representing deviations from the mean value of a feature.

3 DATASETS

The study used a large dataset comprising two types of recordings: those from vinyl records, and those produced digitally described in greater detail in this section. Due to the labelling effort required, only the digital masters' dataset is labelled. A description of how the recordings of vinyl records were leveraged to produce examples of anomalies is given in Section 3.3.

3.1 Vinyl Recordings

Fig. 1. Recording Setup for Vinyl Recordings

A total of 54 vinyl records spanning 31 artists and 60 genres were digitized consistently using a Raspberry Pi 4B with a HifiBerry DAC+ ADC expansion board and a Thorens TD170 record player as shown

in Figure 1. A choice was made to use as few constituent parts in the recording setup as possible in order to reduce the variability and the likelihood of anomalies occurring due to an incorrect setup. These recordings contain the entirety of the discography of The Beatles from 1963-1970.

3.2 Digital Masters

The digital masters' dataset comprises over 12,000 songs from 580 artists and 300 genres. But due to time limitations, only a subset from the Beatles (1963-1970) was used as expanding the amount of training data is directly proportional to training times. An overview of the albums used is given in Table 14.

3.3 Anomaly Examples

The study used expert knowledge to classify representative anomalies captured during the silence before a song starts on a vinyl record. The classification includes crackle, pop, surface noise, flutter/wow, and needle jumps, each associated with specific types of record defects as outlined in Table 1. For this study, a collection of each of these types of anomalies was gathered, for use in the production of labelled data as described in Section 5.1.

Table 1. Anomalies observed in vinyl records and their causes

Anomalies \ Causes	Contaminants	Static Charge	Surface Defect	Warping
Crackle	✓	✓	✓	
Pop	✓	✓	✓	
Continuous Surface Noise	✓		✓	✓
Flutter/Wow				✓
Needle Jumps			✓	✓

The availability for each of these types of anomaly is given in Table 2, for each anomaly an indication of the number of examples available and the lengths of these examples is given. It may be noted that the number of examples is rather low, this was a primary consideration of the model used during this study.

Table 2. Availability of Anomaly Samples

Anomaly Type	Count	Average Length (s)	Minimum Length (s)	Maximum Length (s)	Summed Length (s)
Pop	36	0.03	3.62e-4	0.42	1.16
Hiss	2	1.54	0.28	2.81	3.08
Crackle	27	0.33	0.01	1.63	8.90
Surface Noise	34	6.07	1.25	25.70	206.53
Needle Drop	14	0.59	0.20	1.09	8.22

4 METHODOLOGY

This study employs an investigation around two fundamental research questions geared towards identifying irregularities in the audio of vinyl records. This involves feature selection, metric identification, and the establishment of a reproducible experiment.

4.1 Feature Selection

Identifying deviations in music and abnormal sounds requires an effective feature selection process. This process integrates literature review and analytical methodologies. The initial features set, offering high information gain, is derived from existing literature in the fields of **MIR** and speech classification.

However, the difference between the detection of audio anomalies in vinyl records and the identification of musical compositions in **MIR** calls for a comprehensive examination of feature relevance. Therefore, we evaluate the relevance of these features with respect to anomalies in vinyl records. To navigate this process, we use the **minimal-redundancy maximal-relevance (mRMR)** feature selection method [19], which assists in identifying a relevant, non-redundant set of features.

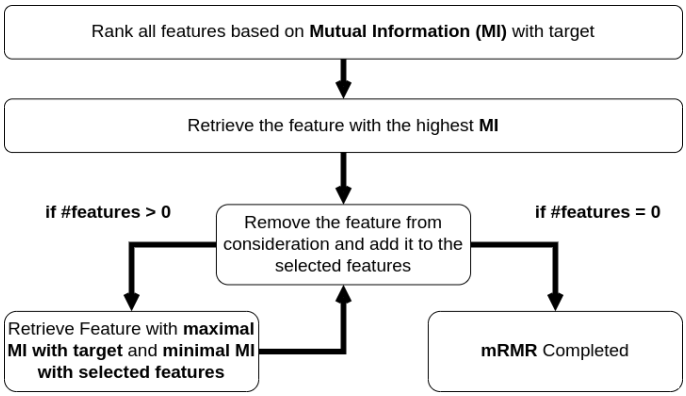

Fig. 2. Schematic overview of mRMR algorithm

The Maximum Relevance Minimum Redundancy (**mRMR**) algorithm operates by maximizing feature relevance and minimizing feature redundancy. Relevance is quantified as a feature's **Mutual Information (MI)** with the target class, while redundancy is quantified as the mutual information between features. The algorithm commences by selecting the feature with the highest mutual information with the target, thereby being maximally relevant. In subsequent iterations, it selects the feature yielding the maximum **Mutual Information (MI)** with the target yet minimum average **Mutual Information (MI)** with previously selected features, thereby ensuring minimum redundancy.

Implementing **mRMR** is beneficial as it identifies a comprehensive yet non-redundant feature set, improving model performance and interpretability by mitigating the inclusion of superfluous, highly correlated features.

The **mRMR** method generates a list of features in order of selection, providing a form of ranking considering both target relevance and feature redundancy. This ranking represents a crucial sequence of feature importance and relevance for distinguishing between music and vinyl record anomalies.

Among the various **mRMR** variants, the **MIQ** (mutual information quotient) variant was selected for this study due to its robust performance and the availability of open-source implementations, despite its low computational efficiency as compared to the **FCQ** variant [7, 27].

4.2 Model Performance

The detection model used in this study is two part, namely the **DAE** which is trained to remove noise from given data, and the detector which leverages the results produced by the **DAE** to make

detections. For this study the relative performance of **LSTM** and **BLSTM** neural networks for use in a **DAE** are compared, in order to determine the relative strengths and weaknesses of these two types of RNN. These two parts are not necessarily sympathetic as an improvement in reconstruction may also lead to a decrease in detection performance, therefore these two parts have been separated in such a way that their individual performance metrics do not influence each other in any way. In this section, the method for determining the performance of the two parts of the detector will be described.

4.2.1 DAE Performance. This study aims to detect anomalies through the use of a **DAE**, which is to be trained using differing features, in order to determine whether a **DAE** detector is able to find anomalies within musical compositions. The ability of autoencoders to construct their own features from given input data is sometimes leveraged to directly reconstruct waveforms or images, however, as the waveforms of which musical compositions are comprised are very complex the choice was made to use features as an input and output of the model instead.

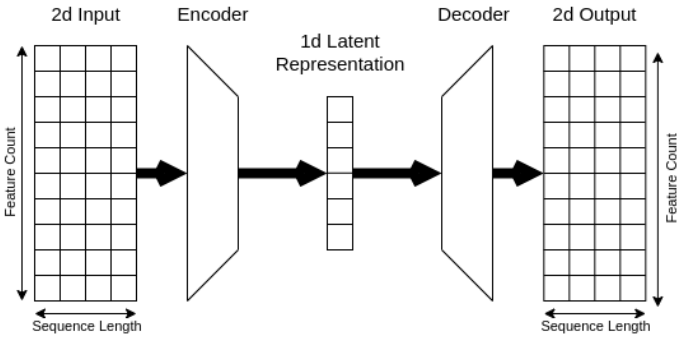

Fig. 3. Schematic overview of Auto-Encoder

The goal of a DAE is to reconstruct some form of distorted, also called "noisy", data back to its undistorted state. This is achieved by giving a DAE distorted data as an input and setting undistorted data as a target, which serves to teach the DAE how to remove distortions from data. Denoising Autoencoders are typically compressive, which indicates that the size of the latent representation is of a lower dimension than the input and target data, a schematic overview of this process is given in Figure 3.

