Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)
IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

Music Genre Classification using Machine Learning

Anirudh Ghildiyal
Graphic Era Deemed to be University
Dept. of Computer Science and Engineering
Dehradun, Uttarakhand, India
a.ghildiyal0506@gmail.com

Komal Singh
Graphic Era Deemed to be University
Dept. of Computer Science and Engineering
Dehradun, Uttarakhand, India

Sachin Sharma
Graphic Era Deemed to be University
Dept. of Computer Science and Engineering
Dehradun, Uttarakhand, India

***Abstract*— The music industry has undergone major changes from its conventional existence and also in the form of music created in last few years. The ever-growing customer base has also increased the market for different music styles. Music not only bring the individuals together, but also provides insight for various cultures. Therefore, it is essential to classify the music according to the genres to satisfy the needs of the people categorically. The manual ranking of music is a repetitive, lengthy task and the duty lies with the listener. The proposed research work has compared few classification models and established a new model for CNN, which is better than previously proposed models. This research work has trained and compared the proposed models on GTZAN dataset, where most of the models were audio file trains, while a few of the models were trained on the spectrogram.**

Keywords*— *Deep Leaning; music genre; GTZAN; convolution neural network; mel-spectogram

I. INTRODUCTION

Genre classification is the way that can classify similar types of data into a single identity (based on its rhythm, instrument played, or harmonic content) and give that identity as its name. The music industry is increasing around the globe and every day new songs are written. Since the classification of such songs every day will become a tired activity, where the technology can be used to cure the music and make classification easier or more efficient by using its rhythms, beats and lyrical composition. A song can be represented in the form of an audio signal. This audio signal has different features [1] such as frequency, spectral roll-off, root-mean-square (RMS) level, bandwidth, zero-crossing rate, spectral centroid, etc. The computer in different format such as wav or mp3 reads this audio signal. Different music streaming apps like Spotify, Wynk, Apple Music etc, use genre classification to recommend songs to its user.

Tremendous research has been done in Music Genre classification. These researches can be mainly classified into two groups based on the type of dataset used. Two known datasets are available, they are FMA and GTZAN dataset.

The machine learning approach is used to solve the challenges present in this paper. Since, the introduction of the first convolution neural network, it gave pace to the field of deep learning like image classification and segmentation, object detection and recognition. CNN [2] is a special kind of neural network that has a gird like topology. This grid can be a linear like time series data or a 2D gird like that of an image. CNN uses a system similar to a multilayer perceptron that reduces the processing requirements. Apart from CNN, support vector machine, artificial neural network, multilayer perceptron and decision tree are used for comparative analysis.

II. LITERATURE REVIEW

As the music industry is growing rapidly with new technological innovations, researchers are also developing their interest in the field of music. Various deep learning solutions are proposed by different authors to classify different types of music.

In [3] the author used a residual neural network (RNN) to train the model on 3 second audio clips, which were extracted from the GTZAN dataset. Overlapping features of different genres were also considered and an accuracy of 94% was achieved. Similarly, the author in [4] compared various algorithm and proposed a new near real-time classification using RNN but with a low accuracy of 64%. The author used the mean and co-variance of MFCC for training their model.

A comparison on various pre-existing model [5], where the author tried to find the best machine learning algorithm for music genre classification was done, where model was trained on MFCCs (Mel-frequency cepstral coefficients) and other features of the songs, where a convolution neural network (CNN) produced the highest accuracy of 88.5%.

Some of authors used convolution neural network (CNN) to classify the genres. There are three different ways to visualize the video – spectrogram, chromagrams and MFCC; different authors used different methods for visualization.

The author [6] compared two classes of model, where CNN architecture (VGG-16) was used for the spectrogram of audio signals and on time and frequency domain of the audio. The dataset used by the author was 10 seconds and the audio clips are extracted from 2.1 million YouTube videos. Author used a new classifier and achieved an accuracy of 65% with AUC of 0.894.

In [7] the author extracted the mel-spectrogram and used it as an input. The author used duplicate convolution layer, where the output was passed through different pooling layers and a statistical analysis was done.

