Musical Genre Classification Using Neural Networks

Rikky L. Francis
Dept. of Comp. Sci. and Eng.
The University of Texas at Arlington
Arlington, TX 76019
rikkylinuz.francis@mavs.uta.edu

Mukka Himaneesh
Dept. of Comp. Sci. and Eng.
The University of Texas at Arlington
Arlington, TX 76019
hxm1524@mavs.uta.edu

Samuel Smitherman
Dept. of Comp. Sci. and Eng.
The University of Texas at Arlington
Arlington, TX 76019
samuel.smitherman@mavs.uta.edu

Abstract

Music genre classification is an interesting as well as challenging problem with large scale music data available today. In this paper we convert audio files into two representations: Mel-frequency cepstral coefficients (MFCCs) and a visual representation Mel-spectrograms. We compared results of different deep learning models on classifying the genre on both representations using public datasets like GTZAN and Free Music Archive (FMA). We initially extracted single feature segment on each audio file for both representations. We then improved the accuracy around 20% by extracting five segments on each audio file on feed-forward and RNN with LSTM models. There are improvements with five segment approach on the other deep learning models used in this paper as well. We were able to achieve accuracy of around 70% on GTZAN dataset and 50% accuracy on the FMA dataset.

1 Introduction

Artificial Neural Networks (ANNs) can be utilized for the purpose of training computers in the task of classification [1]. Particularly for this paper, the classification of music into defined genres will be explored. To do this, four ANN models will be reviewed and compared. This will include a feed-forward neural network, a long-short term memory (LSTM) based recurrent neural network (RNN), a convolutional recurrent neural network (CRNN), and a parallel CNN-RNN. These models were tested by utilizing the GTZAN and FMA datasets, which contained audio files that were converted into Mel-Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) inputs. They were evaluated using overall accuracy, F1-score, recall, and precision. Then, the models were updated, such that inputs from the GTZAN dataset were divided into 5-segments. The updated models were evaluated using the same metrics noted, and they were compared against the models using the original parameters.

2 Related Works

In the following work, Chun Pui Tang, [2] used an LSTM network to improve an RNN to predict the genres of music. They used MFCCs of GTZAN and the LSTM layer as the input layer followed by the hidden layers. They concluded that LSTM has the potential to be a good engine to classify

the music genres. Our RNN with LSTM model idea is based on this paper; we also took this idea to Mel-spectrograms representation.

In their works, Choi, et al., [3] constructed a CRNN for the purposes of musical tagging. They compared a CRNN against three CNNs. Each of the CNNs had a different feature: one used 1D kernels and 2D convolutions, one used 2D kernels and 1D convolutions, and the last used 2D kernels and 2D convolutions. They noted that, of the four, the CNN with 2D kernels and 2D convolutions were faster than the CRNN. However, the CRNN performed the best when passed the same number of parameters.

Additionally, Lin Feng, Shenlan Liu, and Jianing Yao [4] have constructed an ANN where the CNN and RNN blocks were paralleled. In order to allow the RNN layer to work on the raw spectrograms instead of the output from the CNN. Our parallel CNN-RNN model was predominately influenced by this paper and our architecture is similar to theirs with some modifications since our dataset size was smaller.

Lastly, Faiyaz Ahmad [5] split the GTZAN audio data into clips to ensure homogenous content. He then used Short-term Fourier Transform (STFT), Mel-spectrogram, and MFCC for feature extraction. These features were passed into a CNN or VGG16 Neural Network. When evaluating the model, the CNN and VGG16 networks he used each had different RNN features.

3 Problem Solution

To accomplish the noted goal of this paper, datasets containing musical audio files were necessary. With this in mind, two datasets were utilized and converted into Mel-spectrogram and MFCC inputs. These inputs were then applied to each model. The models will then be evaluated to determine their efficacy.

3.1 Datasets and Preprocessing

The datasets utilized for this paper were the GTZAN and the Free Music Archive (FMA) datasets. These datasets had varying number of files, data size, file format, and number of classes; they were both used to initially test the efficacy of each model. The models were converted into Mel-spectrogram and MFCC inputs. In later attempts to improve these models, these inputs were split into five segments.