The performance metric used for the **DAE** is the reconstruction error, defined as the mean absolute error **Mean Absolute Error (MAE)** between the output of the DAE and the target data. The choice was made to use the **Mean Absolute Error (MAE)** as the mean difference is independent of the size of the input and target data, which may vary depending on the number of features used.

4.2.2 Detector Performance. Separate from the process of training the **DAE**, is the process of determining the performance of the detector which leverages the **DAE**. **Reconstruction-Based Novelty Detection** techniques employ thresholds to determine whether or not an anomaly is present in any given location, however, there are several ways in which these thresholds can be applied.

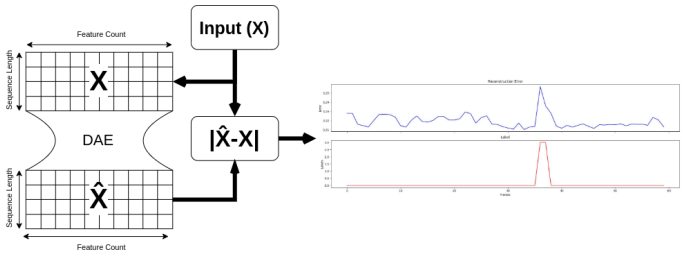

Fig. 4. Schematic overview of Detector

For threshold-based detection the difference between some data **X** which may or may not be noisy is given to the DAE as an input, the absolute difference between the output of the DAE $\hat{\mathbf{X}}$ and the input **X**, dubbed the **Reconstruction Error** $|\hat{\mathbf{X}}\text{-}\mathbf{X}|$, is then used as an indicator for the presence of anomalies. An overview of how the **Reconstruction Error** is obtained for a given input is shown in Figure 4.

The obtained threshold can be leveraged in several ways, for instance, the use of a **Fixed Threshold** or a **Adaptive Threshold**, both of these methods have their own advantages and disadvantages.

The use of a **Adaptive Threshold** comes with many considerations, primarily, the method in which the threshold adapts itself to the data. Whereas a **Fixed Threshold** is a single value determined during model training and is not adjustable.

In this study a **Fixed Threshold** was used, as it reduces the number of variables which influence the results of the study.

4.2.3 Generalization Ability. The problem of **Generalization** is particularly relevant for this application, due to the variability between musical compositions, and within musical compositions themselves. Therefore, a part of this study will be dedicated to determining the ability of the DAE to be generalized.

The generalization ability of the models will be determined by varying the number of songs used during the training step of the models. The **F1-Score** and **Mean Absolute Error (MAE)** loss will be logged for the different number of training songs for both the **LSTM** and **BLSTM** based **DAE**, such that claims can be made about the **Generalization** abilities of both these types of models.

5 EXPERIMENTAL SETUP

For the determination of the efficacy of a **DAE** in the application of anomaly detection on a background of musical data, a methodical approach must be undertaken to underpin the scientific foundations of the study. This methodical approach consists of several components, ensuring reproducible results, the production of representative data through augmentation and the determination of relevant performance metrics for model evaluation.

5.1 Augmentation

As delineated in Section 3.2, a dataset consisting of musical compositions without anomalies has been collected, henceforth referred to as **normal data**. The normal data serves as a target for the **DAE**, as an input for the **DAE** the process of **augmentation** is applied.

The nature of a **DAE** is such that it receives **noisy data as an input** and aims to return **de-noised data as an output**. The process

of augmentation enables the **generation of uniquely noised data** by adding unique generated noise to the normal data that is used as a target.

In order for the model to be able to learn as much as possible about the target, the given inputs for the model are augmented in different ways each epoch, remaining consistent between experiments as described in Section 5.2. Each augmentation is unique, this is achieved by varying the *locations of the augmentations*, the *types of the augmentations* and the *signal-to-noise-ratios* between the augmentations and the normal data.

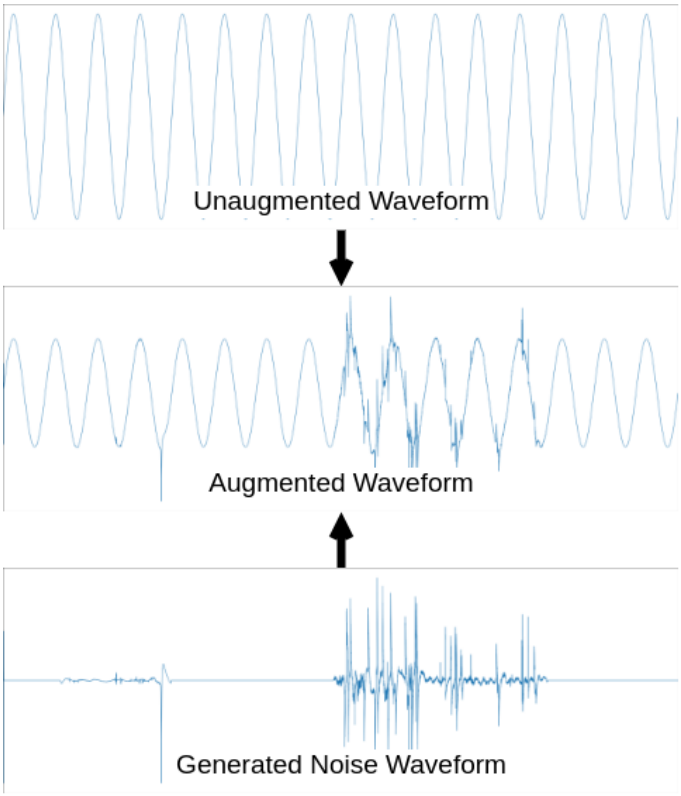

Fig. 5. Schematic overview of Augmentation Process

5.1.1 Augmentation Process. The process of augmentation works as follows, first the length of a target waveform is determined, then a waveform of equal length is constructed from samples of noise, or anomalies. For this process five parameters can be configured, each of these parameters influences the produced augmentation in its own way. The overall process of the augmentation is given in Figure 5.

Two of these parameters are categorical and can either be true or false, these govern whether different types of noise will be interspersed within a single song and whether the samples will be spaced equidistantly. Both of these affect the predictability of the data and are generally set to both be true such that the data is of a more unpredictable nature.

The three other parameters are continuous they are the augmentation fraction, the maximum length of an anomaly and the signal-to-noise ratios to use during augmentation. The augmentation fraction determines what fraction of the original waveform should be augmented, for instance, if set to 1 the entirety of the original waveform is augmented, whereas when set to 0.05 only 5% of the waveform will be augmented.

The maximum length of an anomaly sets the number of seconds any noise sample may at most consists of, and was implemented due to the nature of the different types of anomalies, whereas an example of a "pop" may only last a fraction of a second an example of "surface noise" may last 30 seconds or longer. Therefore to ensure that some types of anomalies are not over-represented in the data a limit was imposed on the length of individual noise samples.

The signal-to-noise ratios speak for themselves, and are an indication of the values of the relative strength of the signal as compared to the noise., For instance, a signal-to-noise ratio of 2.0 indicates that the relative magnitude of the signal is twice as strong as that of the noise during addition.

featuring randomized noise samples, signal-to-noise ratios and spacing between samples is constructed. This noise waveform is then added to the target waveform in order to produce a noisy waveform in which the locations of anomalies are known this process is highlighted in Figure 5.1.

