Music Genre Classification Using Convolutional Neural Network

Nitin Choudhury
Department of Information Technology
Gauhati University
Guwahati, India
nitinnlb@gmail.com

Deepjyoti Deka
Department of Information Technology
Gauhati University
Guwahati, India
deepjyotideka8@gmail.com

Satyajit Sarmah
Department of Information Technology
Gauhati University
Guwahati, India
ss@gauhati.ac.in

Parismita Sarma
Department of Information Technology
Gauhati University
Guwahati, India
pari@gauhati.ac.in

2023 4th International Conference on Computing and Communication Systems (I3CS) | 979-8-3503-2377-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/I3CS58314.2023.10127554

***Abstract*— Music genres are categories that classify music based on its common traditions and customs. These genres can enhance the enjoyment of music by providing listeners with a way to categorize and understand the music. When used constructively, it helps to better understand the art form, to recognize innovation and, above all, to improve the ability to judge quality. The main goal of this work is to study the different behaviors of musical genres based on their spectral representations and create an automated system for classification. Collecting the properly classified music dataset (i.e., GTZAN Music Genre) the feature-map of the data that is extracted is fed to the neural network model for evaluation. Accuracy of training, testing and validation is acquired. Along with that validation losses are reduced to an extent. The evaluation matrix is also computed. After the model is trained, it is deployed to the server along with a Flask-based REST API for easy access and use of the trained model for classification.**

Keywords: *Music, genre, classification, CNN, signal processing, REST API, Flask.*

I. INTRODUCTION

Music is all about the arrangement of sounds in specific patterns. All Human beings in the world seem to be interested and curious about different types of music either for cultural aspects or personal interest. Music can be defined as a style that emphasizes, attenuates, or omits common elements of an organized sound, such as rhythm, volume and pitch.

Music genre is a conventional category and can be defined as a significant ingredient of world tradition. There are some major differences between music format and music style. Experts assume different ways to categorize music into different genres and the mostly used ones are blues, disco, jazz, metal, classical, pop, rock etc. Technology is utilized by music aficionados to differentiate between these genres and to listen to their preferred type of music. Listeners need one listener to move from jazz to rap, but obviously more listeners are needed to inspire the love of music. Sorting music files by genre in the Music Information Retrieval (MIR) is a difficult task. Automatic classification of music genres is important for getting music from large collections. It finds real applications in various areas, such as automatic tagging of unknown music.

Genre classification, which involves categorizing music into various genres, is a concept that helps distinguish between two genres based on their rhythms. Recently, genre classification has become very popular, and many genres are emerging around the world. Today, different online music platforms are using different methodologies to classify music based on its genres. Some well-known platforms are JioSaavn, Spotify etc.

In deep learning, convolutional neural networks are a category of deep neural networks that are widely used for visual image analysis. Thus, CNN can also be used to classify the music in different genres using its spectrographic representations.

The main objective of this research work is to classify music genre using CNN with a better classification accuracy. Here in this work, we played with the sample size of GTZAN dataset which contains 10 genres and 100 music each genre of length 30 seconds duration. By increasing the size of the dataset, we're able to achieve much better classification result compared to previous research works.

This paper is arranged with the following organizations. Section 1 basically introduces the music genre in brief. In the section II, we have done a review of the related works. Section III contains the topics that are used in the concerned study work. The section IV describes the problem statement of the proposed study. In section V, the methodology of the study work is described in brief. In the section VI, the result we have obtained is described and the section VII consists of the conclusion and the future work of the study.

II. LITERATURE REVIEW

The classification of music genres is of interest to many researchers. Tzanetakis and Cook created the GTZAN dataset which is still used today as a criterion for classifying genres. The researchers selected rhythmic and pitch content with timbral texture as feature set and a few classifiers for evaluation [1].

Vishnupriya S. and K. Meenakshi proposed a neural network model and Mel Spectrogram, MFCC as feature vectors to perform classification on the dataset [2].

