**SANDEEP SETTY**
**316506402039**
**M.TECH (CST)**

# MUSIC CLASSIFICATION USING CONVOLUTION NEURAL NETWORK

**ABSTRACT:**

In recent years, the complexity of music production has gradually decreased leads to many of us create music and upload their created music to streaming media. The massive music streaming media has caused people to spend much time trying to find specific music. Therefore, the technique of quick classification of music genres is extremely important in today's society. As machine learning and deep learning technologies maturing, the Convolutional Neural Networks (CNN) are applied to several fields, and various CNN based variants have emerged one after another. The normal genre classification requires relevant professional knowledge to manually extract features from statistic data. Deep learning has been proven to be effective and efficient in statistic data. So as to save lots of the user's time when attempting to find different types of music, we applied CNN's advantages and characteristics in audio to implement a style classification model. Within the pre-processing, we use Librosa is employed to convert the first audio files into their corresponding Mel spectrums. The Mel spectrum is given to the proposed CNN model for training. The bulk voting is applied to the choices made by the ten classifiers, and also the GTZAN dataset is employed.

**KEYWORDS—** GTZAN dataset, Mel-spectrograms, Librosa etc.,

## INTRODUCTION:

Music genre refers to the categorization of music on the idea of interaction between artists, economic process, and culture. It helps to arrange music into collections by indicating similarities between compositions or musicians. Automatic genre classification is non-trivial because it is difficult to tell apart between different genres. Expressive style classification plays an important role in today's world because of the zoom-in music tracks, both online and offline. Many machines learning techniques are used for genre classification. Here convolution neural network is employed for training and classification. The proposed system has two steps are required in music classification. The primary step is to extract the audio features of the input music, and also the second step is to construct a classifier through these features. During this study, we use the Mel spectrum to simulate human perception. This experiment uses the GTZAN

data set as training and verification, which Cook established, the aim is to check the application of machine learning to the classification of music genres. There are 10 genres, each of them with 100 pieces of music, and a complete of 1000 pieces. In recent years, the increase in machine learning and deep learning has resulted in the outstanding of Convolutional neural networks (CNN). Many improvements supported by CNN are emerging in endlessly. The CNN have excellent effects on unchangeable sequences or missing elements. The audio is one in all the information within the unchangeable sequence. If the arrangement or the weather is changed within the audio sequence, the new audio and also the original audio are different files. CNN has been applied to unravel various complex audio problems, for instance, sentiment analysis, feature extraction, genre classification and prediction. The CNN model is additionally widely employed in materials like audio signals and word sorting. Therefore, we propose a way of using Convolutional neural networks for the identification of various music styles.

**EXISTING MODEL AND ITS DRAWBACKS:**

Music recommendations are one in all the important things, like music streaming platforms. Classification of music genres is one in all the important initial stages within the process of music recommendation supported genre. Many music classifications are proposed by extracting audio features that need a not light computing process. This research aims to research and test the performance of style classification supported metadata using three different classifiers, namely Support Vector Machine (SVM) [3][4] with radial kernel base function (RBF), K Nearest Neighbors (K-NN) [5], and Naive Bayes (NB).

Drawbacks: SVM, KNN algorithm isn't suitable for giant data sets. SVM, KNN doesn't perform alright when the information set has more noise i.e., target classes are overlapping. In cases where the number of features for every datum exceeds the quantity of coaching data samples, the SVM will underperform. Because the support vector classifier works by putting data points, above and below the classifying hyper plane there's no probabilistic explanation for the classification. Naive Bayes assumes that every one predictor (or features) is independent, rarely happening in reality. This limits the applicability of this algorithm in real-world use cases.

