Comparison of Deep Learning and Machine Learning in Music Genre Categorization

MSc Research Project
Data Analytics

Saviour Nickolas Derel Joseph Fernandez
Student ID: 21127051

School of Computing
National College of Ireland

Supervisor: Mr Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Saviour Nickolas Derel Joseph Fernandez
Student ID:	21127051
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Mr Vladimir Milosavljevic
Submission Due Date:	01/02/2023
Project Title:	Comparison of Deep Learning and Machine Learning in Music Genre Categorization
Word Count:	4941
Page Count:	17

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:	Saviour Nickolas Derel
Date:	30th January 2023

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):	

Comparison of Deep Learning and Machine Learning in Music Genre Categorization

Saviour Nickolas Derel Joseph Fernandez
21127051

Abstract

Humans invented the concept of musical genres to categorize and describe different types of music. The extensive music libraries that are accessible on the internet are frequently organized using genre hierarchies. Music information retrieval systems would benefit greatly from the addition of automatic musical genre categorization, which may supplement or even take the place of the human user in this process where the genres are categorized manually sometimes. There have been variable degrees of success with different music collections, data formats, learning algorithms, and types of neural networks applied. Music indexing and information retrieval both benefit greatly from automatic musical genre classification. The automated classification of musical genres using a combination of machine learning and deep learning algorithms is provided in this study as an effective and efficient approach. The mp3-formatted audio tracks from the FMA dataset are processed for feature extraction using the Mel-spectrogram method for feature retrieval using the 'librosa' library. Then the array of converted data from the Mel-spectrogram is used by various classification models. Eight machine-learning includes Naïve Bayes, KNN, Random Forest, XG-Boost classifier, SVM, etc, are used along with the CNN and CNN-LSTM neural network models. On final evaluation, the CNN neural classifier performed better in terms of train model accuracy 0.92 and f1 score than any other classification models and this includes a few parameters tuned for better classification results.

1 Introduction

The art of music is the development of aesthetically pleasing or expressive tonal combinations, usually with melody, rhythm, and harmony. We cannot think without audio, which includes music and other sounds. It is also critical to instantly identify the audio signal, and various musical genres evoke the interests of different listeners. Owners of music services are aware that they must regularly provide their clients with more specialized and advanced digital solutions in order to compete in today's competitive market, where sustainability is difficult for even the most successful businesses.

Many companies have taken advantage of this moment to grow their client bases by offering a variety of user-attractive features with Melomaniacs rising around the world. Users are encouraged to utilize their items while on the road by offering a variety of music albums and collections. Automated customizing of music playlists and the availability of a method for switching between them based on particular traits and measurements will surely create major improvements in the knowledge and customization of music services, potentially even increasing consumer happiness. As days progressed, different types of music genres are increasing day-by-day hence the classification of musical libraries must be updated on a regular basis according to their genres respectively which makes the audio-related data enormous daily.

According to Nopthaisong and MarufHasan (2007), due to the administration of music databases through various applications, Music Genre Classification (MGC) is still a hot issue in Music Information Retrieval (MIR). Since different genres have a range of properties, music genre classification helps to categorize the various genres according to their traits by extracting features from the raw data. These retrieved attributes were employed in more recent times by machine learning or deep learning algorithms to categorize the songs based on their traits. The ability to categorize songs would be extremely helpful for music lovers as it would allow them to form playlists of their favorite songs and assist music streaming services in recommending songs to users based on the genres of songs they like. By Pelchat and Gelowitz (2019), Neural network (NN) is a technique of machine learning that is typically successful in extracting critical features from large collections of data and deriving a function or model that reflects those features. In recent years, machine learning has gained a lot of popularity. Depending on the source of data, certain types of machine learning techniques are more appropriate than others for different applications.

Hence in this research, eight machine-learning models including Support Vector Machine(SVM), K-Nearest Neighbors(KNN), Random Forest, XG-Boost, Decision Tree, Naive Bayes, Stochastic Gradient Descent (SGD), Logistic Regression along with deep learning models which include CNN and CNN-LSTM models are being used for MGC. All the models are built according to this research with few parameters being tuned to have a better outcome. For this analysis, a dataset from the Free Music Archive (FMA), which includes mp3 audio recordings and CSV-formatted metadata, was used. Feature extraction is carried out with Mel-spectrogram images and those images are converted to a Numpy array and passed through various ML and DL models for genre classification of each audio track. All these discussed methods are elaborated in Section 3.

1.1 Research Question

- What feature extraction technique can be implemented to extract from the audio file?

- What is the optimal model for music genre classification through extracted features?

1.2 Proposed Solution

- Features are extracted from the audio file using Mel-spectrogram.

- XG-Boost, SVM, KNN, Decision Tree, etc. along with deep learning models like CNN and CNN-LSTM are used for MGC.

2 Related Work

Automatic genre classification through MIR is a most interesting topic in the last decade. Various researchers have followed different techniques and methodologies to achieve their objectives in music classification. Through obtaining this genre classification, stakeholders like Spotify, YouTube, etc, were able to attract more users with the genre-specific musical libraries. From this section, different authors' research will be gathered and relevant information will be used for our research purpose.

2.1 Dataset

There are various datasets related to audio tracks are available over the internet. The GTZAN dataset, which contains 1000 audio files that are each 30 seconds long, is one of the most well-known datasets among academics. MSD dataset has more than 1,000,000 audio tracks and 44,000 plus artists, and Audio-set has more than 2,084,320 mp3 formatted tracks but both datasets have copyright issues. Hence in this study, we will use FMA[1] dataset. According to Defferrard et al. (2016), the dataset has 106,574 audio tracks along with metadata. This dataset is publically available for research purposes hence no copyright permission is needed from the stakeholders. For the computational purpose, fma_small dataset with 8,000 audio tracks will be used from the total audio tracks

2.2 Extracting Features

This section discusses many significant feature extraction methods that were applied in earlier studies. The feature extraction methods used in audio signal processing may be divided into many categories. The frequency and time domains are employed by one of them, digital signal processing, while statistical descriptors are another widely used method. In this project by Atahan et al. (2021), various feature extractions are done using 'librosa' library. Zero Crossing Rate, the signal's frequency value is shown by the spectral centroid, When comparing the peak energy with the valley energy to determine spectral contrast, Noise is indicated by Spectral Flatness. The amount of power transmission, or spectral bandwidth the total energy of the spectrum is determined by spectral rolloff, Chroma STFT, RMSE, MFCC, and a display of the energy signals for 12 distinct pitches, Tonnetz displays the harmonic relationship and Polynomial Features, whereas time to domain frequency transition is offered by Chroma CQT (comparable to FFT). Also, Shah et al. (2022) has done feature extraction is done using Spectral roll-off, Spectral contrast, Zero-Crossing, and Spectral centroid.