Another author [8] used MFCC and compared CNN and Long Short-Term Memory (LSTM) model where CNN model produced a better accuracy. The CNN had five convolution layers of 32 nodes each and one fully connected layer of 128 nodes followed by an output layer of 10 nodes. The LSTM model had five layers, where the first layer had 128 nodes and the rest had 32 nodes.

III. METHODOLOGY

A. Dataset Preperation

There are two well-known datasets available for Music Genre Classification, the FMA dataset [9] which contains details of audio features of 8000 different songs of 8 different genres, the other is the GTZAN dataset which contains 1000 audio files of 10 different genres. The division of each genre can be seen in Table I.

1) In this paper, the GTZAN dataset is used, which was first used by authors in [10]. This dataset has 10 classes, Blues, Classical, Country, Disco, Hip-Hop, Jazz, Pop, Metal, Reggae and Rock. Each class containing 100 audio clips, each clip being 30 seconds long in .wav format and is samples at 22050Hz, 16bit.

TABLE I. NUMBER OF AUDIO CLIPS IN EACH GENRE

S. No	Class	Clips
1	Blues	100
2	Classical	100
3	Country	100
4	Disco	100
5	Hiphop	100
6	Jazz	100
7	Metal	100
8	Pop	100
9	Reggae	100
10	Rock	100
	Total	1000

B. Data Pre-Processing

This dataset contains 100 audio clips for one class which is not enough to get a good accuracy. Either one could use a dataset with more audio clips or pre-process the dataset in such a way that it increases the number of training and testing samples.

1) Spectogram Generation: Each audio file is converted into itsmel-spectrogram then audio clips are loaded using the librosa library and generated the mel-spectrogram for each audio clip. After the spectrogram was generated it was sliced into 64 strips. This increased our samples 64 times. We observed a total of 64000 samples each of dimension 480x10 for training and testing. So, we divided out 64000 images into for 44200 testing, 7000 for validation and 12800 for testing. Fig. 1 shows the original image and the strips generated after processing.

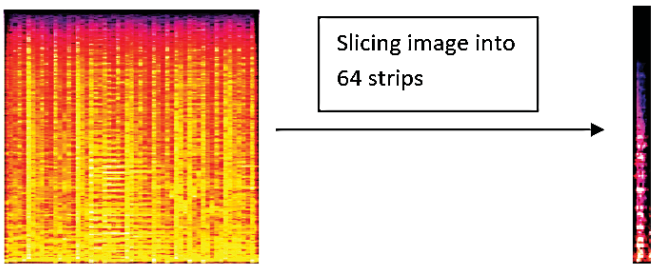

Fig. 1. Pre-processing of images.

2) Feature Extraction: Every audio can be represented in form of an audio signal and this signal has different features. The audio features are extracted that are relevant for solving the problem. These features are divided into 2 sub categories. These definitions are inspired by the work of author in [11].

a) Time Domain Feature:

i. Zero Crossing Rate: The rate at which a signal changes from positive to negative or vice versa is known as zero-crossing rate.

ii. Root Mean Square Energy: Root Mean Square Energy (RMSE) defines how loud a signal is. RMSE is defined as:

$$\sqrt{\frac{1}{N}\sum |x(n)|^2}$$

b) Frequency Domain Features:

i. Mel-Frequency Capstral Coefficient: Set of features (around 10-20) that describe the shape of the audio signal are known as Mel-Frequency Capstral Coefficient.

ii. Chroma Features: It is a representation for a music signal where the entire spectrum is projected onto 12 bits representing the 12 distinct semitones of musical octave.

iii. Spectral Centroid: It is the weighted mean of the frequencies present in the sound, which tells us about the 'center of mass' of the signal.

iv. Spectral Roll-off: The value frequency below which a specified percentage of the total spectral energy lies.

3) Convolution Neural Network: As we can see there are characteristics features even in a 480x10 size image, which are different for every class. Our CNN model is provided with these images as input. A rough architecture of our model is shown in Fig. 2.

4) CNN Model: The training images are passed i.e. the sliced images of the spectogram to our deep neural network for comprising of two sub-networks. The first neural network is a four-layer convolution neural network for extracting features from the images. These extracted features are then passed to the second sub-network for classification. This network is fully connected network containing two fully connected layers. In the end a dense layer is used to predict the genre of the audio. For fine tuning and performance

optimization of our model, Adam optimization algorithm is used.

