National College of Ireland

Music Genre Classification Using Spectrograms and Deep Learning

MSc Research Project
Data Analytics

Kenechukwu Otito Ajufo
Student ID: x19190174

School of Computing
National College of Ireland

Supervisor: Dr. Christian Horn

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Kenechukwu Otito Ajufo
Student ID:	x19190174
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Christian Horn
Submission Due Date:	17/12/2020
Project Title:	Music Genre Classification Using Spectrograms and Deep Learning
Word Count:	5454
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	KOAjufo
Date:	16th December 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Music Genre Classification Using Spectrograms and Deep Learning

Kenechukwu Otito Ajufo
x19190174

Abstract

Automatic music genre classification is useful to instantly structure large collection of music files. This is very relevant today due to the huge amounts of music readily available both online and offline. In this research, the automatic music genre classification problem has been explored and solved using convolutional neural networks (CNN) and convolutional recurrent neural networks (CRNN). Deep Learning is utilized due to its notable success in image recognition. The experiments are performed on the GTZAN dataset. Feature extraction is the first task in audio analysis, in the feature extraction stage the audio snippets are transformed to spectrograms which serve as inputs to the neural networks. The models are then evaluated using Accuracy, Precision, Recall and F1 Score. The results show that these neural networks are able to classify music using these spectrograms. The baseline CNN model has a classification accuracy of 41%, on the other hand the baseline CRNN model obtains a classification accuracy of 66%. The results also reveal that the convolutional recurrent neural networks perform better in most experiments.

Keywords— music genre classification, convolutional neural networks, convolutional recurrent neural networks, deep learning, spectrograms

1 Introduction

1.1 Motivation

Music can be described as an art form, and like art it plays a vital part of human life. It functions as a catalyst for human expression promoting human imagination, interaction, and overall essence.

Music is considerably more accessible on the Internet today. Advancement in technology makes it easy to have large personal collections of music. Creating a solution to assist in organising and structuring large collections of songs cannot be ignored as it is time consuming to manually organize large personal collections. Automatic genre classification is a way to structure music content instantly.

Techniques for analysing audio use numerical features that attempt to model information in music content. Music genre labels are created by human experts with a purpose of structuring the vast music libraries available. Genre categorizations are arguably subjective, as there can be notable intersections because some genres have similar instrumentation and rhythmic structure. However, genre labels serve as ground truth to assess automatic genre classification systems.

1.2 Research Objective

This research aims to compare deep learning algorithms and their ability to automatically classify song snippets into the correct music genre. This will be done in two separate steps: first, the features are extracted from the raw audio, then the extracted features will be used as input into two deep learning algorithms.

Research Question: ***How accurately can Neural Networks like convolutional neural networks and convolutional recurrent neural networks classify music genres?***

The rest of the report is structured as follows: Section 2 introduces the reader to the related work in the literature. Section 3 presents the methodological approach adopted in the research. Section 4 covers the project design. Section 5 provides a detailed description of the stages used to implement the research. Section 6 presents the results and evaluation of the experiments. Section 7 concludes the report and gives recommendation for future research.

2 Related Work

2.1 Overview of Machine Learning Related Tasks in Music

There has been extensive work done in the Music Information Retrieval field. (Liebman; 2020) provides a survey which focuses on Machine Learning and Artificial Intelligence tasks in Music Intelligence. The authors categorize the Artificial Intelligence tasks in Music into three: target task, input type, algorithmic technique. Target task can be described as a precise problem to be solved e.g. Song Identification. Input type is used to express the various ways music can be represented for the target tasks. Finally Algorithmic Technique defines the technique utilized to solve the problem.

The application of Machine Learning to Music data can span across different tasks which can be analytic like Music Recommendation, to synthetic like Music Generation (Sturm et al.; 2019).

Music Recommendation using Machine Learning is an active and widely researched area in Music Information Retrieval (MIR). The Music Recommendation applications seek to learn the listening habits of its users to recommend a range of new items i.e. songs, albums, playlists, artists, genres (Schedl et al.; 2015) to the user.

Another growing research area is Music generation with various researchers exploring the use of Machine Learning to automatically compose music. (Bhave et al.; 2019) experiments the use of Restricted Boltzmann Machine and Long Short-Term Memory to generate music from raw audio files. In addition, there is research work on detecting emotion in music; an instance is the use of Russell's Circumplex Model and Recurrent Neural Network to detect emotions in segments of music by (Grekow; 2020)

Machine Learning has also been applied to tasks of identifying certain parts of a song, cover songs (Serra et al.; 2008), instruments in music (Solanki and Pandey; 2019) or classifying genre and or artists (Nasrullah and Zhao; 2019).

2.2 Music Genre Classification with Machine Learning

(Deepak and Prasad; 2020) attempt to classify music genres using two methods, the first method makes use of a Long Short-Term Memory model (LSTM) and the second involves a combination of two models: LSTM and Support Vector Machine (SVM), the LSTM model acts as an additional feature extractor while the SVM model uses the learned features from the previous model to predict the genres. The raw audio snippets are converted to wav files, then the non-silent portions obtained. In order to fit the models, the non-silent portions are used to derive

Mel-frequency Cepstral Coefficients (MFCCs) using Discrete Cosine Transform (DCT). They report the hybrid combination of the LSTM and SVM model achieves the higher accuracy.

(Sharma et al.; 2018) present an approach of music genre classification using three machine learning models (Decision Trees, Relevance Vector Machines and Support Vector Machines) stacked on each other to maximize the advantages of all three. The features used to train the models can be grouped into two: Perceptual Features and Statistical Spectrum Descriptors. These are acquired from the mathematical representation of audio signals. These features include MFCCs, Rhythm Histograms, Spectral Centroid and Rolloff, Zero Crossing Rate etc. Parameters are tuned using Random Search and the final classification is based on the application of the sum rule of the model's predicted score. The proposed technique obtains accuracy of 87% on the GTZAN[1] dataset.

(Vishnupriya and Meenakshi; 2018) adopts Convolutional Neural Network (CNN) to classify music genres automatically. The process entails two steps, the first step is the feature extraction of audio which is accompanied by the CNN. The audio is represented using two sets of features: MFCCs and Mel Spectrum. The model built with MFCCs features performs significantly better than the model built with Mel Spectrum.

(Liang and Gu; 2020) approach the task of music genre classification using transfer learning to show its effectiveness in music genre classification. Transfer learning is used to deal with data sparsity problems which are prominent when using small freely available datasets. The authors make use of two pretrained Convolutional Neural Networks models and compare with a baseline k Nearest Neighbours model. The models were trained with the MagnaTagATune Dataset[2] and the Million Song Dataset[3] respectively. The dataset contained 100 commercial songs, the audio is sliced into segments then randomly split into train and test sets. The pretrained model built with Million Song Dataset outperforms the baseline model with a higher Receiver Operating Characteristic Curve score. The authors also report Pop and R&B genres are easily misclassified.

(Elbir et al.; 2018) make an effort to classify music genres from the perspective of Short Time Fourier Transform. The authors investigate the influence of window type, size and overlap ratios in short time fourier transform feature-based extraction. Support Vector Machine with linear, polynomial and radial basis function kernels and Random Forest classifiers are used to train the features for classification. From the experiments, the authors conclude that Parzen Window type, Window Size of 512 and Overlap ratio of 256 yield the best accuracy.

(Pelchat and Gelowitz; 2019) propose music genre classification using deep learning. Songs are slashed into three second snippets, the three second snippets are then transformed into 128 by 900-pixel spectrograms. The data is further spliced into train, test and validation sets, and the spectrograms serve as input to a Neural Network with four convolutional layers. The output layer has eight neurons which presents the prediction. They report accuracy of 67% of the model on testing data, furthermore they reveal changing activation function in the convolutional layers from Rectified linear activation function (ReLU) to Exponential Linear Unit (ELU) slightly improves accuracy of the model.

(Wu et al.; 2019) use Independent Recurrent Neural Network under the assumptions that musical signals are sequential data. Their methodology is composed of three steps, the first involves feature extraction of the audio using Scattering Transform; this process reportedly reduces information loss. The next two steps consist of training the data with a five-layer Independent Recurrent Neural Network using Rectified linear activation function (ReLU) and finally classifying the music genres in the output layer using a Softmax activation function. The experiment is carried out using the GTZAN[1] dataset and the proposed method yields 97%.