Due to constraints that became apparent when preprocessing the data (which took several hours to convert to either Mel-spectrogram or MFCC), FMA was only used to verify the results taken from the initial parameters passed to each of the models. As a consequence of these constraints, the FMA Mel-spectrogram and MFCC inputs were not tested by being further divided into five segments. Doing so would increase the number of features to be processed and thus the time needed to process the data.

3.1.1 Datasets

GTZAN Dataset

The GTZAN dataset consists of 1000 audio tracks, each 30 seconds long. Additionally these tracks have a sample frequency of 22050 Hz and 16 bits. They are categorized into 10 genres with 100 tracks for each. The different genres are blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock [6].

FMA Dataset

The FMA dataset was made for the purpose of musical analysis with up to 161 genres consisting of 106,574 untrimmed tracks from 16,341 artists and 14,854 albums in total. It was split into multiple subsets to work with including "small", "medium", "large", and "full". For this project the small dataset, which consists of 8000 audio tracks with 1000 tracks for each of the 8 balanced genres. The different genres are electronic, experimental, folk, hip hop, instrumental, international, pop and rock [7] [8].

3.1.2 Preprocessing

The audio tracks from both datasets were preprocessed into the forms of Mel-Spectrogram and MFCC input types. This preprocessing was done by using the Librosa library [9].

Mel-Spectrograms

Spectrograms are "visual way[s] to represent signal strength of different frequencies over time" [10]. Mel-spectrograms are spectrograms that were mapped to an audio scale known as the "mel scale", which leads to lower frequencies becoming more important than higher frequencies; these lower frequencies are reflective of humans interpret sound [10]. The Mel-Spectrograms can be visualized as shown in Figure 1.

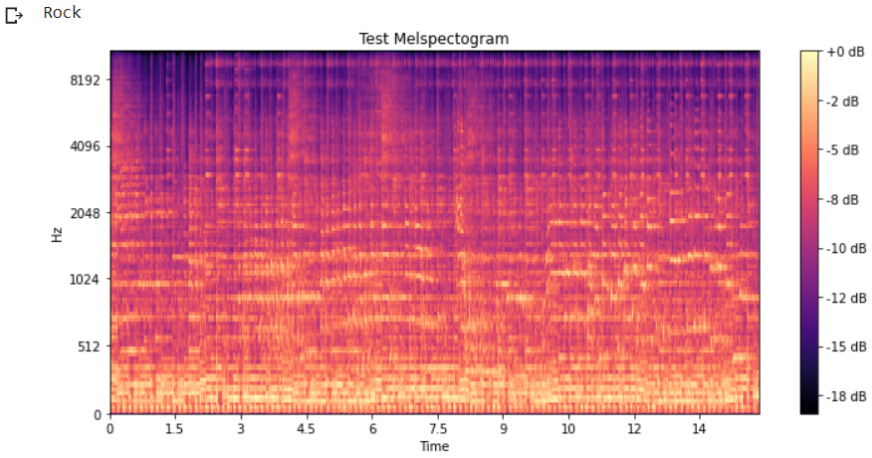

Figure 1: A Mel-Spectrogram for genre Rock.

MFCC

The MFCC inputs are "compressible representation[s] of a mel-spectrogram" [10]. These coefficients are a set of features, which can be utilized to determine the shape of a given frequency spectrum [10]. An example of an MFCC can be visualized as shown in Figure 2.