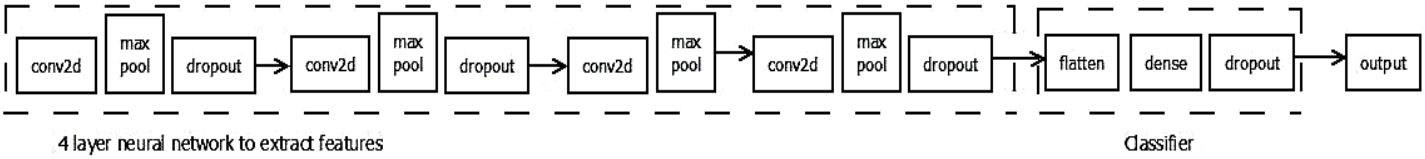

Fig. 2. CNN model.

Each layer of the CNN does the following operations:

a) Convolution: This layer has a set of filters whose parameters need to be learned. The dimension of each filter is smaller than the input image (generally 3x3 size). This filter passes through the image covering all the pixel values in the image. As this filter passes through the image the scaler product of the image and the filter is calculated.

b) Max Pooling: The purpose of pooling layer is to perform down sampling according to the dimension of the given input. Thus, reducing the number of parameters with that activation. Two common pooling methods are average pooling and max pooling.

c) Dropout [11]: It is a technique to prevent our model from overfitting thereby increasing the efficiency. While training our model, in each iteration the weight of some of the neurons are set to zero randomly. Final output is predicted using different combination of neurons. A dropout rate of 0.4 is used, i.e., a neuron weight is set to zero during an iteration, with a probability of 0.4.

5) Other Models: Here we discuss in brief about the rest of our models that we designed. For all these models we used a csv files which contains the hand-crafted features of the audio signals.

a) Artificial Neural Network [12]: It is a computational model that is inspired by the working of a human brain. It is so because the information is travelled similar to the human brain. As the information travels through the neurons of the network, the structure is affected due to which the neural network learns based on the input and output. ANN is non-linear data model, is used when complex relationship between the input and out needs to be found.

b) MLP: It is a type of logistic regressor where we insert an intermediate layer also known as the hidden layer. Nonlinear activation function (tanh or sigmoid) are present in this layer. One can make the architecture deep by inserting as many hidden layers as per the requirement of the user.

c) SVM [13]: Tough SVM is generally used for binary classification. Here we used one-vs-rest method to complete our task. All the 10 classes are trained separately and during testing the class with the highest probability is selected as the predicted class.

d) Decision Tree: This classifier is used for multiclass classification. This type of model can be pictured in form of a binary tree. Starting from the root node to all the internal nodes a set of questions to the dataset (related to its features/attributes) is presented and the nodes are then further split into new nodes having different characteristics. The leaves of the tree represent the classes in which the dataset is split.

IV. Experimental Results and Analysis

1) System Configuration and Dataset

Our proposed models are implemented on a PC having TensorFlow 1.12.0, tflearn 0.3.2, Keras 2.2.4, OpenCV 3.4.3 libraries installed in Python 3.5.2 using an Intel Xeon 3.4GHz processor and 32 GB RAM. For measuring the performance of our deep neural network, the GTZAN dataset is used. The dataset consists of 1000 audio clips, 100 audio clips for each of the class. The time taken for our model was around 30 minutes.

2) Performance Measurement

The three parameters are used to estimate the performance of the model, sensitivity, specificity, and accuracy. Here sensitivity tells us how well our models classify a particular class, and specificity tells us how well our model is classified for non-current class. Accuracy tells us about the overall ratio of correctly detected events. All these parameters are defined as follows:

Sensitivity = TP / (TP + FP)

Specificity = TN / (TN + FP)

Accuracy = (TP + TN) / Total Events

Here sensitivity and specificity are calculated for each class and accuracy for the overall results. For calculating sensitivity and specificity for a particular class say X,

TP for X is all the X instances that are classified as X

TN for X is all the non-X instances that are not classified as X

FP for X is all the non-X instances that are classified as X

FN for X is all the X instances that are not classified as X

We have evaluated the performance of our models on the above-mentioned parameters, by considering the number of samples passed to the model. Table II gives tabulated information about the sensitivity, specificity and accuracy. The ROC curve in Fig. 3 shows the true positive vs the false