In this project report, Modulation Spectrograms (MS) and Harmonic Percussive Source Spectrograms (HPSS) are two novel feature extraction techniques Bakhtyari et al. (2022) introduced. To create the MS, the Fast Fourier Transform (FFT) is applied to the Mel-Spectrogram. In addition, MS finds recurrent patterns in audio signals and transforms the time-to-frequency domain into the frequency-to-frequency domain to distinguish the imaginary portion from the real part before producing the final dimension. HPSS uses the Short Time Fourier Transform (STFT) to separate sound into harmonic and percussive sounds, where the horizontal structure is made by the harmonic sounds which correspond to a certain pitch and the vertical structure is made through the percussive sounds that record the noises made when objects contact.

The Continuous Wavelet Transform (CWT) is used by Xu et al. (2021) to accomplish data preprocessing since he transformed wave impulses into digital signals. The signals

[1] FMA Dataset: `https://github.com/mdeff/fma`

are modified simultaneously in the frequency and time domains during feature extraction, however, Gabor's Short Time Fourier Transformation (STFT) could only provide restricted and fixed-frequency resolution when the frequency varied over time. Due to its advantage of extracting the feature even when the frequency varies over time, the CWT technique is used in this case. Whereas Senac et al. (2017) has created an analysis window for musical studies with a length of 46.44 ms (22050 Hz in 1024 points) and has done several feature extractions. Three major features, namely the Tonality, Dynamics, and Timbre features are employed out of the total eight features. Short-Term Energy signals are used by dynamics characteristics for MGC. To retrieve music information, timbre characteristics employ Zero-Crossing Rate (ZCR), and statistical moments were used to define the spectral distribution. The spectral flatness, which may be thought of as the quantity of information contained in the spectrum, and the spectral Shannon entropy both point to the smoothness or spikiness of the spectrum.
Here the author Han et al. (2018) used MFCC for feature extraction, where the content is quantified. Only 30 seconds of music are used in this approach to intercept and extract speech characteristics. The high-frequency component is subsequently magnified using the voice signal of the pre-emphasis filter differential in order to compensate for the high-frequency section's suppression, emphasis, and high-frequency resonance peak. Chiliguano and Fazekas (2016) converted a sample rate of 22,050 Hz from three seconds of audio clips into a mono channel. From 1,024 sample frames for each segment, a spectrogram driven by a mel-scale with 128 bands is produced. Even here the author used the Mel-spectrogram feature extraction technique before building the models.
Here Mendes (2020) used Fast Fourier Transform (FFT) to compute the frequency domain after the audio files have been pre-processed by conversion to Mel-Spectrogram. The spectrogram is created by dividing the signal's magnitude after the Mel scale is generated when the frequency spectrum contains 128 frequencies. Also, Choi et al. (2016) used various methods of feature extractions like STFT, Mel-spectrogram, and MFCC. But it is clearly evident that Mel's features gave a better outcome and performance on comparing MFCC and STFT.
Based on the above results and performance mel-spectrogram has the ability to acquire more features from the audio files, hence mel feature extraction will be carried out in this research.

2.3 Music Genre Classification - Machine Learning

The author Qi et al. (2022) extracted various features from two different datasets namely, GTZAN and Spotify. The parameters for various machine learning models, such as Decision Trees, Random Forest, and KNN, are being adjusted via grid search. Different models outperformed each other for both audio tracks; for the Spotify dataset, KNN outperformed RF with a model accuracy of 90%. Here by the author Murauer and Specht (2018), 30 seconds long audio segments in 25,000 MP3 files from sixteen different genres were utilized to train the model. Two distinct sets of information are retrieved for the purpose of predicting the musical genre using two distinct methodologies. Three models, ExtraTrees, a random forest variant, XGBoost, which employs extreme gradient boosting, and Deep Neural Network (DNN), were employed for the numerical features. Unexpectedly, the XGBoost model with 1,000 n-estimators and 3 maximum depths beat other models with the loss rate of L=0.82.

The author Bhatia et al. (2021) used MFCC feature extraction and the calculations are made using the Mel scale, which simulates human pitch variation. The first 15 of the 20 waves are preserved in this case due to the high-frequency details, and the 15 cepstral wave vectors serve as frame-independent parameters in matrices for cepstral properties that represent the music. With 80.63% model accuracy, KNN is built and compared with the other models built in various studies of music genre classification. In addition, acoustic music characteristics are retrieved and sent to the KNN algorithm in this research by Prashanthi et al. (2021), which results in a model with a 76% accuracy.

In this research by Chen and Steven (2021), a newly proposed ATMGCM model which is a combination of Random forest and SVM is used for the classification. Cross-entropy loss is the evaluation approach used in this scenario, which calculates the total entropy between the error matrix and the probability distribution to improve the performance of the model. Here the model outperformed Resnet34 with 95.4% model accuracy. Also, from the research of Atahan et al. (2021) SVM and LDA model was built to find the genre classification. With various features being extracted, SVM performed better in terms of classification accuracy at 81.9%. From the above observation, different machine learning models have behaved differently for different types of extracted features. Hence in this study, various types of machine learning models are built to compare the model performance.