Table 3. Parameters used for augmentation

Dataset	Signal-to-noise ratios	Augmentation Fraction	Maximum augmentation length (s)	Regularly Spaced	Interspersed
Training	1.0, 1.5, 2.0, 3.0	1.0	2	No	Yes
Testing	1.0, 1.5, 2.0, 3.0	0.05	0.5	No	Yes
Validation	1.0, 1.5, 2.0, 3.0	0.05	0.5	No	Yes

As described in Section 5.1, there are several parameters for the augmentation which may be adjusted, which each influence the outcome of the experiment. These parameters were kept as consistent as possible, with the exception of augmentation lengths and fractions and are given in Table 3.

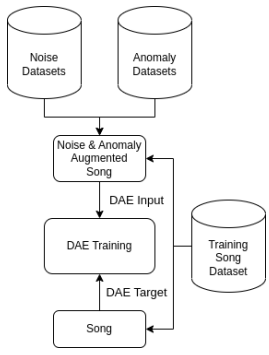

Fig. 6. DAE training methodology

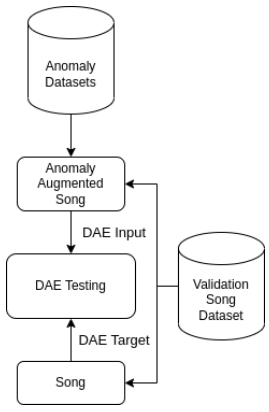

Fig. 7. DAE testing methodology

5.1.2 Augmentation of Training Data. The goal of this study is to create a DAE-based anomaly detector, which is able to detect anomalies other than only those present in the dataset of anomalies as delineated in Section 3.3. To improve the generalization abilities of the detector produced in this study [2], a number of types of noise different from those regularly found on vinyl records are introduced during augmentation.

The "noisy" data used for training is augmented by introducing both the types of anomalies commonly found on records, as well as Gaussian, Brownian, white and pink noise. Through the addition of these types of noises, the DAE will be trained to not only remove the examples of noises gathered for this study but also to remove noises of a more random nature that may not be part of the gathered datasets this process is displayed schematically in Figure 6.

5.1.3 Augmentation of Testing and Validation Data. The testing and validation of the detection model have different requirements for the used data than that used for training, namely, whereas the training data is meant to encompass more examples of noises that are likely to occur on a vinyl record, the testing and validation data must only contain those examples of noises which would occur on vinyl records.

For this reason, the augmentation of testing and validation data is conducted in a similar manner as that of the training data, however, for testing only the anomalies present on vinyl records as described in Section 3.3 are used. Therefore, during the testing and validation stages the performance of the model can be measured against representative data this process is displayed schematically in Figure 7.

5.2 Repeatability

To make claims about the efficacy of different architectures and features, it is of great importance that the data used to train the model is consistent between experiments. The data used for training, testing and validating the model in this study is augmented in a repeatable way with several types of noises and anomalies.

This consistency is often called **Repeatability**, and was achieved through the procedural generation of seeds. To validate that the procedural generation was achieved effectively, three tests were conducted, described as follows.

The first test ensured that the augmentations applied were different between each iteration of the dataset, i.e. between the epochs. The second test ensured that the augmentations were identical between different instances of the experiment.

As augmentation is not the only variable dependent on a seed, as the models also use random values, the third test trained the same model with the same parameters and data twice. It verified that its results were identical between runs, validating the repeatability of the model's training.

5.3 Model Evaluation

The model was evaluated in several ways, as the model consists of multiple stages, the autoencoder training stage followed by the testing stage for the detector, the autoencoder was evaluated by itself once its training was completed the performance of the detector could be measured.

5.3.1 Hyperparameters. The model used in this study may be tuned using several **Hyperparameters**, these are given in Table 4 As an exhaustive search was not feasible within the given time, use was made of **Tree-Structured Parzen Estimator (TPE)** hyperparameter optimization algorithm [3]. A commonly used approach publically available through the python optuna library. For each of the feature sets a total of 200 trials were conducted during the search for optimal **Hyperparameters**.

Table 4. Hyper-Parameters of DAE

Hyper Parameter	Minimum Value	Maximum Value	Condition (optional)
Learning Rate	1e-06	1e-01	
Weight Decay	1e-06	1e-01	Lower than learning rate
Dropout Probability	0.0	0.9	
Compression Ratio	0.5	Feature Count	
Layer Count	1	6	
Sequence Length	3	300	

*5.3.2 Evaluation of the **DAE**.* The performance for the **DAE** will be measured in two ways. Firstly through the **Mean Absolute Error (MAE)** loss, which evaluates the ability of the **DAE** to denoise given input data. Secondly the inference time of the **DAE** will be measured, as this dominates the inference time of the detector. Due to the variability of the lengths of sequences passed to the model, the inference time is used computed relative to the length of these sequences. Thus the inference time is defined as the number of milliseconds for determining the reconstruction error divided by the number of seconds of audio data passed to the model.

5.3.3 Evaluation of the Detector. The generation of augmented labelled data enabled us to measure the model's detection accuracy during testing. Applying our fully trained model to unseen validation data allowed us to assess its accuracy based on comparison with actual labels. Metrics such as **F1-Score**, **Accuracy**, **Precision** and **Recall** are used to evaluate the performance of the anomaly detection model.

6 RESULTS

Through the exploration of features and anomaly detection techniques, it was found that it is possible to detect some types of anomalies within musical composition through the use of advanced reconstruction-based anomaly detection techniques. The use of each available feature is detrimental to the detection of anomalies, therefore a subset of the most useful features has been constructed. Furthermore, the use of reconstruction-based techniques was found to be poorly suited to the problem.

6.1 Feature Selection

Given the findings from our literature review, we undertook an analysis of the relevance and redundancy of the identified features using the mRMR algorithm [19]. The features investigated are listed in Table 5, for each of these features the relevance of the derived "texture" features was also determined.

The results from the mRMR algorithm's computation of mutual information, which indicates the applicability of each feature independently, are presented in Tables 6, 7, 8, 9, 10 and 11.

The ranking of features as produced by the mRMR algorithm is given in Tables 12 and 13. Notably, out of the 10 best features, 9 were the derived texture features, highlighting once more the benefits of these derived features. Furthermore, the spectral centroid, spectral flux, and spectral flatness were particularly relevant for this problem, in line with the justifications given in Table 5.

6.2 Model Performance

The model's performance was evaluated for several sets of features, in order to validate that the **mRMR** feature selection method used is suitable for use with **RNN** based models. The feature sets which were compared were as follows the top 5, top 10, top 30 and top 60 features as selected by the mRMR algorithm, and for validation the commonly used **Mel-frequency Cepstral Coefficients (MFCC)** feature including its means, variances and low energy.

The use of **Mel-frequency Cepstral Coefficients (MFCC)** as a baseline comparison due to its use in comparable studies which use LSTM AEs for novelty detection [16, 17, 21], allowing us to make some claims as to the benefits of the chosen feature selection method.

The performance of the **DAE** as discussed in Section 5.3 are given for both the **LSTM** and **BLSTM** variants in Figure 10. These results highlight the improved performance of the **BLSTM** architecture when used for reconstruction, however, as the size of the model is larger it can be seen that inference times are generally higher.

Furthermore, the inference times of the **DAE** are significantly higher for the **Mel-frequency Cepstral Coefficients (MFCC)** features, this is likely due to the model having an increased complexity due to the complex nature of the features. Whereas the selected features are more readily reconstructed.