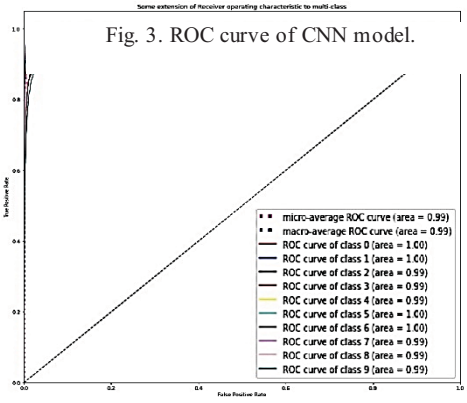

Fig. 3. ROC curve of CNN model.

positive rate. Since the AUC value of each class is great, our model is able to distinguish between each class.

TABLE II. RESULT ANALYSIS OF OUR MODEL.

Class	Sensitivity	Specificity	Accuracy
Blues	0.93	0.93	0.93
Classical	0.92	0.98	0.95
Country	0.87	0.87	0.87
Disco	0.90	0.85	0.88
Hiphop	0.91	0.89	0.90
Jazz	0.91	0.93	0.92
Metal	0.93	0.97	0.95
Pop	0.91	0.90	0.90
Reggae	0.91	0.86	0.89
Rock	0.86	0.86	0.86
Overall Accuracy			0.91

3) Comparative Analysis Measurement

In this paper, we have designed five different models CNN, ANN, SVM, MLP and Decision Tree. The following different audio features (Mel-Frequency Capstral Coefficient, Root Mean Square Energy, Spectral Centroid, Zero Crossing Rate, Chroma Frequencies, Spectral Roll-off) are stored in a csv file and passed it to our ANN, SVM, MLP and Decision Tree models. For our proposed CNN model, we have passed the extracted mel-spectrogram images. The accuracy achieved by our model is given in Table III. Fig. 4 shows the result analysis of our different models.

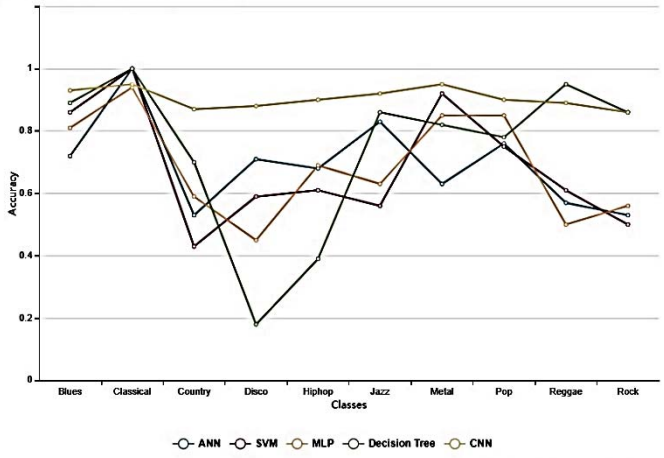

Fig. 4. Result analysis of our different models.

TABLE III. ACCURACY OF OUR DIFFERENT MODELS.

Model	Accuracy
ANN	70%
SVM	68.9%
MLP	68.7%
Decision Tree	74.3%
CNN	91%

It can be seen from the graph that our CNN model and ANN model gave the best results. Out of the 10 genres our models gave best accuracy for blues, classical, pop. Overall, the accuracy achieved by each model is given in the Table III. Our CNN model had advantage on disco and hiphop genre. It can be seen all our model relatively performed worst on country genre. The confusion matrix of our CNN model is shown in Table IV.

There are lot of methods proposed in this work of literature. People have used different types of dataset and some authors created their own dataset. Moreover, some the numbers of genres considered by the authors are different in different works. Some authors have considered 5 or less out of the 10 genres for building their models. Hence, their results cannot be compared with this model. We have only considered only those works where all the 10 genres of the GTZAN dataset are used. Table V demonstrates our model by comparing it with by comparing the performance with other pre-existing models.

TABLE IV. CONFUSION MATRIX FOR CNN MODEL.