(Li and Tzanetakis; 2003) provides a detailed exploration of factors that can affect per-

[1]http://marsyas.info/downloads/datasets.html

[2]http://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset

[3]http://millionsongdataset.com/

formance of the algorithms used in automatic music genre classification. The authors make use of three models, a pairwise Support Vector Machine, Multi Category Proximal Support Vector Machine and Linear Discriminant Analysis. The features used in this experiment consist of Fast Fourier transform (FFT), Mel-Frequency Cepstral Coefficients (MFCCs), Beat and Pitch. The result shows that classification accuracy of the algorithms increases in the following order of categories of features used: Beat, Pitch, MFCC and FFT. Finally, Linear Discriminant Analysis performs best across all the feature categories used.

(Yu et al.; 2020) probe the effect of the attention mechanism of Bidirectional Recurrent Neural Networks (BRNN) on GTZAN[1] and Extended Ballroom[4] datasets. The authors compare Serial Linear and Parallelized Linear attention-based mechanism models. Audio spectrograms derived from Short-Time Fourier Transform (STFT) are used as inputs for the models. The results demonstrate the effectiveness of the approach as BRNN models without attention do not perform as well as the BRNN models with attention mechanisms.

(Tzanetakis and Cook; 2002) present a method to automatically classify audio signals using timbral and non timbral text features. The authors utilize a Gaussian Classifier model, Gaussian Mixture model and kNN (k Nearest Neighbors) to compare model performance. The results prove that the models make misclassifications similar to humans. In conclusion they report that timbral texture features perform better than non timbral texture features.

(Sigtia and Dixon; 2014) proffer an improved music feature learning using Deep Neural Network whilst investigating the usefulness of Rectified linear activation function (ReLU) against Sigmoid activation function. The Deep Neural Network model extracts the latent features from the audio, then a Random Forest model is used to predict the genre labels. The deep neural network model is optimized with a Hessian-Free Optimization technique. The experiment is conducted using two datasets, GTZAN[1] for training and ISMIR[5] data for testing. The best accuracy is achieved using ReLU activation with a large number of hidden layers.

(Zhang et al.; 2016) presents a method to classify music genres using Convolutional Neural Network. The authors combine Max and Average Pooling Layers and use shortcut connections inspired by residual learning method in an attempt to improve the performance in the state-of-the-art. For the experiment the model is built using STFT; spectrogram images as inputs, this is followed by stacked CNN modules with Rectified linear activation functions (ReLU) to learn features in the spectrograms and a fully connected layer for classification. The report gives an improvement in the state-of-the-art with an overall accuracy of 87.4%.

The Paper by (Shakya et al.; 2017) addresses classification of music based on genre and mood. The models used for the experiments are Support Vector Machine (SVM) and Artificial Neural Network (ANN). The input to the models are 20ms and 40ms temporal audio features MFCCs, Spectral Rolloff and Centroid, Zero Crossing Rate etc. The GTZAN[1] dataset is used for music genre classification while the FMA dataset[6] is used for mood classification. The authors present MFCCs as the best feature and report the smaller frames have made the models achieve higher accuracy as in large frames the signal varies excessively, finally the ANN model performs better than the SVM in both tasks.

2.3 Conclusion

The literature highlights three major ways audio data can be used as input for training of machine learning models: representing the audio signals as timbral and acoustic features, extracting spectrograms from the audio signals and finally using raw audio. However (Wyse; 2017) argues that spectrograms retain more information than hand crafted features. Furthermore, spectrograms have lower dimension than raw audio.

[4]http://anasynth.ircam.fr/home/media/ExtendedBallroom

[5]http://ismir2004.ismir.net/genre_contest/index.html

[6]https://github.com/mdeff/fma

(Sturm; 2014) addresses the evaluation of music genre recognition systems. The article provides an extensive survey which cuts across several published research in the field. The author reports majority of research work involves datasets which primarily consist of Western Music. It also reveals that model's mean accuracy and contingency tables are the most used features of merit to assess the performance of music genre recognition systems.

The GTZAN dataset is widely used in the literature. (Sturm; 2012) carries out a study on the dataset. The author critically analyses it, highlighting the composition, the issues with integrity, the mislabelling. The analysis reveals several mislabelling and replications of the music tracks, which make them question the validity of the results in the literature since the replications can inflate the accuracy of some Machine Learning algorithms like kNN (k Nearest Neighbours) and in other algorithms like Boosted Trees reduce accuracy.

Based on the rigorous analysis of the related works, I have decided to focus on utilizing two Deep Learning algorithms (CNN, CRNN) to automatically classify music genres in the GTZAN dataset using Spectrograms.

3 Methodology

There are three major methodological processes that direct the execution of data mining applications in the industry, these are CRoss-Industry Standard Process for Data Mining (CRISP-DM), Knowledge Discovery in Databases (KDD) and Sample, Explore, Modify, Model, Assess (SEMMA), (Azevedo and Santos; 2008). This research work adopts the KDD process, which is an iterative process that is beneficial in uncovering knowledge in data. It involves the following stages: Data Selection, Data Preparation and Preprocessing etc. Figure 1 shows the flow of the methodology and the stages are explained in the remaining sections.

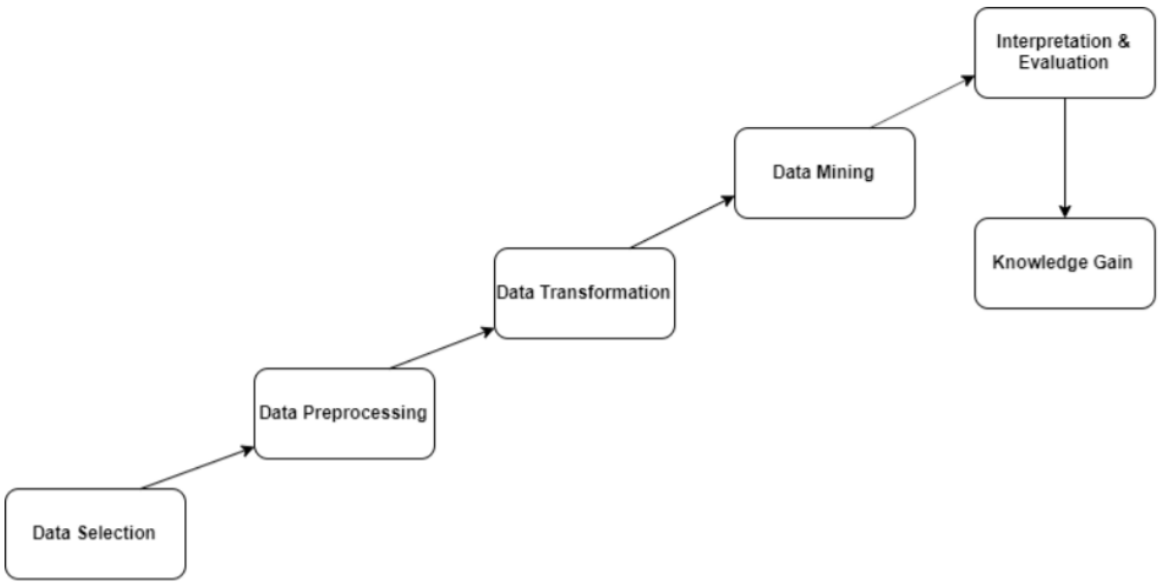


Figure 1: Knowledge Discovery in Databases (KDD) Process Flow

3.1 Data Selection

The experiments are implemented on the GTZAN Dataset[1]. The dataset is first used by (Tzanetakis and Cook; 2002) in which they present a method to classify audio signals based on non-timbral and timbral features. The dataset is primarily made up of 1000 audio songs that are 30 seconds in length. The songs are recorded as 22050Hz Mono 16-bit audio files in wav file format.

3.2 Data Preparation and Preprocessing

In this stage of the KDD methodology process, there are steps undertaken to further understand the dataset. The steps are discussed extensively below.

3.2.1 Exploratory Data Analysis

Exploratory Data Analysis provides a deep understanding of the data. Audio is unstructured data therefore the Exploratory Data Analysis performed is done to visually observe and compare the varying waveforms each genre possesses. The audio is equally distributed across 10 genres, with 100 songs in each category. The Waveform plot of One Bourbon, One Scotch, One Beer by John Lee Hooker (**blues00000**[7]) file in the dataset is shown in figure 2.