2.4 Music Genre Classification - Deep Learning

In this research, Choi et al. (2016) proposed an automatic music labeling technique called CB using fully convolutional neural networks (FCN). Different models with sub-sampling layers and 2D convolutional layers have been assessed. Presented was a 4-layer FCN topology with 2 max-pooling and 4 convolutional layers. It was shown that deeper networks with more layers outperformed the 4-layer design, and the deeper network profited from more training data. Chang et al. (2018) classified the audio files into different genres using the CNN method based on the rhythms in the audio stream. This model, which makes use of ReLU and the MaxPool activation function, converts the audio signals into bins of 128 frequency and 599 frames in Mel-spectrograms. Sainath et al. (2015) has combined Deep Neural Networks (DNN) with LSTM and CNN into a single model and evaluated with traditional LSTM model, here the newly proposed model has produced 6% higher model accuracy than the LSTM.

The author Tao et al. (2019) proposed a modified LSTM model to examine the integration of both the customer and the music utilizing the sequential data and temporal context. The Adam optimizer updates the learning rate of the model to improve accuracy. The author Fulzele et al. (2018) offers a hybrid classification method that improved accuracy by combining LSTM and SVM. Its accuracy, at 89%, was much greater than any of their individual accuracy rates. Irene et al. (2019) employs RNN for sequence modeling and to learn the audio descriptions CNN was used. The vanishing gradient-related issues and the bursting gradient phenomenon that are present in RNNs are satisfactorily addressed by LSTM.

To sum up the above findings, for this study CNN and CNN-LSTM models will be built for the classification of genres.

3 Methodology

Understanding the study's objectives is covered in the first stage of the investigation, which involves converting the gathered data into a machine-learning problem. This study aim to classify various music in accordance with their respective genres, which makes it easier for the users to listen to the songs of their interest through the classification approaches discussed above in section 2.

Here the Knowledge Discovery in Databases (KDD) approach as described by Peng et al. (2009) will be used in this study to extract significant information from large raw data and make the most of it. The approaches used in this research are depicted in Figure 1.

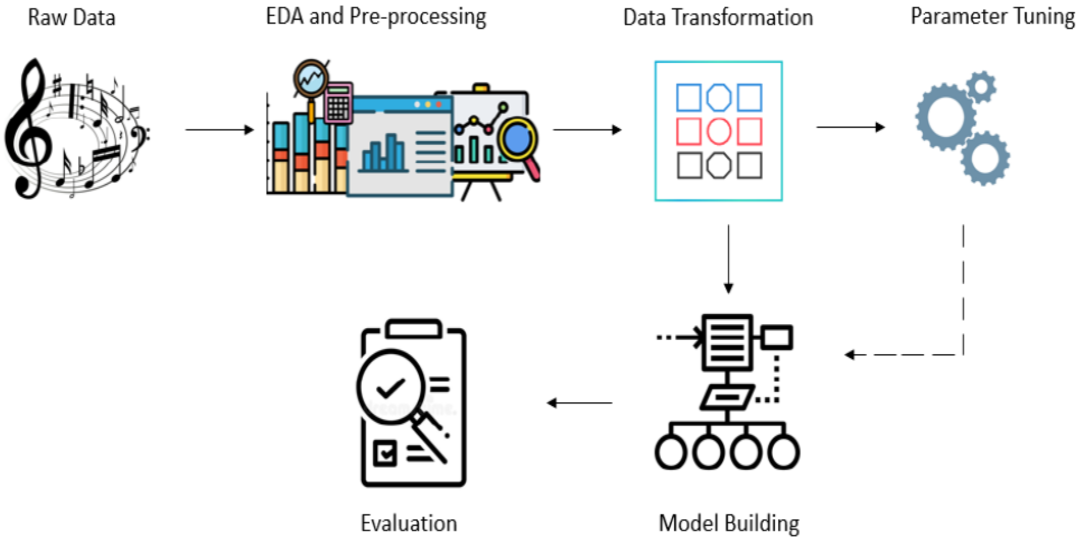

Figure 1: KDD Methodology

3.1 Dataset

For this research, FMA dataset which is an open dataset easily available on Github is been used. This dataset consists of features CSV, tracks CSV, genre CSV, also with 106,574 audio tracks. All the tracks available in this dataset are 30 seconds long with high-quality features with metadata information. Here for this study fma_small, audio data is been used which consists of 8,000 tracks with an equal split across various genres.

3.2 Datas Pre-processing

Here the subset of 8,000 audio tracks is taken from the total numbers and data cleaning is been done, 8 different genres with 1000 tracks each are been taken to have the balance between the genres. A minimum threshold of 28 seconds is fixed for all the audio tracks, if the length of the audio is less than the threshold value then those tracks are been removed. The below Figure 2 shows the equal split of audio tracks across various genres.

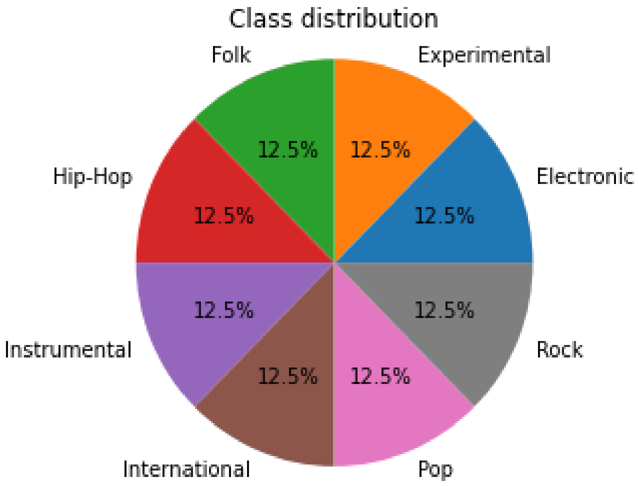

Figure 2: Genres Classification

As discussed in section 2, Mel-spectrogram has the ability to give a better feature when comparing the rest of the feature extraction techniques. With Fast Fourier Transform (FFT), the time-frequency domain is carried out for 512 hop size and 2048 window size of 22,050 sample rate. Finally, the spectrogram is generated with the Mel scales. The below Figure 3 represents the Mel-spectrogram for the folk audio track.