**PROPOSED MODEL:**

To overcome the all the disadvantage in existing system we'll use the Convolutional neural network. The Convolutional Neural Networks (CNN) is applied to several fields, and various CNN based variants have emerged one after another. We applied CNN’s advantages and characteristics in audio to implement a genre classification model. Convolutional neural networks are basically composed of a Convolutional layer, pooling layer, and fully connected layer. The principle of the Convolutional layer is to get the local features of the audio or picture through a window with a specified size (Convolutional kernel) by sliding up and down sequentially. Next, through the Activation function, the feature map is generated because the input of the subsequent layer. The function of the pooling layer is to cut back the dimensions of the input audio or picture to cut back the dimension of every Feature map and

retain important features. The fully connected layer may be used as a general neural network, which is able to classify after receiving the feature information of the previous Convolutional layer and pooling layer. Neurons within the fully connected layer are only connected to the pixels of the previous layer of kernel, and also the weight of every link is that the same and shared within the same layer. Within the pre-processing, we use Librosa is used to convert the primary audio files into their corresponding Mel spectrums. The Mel spectrum is given to the proposed CNN model for training. The majority voting is applied to the alternatives made by the ten classifiers, and also the GTZAN dataset is used.

**METHODOLOGY:**

The expressive style classification using convolution neural network. During this we are having two phases. Within the first phase music file is converted into the wave file so it converted into Mel-spectrograms. Within the second phase spectrograms are given to convolution three layers like Convolutional layer, pooling layer and fully connected layer. The output from these layers is given to the classification using softmax function. Refer to fig2.

**A. MEL-SPECTROGRAMS:**

In sound processing, the Mel-frequency cepstrum (MFC) may be a representation of the short-term power spectrum of a sound, supported a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency.

1. We took samples of music file to digitally represent an audio signal.

2. We mapped the audio signal from the time domain to the frequency domain using the fast Fourier transform, and that we performed this on overlapping windowed segments of the audio signal.

3. We converted the y-axis (frequency) to a log scale and also the color dimension (amplitude) to decibels to create the spectrogram.

4. We mapped the y-axis (frequency) onto the Mel scales to create the Mel-spectrogram.

**B. CNNMODEL:**

In the example model below (refer fig 1), a Convolutional Layer unit is that the portion that learns the interpretation invariant spatial patterns and their spatial hierarchies. The Max Pooling Layer halves the scale of the feature maps by down sampling them to the max value inside a window. Why down sample? Because otherwise it might lead to a generous number of parameters and your computer would widen and finally that the model would