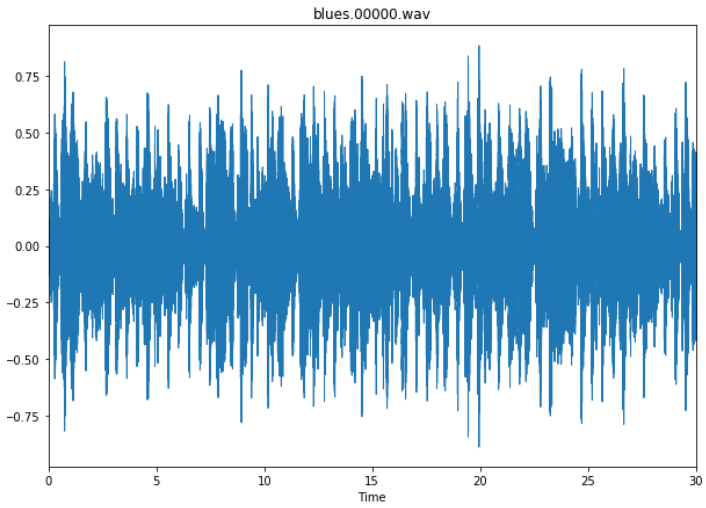


Figure 2: Blues0000 Waveform Plot - Time in seconds is shown on the x-axis. The waveform plot lays out the audio signal's Amplitude against time

3.2.2 Audio Processing

Feature extraction is influential to the performance of the Machine Learning algorithms. The raw audio is processed using Librosa[8], a Python library. Mel Spectrogram images are generated at this stage and serve as input to the Machine Learning models. In figure 3, the Spectrogram image for (**blues0000**[7]) is also displayed.

[7]https://music.apple.com/ng/album/one-bourbon-one-scotch-one-beer/1443903193?i=1443903656
[8]https://librosa.org/doc/main/index.html

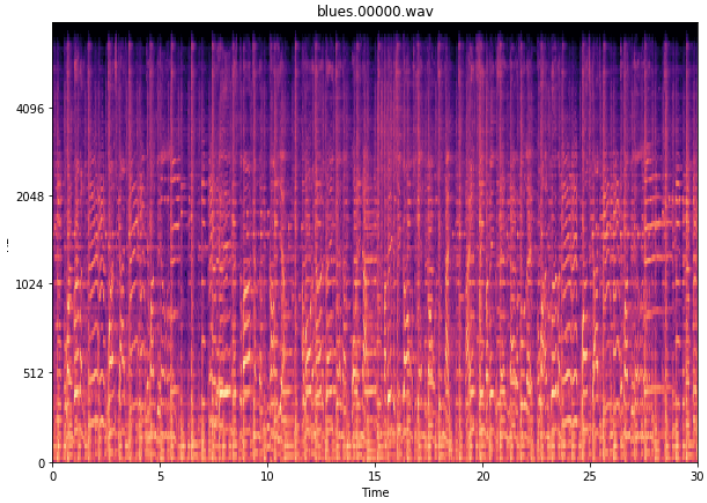


Figure 3: Blues0000 Spectrogram Plot - Time is on the x-axis and the audio signal's frequencies is on the y-axis. The colours represents the sound's Amplitude (Energy), The horizontal bands correspond to particular notes, lastly the vertical stripes represent beats and their horizontal spread the rhythm

3.3 Data Mining Techniques

There is a considerable amount of research done on music genre classification, this research explores the task of classifying audio signals automatically using two deep learning models: Convolutional Neural Network and Convolutional Recurrent Neural Network.

1. **Convolutional Neural Network (CNN)**
   Convolutional Neural Networks (CNN) are neural networks which are widespread models used in image classification, image recognition, object detection, and face recognition tasks. The CNN models have convolutional layers that treat image data as spatial and which allows it to learn features from the image data. CNN are widely adopted in the literature, this is attributed to the success they have obtained in Computer Vision. Spectrogram images are a representation of audio signals since CNN models function well on image data, it is an ideal choice.

2. **Convolutional Recurrent Neural Network (CRNN)**
   Convolutional Recurrent Neural Network (CRNN) combines a convolutional neural network (CNN) and a recurrent neural network (RNN) to harness the benefits of the two models. The architecture of the CRNN model consists of a CNN unit and a RNN unit. The CNN model extracts the features from the spectrograms, while the RNN model learns long-term context in audio signals. The CRNN model is chosen due to the research done by (Choi et al.; 2017). The results from their experiment reveal the accuracy of CRNNs are higher than CNNs. This research attempts to examine if similar results can be replicated.

4 Design Specification

This section provides the architecture for the implementation of the research objectives. The project implementation takes on a two tier architecture. This is shown in figure 4 below. The first layer is the Data Layer. This layer presents the techniques used in the collection and

processing of the audio data and also the data mining algorithms utilized in classifying the audio data. The second layer is the Presentation Layer which displays the evaluation and results of the data mining algorithms, the results are presented in confusion matrices at this stage. The project is implemented in Python, using Keras and TensorFlow deep learning libraries.

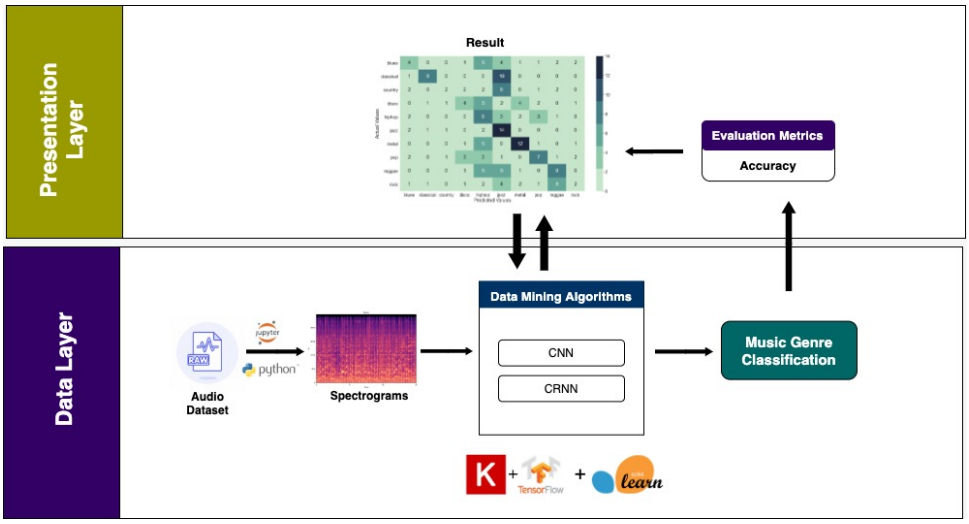


Figure 4: Project Architecture

5 Implementation

This section reviews the implementation process. The first step is retrieving the dataset from the data source. The remaining steps are described in subsequent subsections.

5.1 Data Preparation & Processing

1. **Audio Processing**
   The dataset contains 1,000 audio songs across 10 genres, however, the jazz.0054 file is corrupted and not used in the experiment. Raw audio are used to generate Mel Spectrograms. This process is accomplished with Librosa. The Mel Spectrogram images are generated at the sampling rate of 22050Hz, Fast Fourier transform (FFT) window length of 2048 and a hop length of 1024. Finally, the images are uniformly resized to 128x660 pixels. The spectrogram images are roughly identical to the spectrogram used as inputs in the article by (Zhang et al.; 2016). Figure 3 displays a sample spectrogram image used for training the model.

2. **Label Transformation**
   The ten genre labels are transformed to one hot encoded values this is done with to_categorical function in Keras.

3. **Train-Test Split**
   The dataset is split into Training and Testing set at 80:20.

5.2 Data Mining Techniques

Three Deep Learning algorithms are applied on the spectrogram images with two classification tasks. The first classification task, the model classifies ten genres. In the second classification

task, the genres are grouped. Disco, HipHop, Pop, Reggae, Country and Rock are grouped together under a class called Pop Style, likewise Blues, Jazz and Classical are grouped under an Eclectic Style class.

5.2.1 Multi-Class Classification

1. **Convolutional Neural Network (CNN)**
   The CNN model has a 2D Convolution input layer, with 32 filters, and a kernel size of 3 by 3, it receives the input training set tensor, it is followed by a 2D Max Pooling layer with pool size of 2 by 4. The next layer is another 2D Convolution layer, with 64 filters and another 2D Max Pooling layer with pool size set to 2 by 4. Finally, one Flatten layer flattens the output of the preceding layer, then there are two Dense layer with the last Dense producing the output of the classification. All layers have Rectified linear activation function (ReLU) asides the Output layer which has Softmax Activation. The model is compiled with the Categorical Cross-Entropy loss function, Adam Optimizer and Metrics set to Accuracy. The model is trained for twenty five epochs with a batch size of 20. The structure of the model is shown in figure 11.