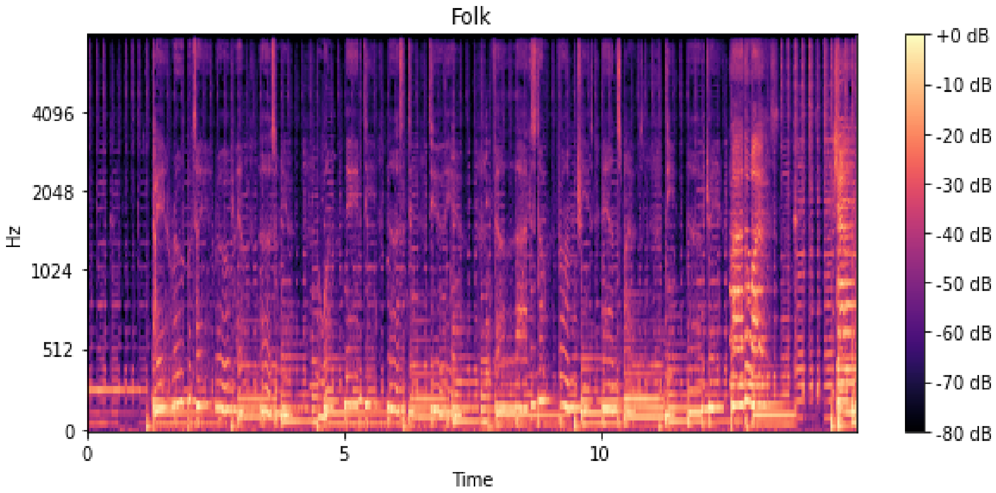

Figure 3: Genres Classification

3.3 Classification Models

From section 2 it is observed that various machine learning models have performed differently with their own boon and burdens. As a result, the model performances will be assessed using eight distinct machine learning models. The several machine learning algorithms that were employed in this study are described in the sections that follow.

- Naive Bayes
  - The Naive Bayes Gaussian classification model which assumes a normal distribution for the features is used.
  - If continuous values are utilized, it shows that the model anticipates that the predictor values will be samples from the Gaussian distribution.
- Stochastic Gradient Descent
  - By merging several binary classifiers in a ”one versus one” (OVO) approach, it provides multi-class classification.
  - A binary classifier is trained for each of the K classes that can distinguish it from the other K-1 classes.
  - Additionally, the fit parameters function allows both weighted classes and weighted instances.
- KNN
  - As discussed by Qi et al. (2022) Prashanthi et al. (2021), With simple math measuring the separation between points on a graph, KNN conveys the concept of similarity.
  - In Euclidean space, the two points are determined by the distance of the straight line.
- Decision Tree
  - The ID3 algorithm is the fundamental method utilized in decision trees.
  - To show every potential result for a certain input, it employs a branching technique.
  - The top-down, greedy construction of decision trees is characteristic of the ID3 algorithm.
- Random Forest
  - Combination of several decision trees with the goal of producing an uncorrelated forest of trees whose forecast by committee is more accurate than that of any one individual tree.
  - Each individual tree is built using bagging and feature randomization.
- Support Vector Machine
  - As per the research by Zhuang et al. (2020), a subset of the target points is used in the decision function in SVM, which performs well in high-dimensional domains.
  - In an N-dimensional space, the hyperplane that best classifies the data points is to be found using the SVM method.
- Logistic Regression

  - In place of using OLS to fit the model and calculate the coefficients, logistic regression iteratively fits the model using the approach of maximum likelihood.
  - Logistic regression is a statistical technique used to predict the connection between predictors and a predicted variable.

- XGB Cross Gradient Booster
  - As mentioned by Murauer and Specht (2018), a single machine performs concurrent calculations using a linear model and a tree-learning technique.
  - Additionally, it incorporates enhanced regularization, which enhances the model's generalization skills, and additional features for performing cross-validation and determining the relevance of features.

- Convolutional Neural Network
  - As used by Chang et al. (2018), CNN uses convolutional, pooling, and fully connected layers.
  - The sizes of the features are down-sampled to accelerate computation, and dropout layers are utilized to prevent overfitting.
  - Kernel size of 2x2 RelU and sigmoid activation functions are used.

- CNN-LSTM
  - As discussed by Mendes (2020) Fulzele et al. (2018), CNN with additional two layers for the LSTM model is been used.
  - The dense layers receive the output of the LSTM layers and use it to create a higher-level feature representation that can be readily divided into other classes as necessary.

4 Design Specification

The complete functioning of the music genre classification technique and methodology is explained in the Figure 4. The design architecture will have two layers. Initially, the raw data which is in CSV and audio track data is collected and fed into python to initiate the further process of music genre classification. Data cleaning and pre-processing are done on the data to extract the features through Mel-spectrogram and converted into an array of data. Then the processed data is passed into various machine learning and deep learning models to have a final look at the model's accuracy. Sklearn, which has various functions to build machine learning is used along with Tensor and Keras for deep learning models. The model with lesser loss and higher accuracy is selected for the music genre classification in order to have a better-classified genre of music, which can be used by the application stack-holders to have segregated libraries of music for different users.