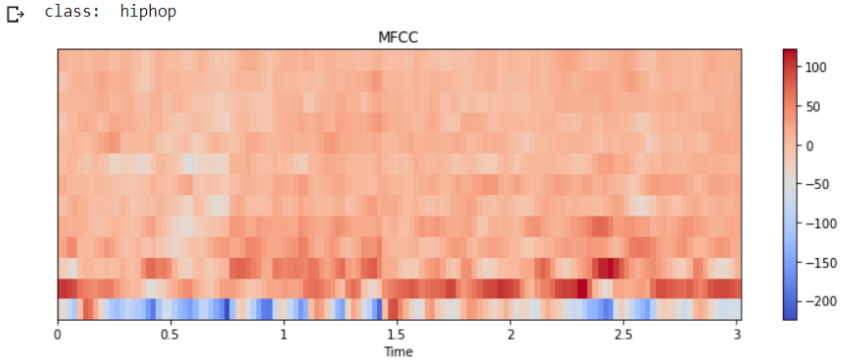

Figure 2: A MFCC for genre HipHop.

Segmentation

After the model was evaluated with the Mel-spectrogram and MFCC inputs, the inputs were then divided into five segments. This was particularly done for the GTZAN dataset. The FMA data set was not segmented because it already had 8000 mel-spectrogram and MFCC inputs. This was in addition to the time constraints noted previously.

With this in mind, the segments were determined by multiplying the sample frequency of GTZAN (22050) with by the length of each song (30 seconds). This was then divided by a hop-length of 512, which was further divided amongst the five segments. As a result of this segmentation, more input vectors were created for the models. Instead of 1000 MFCC and 1000 Mel-spectrogram inputs,

shaped approximately (1000, 640, 20) and (1000, 640, 128) respectively, this segmentation allowed for 5000 MFCC and 5000 Mel-spectrogram inputs, shaped approximately (5000, 259, 20) and (5000, 259, 128) respectively. This allowed for more features to be utilized by the ANN models when classifying the musical genre of each track [5].

3.2 Experiments

The experiments conducted for this project relied on the GTZAN and FMA dataset, which both consist of various audio files noted above. The audio files in GTZAN are comprised of WAV files, and the audio files in FMA are MP3. Using Librosa [9], these audio files were converted into Mel-spectrograms and mel-frequency cepstral coefficients. These were then passed as inputs into each of the ANN models which were trained to classify the audio files into one of the respective target genres.

The ANN models were created using Keras [11]. The accuracy was determined using Keras' evaluation method. Additionally, the F1-score, precision, and recall were collected using the sci-kit learn API [12] [13]. After using these metrics to determine the overall effectiveness of each model, the data was processed again applying five segments to the Mel-spectrogram and MFCC inputs, as noted in the Segmentation section. The models were trained and tested with the new inputs, using the same metrics to verify any improvement.

3.3 Models

The models utilized for this paper were a feed-forward neural network, an LSTM based RNN, a CRNN, and a parallel CRNN. These models were developed iteratively with the feed-forward neural network being the initial design which was used as a base model for comparison. The ANN architectures that followed built upon that feed-forward neural network to test the efficacy of each enhancement.

3.3.1 Fully Connected Neural Network

A fully connected neural network connects each neuron in one layer with every other neuron in another layer. Source [14] shows the difference between the deeper and wider dataset. It states how to declare the number of neurons respectively. It also states that fully connected networks are prone to an overfitting problem. This can be overcome by using dropouts and batch normalization on each layer.

In this model, dense layers were used to implement the fully connected network for classifying genres on MFCCs and Mel-Spectrogram features. The inputs were the coefficients in the case of MFCCs and the image data arrays in the case of Mel-Spectrograms. The inputs were then passed to a series of four dense layers, and at the end of each layer dropouts and batch normalizations were used to reduce the overfitting. The dense layers have the RELU activation function and softmax activation for the output layer.

The model used the RMSprop optimizer with a learning rate of 0.0075. The loss function used was sparse-categorical cross entropy. The best accuracy was when single segment features used was when MFCCs were used on both datasets. But when 5 segment features were used the Mel-Spectrograms performed slightly better than MFCCs on GTZAN dataset. Below is the architecture diagram for the fully connected model.