2. **Convolutional Recurrent Neural Network (CRNN)**
   This model is similar to the model structure in the work done by (Nasrullah and Zhao; 2019). The Convolutional Unit consist of four 2D Convolution layers, there are 2D Max Pooling layers, Batch Normalization and Dropout layers after each Convolution layer. The Recurrent Unit contains two Gated Recurrent Unit (GRU) layers with a final Dense output layer for classification. All four 2D Convolution layers have Exponential Linear Unit activation function (ELU). The Dense Output layer has Softmax Activation. The model is compiled and trained with the same settings as the CNN model. Figure 12 lays out the pictorial representation of the model's architecture.

5.2.2 Binary Classification

In the binary classification the architecture of the model remain the same. Similar to the multi-class classification, the models for the Binary Classification are compiled with Categorical Cross-Entropy loss function. Adam Optimizer and Metrics set to Accuracy also stay the same. The models are trained with the same number of epochs and batch size.

5.3 Hyperparameter Optimization

The baseline Multi-Class Classification models are optimized using **Random Search**. Random Search Optimization uses probability to find a good solution quickly, sacrificing a guaranteed optimal values in the search space (Zabinsky; 2011). Three sets of experiments are performed on each baseline model. The hyperparameters in the first experiment are the units in each layer and the learning rates for the Adam Optimizer, in the second experiment, the hyperparameters are the activation functions and the learning rates for the Adam Optimizer. In the last experiment both the units and activation functions are tuned. The objective of the hyperparameter optimization is to maximize Validation Accuracy. The experiments are implemented using Keras Tuner[9] on Google Colaboratory[10].

[9]https://keras-team.github.io/keras-tuner/
[10]https://colab.research.google.com/

6 Evaluation

The results of the implementation are discussed in this section. The main objective of the research project is to classify music genres automatically. A variety of evaluation metrics are used in the literature. In this research, Accuracy, Precision, Recall and F1 Score are used for the purpose of assessing the models implemented. These metrics are discussed below. In addition to the results, the Contingency Matrices for Models are shown.

1. Accuracy: This is a metric used to assess Machine Learning classification models. Accuracy is the ratio of correct predictions from model to the total number of predictions from the model. The formula used to calculate Accuracy is displayed in equation 1.

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + TrueNegative + FalseNegative} \quad (1)$$

2. Precision: This metric assesses the classification model's ability to identify relevant cases. Precision is calculated by dividing the true positives with the addition of the true positives and false positives. The formula is shown in equation 2.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (2)$$

3. Recall: Recall displays the percentage of total relevant cases correctly classified by the model. Recall is the ratio of True Positives by the addition of True Positives and False Negatives. The formula is shown in equation 3.

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (3)$$

4. F1 Score: This metric balances the model's Precision and Recall. The F1 Score awards equal weights to the Precision and Recall. The F1 Score equation is displayed below.

$$F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

6.1 Multi-Class Classification

There are ten genre classes: Blues, Classical, Country, Disco, HipHop, Jazz, Metal, Pop, Reggae, and Rock used in the multi-class classification experiments.

Table 1: Multi-Class Classification Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN	41	44	40	41
CRNN	**66**	**68**	**66**	**65**

The metrics of the two baseline models are shown in table 1.

- The baseline CNN model has an accuracy of 41%, with the corresponding Precision, Recall and F1 Score of 44%, 40% and 41%. From figure 5, it can be observed that the baseline CNN model correctly classifies 5 Blues, 13 Classical, 7 Country, 5 Disco, 13 Hiphop, 6 Jazz, 12 Metal, 9 Pop, 7 Reggae and 4 Rock songs in the test set.

- The baseline CRNN model has an accuracy of 66%, with Precision, Recall and F1 Score of 68%, 66% and 65% respectively. In figure 6, it can be deduced that the baseline CRNN model correctly classifies 15 Blues, all 20 Classical, 6 Country, 7 Disco, 18 Hiphop, 10 Jazz, 14 Metal, 15 Pop, 11 Reggae and 16 Rock songs in the test set.

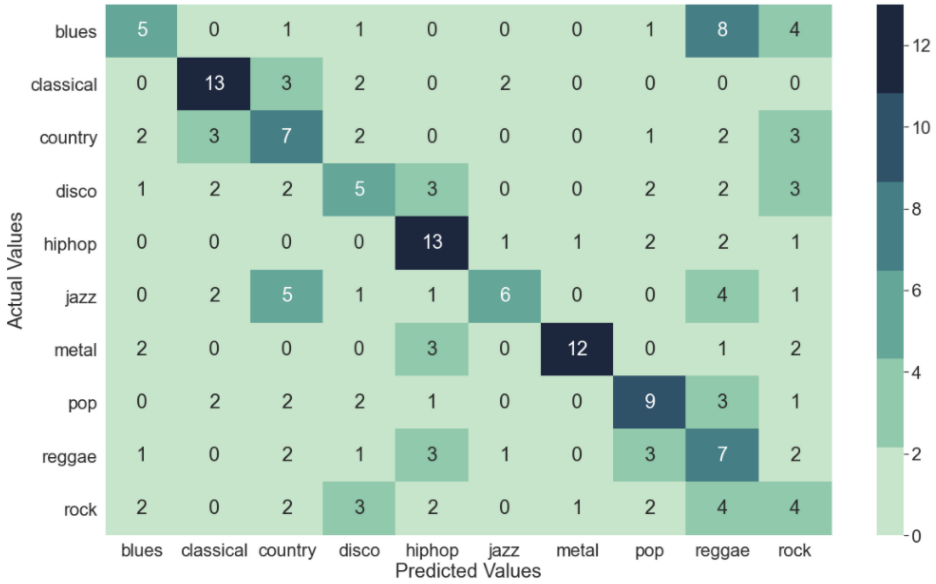


Figure 5: Baseline CNN Model Confusion Matrix

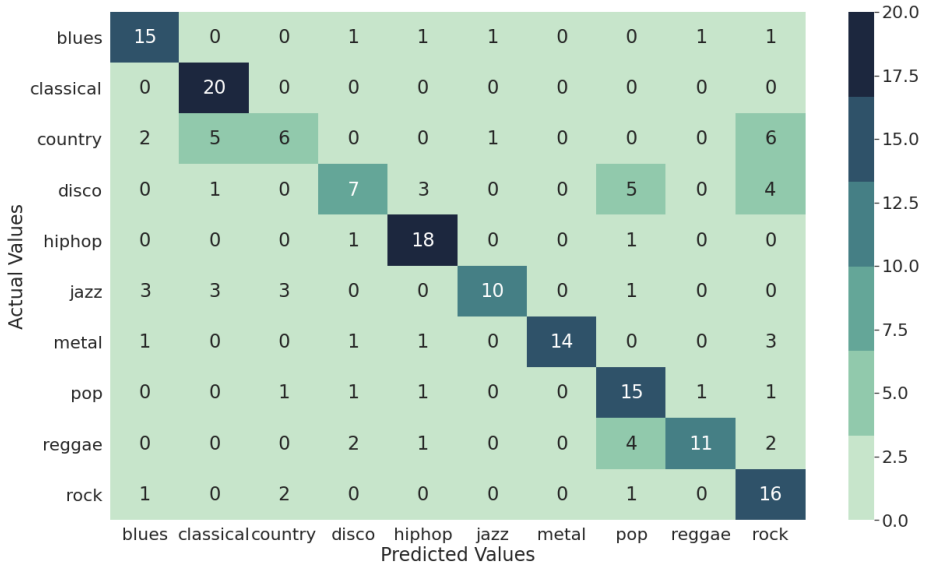


Figure 6: Baseline CRNN Model Confusion Matrix

6.2 Binary Classification

The table below gives a description of the metrics for the Binary Classification experiment. The CRNN model performs best in this category, this is followed closely by the CNN model. There are two classes used for the binary classification experiments. Disco, HipHop, Pop, Reggae,

Country and Rock are grouped together under a class called **Pop Style**, likewise Blues, Jazz and Classical are grouped under a class named **Eclectic Style**

Table 2: Binary Classification Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN	83	80	80	80
CRNN	**86**	**84**	**85**	**84**

Figure 7: Binary CNN Model Confusion Matrix

Figure 7 displays the confusion matrix of the CNN Model. The Model classifies 43 songs in the Eclectic class and 124 songs in the Pop-Style class correctly. The model incorrectly predicts 17 songs as Pop-Style and 16 songs as Eclectic.