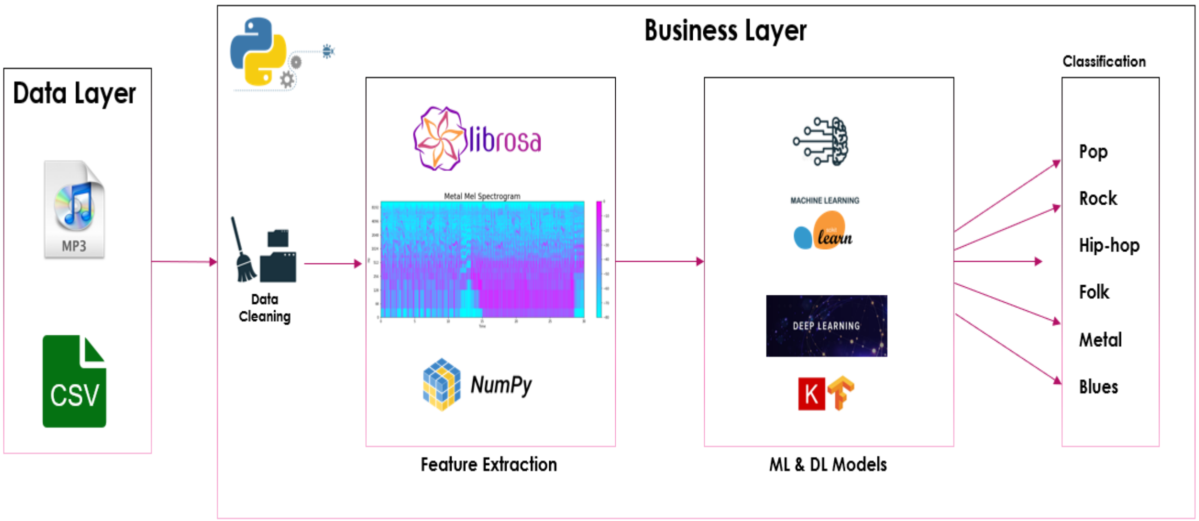

Figure 4: Design Architecture

5 Implementation

The process of creating a music genre categorization is explained in this section. The above-discussed methodologies from section 3 are implemented and discussed in this section.

- Data Cleaning
  - The FMA small dataset with 8,000 mp3-formatted audio tracks 30 seconds long is been used, along with the track's CSV file.
  - Data cleaning is done as per the discussed methods from the section 3.2
- Feature Extraction
  - Mel-spectrogram is used for feature extraction, as described in section 2.2.
  - All the tracks from the FMA small dataset are converted to mel-spectrogram images to obtain the Numpy array.
  - The temporary memory is been cleared after each Numpy array conversion to have the best usage.
- Data Split
  - The converted arrays have a split of 15% for the test and the remaining 85% for the train.
  - In order to assess the trained deep learning and machine learning models, a test dataset is used at the conclusion.
  - From the 85% of the training dataset, 15% of the array is allocated for validation.
  - To sum up, the total dataset has a split of 70% train, 15% test, and 15% validation.

5.1 Machine Learning Classification Models

As discussed in section 3.3, different types of machine learning models were built to evaluate the classification strength. The largest number of iterations in the training set for the Stochastic Gradient Descent model is 5000 max iter, and the random state for rearranging the data is set to 0. KNeighborsClassifier was set with 19 n_neighbors which tells the number of neighbors required for each sample. Then the Random Forest was set with 1000 n_estimators, the max deep of the trees was considered as 10 along with the randomness of bootstrapping was set to random_state to 0. The Support Vector Machine was built with "ovo" decision_function_shape, here the ovo represents the one vs one which is always used as a multi-class strategy to train the model. For the Logistic Regression, the random_state was considered as 0, then 'lbfgs' solver is used which handles the multinomial loss along with multi_class is chosen as 'multinomial' where the multinomial loss fit across the full probability distribution is the loss that is minimized. The XG-Boost classifier is set with 1000 n_estimators along with a 0.05 learning rate. Last but not least, the classification is performed using the Decision Tree and Naive Bayes models, both of which have their default settings for their parameters.

5.2 Deep Learning Classification Models

5.2.1 Convolutional Neural Network

The CNN has the input shape of (128, 128, 1). A 2D kernel is added to the convolutional layers with the express purpose of extracting certain features from the input. Batch normalization layers are also included in the layers for scaling and normalizing. AveragePooling2D layers are utilized for downsizing, together with four dense layers with a RelU activation function and the fifth dense layer with a softmax activation function. To control the overfitting, the dropout layer with 0.5 and 0.2 rates are included. The below Figure 5 represents the CNN structure.

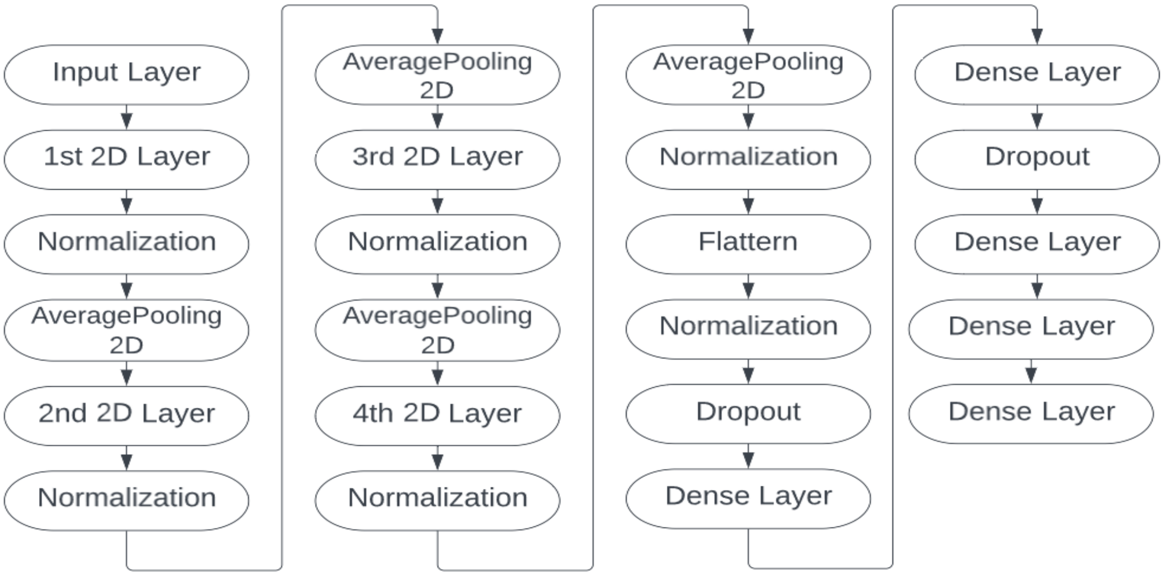

Figure 5: CNN Structure

5.2.2 Convolutional Neural Network - LSTM

With the exception of the inclusion of two LSTM layers for better model performance, the CNN-LSTM model is constructed very similarly to the CNN 5.2.1 in this case. The first layer is bi-directional, allowing the neural network to receive sequential input in both ways. The below Figure 6 represents the CNN structure.