The performance of the detector is given for both the **LSTM** and **BLSTM** architectures and shows the **Mel-frequency Cepstral Coefficients (MFCC)** features outperforming the selected features in most metrics. Particularly in the **F1-Score**, this result is unexpected and the opposite of the **Mean Absolute Error (MAE)** loss for the same features.

7 DISCUSSION

The results of this study are not in line with initial expectations, the relation between the efficacy of the DAE and the F1-Score of the detector is expected to be proportional. However, as the reconstruction error was reduced through the use of features selected through mRMR the F1-Score was reduced in turn.

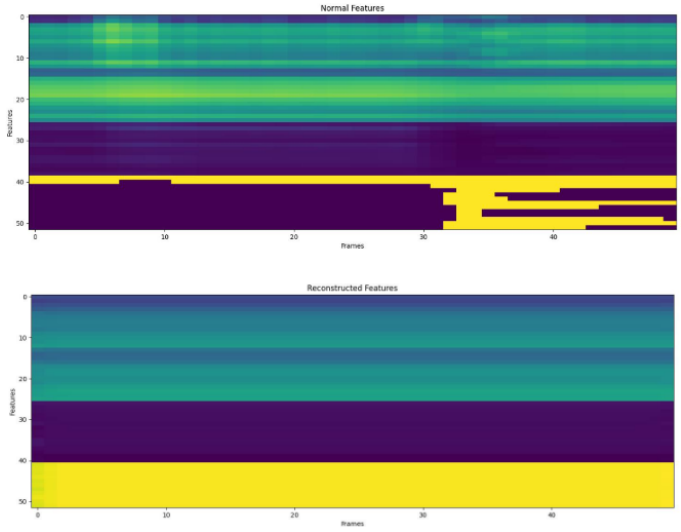

Fig. 8. Reconstruction example of best performing model and features

Figure 8 shows an example of reconstruction through the best-performing model and features. The figure highlights the inability of the model to reconstruct the complex patterns of the features of musical compositions. Comparable results can be observed in the reconstruction of the other feature sets.

This shows that whilst the selected feature may be highly informative with regard to the detection of anomalies, the model is not able to reconstruct these features. This lack of ability of the model may be explained due to the simplistic architecture of the used model which makes it difficult to draw conclusions on the applicability of the selected features.

8 CONCLUSION

This study has examined the applicability of features selected through the **mRMR** algorithm and the effectiveness of the models utilizing these features in the domain of musical anomaly detection. We found that, while the selected features demonstrated a heightened performance for the **DAE** used, no substantial enhancement was observed in the **F1-Score**, a measure of detection efficacy. It seems that while these features are highly reconstructable, their aggregation may not be conducive to anomaly detection via a thresholding mechanism. Nonetheless, it is worth noting that the inference times using these features were significantly lower compared to the direct usage of **Mel-frequency Cepstral Coefficients (MFCC)**.

This suggests that the **mRMR**-selected features are indeed more amenable for reconstruction, as evidenced by lower inference times and **Mean Absolute Error (MAE)** loss compared to the direct use of **Mel-frequency Cepstral Coefficients (MFCC)**. However, considering the observed inefficiency in detection, alternative feature selection methods that take into account the temporal context of the features could potentially yield better results, as the current approach does not factor in preceding and succeeding samples.

In the comparison between **LSTM** and **BLSTM** models, no substantial difference was found in terms of **F1-Score** or detection performance. However, the **BLSTM** model consistently outperformed the **LSTM** in terms of reconstruction, registering a lower **Reconstruction Error** across all selected feature sets.

Overall, these findings underscore the importance of judicious feature selection and model choice in enhancing both the speed and performance of musical anomaly detection. Further research might focus on refining the feature selection process, possibly by incorporating temporal context and testing the applicability of more sophisticated model architectures for improved reconstruction ability.

9 FUTURE RESEARCH

This research has highlighted the viability of detecting anomalies within musical composition through the use of advanced reconstruction-based anomaly detection techniques, specifically **DAEs**, in combination with features selected using the **mRMR** algorithm. Despite these advances, however, certain discrepancies between the efficacy of the **DAE** and the detection performance, as measured by the **F1-Score**, suggest that further research is needed to refine these techniques.

Future studies might consider the construction of more complex and configurable **LSTM** and **BLSTM** based models that allow for the adjustment of the number of hidden layers and their sizes. Such advancements could potentially offer a more nuanced reconstruction of the selected features, thus addressing the limitations observed in the current model's performance.

Moreover, the adoption of the Next Prediction BLSTM DAE (NP-BLSTM-DAE) method [17], where a current sample is used to predict the subsequent sample, may offer improvements. Given the temporal nature of music, this approach could provide enhanced reconstruction capabilities by better capturing the inherent temporal dependencies in musical compositions.

In addition, the introduction of an adaptive threshold could be beneficial. By adjusting the threshold according to the data, the model might exhibit more robust performance when dealing with unseen data.

Advancements could also be made through the introduction of an adversarial stage to improve reconstruction performance. The benefits of this adversarial approach were highlighted in previous research [21], and the incorporation of such a stage might aid in addressing the observed inefficiencies in detection.

To better ascertain the generalizability of the model, future research could utilize recordings made using different recording setups. By analyzing how well the model generalizes across various recording conditions, the robustness of the anomaly detection method could be evaluated.

Moreover, expanding the diversity of the training and testing datasets to include more genres could provide insight into the model's ability to generalize across not only songs from the same artist but also songs from different genres.

Finally, future research might explore the implementation of LSTM/BLSTM variants of the mRMR algorithm, such as the random forest-based variants produced in [27]. Such an approach could potentially offer a more temporally aware feature selection process, which may result in improved anomaly detection performance.

REFERENCES

[1] 2018. How To Grade Items. https://support.discogs.com/hc/en-us/articles/360001566193-How-To-Grade-Items

[2] Guozhong An and Guozhong An. 1996. The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation* (1996). https://doi.org/10.1162/neco.1996.8.3.643

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for Hyper-Parameter Optimization. *NIPS* (2011). https://doi.org/null

[4] Sebastian Böck and G. Widmer. 2013. MAXIMUM FILTER VIBRATO SUPPRESSION FOR ONSET DETECTION Sebastian Bock and Gerhard Widmer Department of Computational Perception. *null* (2013). https://doi.org/null

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. 2009. Anomaly detection: A survey. *Comput. Surveys* (2009). https://doi.org/10.1145/1541880.1541882

[6] Kukjin Choi, Jihun Yi, Changhwa Park, and Sungroh Yoon. 2021. Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines. *IEEE Access* (2021). https://doi.org/10.1109/access.2021.3107975

[7] Chris Ding and Hanchuan Peng. 2003. Minimum redundancy feature selection from microarray gene expression data. *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003* (2003). https://doi.org/10.1109/csb.2003.1227396

[8] Shlomo Dubnov. 2004. Generalization of spectral flatness measure for non-Gaussian linear processes. *IEEE Signal Processing Letters* (2004). https://doi.org/10.1109/lsp.2004.831663

[9] Theodoros Giannakopoulos and Aggelos Pikrakis. 2014. Chapter 4 audio features. *null* (2014). https://doi.org/10.1016/b978-0-08-099388-1.00004-2

[10] Christopher Harte, Mark Sandler, and Martin Gasser. 2006. Detecting harmonic change in musical audio. *AMCMM '06* (2006). https://doi.org/10.1145/1178723.1178727