Changsheng Xu and others did perform Support Vector Machine (SVM) based classification for the task [3].

In the feature vectors they have used calculated parameters like LPC, MFCC, ZCR, STE and beat spectrum.

For classification they had selected Multi-Layer SVM learning approach.

Matthew Creme Charles Burlin and Raphael Lenaine of Stanford University applied four methods to classify music [4]. These methods were support vector machines, neural networks, decision trees, and K-nearest neighbors. They applied PCA, calculated MFCC as feature vectors for the classification.

Zhengxin Qi and the team [5] compared and analyzed the feasibility, performance and understandability of the feature vectors that are used to describe the music by classifying its genre using machine learning techniques.

Kumaraswamy and the team used KNN and CNN based classification approach for the genre classification task and Mel spectrogram and MFCC as feature vectors [6].

Tao demonstrates limited use of Boltzmann machines and generates more data from the original GTZAN dataset, which outperforms typical multi-layer neural networks [7].

Archit Rathore and Margaux Dorido calculated MFCC, zero crossing rate, spectral centroid, chroma frequency, spectral roll-off as feature vectors and used different classifiers for the task of which SVM with Polynomial Kernel performed quite well for them [8].

Kostrzewa and team used a combination of several different deep neural networks as base models to perform the classification task. A series of experiments were carried out for 1-D and 2-D CRNN, RNN with LSTM and also for both the CNN [9].

III. MUSIC GENRE CLASSIFICATION

A. Music Genre

Genre is one of the most important features to identify a music [1]. Genre can describe the behavior of a music. Different genre has different impact on people. The results of the classification process can aid in the investigation of socio psychological aspects of how people perceive similarities in music and form musical communities. [4]. Music can be categorized into various genres in a multitude of ways. The most popular music genres are blues, classic, country, disco, hip hop, jazz, metal, pop, reggae and rock. However, it is a difficult task to identify the genre of a music. It is not that straight forward task.

B. Mel-frequency Cepstral Coefficient (MFCCs)

Mel-frequency cepstral coefficients (MFCCs) are a limited set of features that succinctly characterize the spectral envelope's general shape for a signal [10]. The MFCC method involves breaking an audio signal into overlapping frames, computing the power spectrum via a Fourier transform, transforming the power spectrum into the mel frequency scale, and applying a set of mel-scale filter banks to extract a series of mel-scale filter-bank energies. These features are then transformed via a discrete cosine transform to obtain a set of coefficients that capture the most important characteristics of the audio signal. MFCCs have proven effective in numerous audio-related applications, such as speech recognition, speaker identification, and music genre classification. In MIR, it is often used to describe the quality given to a sound by its overtones. MFCCs are a set of features that describe the short-term power spectrum characteristics of sound, and they have been widely utilized in advanced sound categorization and recognition techniques. This feature is a large part of the final feature vector of a music.

C. Convolutional Neural Network (CNN)

A convolutional neural network (CNN) is a type of artificial neural network that is primarily used for visual analysis of images.

Convolutional neural networks (CNNs) utilize a structure that is similar to multilayer perceptrons, but with a design that aims to reduce computational demands. The CNN layer is composed of an input layer, an output layer, and a hidden layer consisting of multiple convolutional layers, pooling layers, fully connected layers, and regularization layers. By relaxing constraints and enhancing processing efficiency, CNNs can enable faster and more effective image processing, resulting in a more efficient, easier to learn, and more streamlined image processing system.

D. CNN over LSTM model

Both Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) can be used for music genre classification. However, CNNs are often preferred over LSTMs for this task because they are better suited for analyzing the temporal features of audio data, which are important for distinguishing between different genres.

CNNs are particularly good at detecting local patterns in data, which can be useful for identifying short-term patterns in audio signals, such as the rhythm and timbre of a song. On the other hand, LSTMs are better suited for modeling longer-term dependencies in data, which may not be as relevant for music genre classification.