	Bl	Cl	Co	Di	Hi	Ja	Me	Po	Re	Ro
Bl	93	1	1	0	2	0	0	0	0	2
Cl	0	98	0	0	1	0	0	0	0	0
Co	2	2	87	1	0	2	0	1	1	4
Di	1	1	2	85	2	1	1	2	2	3
Hi	1	0	1	1	89	1	2	2	2	1
Ja	1	3	1	0	0	93	0	0	0	1
Me	0	0	0	0	1	0	97	0	0	2
Po	0	1	1	2	1	1	0	90	1	2
Re	1	1	2	2	2	1	0	2	86	2
Ro	1	1	3	1	1	1	4	1	1	86

TABLE V. COMPARATIVE ANALYSIS WITH OTHER MODELS.

Model	Model Used	Accuracy %
[3]	Residual Neural Network	94.0
[14]	Compressive Sampling	92.7
[8]	Convolution Neural Network	90.7
[5]	Convolution Neural Network	88.5
[16]	Support Vector Machine	84.4
[15]	AM-FM	84.3
[7]	Convolution Neural Network	65
[4]	Residual Neural Network	64
Proposed Model	Convolution Neural Network	91

V. CONCLUSION

The proposed research work has utilized the GTZAN dataset and produced multiple models to complete this task in this piece of music classification. The proposed model has used multiple inputs for various models along with the audio mel-spectrogram and transferred this to our CNN, and various sound file characteristics stored in the ANN, SVM, MLP and Decision Tree csv archives 91%, equivalent to the human understanding of genre with highest accurate achievement. Since, some styles were quite distinctive and some rather distinctive such as the country and the rock genre were confused with other styles, although traditional and blues were easily identified.

REFERENCES

[1] McKinney, Martin, and Jeroen Breebaart. "Features for audio and music classification." (2003).
[2] O'Shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." *arXiv preprint arXiv:1511.08458* (2015).

[3] Bisharad, Dipjyoti, and Rabul Hussain Laskar. "Music Genre Recognition Using Residual Neural Networks." In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pp. 2063-2068. IEEE, 2019.
[4] Zhang, Scott, Huaping Gu, and Rongbin Li. "MUSIC GENRE CLASSIFICATION: NEAR-REALTIME VS SEQUENTIAL APPROACH." (2019).
[5] Chillara, Snigdha, A. S. Kavitha, Shwetha A. Neginhal, Shreya Haldia, and K. S. Vidyullatha. "Music Genre Classification using Machine Learning Algorithms: A comparison." (2019).
[6] Bahuleyan, Hareesh. "Music genre classification using machine learning techniques." *arXiv preprint arXiv:1804.01149* (2018).
[7] Yang, Hansi, and Wei-Qiang Zhang. "Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks." In *INTERSPEECH*, pp. 3382-3386. 2019.
[8] Gessle, Gabriel, and Simon Åkesson. "A comparative analysis of CNN and LSTM for music genre classification." (2019).
[9] Defferrard, Michaël, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. "Fma: A dataset for music analysis." *arXiv preprint arXiv:1612.01840* (2016).
[10] George, Tzanetakis, Essl Georg, and Cook Perry. "Automatic musical genre classification of audio signals." In *Proceedings of the 2nd international symposium on music information retrieval. Indiana*. 2001.
[11] Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The journal of machine learning research* 15, no. 1 (2014): 1929-1958.
[12] Grossi, Enzo, and Massimo Buscema. "Introduction to artificial neural networks." *European journal of gastroenterology & hepatology* 19, no. 12 (2007): 1046-1054.
[13] Weston, Jason, and Chris Watkins. "Support vector machines for multi-class pattern recognition." In *Esann*, vol. 99, pp. 219-224. 1999.
[14] Chang, Kaichun K., Jyh-Shing Roger Jang, and Costas S. Iliopoulos. "Music Genre Classification via Compressive Sampling." In *ISMIR*, pp. 387-392. 2010.
[15] Hamel, Philippe, and Douglas Eck. "Learning features from music audio with deep belief networks." In *ISMIR*, vol. 10, pp. 339-344. 2010.
[16] Zlatintsi, Athanasia, and Petros Maragos. "Comparison of different representations based on nonlinear features for music genre classification." In *2014 22nd European Signal Processing Conference (EUSIPCO)*, pp. 1547-1551. IEEE, 2014.