[11] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jianhua Tao, and Lianhong Cai. 2002. Music type classification by spectral contrast feature. *Proceedings. IEEE International Conference on Multimedia and Expo* (2002). https://doi.org/10.1109/icme.2002.1035731

[12] Anssi Klapuri and Manuel Davy. 2006. Signal Processing Methods for Music Transcription. *null* (2006). https://doi.org/10.1007/0-387-32845-9

[13] Alexander Lerch. 2012. An introduction to audio content analysis. *null* (2012). https://doi.org/10.1002/9781118393550

[14] Beth Logan. 2000. Mel frequency cepstral coefficients for music modeling. *International Society for Music Information Retrieval Conference* (2000). https://doi.org/null

[15] Yen-Cheng Lu, Chih-Wei Wu, Chang-Tien Lu, and Alexander Lerch. 2016. An Unsupervised Approach to Anomaly Detection in Music Datasets. *null* (2016). https://doi.org/10.1145/2911451.2914700

[16] Erik Marchi, Fabio Vesperini, Florian Eyben, Stefano Squartini, and Björn Schuller. 2015. A novel approach for automatic acoustic novelty detection using a denoising autoencoder with bidirectional LSTM neural networks. *null* (2015). https://doi.org/10.1109/icassp.2015.7178320

[17] Erik Marchi, Fabio Vesperini, Stefano Squartini, and Björn Schuller. 2017. Deep Recurrent Neural Network-Based Autoencoders for Acoustic Novelty Detection. *Computational Intelligence and Neuroscience* (2017). https://doi.org/10.1155/2017/4694860

[18] Timothy J. O'Shea, T. Charles Clancy, and Robert W. McGwier. 2016. Recurrent Neural Radio Anomaly Detection. *arXiv: Learning* (2016). https://doi.org/null

[19] Hanchuan Peng, Fuhui Long, and Chris Ding. 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2005). https://doi.org/10.1109/tpami.2005.159

[20] Marco A. F. Pimentel, David A. Clifton, Lei Clifton, and Lionel Tarassenko. 2014. Review: A review of novelty detection. *Signal Processing* (2014). https://doi.org/10.1016/j.sigpro.2013.12.026

[21] Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. 2017. Acoustic novelty detection with adversarial autoencoders. *null* (2017). https://doi.org/10.1109/ijcnn.2017.7966273

[22] Ellen Rushe and Brian Mac Namee. 2019. Anomaly Detection in Raw Audio Using Deep Autoregressive Networks. *null* (2019). https://doi.org/10.1109/icassp.2019.8683414

[23] Eric D. Scheirer and Malcolm Slaney. 1997. Construction and evaluation of a robust multifeature speech/music discriminator. *IEEE International Conference on Acoustics, Speech, and Signal Processing* (1997). https://doi.org/10.1109/icassp.1997.596192

[24] Bernhard Schölkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor, and John Platt. 1999. Support Vector Method for Novelty Detection. *NIPS* (1999). https://doi.org/null

[25] George Tzanetakis and Perry R. Cook. 2002. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* (2002). https://doi.org/10.1109/tsa.2002.800560

[26] M. Wohlfeil. 2022. Vinyl Strikes (Not Once But Twice): The Non-Digital Future of Listening to Music?: An Abstract. *Developments in Marketing Science: Proceedings of the Academy of Marketing Science* (2022), 571–572. https://doi.org/10.1007/978-3-030-95346-1_186

[27] Zhenyu Zhao, Zhenyu Zhao, Radhika Anand, and Mallory Wang. 2019. Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform. *International Conference on Data Science and Advanced Analytics* (2019). https://doi.org/10.1109/dsaa.2019.00059

GLOSSARY

Accuracy
A metric used to assess the performance of classification models. Accuracy is defined as the proportion of true results (both true positives and true negatives) in the total number of cases examined. It ranges from 0 to 1, where 1 indicates perfect accuracy. 6

Adaptive Threshold
An adaptive threshold, in the context of reconstruction-based anomaly detection, is a threshold value that dynamically adjusts based on the underlying data or some other factors. Unlike a fixed threshold, an adaptive threshold can change to better accommodate variations in the data or specific conditions. This can be particularly useful when the data is highly variable, or the distinction between normal and anomalous data changes over time. The adaptive threshold can be based on the statistical properties of the data, predictive models, feedback loops, or other methods. 4

BLSTM
A Bidirectional Long-Short Term Memory network extends LSTM by presenting the network with both past and future data. 1, 2, 4, 7, 8

Classification-Based Novelty Detection
This technique involves training a model on a set of known categories or classes, and anything that doesn't fit these categories is considered as a novel or an anomaly. It's applicable when known classes exist and there is enough data to train a classifier. 1, 2

Clustering-Based Novelty Detection
This approach uses clustering algorithms to group similar data together and identify the normal data clusters. Data points that don't belong to any of these clusters or belong to small and sparse clusters are considered anomalies. 1, 2

DAE
A Denoising Autoencoder is a type of neural network that is trained to use its hidden layer to encode robust representations by reconstructing the input from a noisy version of itself. 1–4, 6–8, 17

Distance-Based Novelty Detection
This technique defines an anomaly based on the distance of a data point from the rest of the data. Data points that are far away from others are considered anomalies. The distance measure can be Euclidean, Manhattan, or any other distance metric. 1, 2

F1-Score
A measure used to evaluate the performance of binary classification models, although it can be extended for multi-class problems. The F1-Score is the harmonic mean of precision and recall, and ranges from 0 to 1, with 1 being the best possible score. A higher F1-Score indicates a more accurate and robust model. 4, 6–8

FCQ
Feature Correlation Quotient (FCQ) is a term often used in feature selection to denote the degree of correlation between features. The aim in many algorithms is to minimize the FCQ to ensure that the selected features provide non-redundant information. 3

Fixed Threshold
In the context of reconstruction-based anomaly detection, a fixed threshold is a predetermined value used to decide whether a given data point is an anomaly based on its reconstruction error. The reconstruction error is compared with the fixed threshold: if the error is less than or equal to the threshold, the data point is considered normal; if the error exceeds the threshold, the data point is flagged as an anomaly. The key characteristic of a fixed threshold is that it remains constant, regardless of changes in the data or other conditions. 4

Generalization
In the context of machine learning, generalization refers to the model's ability to adapt properly to new, previously unseen data, drawn from the same distribution as the one used to create the model. A model that generalizes well will be able to make accurate predictions on unseen data after being trained on a subset of that data. The goal of a good machine learning model is to generalize well from the training data to any data from the problem domain. This allows us to make predictions in the future on data the model has never seen. 4

Hyperparameters
In machine learning, hyperparameters refer to the parameters of the model that are set prior to the start of the learning process. Unlike other parameters, hyperparameters cannot be learned directly from the data in the standard training process and must be predefined. These can include learning rates, regularization parameters, the number of layers in a deep neural network, the number of clusters in a k-means clustering algorithm, etc. The choice of hyperparameters can significantly influence the performance of the model. 6

LSTM
A Long Short-Term Memory network is an artificial recurrent neural network architecture used in the field of deep learning, capable of learning long-term dependencies. 2, 4, 7, 8

Mean Absolute Error (MAE)
A metric used to quantify the difference between predicted and actual values in regression problems. It is calculated as the average of the absolute differences between the predicted and actual values. It provides a measure of how far off the predictions are on average, with a value of 0 indicating no error. 4, 6, 7