Additionally, CNNs can handle high-dimensional data efficiently, which is important for processing the complex spectrogram or mel-spectrogram representations of audio data. LSTMs, on the other hand, can suffer from the vanishing gradient problem when trained on long sequences of data, which can make it difficult to learn meaningful representations of the audio data.

Overall, while LSTMs can be useful for some aspects of music analysis, such as generating new music or predicting future notes in a sequence, CNNs are generally more suitable for music genre classification because they are better at capturing the short-term temporal features that are important for distinguishing between different genres.

E. Representational State Transfer (Rest) API

REST API is basically used for creation of web server. It works by defining a few constraints and is mostly used architectural style. REST APIs are simple and flexible ways to access web services without processing.

REST can work over less bandwidth, a simple and mostly preferred Simple Object Access Protocol. It is a suitable system for working on Internet used to obtain or provide some information from a web service.

REST API communicates only through HTTP request. Client sends request to the server as an URL HTTP PUT, GET, DELETE command. The server then returns the response as a resource such as HTML, XML, image, or JSON. Despite all people prefer to use JSON for web services applications over all.

F. Problem Statement

As mentioned earlier, music genre classification is not a simple and straight forward task. Therefore, using neural network can be helpful to automate the classification task effectively and efficiently. Convolutional neural networks are a category of deep neural networks that are widely used for visual image analysis. Here, we approached the statement to classify music genre based on its spectrographic representation using CNN.

IV. METHODOLOGY

In the proposed work, the workflow consists of collecting properly classified music dataset based on its genre, preprocessing of the data and feature vector extraction from the preprocessed data, model creation and classification and finally the deployment of the model. The flow diagram of the work is as shown below (Fig 1)—

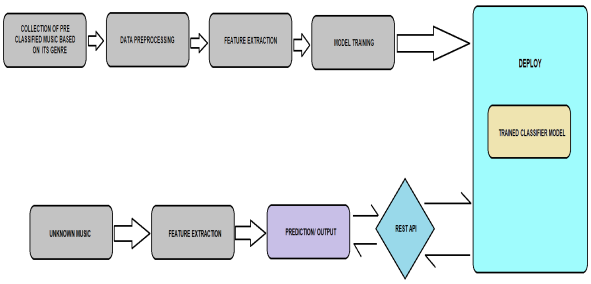

Fig 1. Methodology Flow Diagram

A. Music Dataset

The most commonly used public dataset to evaluate in machine hearing studies of music genre recognition is the GTZAN dataset. In this dataset, the music files were collected from 2000 to 2001 from a variety of sources that includes radio and microphone recordings that represent different recording conditions.

GTZAN classification was performed using a genre collection record. The dataset was obtained from the widely-used software framework called MARSYAS (Music Analysis, Audio Signal Acquisition and Synthesis). MARSYAS is an audio signal processing opensource software mainly used in information retrieval from music. The creator of this MARSYAS was George Tzanetakis who himself designed the framework. This platform is important for both academic and industrial projects.

The dataset is a collection of 1000 audio tracks with 30 second duration clips. The collection consists of 10 genres as listed in column 1 of table 1 below. Each of these genres comprised of 100 (hundred) tracks. These tracks are 22050 Hz, monaural, 16-bit audio files in .wav format. The data representation is as shown in Table I—

TABLE I. TABLE FOR DATA DISTRIBUTION

Genre	Music Count
reggae	**100**
classical	**100**
hip-hop	**100**
disco	**100**
country	**100**
jazz	**100**
metal	**100**
pop	**100**
Blues	**100**
rock	**100**

B. Data Pre processing

A total of 1000 tracks for feeding a neural network is not sufficient for a better classification. Therefore, the number of samples must be increased. In order to increase the number of samples, each track is split into 10 equal sized tracks. Now total numbers of samples are 10000. Each genre contains 1000 number of samples, each of a duration of 3 seconds. Now the data distribution is as follows (Table II)—