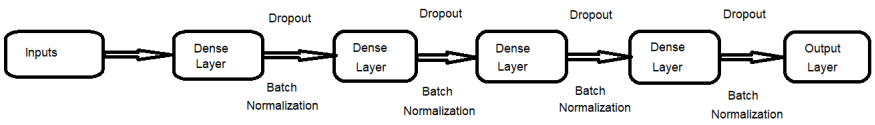

Figure 1: An architecture flow diagram for the Fully Connected Neural Network model.

3.3.2 LSTM based RNN

LSTM is a type of RNN, which works well in sequential data such as music and speech. The typical RNN has the limitation of making use of the previous context in data causing problems in learning

long-term dependencies. The LSTM avoids this problem by using memory blocks to model temporal dependencies [15].

In this model, LSTM RNN based MFCC and Mel-Spectrogram features were used for Music genre classification. The model was built using Keras framework as follows, the inputs were either MFCCs or Mel-Spectrogram. The input was then passed to two layers of LSTM with 64, 64 nodes respectively. The output of the LSTM layer was then passed to the dense layer with RELU as activation function. The dropout layer was added for about 0.3 percent. Then it was finally passed to the output layer with softmax as the activation function.

The model was optimized with RMSprop optimizer with a learning rate of 0.00075 and the loss function used was categorical cross entropy. The best test accuracy was achieved when MFCCs were used in both datasets. An architecture flow diagram can be seen for this model in Figure 1.

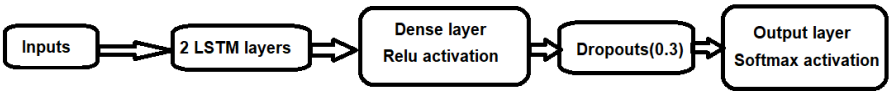

Figure 1: An architecture flow diagram for the LSTM RNN model.

3.3.3 CRNN

The third model designed was CRNN. A CRNN consists of a CNN which replaces the final convolutional layer with an RNN layer [3]. The LSTM architecture was again chosen as the RNN. This model was set up to not use any dropout layers; this was done to get a base understanding of the accuracy of this type of architecture.

More specifically, this model was comprised of three convolutional layers, followed by two LSTM layers, followed by two dense layers, and finally an output layer which uses softmax activation. The model was optimized by using RMSProp with a learning rate of 0.00075. The model was trained using 75 epochs and a batch size of 32 for the single segment GTZAN and FMA. The updated model which had five segments for GTZAN utilized a batch size of 128. The basic premise of the architecture can be seen in figure 2.

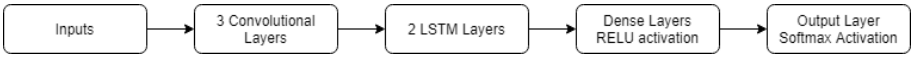

Figure 2: An architecture flow diagram for the CRNN model.

3.3.4 Parallel CNN - RNN

The way this model works is by passing the input Mel-spectrogram and MFCC through both a CNN and RNN layer which run in parallel [4]. It concatenates the output of these layers. The concatenation of these layers is then sent through a dense layer with softmax activation to perform classification of genres.

The convolutional block of the model consists of 2D convolution layer followed by a 2D Max pooling layer. There are 5 blocks of Convolution Max pooling layers. The kernel size is (3,1) for all 5 blocks. The filter sizes are 16 for the first block, 32 for the second block and 64 for the remaining 3 blocks. RELU activation is applied after each convolution. The final output is flattened and is a tensor of shape (None, 256).

The recurrent block starts with 2D max pooling layer of pool size (4,2) to reduce the size of the spectrogram before the LSTM operation. This feature reduction is done primarily to speed up processing. The reduced image is sent to a bidirectional GRU with 64 units. The output from this layer is a tensor of shape (None, 128).

The outputs from the convolutional and recurrent blocks are then concatenated resulting in a tensor of shape (None, 384). Finally, we have a dense layer with softmax activation. The model was trained using RMSProp optimizer with a learning rate of 0.00001 and the loss function is categorical cross entropy. The model was trained for 50 epochs. Figure 3 shows the model architecture.