Mel-frequency Cepstral Coefficients (MFCC)
A type of feature used in signal processing and machine learning for audio analysis. MFCCs are derived from a type of cepstral representation of the audio clip (a nonlinear "spectrum-of-a-spectrum"). The mel-frequency part indicates that the frequency bands are positioned logarithmically (on the mel scale), which approximates the human auditory system's response more closely than linearly-spaced frequency bands. They are commonly used in speech and music processing, as well as in many other machine-learning tasks related to audio. 2, 7

MIQ
Mutual Information Quotient (MIQ) is a term often used in the context of feature selection or filtering methods, where the aim is to maximise the mutual information between the selected features and the target class while minimizing the mutual information between the features themselves. 3

MIR
Music Information Retrieval is the interdisciplinary science of retrieving information from music. 2, 3, 12

mRMR
Minimum Redundancy Maximum Relevance (mRMR) is a feature selection method that aims to select features that are highly correlated with the class but uncorrelated with each other. It balances the importance of mutual information (relevance) and redundancy of the features. 3, 7, 8

Mutual Information (MI)
Mutual Information (MI) is a metric in Information Theory that quantifies the statistical dependence between two variables or sets of variables. It represents the amount of information that can be obtained about one variable by observing another variable. 3

Precision
A measure used to evaluate the performance of classification models. Precision is defined as the number of true positives divided by the sum of true positives and false positives. It provides an understanding of the reliability of positive predictions. A precision of 1 indicates that all positive predictions were correct. 6

Recall
A measure used to evaluate the performance of classification models. Recall (or sensitivity or true positive rate) is defined as the number of true positives divided by the sum of true positives and false negatives. It provides an understanding of the model's ability to identify all relevant instances. A recall of 1 indicates that all relevant instances were identified. 6

Reconstruction Error
Reconstruction error, in the context of reconstruction-based anomaly detection techniques, is a measure of the difference between the original input data and the same data after being processed (e.g., compressed and then decompressed) by a model, such as an autoencoder. A high reconstruction error indicates a large discrepancy between the original and reconstructed data and is typically used as an indicator of an anomaly. In other words, if the model is trained primarily on normal data, it should have a low reconstruction error on similar data. Conversely, it should have a high reconstruction error on anomalous data, which significantly deviates from the norm. 4, 8

Reconstruction-Based Novelty Detection
In this approach, a model is trained to reconstruct normal data. The model typically performs poorly when trying to reconstruct anomalies. Therefore, data points with high reconstruction errors are considered anomalies. Techniques such as autoencoders can be used for this purpose. 1, 2, 4

Repeatability
In the context of machine learning, repeatability refers to the consistent reproduction of results in an experiment or analysis. As machine learning models often incorporate elements of randomness during training (e.g., initialization of weights, data shuffling, splitting of training and testing data), repeatability becomes critical when comparing models, fine-tuning hyperparameters, and confirming the robustness of results. Achieving repeatability often involves setting a specific seed for the random number generator to ensure consistent randomness between runs. Additionally, repeatability also includes the ability for other researchers to replicate results using described methods and provided code/data, promoting good scientific practice. 6

RNN
A Recurrent Neural Network is a type of artificial neural network where connections between nodes form a directed graph along a temporal sequence, allowing it to exhibit temporal dynamic behaviour. 1, 2, 4, 7

Statistical Novelty Detection
This technique is based on statistical models and it assumes that normal data follows a certain statistical distribution. Any data point that deviates significantly from this distribution is considered novel or an anomaly. 1, 2

Tree-Structured Parzen Estimator (TPE)
The Tree-Structured Parzen Estimator (TPE) is a sequential model-based optimization (SMBO) approach used for hyperparameter tuning in machine learning. It constructs two probabilistic models based on good and bad hyperparameter settings observed so far in the optimization process. These models, which estimate the conditional probability of a score given the hyperparameters, are used to explore and exploit the hyperparameter space. The approach is called 'tree-structured' because a tree structure is used to partition the space of hyperparameters according to the regions they affect the most. TPE was introduced by James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl in the paper titled "Algorithms for Hyper-Parameter Optimization" presented at the NIPS 2011 conference [3]. 6

A FEATURE DESCRIPTION

Table 5. Feature Descriptions

Feature	Dimensions	Description
Chroma STFT	12 bins	The spectrum of an audio signal mapped into 12 bins representing 12 distinct semitones of the musical octave, particularly relevant when dealing with musical compositions.
Tonnetz	6 bins	Tonal Centroid features, A projection of chroma features onto a 6-dimensional basis representing the perfect fifth, minor third, and major third each as two-dimensional coordinates [10]. Of particular relevance for distinguishing musical compositions from anomalies as anomalies will not map neatly into musical notes.
MFCC	13 bins	Mel Frequency Cepstral Coefficients, A concise description of the whole spectral envelope scaled to mel scale, which is a perceptual scale of pitches which sound to the human ear to be separated by an equal distance. Its relevance for use with music has been established extensively in the field of MIR [14].
Spectral Flux	16 bins	Spectral flux measures the spectral change between two successive frames and is computed as the squared difference between normalized magnitudes of the spectra of the two successive windows [9].
Pitch	12 bins	The pitch of an audio sample, the degree of highness or lowness of a tone. The dimension of the pitches vastly exceeds other features, therefore principal component analysis (PCA) was used to scale the pitches down into 12 bins.
Spectral Centroid	1 bin	The mean of all frequencies present in a signal, the centroid of the spectrum [12].
Spectral Contrast	1 bin	The contrast between the energy present in the highest frequency bands and the energy present in the lower frequency bands [11]
Spectral Rolloff	1 bin	In digital signal processing, Spectral Rolloff is a critical measure that captures the frequency below which a certain portion of the total spectral energy is located. Its role in anomaly detection is crucial as it identifies unusual energy distributions across the frequency spectrum often indicative of audio anomalies.
Spectral Flatness	1 bin	The spectral flatness, also called tonality coefficient, quantifies how noise-like a signal is, as opposed to being tone-like[8]. This measure is particularly useful as the music is tone-like, whereas the anomalies are not.
Zero Crossing Rate	1 bin	Defined as the number of times the zero-point is crossed by a waveform, of particular use as the deflection of a needle on a record causes additional zero-crossings to occur at that point in time.
Pulse Curve	1 bin	The onset strength of the spectral flux, indicates the absolute amount with which the frequency changes at a given point [4]. This is of particular relevance as it is known that an anomaly presents itself as a change throughout the entire spectrum, which is particularly noticeable in this feature.