TABLE II. DATA DISTRIBUTION AFTER PREPROCESSING

Genre	Music Count
reggae	**1000**
classical	**1000**
hip-hop	**1000**
disco	**1000**
country	**1000**
Jazz	**1000**
Metal	**1000**
Pop	**1000**
Blues	**1000**
Rock	**1000**

C. Feature Vector Extraction

After splitting the data samples, feature vector is extracted using the Librosa library in python. Librosa is a python library specifically used for music/audio processing/analysis. It offers the necessary components to build a music information retrieval system which involves extracting a feature vector from each audio file, known as the MFCCs i.e., Mel-Frequency Cepstral Coefficients.

MFCCs are derived from the power spectrum of a sound signal and are used to represent the spectral envelope of a sound in a compressed form, which can then be used for tasks such as speech recognition, music genre classification and speaker identification. It basically models the characteristics of the human voice. This feature is the largest part of the final feature vector. The feature implementation steps are described below—

- Dividing the original signal into several equal short frames. The primary aim of doing this step is to keep the audio signal constant. Calculating the periodogram for each short frame of a signal can provide an estimate of the power spectrum for that frame. This allows us to analyze the frequency content of the signal over time.
- Pushing the power spectra into the mel filter bank and collecting the energy in each filter to sum it up. Then energy existing in the various frequency regions will be known. Following equation is generally used to compute Mel values of frequency spectrum.

$$M(f) = F^{-1}[\log (F(f))]$$

The spectral view of MFCC feature vector for each music genre is as shown below (Fig 2)—

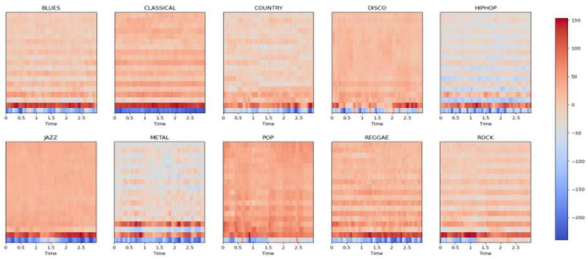

Fig 2. Spectral visualization of MFCC feature vector of each genre

D. Model Set up and Classification

In the proposed work, a sequential convolutional neural network model is used. The network model consists of three main types of layers to build the convolutional network architecture— Convolutional Layer, Max-Pooling Layer and Fully Connected Layer. ReLU activation function is used in the convolutional and the max-pooling layers and Softmax activation is used for the final output layer. In the network, Sparse Categorical Entropy is used as loss function and Adam optimizer is used for optimization of the model. The model architecture is as mentioned in Table III—

TABLE III. MODEL ARCHITECTURE

Layer	Type	Neurons	Kernal Size
1	Convolution (ReLU)	32	3x3
2	Max Pool	32	3x3
3	Batch Normalization	-	-
4	Convolution (ReLU)	32	3x3
5	Max Pool	32	3x3
6	Batch Normalization	-	-
7	Convolution (ReLU)	32	2x2
8	Max Pool	32	2x2
9	Batch Normalization	-	-
10	Flatten	-	-
11	Dense (ReLU)	64	-
12	Dropout	-	-
13	Dense (Output) (Softmax)	10	-

In the proposed work, supervised learning approach is used for classification. The model training parameters are mentioned in Table IV—

TABLE IV. MODEL TRAINING PARAMETERS

Parameters	Values
Training Size	**8000**
Test Size	**1000**
Validation Size	**1000**
Batch Size	**32**
Epochs	**100**
Learning Rate	**0.0001**
Total Classes	**10**
Dropout	**0.3**

V. RESULTS

On training the model, training and validation accuracy is calculated. Along with it, training loss and validation loss is also calculated with respect to epochs. The accuracy is calculated with the formula that is mentioned below—

$$\text{Accuracy} = \frac{Correctly\ Classified}{Total\ no\ of\ Musics} \text{ x } 100$$

On training, a highest of 97.43% training accuracy is obtained along with a validation accuracy of 78.64%. Along with that, training loss is minimized to 0.08 and validation loss is minimized to 0.83. On the test dataset, a 80.50% of test accuracy is obtained.