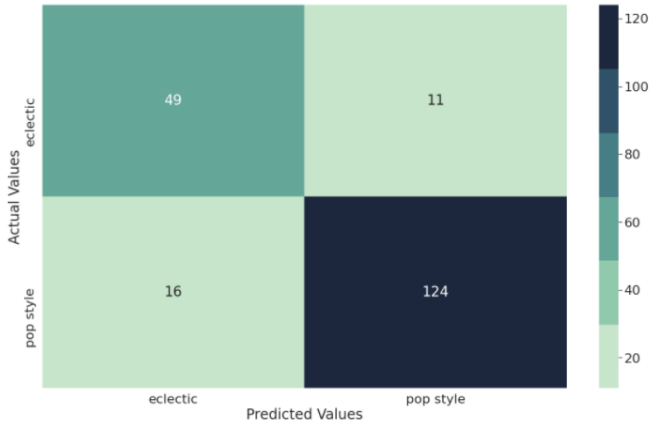


Figure 8: Binary CRNN Model Confusion Matrix

The confusion matrix of the CRNN Model is shown in figure 8. Here, the model correctly

predicts 49 songs as Eclectic and 124 songs as Pop-Style, the model also classifies 16 songs as Eclectic and 11 songs as Pop-Style incorrectly.

6.3 Hyperparameter Optimization

6.3.1 Unit Parameter Tuning

The Units at each layer are tuned to find the optimum parameter that yields the highest validation accuracy. Table 3 is explained below:

- The best CNN Model's Accuracy is 59%. Similarly the Model's Precision, Recall and F1 Score are 63%, 59%, 59% respectively. The best model is the model with 32 units in the first 2D Convolutional layer, 96 units in the second 2D Convolutional layer, Dense layer with 112 units and a learning rate of 0.00021. Figure 13a displays the confusion matrix.
- The best CRNN Model has an Accuracy of 65%. The Model's Precision is 68%, Recall is 64% and F1 Score is 64%. The best model is the model with 128 units in the first 2D Convolutional layer, 48 units in both the second and third 2D Convolutional layers, 80 units in the last Convolutional layer. In the Recurrent unit, the first GRU layer has 112 units and the final GRU layer has 48 units. The learning rate is 0.00028. The confusion matrix for the model in this experiment is presented in figure 13b.

Table 3: Unit Parameter Tuning Classification Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN	59	63	59	59
CRNN	**65**	**68**	**64**	**64**

6.3.2 Activation Function Parameter Tuning

The Activation Functions in each layer are tuned to discover the optimum activation functions that present the best validation accuracy. The confusion matrices are presented in Appendix C. Table 4 is discussed below:

- The best CNN Model has an Accuracy, Precision, Recall and F1 Score of 48%, 51%, 48% and 47%. The best CNN Model has Hyperbolic Tangent (Tanh) activation function in the first and second Convulational layers, Exponential Linear Unit (ELU) functions is present in the final Dense layer. The optimal learning rate is 0.00102.
- The best CRNN Model's Accuracy is 64%, Precision is 68%, Recall is 64% and F1 Score is 65%. The best model consists of Exponential Linear Unit (ELU) functions in the first Convulational layer, and the first and second Gated Recurrent Unit layers. Rectified Linear Activation (ReLU) function is used in the second and third Convulational layer. The last Convulational layer has a Hyperbolic Tangent (Tanh) activation function. The learning rate for the model is 0.000454.

Table 4: Activation Function Parameter Tuning Classification Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN	48	51	48	47
CRNN	**64**	**68**	**64**	**65**

6.3.3 Activation Function & Unit Parameter Tuning

Table 5: Activation Function & Unit Parameter Tuning Classification Results

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNN	**70**	**70**	**70**	**70**
CRNN	66	66	66	65

Table 5 presents the results from both hyperparameter tuning of both the units and activation functions. The CNN model produces the best results in this category with an Accuracy of 70%. The following describes the best parameters for both Models:

- The best CNN model has in its first Convolutional layer 128 units and a Hyperbolic Tangent Activation Function (Tanh). The second layer has an Exponential Linear Unit (ELU) activation function and 48 units. The Dense layer has 112 units and an Exponential Linear Unit (ELU) activation function. The optimal learning rate is 0.000284.
- The best CRNN model has 96 units and Rectified Linear Activation (ReLU), 48 and Sigmoid activation function, 64 and a Hyperbolic Tangent Activation Function (Tanh) in the first, second, third Convolutional layers respectively. In the fourth Convolutional layers there are 80 units and a Hyperbolic Tangent Activation Function (Tanh). The Recurrent Unit contains 112 units in the first Gated Recurrent Unit and 96 units in the second Gated Recurrent Unit. The best learning rate is 0.000493.

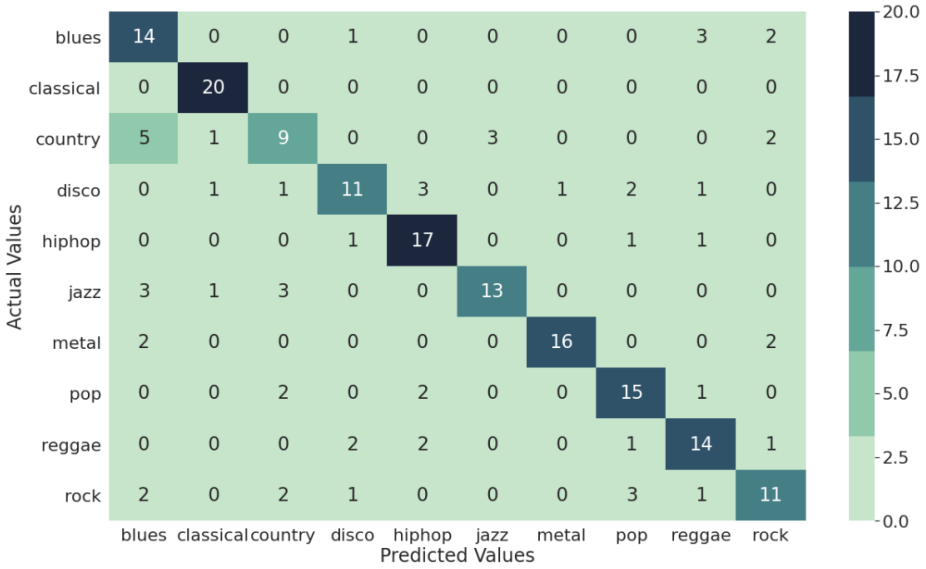


Figure 9: CNN Model Confusion Matrix - Activation Function & Unit Parameter Tuning

Figure 9, displays the confusion matrix of the best CNN model in the experiment where all parameters are tuned. The CNN model classifies 14 Blues, 20 Classical, 9 Country, 11 Disco, 17 Hiphop, 13 Jazz, 16 Metal, 15 Pop, 14 Reggae and 11 Rock songs correctly in the test set.

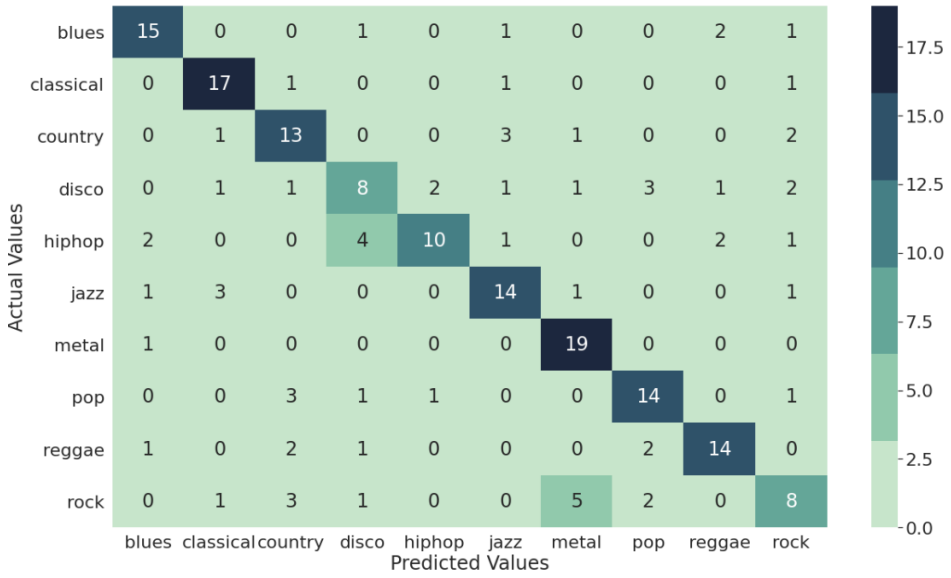


Figure 10: CRNN Model Confusion Matrix - Activation Function & Unit Parameter Tuning

Figure 10, shows the confusion matrix of the best CRNN model from the Activation Function & Unit Parameter Tuning. The best CRNN model classifies 15 Blues, 17 Classical, 13 Country, 8 Disco, 10 Hiphop, 14 Jazz, 19 Metal, 14 Pop, 14 Reggae and 8 Rock songs correctly in the test set.