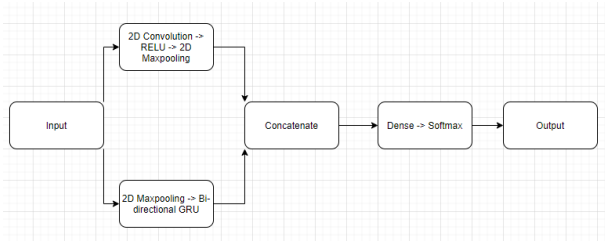

Figure 3: An architecture flow diagram for the Parallel CNN and RNN model.

4 Evaluation

4.1 Fully Connected Neural Network

For this model, the data was split into 56% for training, 19% for validation and 25% for testing data. For single segment GTZAN Mel-spectrograms, the model achieved 46.50% accuracy on the test data; it also had a weighted average F1 score of 40%. The five segment GTZAN Mel-spectrograms, the model achieved 73.04% accuracy on the test data; it had a weighted average F1 score of 72%.

Table 1: GTZAN: Single Segment Mel-spectrogram vs Five Segment Mel-spectrogram

	Precision	Recall	F1-score
Blues	0.27	0.15	0.19
Classical	0.43	0.95	0.59
Country	0.38	0.25	0.30
Disco	0.19	0.40	0.26
Hip Hop	0.43	0.15	0.22
Jazz	1.00	0.35	0.52
Metal	0.76	0.80	0.78
Pop	0.88	0.35	0.50
Reggae	0.36	0.20	0.26
Rock	0.31	0.55	0.39

	Precision	Recall	F1-score
Blues	0.70	0.66	0.68
Classical	0.87	0.98	0.92
Country	0.56	0.76	0.65
Disco	0.78	0.31	0.44
Hip Hop	0.89	0.68	0.77
Jazz	0.85	0.91	0.88
Metal	0.81	0.91	0.86
Pop	0.73	0.88	0.80
Reggae	0.68	0.70	0.69
Rock	0.55	0.55	0.55

The data for the MFCC inputs was split in the same way as Mel-spectrograms. For the single segment GTZAN MFCCs, the model achieved 52.39% on the test data; it also had a weighted F1 score of 0.47. For the five segment GTZAN MFCCs, the model achieved 70.95% accuracy on the test data; it had a weighted average F1 score of 71%.

Table 2: GTZAN: Single Segment MFCC vs Five Segment MFCC

	Precision	Recall	F1-score
Blues	0.47	0.37	0.41
Classical	0.79	0.79	0.79
Country	0.44	0.37	0.40
Disco	0.29	0.11	0.15
Hip Hop	0.58	0.39	0.47
Jazz	0.29	0.79	0.43
Metal	0.78	0.95	0.86
Pop	0.64	0.84	0.73
Reggae	0.33	0.11	0.17
Rock	0.29	0.21	0.24

	Precision	Recall	F1-score
Blues	0.58	0.83	0.68
Classical	0.68	0.96	0.80
Country	0.75	0.69	0.72
Disco	0.75	0.43	0.54
Hip Hop	0.73	0.56	0.63
Jazz	0.95	0.80	0.87
Metal	0.79	0.91	0.85
Pop	0.90	0.74	0.81
Reggae	0.64	0.71	0.68
Rock	0.49	0.48	0.49

For FMA Mel-spectrograms, the data was split into 80% for training, 10% for validation and 10% for testing data. The model achieved 34.87% accuracy on the test data; it had a weighted F1 score of 40%. For FMA MFCCs, the data was split into 56% for training, 19% for validation and 25% for testing data. The model achieved 43.39% accuracy on the test data; it had a weighted F1 score of 44%.

4.2 LSTM based RNN

For this model, the data was split into 56% for training, 19% for validation and 25% for testing data. For single segment GTZAN Mel-spectrograms, the model achieved 48% accuracy on the test data; it also had a weighted average F1 score of 43%. The five segment GTZAN Mel-spectrograms, the model achieved 70% accuracy on the test data; it had a weighted average F1 score of 71%.