B FEATURE IMPORTANCE

Table 6. Relevance of Mel-frequency cepstral coefficients

Information Gain	Normal	Mean	Variance	Low Energy
MFCC Bin 0	0.024	0.138	0.015	0.037
MFCC Bin 1	0.010	0.065	0.022	0.017
MFCC Bin 2	0.008	0.008	0.022	0.021
MFCC Bin 3	0.005	0.007	0.022	0.022
MFCC Bin 4	0.005	0.005	0.022	0.017
MFCC Bin 5	0.008	0.018	0.020	0.018
MFCC Bin 6	0.005	0.010	0.021	0.019
MFCC Bin 7	0.007	0.014	0.021	0.019
MFCC Bin 8	0.0 13	0.024	0.018	0.023
MFCC Bin 9	0.015	0.014	0.021	0.021
MFCC Bin 10	0.011	0.017	0.023	0.025
MFCC Bin 11	0.015	0.010	0.025	0.019
MFCC Bin 12	0.023	0.012	0.021	0.023

Table 7. Relevance of Spectral Flux

Information Gain	Normal	Mean	Variance	Low Energy
Spectral Flux Bin 0	0.043	0.056	0.013	0.010
Spectral Flux Bin 1	0.008	0.016	0.008	0.002
Spectral Flux Bin 2	0.006	0.012	0.004	0.002
Spectral Flux Bin 3	0.006	0.010	0.000	0.002
Spectral Flux Bin 4	0.007	0.011	0.000	0.003
Spectral Flux Bin 5	0.003	0.008	0.000	0.002
Spectral Flux Bin 6	0.002	0.003	0.000	0.001
Spectral Flux Bin 7	0.005	0.011	0.000	0.002
Spectral Flux Bin 8	0.004	0.009	0.000	0.002
Spectral Flux Bin 9	0.003	0.005	0.000	0.001
Spectral Flux Bin 10	0.007	0.015	0.000	0.002
Spectral Flux Bin 11	0.012	0.023	0.001	0.005
Spectral Flux Bin 12	0.004	0.011	0.000	0.003
Spectral Flux Bin 13	0.003	0.006	0.000	0.003
Spectral Flux Bin 14	0.006	0.016	0.000	0.005
Spectral Flux Bin 15	0.006	0.013	0.000	0.005

Table 8. Relevance of Chroma

Information Gain	Normal	Mean	Variance	Low Energy
Chroma Bin 0	0.011	0.011	0.009	0.011
Chroma Bin 1	0.029	0.011	0.013	0.012
Chroma Bin 2	0.022	0.009	0.017	0.011
Chroma Bin 3	0.048	0.043	0.015	0.006
Chroma Bin 4	0.022	0.016	0.014	0.006
Chroma Bin 5	0.041	0.024	0.015	0.003
Chroma Bin 6	0.023	0.019	0.014	0.006
Chroma Bin 7	0.041	0.030	0.010	0.013
Chroma Bin 8	0.024	0.009	0.012	0.003
Chroma Bin 9	0.032	0.020	0.012	0.003
Chroma Bin 10	0.023	0.025	0.008	0.005
Chroma Bin 11	0.029	0.022	0.009	0.004

Table 9. Relevance of Tonnetz

Information Gain	Normal	Mean	Variance	Low Energy
Tonnetz Bin 0	0.040	0.093	0.000	0.033
Tonnetz Bin 1	0.032	0.085	0.000	0.032
Tonnetz Bin 2	0.023	0.065	0.000	0.029
Tonnetz Bin 3	0.004	0.020	0.000	0.015
Tonnetz Bin 4	0.028	0.066	0.000	0.032
Tonnetz Bin 5	0.021	0.048	0.000	0.015

Table 10. Relevance of Pitch

Information Gain	Normal	Mean	Variance	Low Energy
Pitch Bin 0	0.000	0.000	0.000	0.000
Pitch Bin 1	0.010	0.008	0.007	0.000
Pitch Bin 2	0.007	0.005	0.008	0.001
Pitch Bin 3	0.005	0.005	0.006	0.014
Pitch Bin 4	0.013	0.015	0.003	0.018
Pitch Bin 5	0.008	0.006	0.006	0.001
Pitch Bin 6	0.002	0.001	0.002	0.021
Pitch Bin 7	0.020	0.030	0.003	0.046
Pitch Bin 8	0.002	0.002	0.002	0.057
Pitch Bin 9	0.069	0.079	0.002	0.183
Pitch Bin 10	0.059	0.060	0.000	0.158
Pitch Bin 11	0.019	0.020	0.000	0.156

Table 11. Relevance of 1D Features

Information Gain	Normal	Mean	Variance	Low Energy
Spectral Centroid	0.140	0.218	0.011	0.041
Spectral Contrast	0.051	0.203	0.005	0.048
Spectral Rolloff	0.051	0.203	0.005	0.048
Spectral Flatness	0.056	0.118	0.012	0.065
Zero Crossing Rate	0.012	0.073	0.003	0.010
Pulse Curve	0.001	0.001	0.000	0.000

Table 12. Top 60 mRMR features (Part 1)

Feature Name	Bin	Type	Order	Score (MIQ)
Spectral Centroid	0	Mean	1	0.218
Spectral Flux	0	Normal	2	3.970
Spectral Flatness	0	Low Energy	3	1.859
Pitch	10	Low Energy	4	1.391
Pitch	2	Variance	5	1.086
Spectral Flatness	0	Mean	6	1.859
Pitch	11	Low Energy	7	1.333
Pitch	9	Mean	8	1.347
MFCC	0	Low Energy	9	1.180
Pitch	9	Low Energy	10	1.080
Pitch	7	Variance	11	1.013
Chroma	0	Low Energy	12	1.052
MFCC	12	Normal	13	1.097
MFCC	11	Variance	14	1.067
Chroma	2	Variance	15	1.115
Spectral Flatness	0	Normal	16	1.162
Tonnetz	0	Mean	17	1.198
Pitch	3	Variance	18	1.072
Pitch	10	Mean	19	1.131
Pitch	8	Low Energy	20	1.175
Spectral Flux	15	Mean	21	1.184
Spectral Flux	0	Mean	22	1.130
Spectral Contrast	0	Mean	23	1.135
MFCC	0	Mean	24	1.021
Spectral Flatness	0	Variance	25	1.002
Pitch	11	Normal	26	0.934
Spectral Flux	11	Normal	27	0.952
Pitch	7	Mean	28	0.978
Tonnetz	0	Mean	29	0.983
Spectral Centroid	0	Normal	30	0.913

Table 13. Top 60 mRMR features (Part 2)

Feature Name	Bin	Type	Order	Score (MIQ)
Spectral Centroid	0	Variance	31	0.909
Pitch	9	Normal	32	0.900
Spectral Flux	14	Mean	33	0.869
Chroma	5	Variance	34	0.872
Chroma	7	Low Energy	35	0.884
Pitch	7	Low Energy	36	0.882
Pitch	10	Normal	37	0.889
Pitch	1	Variance	38	0.847
Spectral Flux	11	Mean	39	0.863
Spectral Flux	0	Variance	40	0.829
Pitch	5	Variance	41	0.837
Pitch	4	Low Energy	42	0.851
Tonnetz	4	Mean	43	0.853
Spectral Rolloff	0	Mean	44	0.847
MFCC	10	Low Energy	45	0.795
Pitch	11	Mean	46	0.803
Spectral Flux	4	Mean	47	0.794
Chroma	8	Variance	48	0.772
Spectral Flux	4	Normal	49	0.765
Tonnetz	2	Mean	50	0.766
MFCC	1	Variance	51	0.749
Chroma	2	Low Energy	52	0.740
Pitch	6	Low Energy	53	0.728
Chroma	3	Variance	54	0.730
Spectral FLux	2	Variance	55	0.708
Chroma	3	Normal	56	0.712
Spectral Flux	7	Mean	57	0.707
Chroma	6	Variance	58	0.693
MFCC	12	Low Energy	59	0.698
Spectral Contrast	0	Low Energy	60	0.700