6.4 Discussion

6.4.1 Multi-Class Classification

- Rock, Jazz, Disco and Blues are the genres with the least accurate predictions from the baseline CNN model. The model misclassified Rock as Blues, Country, Metal, Disco. Pop. Disco as Blues, Classical, Country. Blues as Disco, Country, Pop. Metal, Hiphop and Classical are the genres with the most accurate predictions from the model.
- The genres with the highest correct predictions from the baseline CRNN model are Blues, Classical, Hiphop, Pop, Metal and Rock. Country and Disco genres have the least correct predictions. The model misclassified Country as Blues, Classical, Jazz and Rock. It also misclassified Disco as Classical, Hiphop, Pop and Rock.

6.4.2 Binary Classification

The high Accuracy in the binary classification experiment can be attributed to the class imbalance. Class imbalance is a strong factor that can affect performance of machine learning classification algorithms (Zheng and Jin; 2020). The model is trained with a Pop Style class that has twice the amount Eclectic Values in the dataset. This makes the model biased to the Pop Style class. The Accuracy of the models in the Binary Classification experiments are reasonably good, hence, no hyperparameter optimization is done.

6.4.3 Hyperparameter Optimization

Hyperparameter Tuning increases the performance of the Baseline CNN Model. Another interesting thing noticed is that in the hyperparameter tuning of both the Unit and Activation

Functions, the Baseline CNN model performs better. In the case of the CRNN model, the performance of the in both hyperparameter tuning experiment does not yield better performance results.

The Best CNN model performance is obtained when both the Units and Activation functions are tuned, while the performance of the CRNN plateaus.

6.4.4 Summary

In summary, it is important to note the feature extraction is crucial to obtaining high accuracy as reported in the literature. To improve on accuracy, the audio data can be sliced into smaller frames (3-20 seconds), in large frames the audio signal can vary too much. Furthermore, the spectograms can be transformed taking careful considerations of the rhythmic and pitch of the musical instruments of each genre.

7 Conclusion and Future Work

The research sought out to classify music genres using Deep Learning and Spectrograms. The extensive experiments performed on the GTZAN dataset shows that CNN and CRNN models can automatically classify music genres using Spectrograms. The CRNN model performs better across all experiments, this is consistent with the results in the literature. The models make inaccurate predictions identical to what humans could make, for instance Blues is misclassified for Classical and Jazz, Rock for Metal, Hiphop and Pop is misclassified as Disco and Reggae, Reggae as Disco and Pop.

This research can be extended in various ways: music genre and artist classification is the foundation on which music recommendation can be made, with the appropriate user listening dataset, this research can be used to build a content based music recommendation system. Future work can also focus on classifying music based on mood which can be used for Music therapy, finally, Transfer Learning can be explored, this can be useful in leveraging the potential of this method to learn features that can be beneficial for Music Information Retrieval (MIR) tasks.

8 Acknowledgement

Foremost, I would like to convey my gratitude to my supervisor Dr. Christian Horn for the support over the period of my MSc. research project, I am grateful for his patience, inspiration, and vast knowledge. His guidance assisted me through the research and writing of this report. I cannot imagine having a better supervisor.

My heartfelt thanks also go to my family: my mom and dad, Mr and Mrs Ajufo for this opportunity. And last but not the least, my friends for the emotional support through my Masters.

References

Azevedo, A. and Santos, M. F. (2008). KDD, SEMMA and CRISP-DM: A Parallel Overview, *in* A. Abraham (ed.), *IADIS European Conference on Data Mining 2008, Amsterdam, The Netherlands, July 24-26, 2008. Proceedings*, IADIS, pp. 182–185.

Bhave, A., Sharma, M. and Janghel, R. R. (2019). Music Generation Using Deep Learning", *in* J. Wang, G. R. M. Reddy, V. K. Prasad and V. S. Reddy (eds), *Soft Computing and Signal*

Processing, Springer Singapore, Singapore, pp. 203–211. DOI:https://doi.org/10.1007/978-981-13-3393-4_21.

Choi, K., Fazekas, G., Sandler, M. and Cho, K. (2017). Convolutional recurrent neural networks for music classification, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 2392–2396. DOI:https://doi.org/10.1109/ICASSP.2017.7952585.

Deepak, S. and Prasad, B. (2020). Music classification based on genre using lstm, *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, pp. 985–991. DOI:https://doi.org/10.1109/ICIRCA48905.2020.9182850.

Elbir, A., İlhan, H. O., Serbes, G. and Aydın, N. (2018). Short Time Fourier Transform based music genre classification, *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, IEEE, pp. 1–4. DOI:https://doi.org/10.1109/EBBT.2018.8391437.

Grekow, J. (2020). Static Music Emotion Recognition Using Recurrent Neural Networks, *in* D. Helic, G. Leitner, M. Stettinger, A. Felfernig and Z. W. Raś (eds), *Foundations of Intelligent Systems*, Springer International Publishing, Cham, pp. 150–160. DOI:http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-030-59491-6_14.

Li, T. and Tzanetakis, G. (2003). Factors in automatic musical genre classification of audio signals, *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*, IEEE, pp. 143–146. DOI:https://doi.org/10.1109/ASPAA.2003.1285840.

Liang, B. and Gu, M. (2020). Music Genre Classification Using Transfer Learning, *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, IEEE, pp. 392–393. DOI:https://doi.org/10.1109/MIPR49039.2020.00085.

Liebman, E. (2020). Related Work and a Taxonomy of Musical Intelligence Tasks, *Sequential Decision-Making in Musical Intelligence*, Springer International Publishing, Cham, pp. 143–196. DOI:https://doi.org/10.1007/978-3-030-30519-2_8.

Nasrullah, Z. and Zhao, Y. (2019). Music Artist Classification with Convolutional Recurrent Neural Networks, *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8. DOI:https://doi.org/10.1109/IJCNN.2019.8851988.

Pelchat, N. and Gelowitz, C. M. (2019). Neural Network Music Genre Classification, *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, IEEE, pp. 1–4. DOI:https://doi.org/10.1109/CCECE.2019.8861555.

Schedl, M., Knees, P., McFee, B., Bogdanov, D. and Kaminskas, M. (2015). Music Recommender Systems, *in* F. Ricci, L. Rokach and B. Shapira (eds), *Recommender Systems Handbook*, Springer US, Boston, MA, pp. 453–492. DOI:https://doi.org/10.1007/978-1-4899-7637-6_13.

Serra, J., Gómez, E., Herrera, P. and Serra, X. (2008). Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification, *IEEE Transactions on Audio, Speech, and Language Processing* **16**(6): 1138–1151. DOI:https://doi.org/10.1109/TASL.2008.924595.

Shakya, A., Gurung, B., Thapa, M. S., Rai, M. and Joshi, B. (2017). Music Classification Based on Genre and Mood, *in* J. K. Mandal, P. Dutta and S. Mukhopadhyay (eds), *International Conference on Computational Intelligence, Communications, and Business Analytics*, Springer Singapore, Singapore, pp. 168–183. DOI:https://doi.org/10.1007/978-981-10-6430-2_14.

Sharma, S., Fulzele, P. and Sreedevi, I. (2018). Novel hybrid model for music genre classification based on support vector machine, *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, pp. 395–400. DOI:https://doi.org/10.1109/ISCAIE.2018.8405505.

Sigtia, S. and Dixon, S. (2014). Improved music feature learning with deep neural networks, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6959–6963. DOI:https://doi.org/10.1109/ICASSP.2014.6854949.

Solanki, A. and Pandey, S. (2019). Music instrument recognition using deep convolutional neural networks, *International Journal of Information Technology* pp. 1–10. DOI:https://doi.org/10.1007/s41870-019-00285-y.

Sturm, B. L. (2012). An Analysis of the GTZAN Music Genre Dataset, MIRUM '12, Association for Computing Machinery, New York, NY, USA, p. 7–12. DOI:https://doi.org/10.1145/2390848.2390851.