Table 3: GTZAN: Single Segment Mel-spectrogram vs Five Segment Mel-spectrogram

	Precision	Recall	F1-score
Blues	0.29	0.25	0.27
Classical	0.71	0.85	0.77
Country	0.27	0.20	0.23
Disco	0.22	0.25	0.23
Hip Hop	0.60	0.15	0.24
Jazz	0.36	0.40	0.38
Metal	0.57	0.60	0.59
Pop	0.52	0.75	0.61
Reggae	0.65	0.55	0.59
Rock	0.30	0.40	0.34

	Precision	Recall	F1-score
Blues	0.65	0.63	0.64
Classical	0.88	0.88	0.88
Country	0.65	0.57	0.61
Disco	0.63	0.60	0.61
Hip Hop	0.67	0.76	0.71
Jazz	0.86	0.82	0.84
Metal	0.85	0.87	0.86
Pop	0.77	0.84	0.80
Reggae	0.65	0.73	0.69
Rock	0.44	0.38	0.40

The data for the MFCC inputs was split in the same way as Mel-spectrograms. For the single segment GTZAN MFCCs, the model achieved 53.20% on the test data; it also had a weighted F1 score of 51%. For the five segment GTZAN MFCCs, the model achieved 72.64% accuracy on the test data; it had a weighted average F1 score of 73%.

Table 4: GTZAN: Single Segment MFCC vs Five Segment MFCC

	Precision	Recall	F1-score
Blues	0.55	0.32	0.40
Classical	0.87	0.68	0.76
Country	0.38	0.63	0.47
Disco	0.57	0.21	0.31
Hip Hop	0.44	0.61	0.51
Jazz	0.50	0.53	0.51
Metal	0.83	0.53	0.65
Pop	0.57	0.84	0.68
Reggae	0.62	0.56	0.59
Rock	0.23	0.26	0.24

	Precision	Recall	F1-score
Blues	0.75	0.67	0.71
Classical	0.92	0.91	0.92
Country	0.57	0.83	0.67
Disco	0.63	0.70	0.67
Hip Hop	0.69	0.72	0.71
Jazz	0.80	0.78	0.79
Metal	0.91	0.76	0.83
Pop	0.78	0.90	0.84
Reggae	0.73	0.62	0.67
Rock	0.58	0.41	0.48

For FMA Mel-spectrograms, the data was split into 80% for training, 10% for validation and 10% for testing data. The model achieved 39.25% accuracy on the test data; it had a weighted F1 score of 44%. For FMA MFCCs, the data was split into 56% for training, 19% for validation and 25% for testing data. The model achieved 51.95% accuracy on the test data; it had a weighted F1 score of 50%.

4.3 CRNN

For this model, the data was split such that there was a 55% training set, 30% test set, and 15% validation set. For the single segment GTZAN Mel-spectrograms, this model achieved approximately 39% accuracy on the training data; it also had a weighted average F1 score of 38%. Lastly, for the five segment GTZAN Mel-spectrograms, the accuracy was approximately 57.9439%. The five segment GTZAN Mel-spectrogram inputs had a weighted average F1 score of 58%. The F1 scores, precision, and recall based on Mel-spectrograms for each individual genre can be seen in Table 5.

The data for the MFCC inputs was split in the same way as above. For the single segment GTZAN MFCCs, the model reached an approximate 32% accuracy. It had a weighted average F1 score of approximately 30%. Finally, the five segment GTZAN MFCC had an accuracy of approximately 37%

Table 5: GTZAN: Single Segment Mel-spectrogram vs Five Segment Mel-spectrogram

	Precision	Recall	F1-score
Blues	0.31	0.17	0.22
Classical	0.79	0.87	0.83
Country	0.28	0.23	0.25
Disco	0.22	0.23	0.23
Hip Hop	0.06	0.03	0.04
Jazz	0.27	0.60	0.37
Metal	0.52	0.53	0.52
Pop	0.44	0.37	0.40
Reggae	0.38	0.57	0.45
Rock	0.30	0.30	0.30