The accuracy and loss graphs are shown below in Fig 3 and Fig 4 respectively—

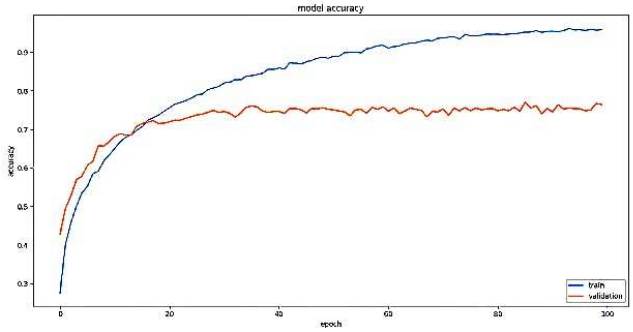

Fig 3. Training Accuracy vs Validation Accuracy

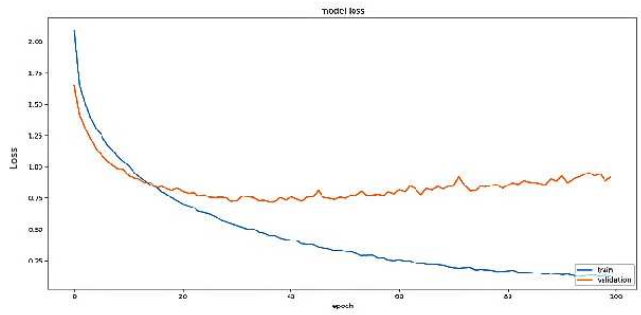

Fig 4. Training Loss vs Validation Loss

The confusion matrix is calculated as shown below in Fig 5—

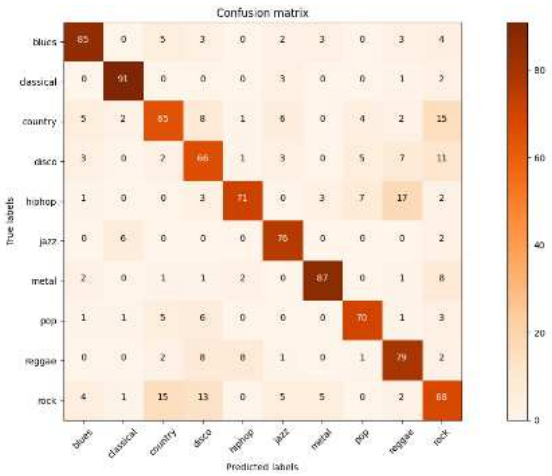

Fig 5. The Confusion Matrix

The different performance parameter map is shown in Table V below—

TABLE V. TABLE FOR DATA DISTRIBUTION

Label	Precision	Recall	F1-Score	Support
Blues	0.77	0.73	0.75	51
Classical	0.84	0.91	0.87	45
Country	0.65	0.67	0.66	48
Disco	0.58	0.67	0.62	48
Hip-hop	0.82	0.81	0.81	57
Jazz	0.84	0.88	0.86	49
Metal	0.87	0.84	0.85	49
Pop	0.83	0.88	0.85	56
Reggae	0.74	0.63	0.68	51
Rock	0.67	0.63	0.65	46

Now the trained classifier model is deployed in Heroku server. A REST API is also developed so that the classifier can be accessed and used easily. A MP3 to WAV converter is also developed and used for converting MP3 codec to WAV codec as Librosa was failed to load some MP3 codecs.