Sturm, B. L. (2014). A Survey of Evaluation in Music Genre Recognition, *in* A. Nürnberger, S. Stober, B. Larsen and M. Detyniecki (eds), *International Workshop on Adaptive Multimedia Retrieval: Semantics, Context, and Adaptation*, Springer International Publishing, Cham, pp. 29–66. DOI:https://doi.org/10.1007/978-3-319-12093-5_2.

Sturm, B. L., Ben-Tal, O., Monaghan, Ú., Collins, N., Herremans, D., Chew, E., Hadjeres, G., Deruty, E. and Pachet, F. (2019). Machine learning research that matters for music creation: A case study, *Journal of New Music Research* **48**(1): 36–55. DOI:https://doi.org/10.1080/09298215.2018.1515233.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals, *IEEE Transactions on speech and audio processing* **10**(5): 293–302. DOI:https://doi.org/10.1109/TSA.2002.800560.

Vishnupriya, S. and Meenakshi, K. (2018). Automatic Music Genre Classification using Convolution Neural Network, *2018 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, pp. 1–4. DOI:https://doi.org/10.1109/ICCCI.2018.8441340.

Wu, W., Han, F., Song, G. and Wang, Z. (2019). Music Genre Classification Using Independent Recurrent Neural Network, *2018 Chinese Automation Congress (CAC)*, IEEE, pp. 192–195. DOI:https://doi.org/10.1109/CAC.2018.8623623.

Wyse, L. (2017). Audio Spectrogram Representations for Processing with Convolutional Neural Networks, *arXiv preprint arXiv:1706.09559* .

Yu, Y., Luo, S., Liu, S., Qiao, H., Liu, Y. and Feng, L. (2020). Deep attention based music genre classification, *Neurocomputing* **372**: 84–91. DOI:https://doi.org/10.1016/j.neucom.2019.09.054.
URL: *http://www.sciencedirect.com/science/article/pii/S0925231219313220*

Zabinsky, Z. B. (2011). *Random Search Algorithms*, Wiley.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470400531.eorms0704*

Zhang, W., Lei, W., Xu, X. and Xing, X. (2016). Improved Music Genre Classification with Convolutional Neural Networks, *Interspeech 2016*, pp. 3304–3308.
URL: *http://dx.doi.org/10.21437/Interspeech.2016-1236*

Zheng, W. and Jin, M. (2020). The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study, *SN Computer Science* **1**(2): 1–13. DOI:`https://doi.org/10.1007/s42979-020-0074-0`.

A Appendix A: Model Architecture

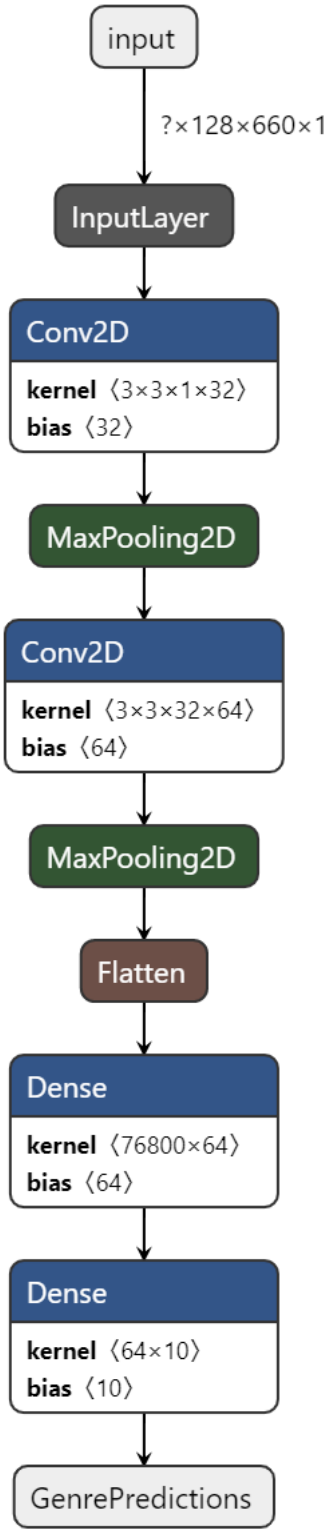


Figure 11: CNN Model Architecture

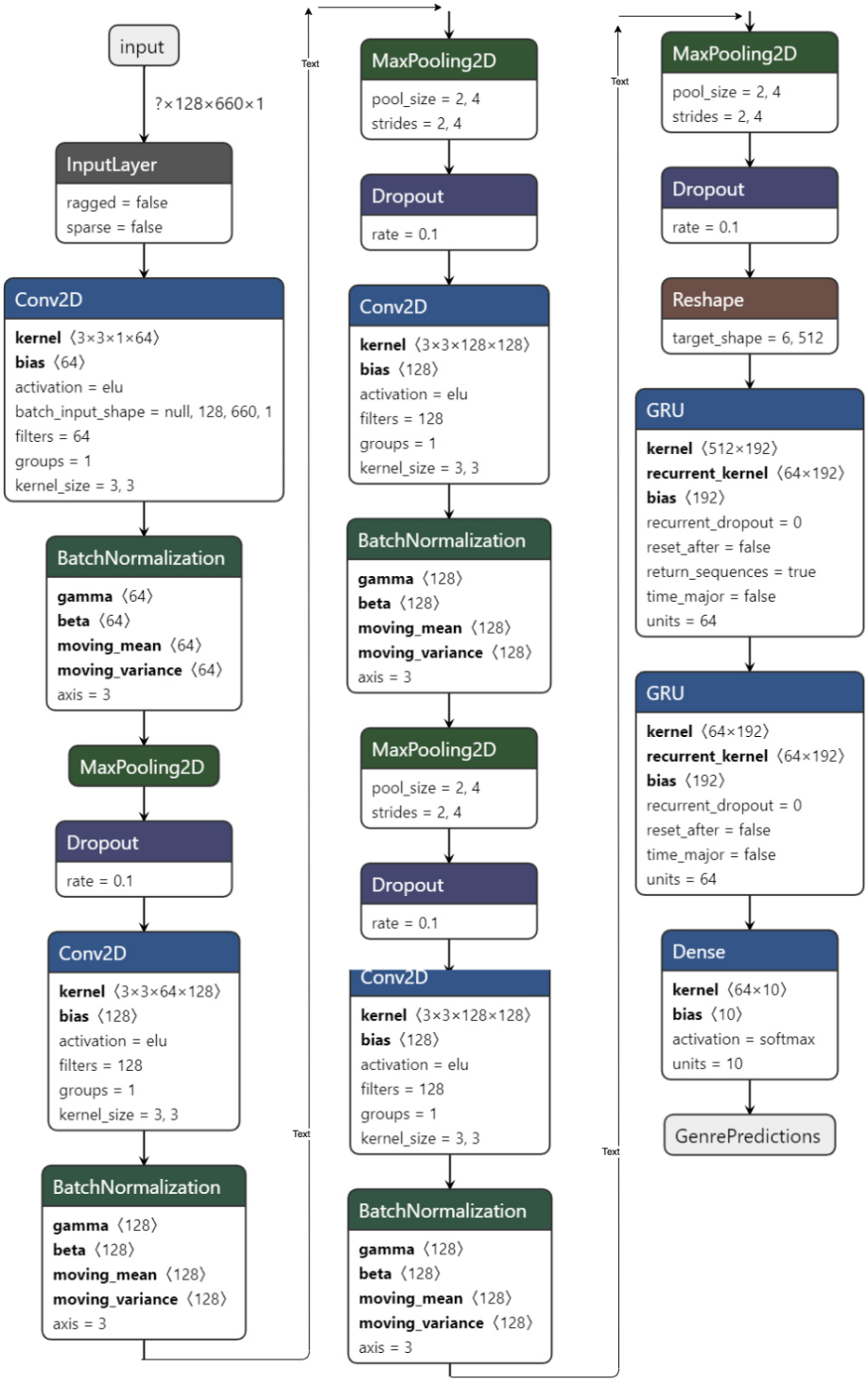
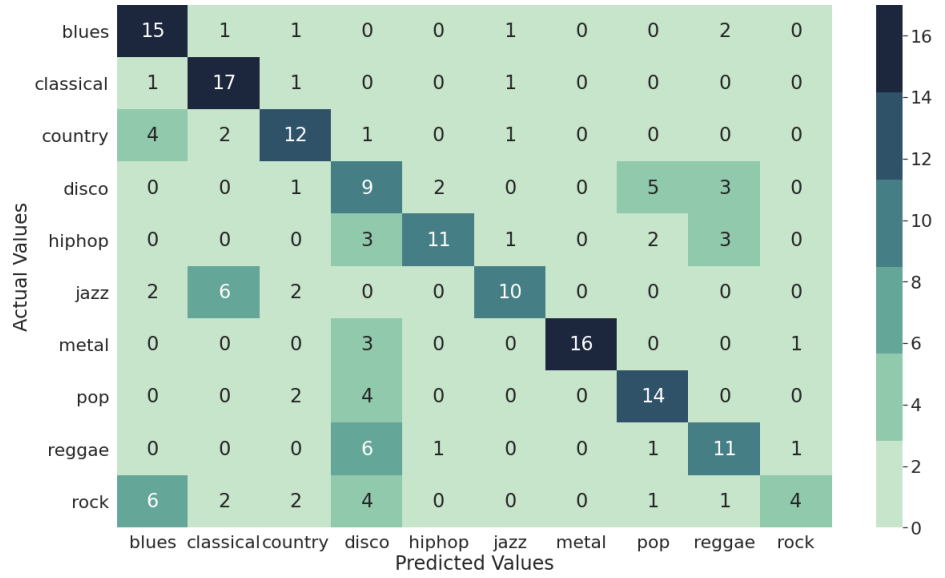