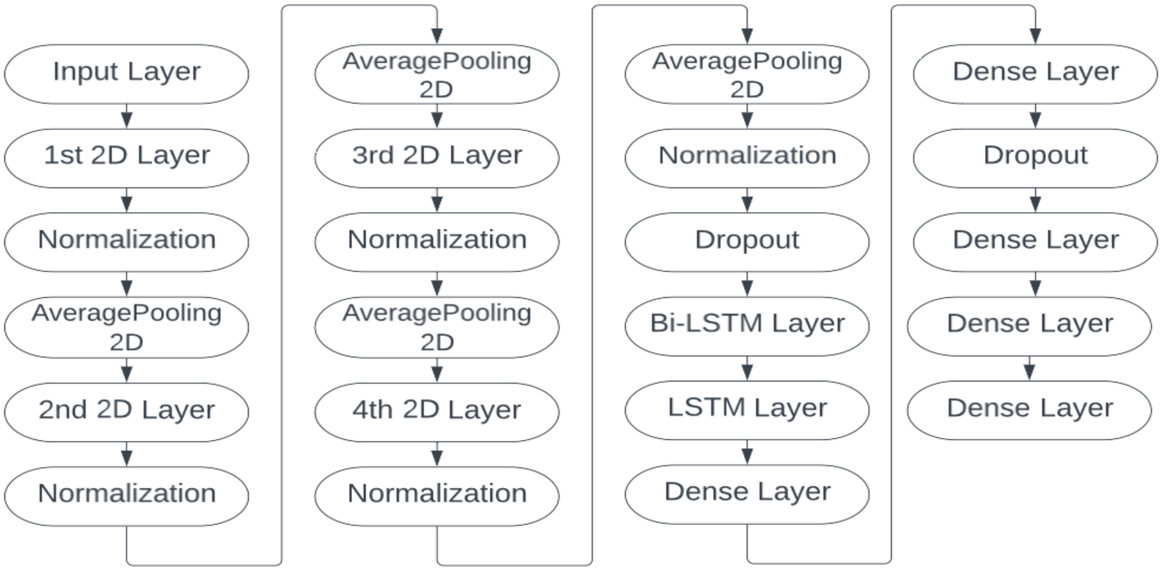

Figure 6: CNN-LSTM Structure

5.3 Parameter Tuning

In developing the machine-learning and deep-learning models, their parameters are tuned accordingly to the research to acquire optimal solutions through genre classification. In the machine-learning model, the parameter for the Support Vector Machine is set to 'ovo' as it supports multi-class classification efficiently, then the XGB classifier learning rate is set to 0.05 which is an ideal rate for the classifier to give a better outcome. For the deep-learning models, dropout layers are added to the neural network architecture to prevent the overfitting of the model. The Adam optimizer which supports multi-class problems is used with a 0.00004 learning rate and 28 epochs.

6 Evaluation

In this section, the evaluation of various machine-learning models is carried out with classification accuracy along with precision score, recall score, loss, and F1 score for CNN and CNN-LSTM models.

6.1 ML Classification models

Here different machine-learning models are validated according to their model accuracy and f1-score. The below Table 1 represents the model accuracy of each model from the test dataset. It is clear that the XG-Boost classifier has performed better on comparing the other models with 0.66 model accuracy. Also, for the XGB classifier, the f1- score resulted in 0.66 accuracies.

Table 1: ML Model Accuracy

Naive Bayes	SGD	KNN	DT	RF	SVM	LR	XGB
0.31	0.51	0.45	0.39	0.56	0.62	0.54	0.66

The below correlation Figure 7 represents the classification of genres as per the XBG classifier. It is seen that the folk genres have been classified more accurately followed by hip-hop, instrumental, etc. Also, pop genres have been misclassified at a higher rate than all the other genres.