C DATASETS

Table 14. Beatles albums used during this study

Title	Artist	Year	Genres	Song Count
Please Please Me	The Beatles	1963	Beat, Rock & Roll, Pop	14
With The Beatles	The Beatles	1963	Rock & Roll, Pop	14
A Hard Day's Night	The Beatles	1964	Rock, Pop Rock, Pop	14
Beatles for Sale	The Beatles	1964	Folk Rock, Rock & Roll, Pop Rock	14
Help!	The Beatles	1965	Folk Rock, Pop Rock	14
Rubber Soul	The Beatles	1965	Folk Rock, Rock, Pop	14
Revolver	The Beatles	1966	Rock, Pop	14
Magical Mystery Tour	The Beatles	1967	Psychedelic Rock	11
Sgt. Pepper's Lonely Hearts Club Band	The Beatles	1967	Pop, Art Rock, Rock	13
The Beatles	The Beatles	1968	Rock, Pop	30
Abbey Road	The Beatles	1969	Rock	17
Yellow Submarine	The Beatles	1969	Pop Rock, Psychedlic Rock	13
Let It Be	The Beatles	1970	Rock, Blues	12

D RESULTS

Fig. 9. Performance of the Detector

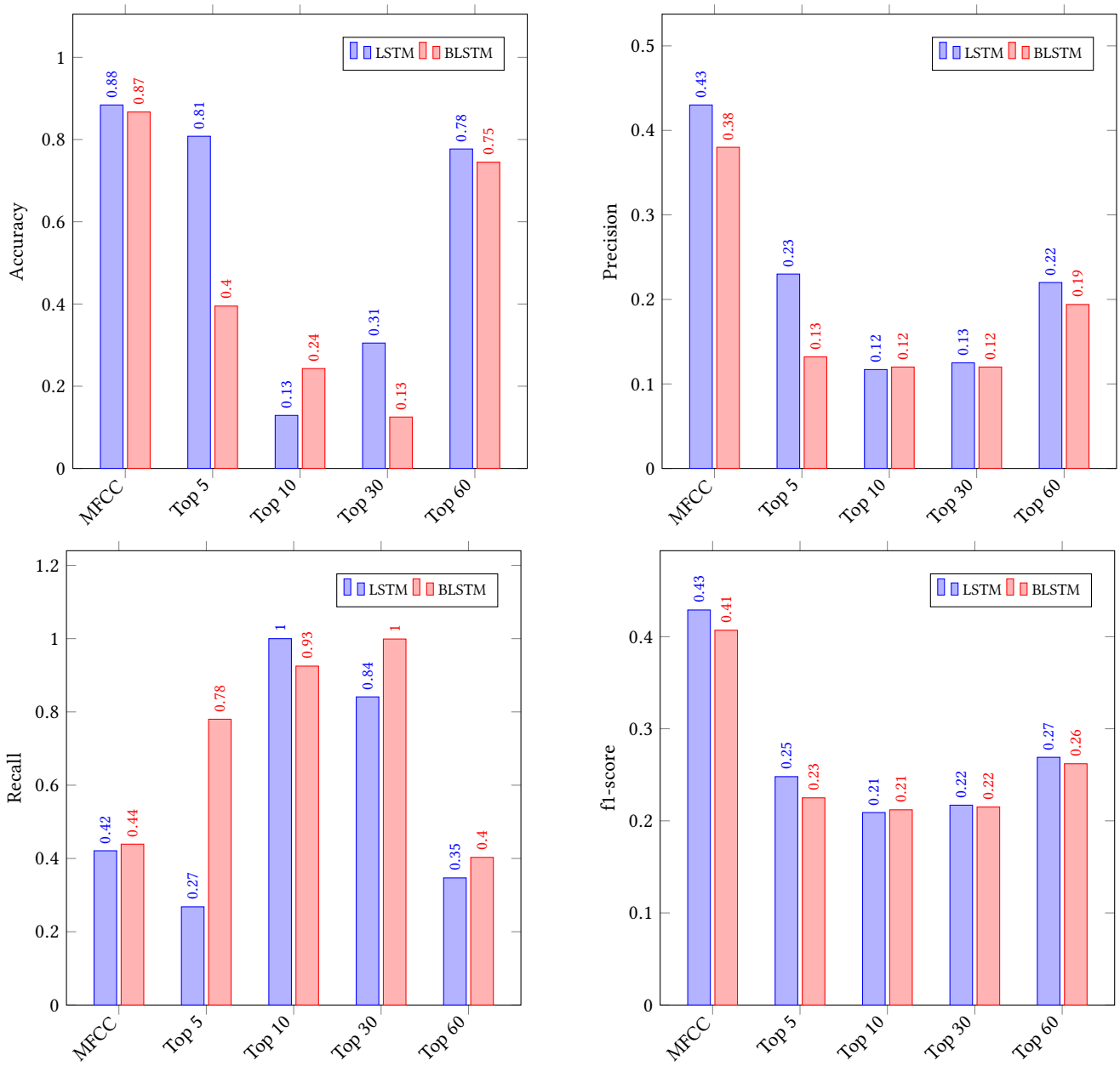

Fig. 10. Performance of the **DAE**

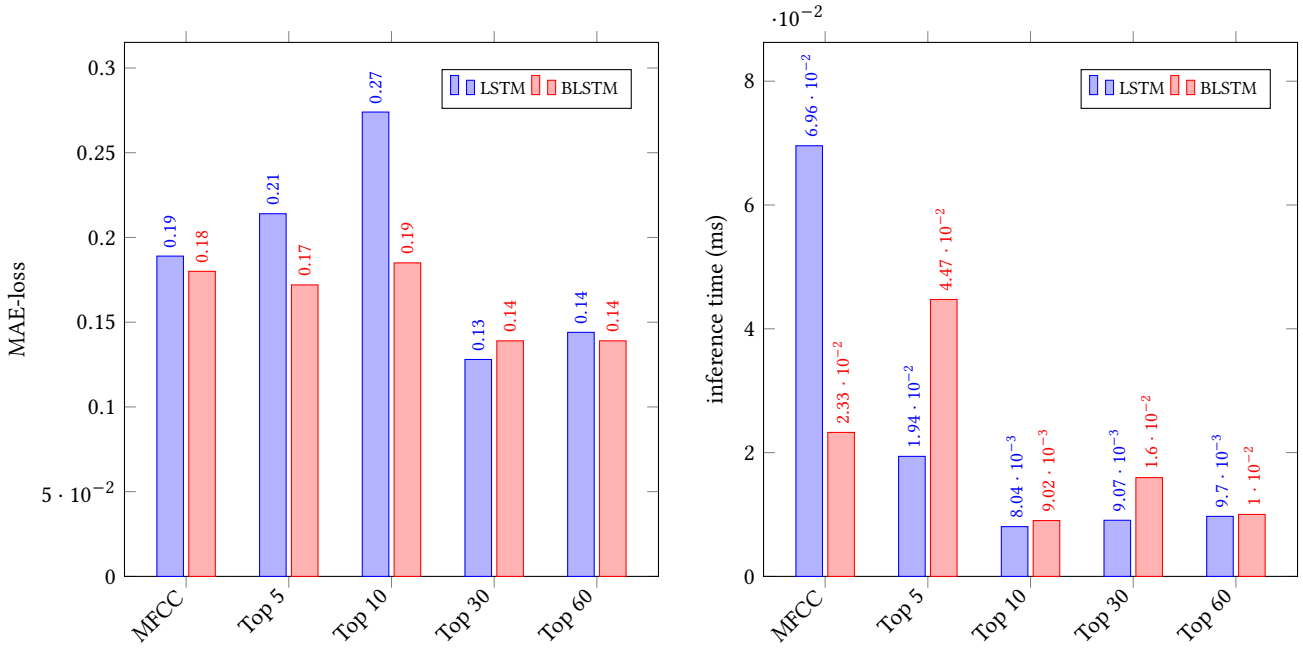

Fig. 11. Generalization Performance