	Precision	Recall	F1-score
Blues	0.43	0.38	0.40
Classical	0.78	0.91	0.84
Country	0.37	0.49	0.42
Disco	0.58	0.59	0.59
Hip Hop	0.70	0.53	0.60
Jazz	0.48	0.67	0.56
Metal	0.83	0.71	0.77
Pop	0.74	0.64	0.69
Reggae	0.56	0.52	0.54
Rock	0.45	0.35	0.39

and a weighted average F1 score of 36%. The F1 scores, precision, and recall for each individual genre based on MFCC inputs can be seen in Table 6.

Table 6: GTZAN: Single Segment MFCC vs Five Segment MFCC

	Precision	Recall	F1-score
Blues	0.06	0.03	0.04
Classical	0.61	0.63	0.62
Country	0.26	0.20	0.23
Disco	0.22	0.23	0.23
Hiphop	0.32	0.23	0.27
Jazz	0.34	0.57	0.42
Metal	0.41	0.80	0.55
Pop	0.41	0.23	0.30
Reggae	0.29	0.13	0.18
Rock	0.11	0.13	0.12

	Precision	Recall	F1-score
Blues	0.19	0.33	0.24
Classical	0.83	0.80	0.81
Country	0.28	0.50	0.36
Disco	0.31	0.37	0.34
Hiphop	0.38	0.20	0.26
Jazz	0.42	0.37	0.39
Metal	1.00	0.13	0.24
Pop	0.65	0.67	0.66
Reggae	0.23	0.27	0.08
Rock	0.11	0.07	0.08

For the FMA dataset, the split between training, validation, and test remained the same as noted for the GTZAN dataset. Mel-spectrograms, the model managed approximately 32.8053%. Its weighted average F1 score of 29%. The FMA MFCC, had an approximate accuracy of 33.3889% and an approximate weighted average F1 score of 31%.

4.4 Parallel CNN-RNN

For this model, the data was split into 56% for training, 19% for validation and 25% for testing data. For single segment GTZAN Mel-spectrograms, the model achieved 49.5% accuracy on the test data; it also had a weighted average F1 score of 48%. The five segment GTZAN Mel-spectrograms, the model achieved 60.24% accuracy on the test data; it had a weighted average F1 score of 59%. The full F1 score, precision, and recall values are shown in Table 7.

The data for the MFCC inputs was split in the same way as Mel-spectrograms. For the single segment GTZAN MFCCs, the model achieved 40.79% on the test data; it also had a weighted F1 score of 38%. For the five segment GTZAN MFCCs, the model achieved 49.59% accuracy on the test data; it had a weighted average F1 score of 48%. More detailed F1 score, precision, and recall are shown in Table 8.

For FMA Mel-spectrograms, the data was split into 80% for training, 10% for validation and 10% for testing data. The model achieved 44% accuracy on the test data; it had a weighted F1 score of 39%. For FMA MFCCs, the data was split into 56% for training, 19% for validation and 25% for testing data. The model achieved 44.74% accuracy on the test data; it had a weighted F1 score of 44%.

Table 7: GTZAN: Single Segment Mel-spectrogram vs Five Segment Mel-spectrogram

	Precision	Recall	F1-score
Blues	0.27	0.15	0.19
Classical	0.56	0.45	0.50
Country	0.62	0.80	0.70
Disco	0.56	0.45	0.50
Hip Hop	0.44	0.35	0.39
Jazz	0.31	0.45	0.37
Metal	0.55	0.30	0.39
Pop	0.47	0.40	0.43
Reggae	0.52	0.85	0.64
Rock	0.60	0.75	0.67

	Precision	Recall	F1-score
Blues	0.58	0.60	0.59
Classical	0.52	0.62	0.57
Country	0.83	0.90	0.87
Disco	0.68	0.48	0.56
Hip Hop	0.53	0.43	0.48
Jazz	0.58	0.42	0.49
Metal	0.41	0.43	0.42
Pop	0.50	0.46	0.48
Reggae	0.70	0.86	0.77
Rock	0.66	0.81	0.73