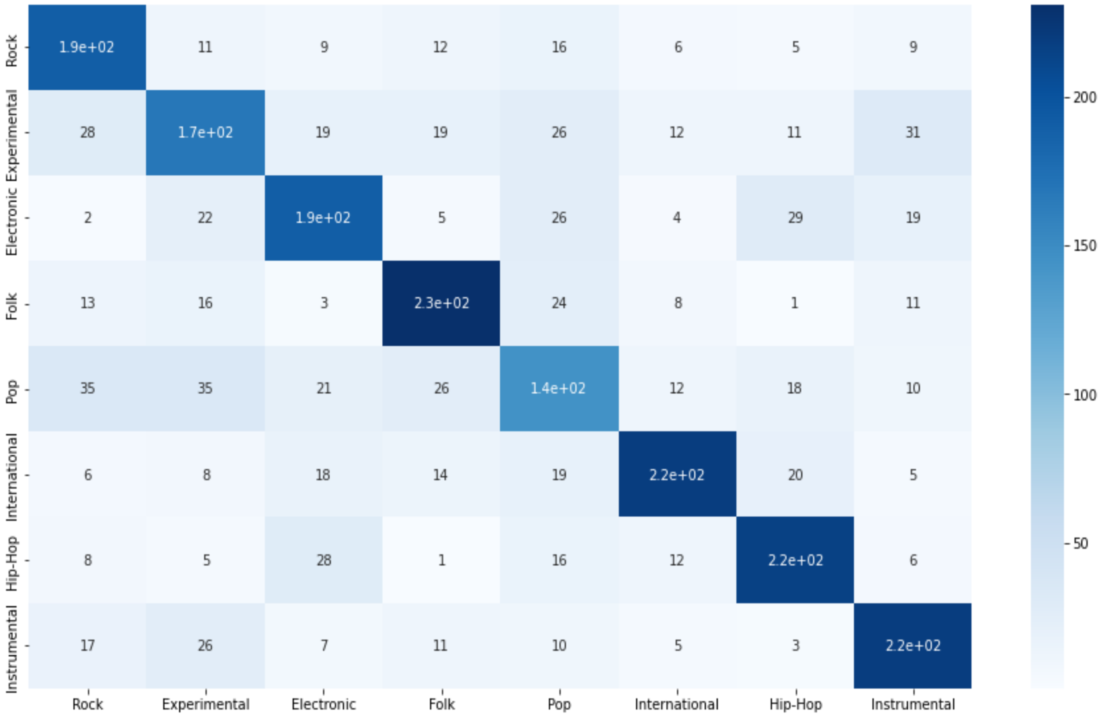

Figure 7: Multiclass Correlation Matrix

6.2 DL Classification models

Here for the neural models, the evaluation is done based on precision, f1-score, loss, accuracy, and recall. The below Figure 8 shows the f1-score and accuracy produced by the Convolutional Neural Network model. It is seen that the training accuracy creeps towards 90% with the increase in the number of epochs also for the validation dataset the model accuracy fluctuates around 60% after the 10th epoch. Meanwhile, on the other hand, the f1- score for the training dataset works well, and for the validation data, it ranges between 65% and 70%. In both cases, the mode performance looks to be consistent for both the train and validation datasets.

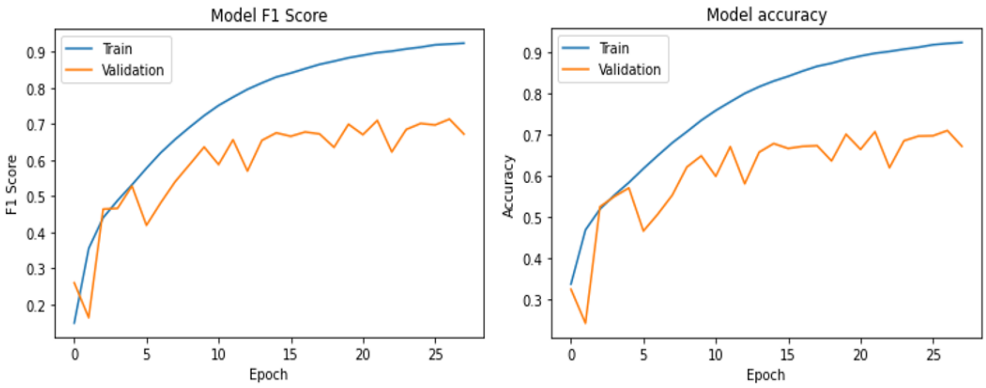

Figure 8: CNN - Accuracy and F1 Score

The below Figure 9, shows the f1-score and accuracy for the CNN-LSTM model. The training and validation datasets' accuracy and f1 scores appear to be virtually identical to those of the CNN 6.2. Training accuracy has increased with increased epochs, and validation accuracy has moved around 60% to 70% after the 10th epochs.

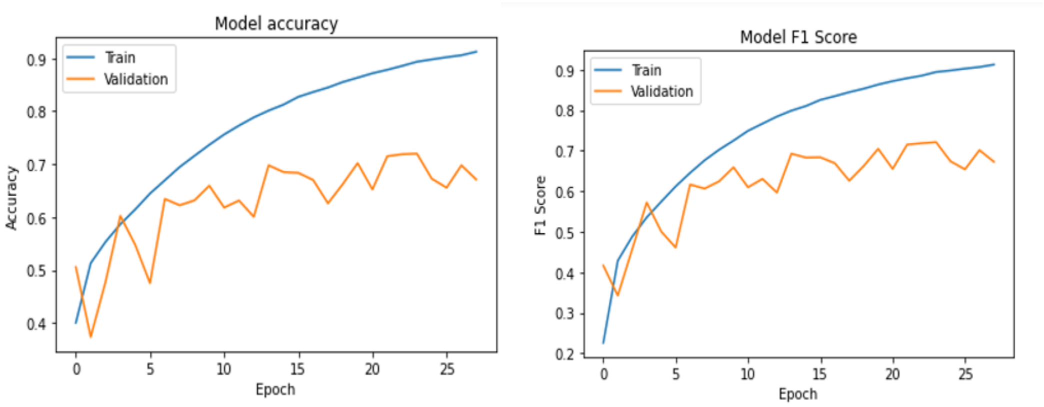

Figure 9: CNN-LSTM - Accuracy and F1 Score

The below Table 2 represents the test data score for all the metrics of both CNN and CNN-LSTM models. It is seen that CNN and CNN-LSTM have similar accuracy and f1-score, which means that both models have similar characteristics in genre classification. But CNN has a lesser logarithmic loss rate of 1.45 than the CNN-LSTM which has a 1.50 loss rate, it is understood that with lesser logarithmic loss the model performance and accuracy will be increased. Hence keeping this fact, CNN is considered to be the best model among the two neural network models.

Table 2: DL Model Accuracy

Models	Accuracy	Loss	Precision	Recall	F1
CNN	0.67	1.45	0.70	0.65	0.67
CNN-LSTM	0.67	1.50	0.69	0.65	0.67

6.3 Discussion

In Section 2.3, the author Shah et al. (2022) used multiple features and built the XG-Boost model to achieve a model accuracy of 0.71 also Murauer and Specht (2018) has a lesser log loss of 0.82 model performance for XG-Boost. But in this research, XG-Boost has a train and test data accuracy of 0.66 with Mel-spectrogram features being extracted. As discussed by Atahan et al. (2021), SVM has performed with 0.82 model accuracy with MFCC features, and here in this study SVM has performed with 0.62 test data accuracy. It is evident that feature extraction has played a significant role in model performance, where both the top-performing model from this research XG-Boost and SVM have performed less. According to Fulzele et al. (2018), SVM combined with LSTM has achieved 0.89 model accuracy, in this research the LSTM model has 0.91 train accuracy and 0.67 test accuracy. The model performance could have been more with an increased learning rate.