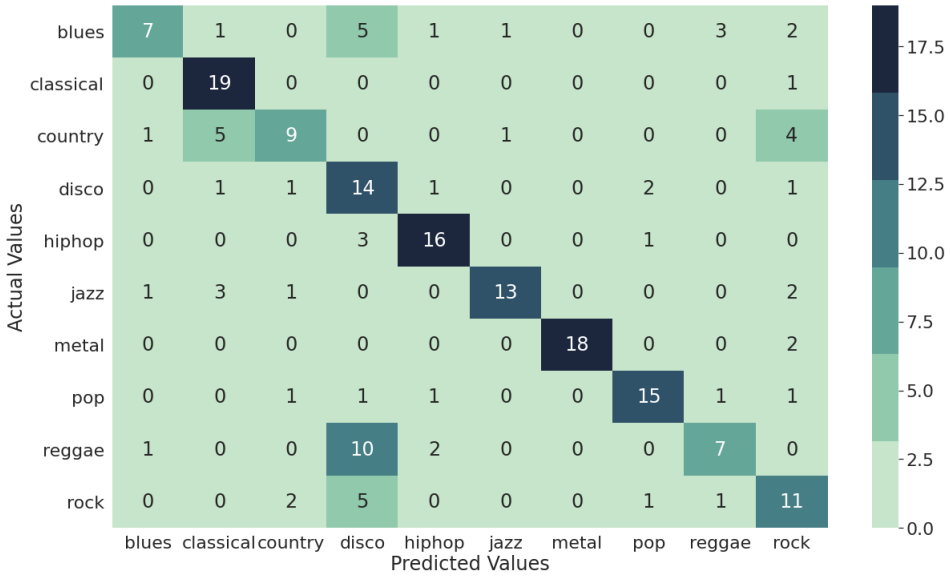
Figure 12: CRNN Model Architecture

The CNN and CRNN models' architecture is shown in figures 11 and 12 respectively. It gives a diagrammatic representation both models.

B Appendix B: Unit Parameter Tuning Confusion Matrices



(a) CNN Model Confusion Matrix



(b) CRNN Model Confusion Matrix

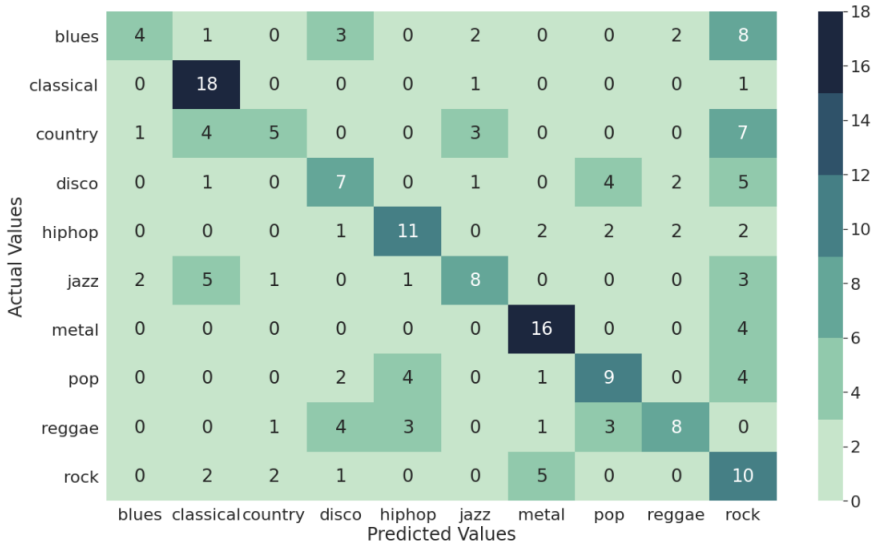
Figure 13: Unit Parameter Tuning Confusion Matrices

Figures 13a and 13b reveal the predictions for the best CNN and CRNN Models in the Unit Parameter Tuning experiments.

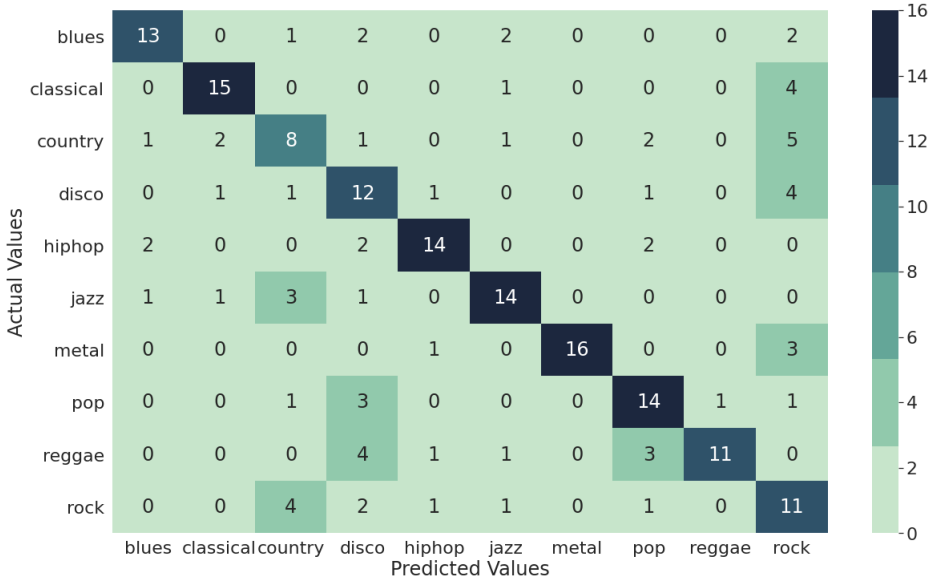
In figure 13a, the best CNN Model correctly predicts 15 Blues, 17 Classical, 12 Country, 9 Disco, 11 Hiphop, 10 Jazz, 16 Metal, 14 Pop, 11 Reggae and 4 Rock songs in the test set.

Additionally, figure 13b shows that the best CRNN Model correctly classifies 7 Blues, 19 Classical, 9 Country, 14 Disco, 16 Hiphop, 13 Jazz, 18 Metal, 15 Pop, 7 Reggae and 11 Rock songs in the test set.

C Appendix C: Activation Function Parameter Tuning Confusion Matrices



(a) CNN Model Confusion Matrix



(b) CRNN Model Confusion Matrix

Figure 14: Activation Function Parameter Tuning Confusion Matrices

Figure 14a provides a clear description of the predictions the model makes. In the Activation Function Parameter Tuning experiment, the best CNN model classifies 4 Blues, 18 Classical, 5 Country, 7 Disco, 11 Hiphop, 8 Jazz, 16 Metal, 9 Pop, 8 Reggae and 10 Rock songs correctly in the test set. Additionally, figure 14b, displays the confusion matrix of the best CRNN model in the Activation Function Parameter Tuning experiment. the CRNN model classifies 13 Blues, 15 Classical, 8 Country, 12 Disco, 14 Hiphop, Jazz & Pop, 16 Metal, 11 Reggae and 11 Rock songs correctly in the test set.