Table 8: GTZAN: Single Segment MFCC vs Five Segment MFCC

	Precision	Recall	F1-score
Blues	0.36	0.16	0.22
Classical	0.43	0.48	0.45
Country	0.79	0.76	0.78
Disco	0.40	0.16	0.23
Hiphop	0.42	0.44	0.43
Jazz	0.20	0.40	0.26
Metal	0.00	0.00	0.00
Pop	0.29	0.32	0.30
Reggae	0.42	0.84	0.56
Rock	0.62	0.52	0.57

	Precision	Recall	F1-score
Blues	0.64	0.56	0.60
Classical	0.57	0.47	0.52
Country	0.82	0.92	0.86
Disco	0.39	0.38	0.39
Hiphop	0.44	0.38	0.41
Jazz	0.34	0.34	0.34
Metal	0.33	0.13	0.18
Pop	0.37	0.38	0.37
Reggae	0.49	0.83	0.62
Rock	0.47	0.56	0.51

5 Conclusion

In this paper, we have compared the music genre classification on two representation Mel-spectrograms and MFCCs. Based on the results obtained from all the four models as shown in evaluation section, we can conclude that using Mel-spectrograms performs better in most of the models. The five segment features models give the best accuracy and F1 score in all our models. Thus, five segment feature extraction is better than the single whole segment feature on each audio file. The fully connect neural network model gives the highest accuracy of 73.04% and F1 score of 72% on five segment Mel-spectrograms. The LSTM based RNN model gives the highest accuracy of 72.64% and F1 score of 73% on five segment MFCCs.

References

[1] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[2] Chun Pui Tang, Ka Long Chui, Ying Kin Yu, Zhiliang Zeng, and Kin Hong Wong. Music genre classification using a hierarchical long short term memory (lstm) model. In *Third International Workshop on Pattern Recognition*, volume 10828, page 108281B. International Society for Optics and Photonics, 2018.

[3] Keunwoo Choi, György Fazekas, Mark Sandler, and Kyunghyun Cho. Convolutional recurrent neural networks for music classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2392–2396. IEEE, 2017.

[4] Lin Feng, Shenlan Liu, and Jianing Yao. Music genre classification with paralleling recurrent convolutional neural network. *arXiv preprint arXiv:1712.08370*, 2017.

[5] Sahil Faiyaz Ahmad. Music genre classification using spectral analysis techniques with hybrid convolution-recurrent neural network. In *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, pages 149–154. IJITEE, 2019.

[6] George Tzanetakis, Georg Essl, and Perry Cook. Automatic musical genre classification of audio signals, 2001.

[7] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[8] Michaël Defferrard, Sharada P. Mohanty, Sean F. Carroll, and Marcel Salathé. Learning to recognize musical genre from audio. In *The 2018 Web Conference Companion*. ACM Press, 2018.

[9] Brian McFee, Vincent Lostanlen, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Jack Mason, Dan Ellis, Eric Battenberg, Scott Seyfarth, Ryuichi Yamamoto, Keunwoo Choi, viktorandreevichmorozov, Josh Moore, Rachel Bittner, Shunsuke Hidaka, Ziyao Wei, nullmightybofo, Darío Hereñú, Fabian-Robert Stöter, Pius Friesch, Adam Weiss, Matt Vollrath, and Taewoon Kim. librosa/librosa: 0.8.0, July 2020.

[10] Gabriel Gessle and Simon Åkesson. A comparative analysis of cnn and lstm for music genre classification, 2019.

[11] François Chollet et al. Keras. `https://keras.io`, 2015.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[13] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.

[14] Sh Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378, 10 2019.

[15] Jia Dai, Shan Liang, Wei Xue, Chongjia Ni, and Wenju Liu. Long short-term memory recurrent neural network based segment features for music genre classification. In *2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5, 2016.