On looking at the overall results from all the machine-learning and deep-learning models from this research, the test model accuracy is 0.66 for the XG-Boost classifier and 0.67 for both the CNN and CNN-LSTM. From this, it can be seen that the accuracy of the neural network models has outperformed that of the machine-learning models. And out of both the neural models, CNN takes a slight edge up over the CNN-LSTM model with a lesser logarithmic loss difference of 0.05. Even with additional LSTM layers, CNN-LSTM has performed similarly to the CNN model, although the model performance could have been higher with increased epochs and learning rate.

7 Conclusion and Future Work

The music business has changed significantly over the past ten years in many areas, and new musical genres have emerged with the introduction of several new instruments and musicians from diverse cultures throughout the globe. In parallel space, many machine-learning models have also been introduced by infusing novel techniques with iterative research. In this research, a public audio dataset called the FMA dataset is used and the Mel-spectrogram feature extraction technique is been implemented to study the information of each mp3 audio track. Two neural network models including CNN-LSTM and CNN, along with eight different machine-learning are used for the music genre classification. The test data model performance went at a phase of 0.61 for SVM, 0.66 for XG-Boost, and 0.67 for both the CNN and CNN-LSTM models. But with a difference of 0.05 lesser logarithmic loss rate, CNN overtook the CNN-LSTM model and came on top. But it is seen that both neural models have performed neck-to-neck and with higher computational power, LSTM could have come on top over CNN if it had a higher learning rate and epochs rate.

For future work, data augmentation techniques can be applied to the Mel-spectrogram images. By creating additional data points from existing data, data augmentation artificially boosts the volume of data. Also, it has the benefit of reducing over-fitting, preventing scarcity of data, and more than that it can boost the model performance. By replacing the traditional CNN model with VGG-16 and Residual Network (ResNet) models, it can produce higher model accuracy.

8 Acknowledgment

My sincere thanks to Prof. Vladimir Milosavljevic for giving me the knowledge and independence to work on my thesis through his guidance over the past weeks to complete this thesis successfully.

References

Atahan, Y., Elbir, A., Keskin, A. E., Kiraz, O., Kirval, B. and Aydin, N. (2021). Music genre classification using acoustic features and autoencoders, Institute of Electrical and Electronics Engineers Inc.

Bakhtyari, M., Davoudi, S. and Mirzaei, S. (2022). Evaluating various feature extraction methods and classification algorithms for music genres classification, Institute of Electrical and Electronics Engineers Inc.

Bhatia, J. K., Singh, R. D. and Kumar, S. (2021). Music genre classification, Institute of Electrical and Electronics Engineers Inc.

Chang, S.-H., Abdul, A., Chen, J. and Liao, H.-Y. (2018). A personalized music recommendation system using convolutional neural networks approach, *2018 IEEE International Conference on Applied System Invention (ICASI)*, pp. 47–49.

Chen, C. and Steven, X. (2021). Combined transfer and active learning for high accuracy music genre classification method, Institute of Electrical and Electronics Engineers Inc., pp. 53–56.

Chiliguano, P. and Fazekas, G. (2016). Hybrid music recommender using content-based and social information, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2618–2622.

Choi, K., Fazekas, G. and Sandler, M. (2016). Automatic tagging using deep convolutional neural networks.
URL: *https://arxiv.org/abs/1606.00298*

Defferrard, M., Benzi, K., Vandergheynst, P. and Bresson, X. (2016). Fma: A dataset for music analysis.
URL: *https://arxiv.org/abs/1612.01840*

Fulzele, P., Singh, R., Kaushik, N. and Pandey, K. (2018). A hybrid model for music genre classification using lstm and svm, *2018 Eleventh International Conference on Contemporary Computing (IC3)*, pp. 1–3.

Han, H., Luo, X., Yang, T. and Shi, Y. (2018). Music recommendation based on feature similarity, *2018 IEEE International Conference of Safety Produce Informatization (IICSPI)*, pp. 650–654.

Irene, R. T., Borrelli, C., Zanoni, M., Buccoli, M. and Sarti, A. (2019). Automatic playlist generation using convolutional neural networks and recurrent neural networks, *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5.

Mendes, J. (2020). *Deep learning techniques for music genre classification and building a music recommendation system*, Master's thesis, Dublin, National College of Ireland. **URL:** *https://norma.ncirl.ie/4455/*

Murauer, B. and Specht, G. (2018). Detecting music genre using extreme gradient boosting, Vol. 2018-January, Association for Computing Machinery, pp. 1923–1927.

Nopthaisong, C. and MarufHasan, M. (2007). Automatic music classification and retreival: Experiments with thai music collection.

Pelchat, N. and Gelowitz, C. M. (2019). Neural network music genre classification, *2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE)*, pp. 1–4.

Peng, Z., Yang, B. and Ren, H. (2009). Research on kdd process model and an improved algorithm, pp. 113–115.

Prashanthi, V., Kanakala, S., Akila, V. and Harshavardhan, A. (2021). Music genre categorization using machine learning algorithms, Institute of Electrical and Electronics Engineers Inc.

Qi, Z., Rahouti, M., Jasim, M. A. and Siasi, N. (2022). Music genre classification and feature comparison using ml, Association for Computing Machinery, pp. 42–50.

Sainath, T. N., Vinyals, O., Senior, A. and Sak, H. (2015). Convolutional, long short-term memory, fully connected deep neural networks, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584.

Senac, C., Pellegrini, T., Mouret, F. and Pinquier, J. (2017). Music feature maps with convolutional neural networks for music genre classification, Vol. Part F130150, Association for Computing Machinery.

Shah, M., Pujara, N., Mangaroliya, K., Gohil, L., Vyas, T. and Degadwala, S. (2022). Music genre classification using deep learning, Institute of Electrical and Electronics Engineers Inc., pp. 974–978.

Tao, Y., Zhang, Y. and Bian, K. (2019). Attentive context-aware music recommendation, *2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC)*, pp. 54–61.

Xu, K., Alif, M. A. and He, G. (2021). A novel music genre classification algorithm based on continuous wavelet transform and convolution neural network, Association for Computing Machinery, pp. 1269–1273.

Zhuang, Y., Chen, Y. and Zheng, J. (2020). Music genre classification with transformer classifier, Association for Computing Machinery, pp. 155–159.