The deployed model output is as follows (Fig 6)—

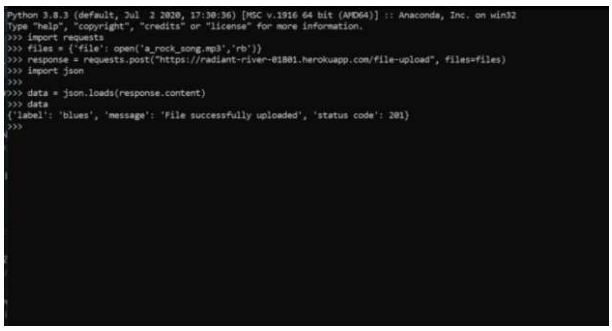

Fig 6. The API output

In the previous work by S. Vishnupriya and K. Meenakshi [2], they have also followed similar kind of methodology and they achieved a training accuracy of 76% in 100000 iterations and a batch size of 64 using learning rate of 0.001.

They have used a total of 1000 music of which 800 for training and 200 for testing. In our work, by increasing the size of the dataset and modifying some basic parameters we are able to achieve better result in comparison with the other studies conducted with similar methodology.

VI. CONCLUSION AND FUTURE WORK

This work provides a CNN based music genre classifier which is more accurate than the other Neural Network based classifiers. As already discussed, the feature vector was created using Librosa from python library and thus we were able to get some prominent feature values from the sample set. The training, validation and testing accuracies are obtained as 97.43%, 78.64% and 80.50% using the MFCC feature vector respectively. Therefore, this approach shows promise for categorizing a large database of songs into their respective genres.

The future work will include the effect of different music genre on human brain. Thus, further classification can be done and can be used for improving logical thinking ability, reducing stress etc. A robust music genre classification tool can be used for meditation of music. So, this work has importance both in medical and entertainment fields.

REFERENCES

[1] Tzanetakis, George & Cook, Perry. (2002). Musical Genre Classification of Audio Signals. IEEE Transactions on Speech and Audio Processing. 10. 293-302.

[2] S. Vishnupriya and K. Meenakshi, "Automatic Music Genre Classification using Convolution Neural Network," 2018 International Conference on Computer Communication and Informatics (ICCCI), 2018, pp. 1-4, doi: 10.1109/ICCCI.2018.8441340.

[3] Xu, Changsheng & Maddage, Namunu & Shao, Xi & Cao, Fang & Tian, Qi. (2003). Musical genre classification using support vector machines. 5. V - 429. 10.1109/ICASSP.2003.1199998.

[4] Matthew Creme, Charles Burlin, Raphael Lenain: Music Genre Classification

[5] Zhengxin Qi, Mohamed Rahouti, Mohammed A. Jasim, and Nazli Siasi. 2022. Music Genre Classification and Feature Comparison using ML. In 2022 7th International Conference on Machine Learning Technologies (ICMLT) (ICMLT 2022). Association for Computing Machinery, New York, NY, USA, 42–50. https://doi.org/10.1145/3529399.3529407

[6] Kumaraswamy, Balachandra & Shukla, Tushar & Swati, & Satyam, Kumar. (2021). Music Genre Classification for Indian Music Genres. International Journal for Research in Applied Science and Engineering Technology. 9. 10.22214/ijraset.2021.37669.

[7] T. Feng. Deep learning for music genre classification. 2014.

[8] Archit Rathore, Margaux Dorido: Music Genre Classification.

[9] Kostrzewa, D., Kaminski, P., Brzeski, R. (2021). Music Genre Classification: Looking for the Perfect Network. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M.A. (eds) Computational Science – ICCS 2021. ICCS 2021. Lecture Notes in Computer Science(), vol 12742. Springer, Cham. https://doi.org/10.1007/978-3-030-77961-0_6

[10] Brian McFee, Alexandros Metsai, Matt McVicar, Stefan Balke, Carl Thomé, Colin Raffel, Frank Zalkow, Ayoub Malek, Dana, Kyungyun Lee, Oriol Nieto, Dan Ellis, Jack Mason, Eric Battenberg, Scot