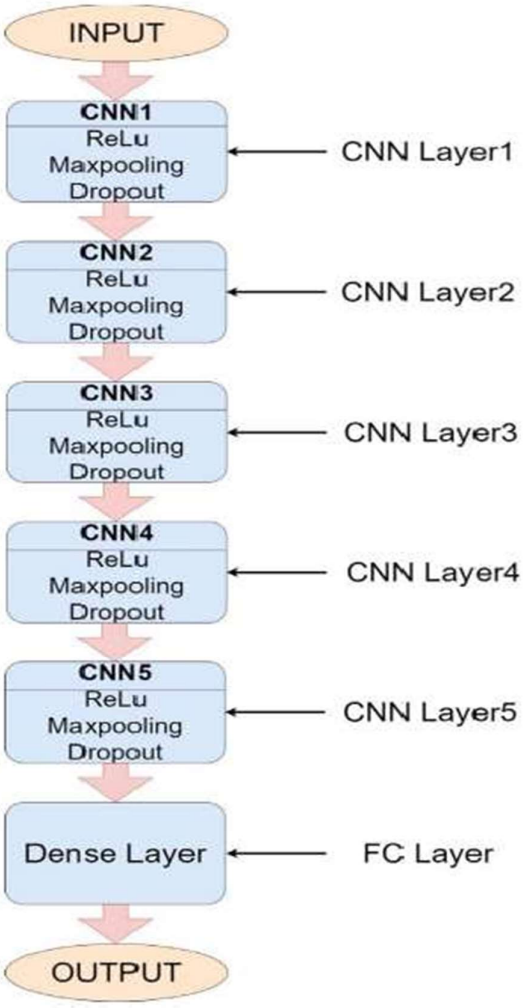


Fig 1: Proposed CNN model

**DATASET:**

For this project, the dataset that we'll be working with is GTZAN Genre Classification dataset which consists of 1,000 audio tracks, each 30 seconds long. It contains 10 genres , each represented by 100 tracks. the ten genres are as follows: Blues, Classical, Country, Disco, Hip-hop, Jazz, Metal, Pop, Reggae, Rock. The dataset has the subsequent folders: Genres original a group of 10 genres with 100 audio files each, all having a length of 30 seconds (the famous GTZAN dataset, the MNIST of sounds). Images original — a visible representation for every audio file. a technique to classify data is thru neural networks because NN's usually soak up some variety of image representation.2 CSV files — Containing features of the audio files. One file has for every song (30 seconds long) a mean and variance computed over multiple features which will be extracted from an audio file. The opposite file has the identical structure, but the songs are split before into 3 seconds audio files.

**MODEL ARCHITECTURE:**

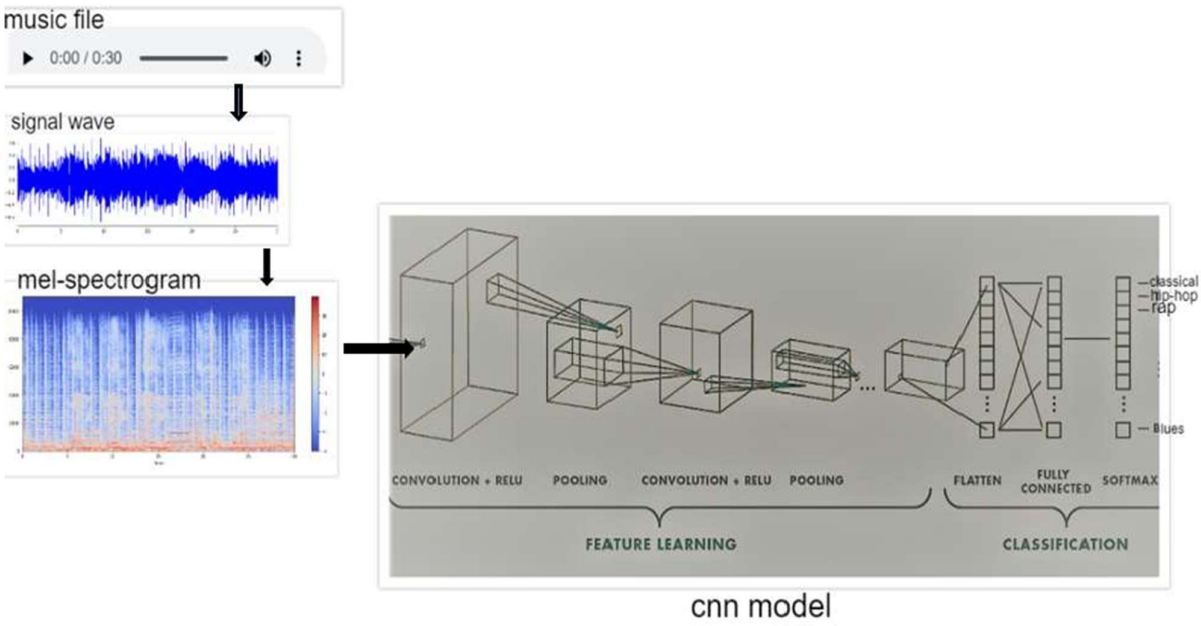


Fig 2: model architecture

**RESULT:**

First of all, this research splits the GTZAN dataset into 70% for training set and 30% for test set, and convert all the audio in the dataset into their respective MFCC and send them to the proposed CNN model for training. Librosa is a tool for audio signal processing, we use it for audio conversion in preprocessing to obtain the spectrogram we need. Fig 3.1explains model summary. The experiment was performed under a Google's cloud servers, it is equipped with 64GB of memory for training actions. The total number of iterations performed in the experiment is 20,338, batch size is set to 128, epochs is set to 600, and the experiment execution time is 45 minutes. ADAM is an optimizer for controlling the learning rate, it can iteratively update neural network weights and optimize the objective function based on training data. Therefore, we also used ADAM into our architecture. Fig 3.2 presents model validation graph of Train and test data by taking 5min time travel and price with respective axis. The architecture proposed in this study has an accuracy rate of 93.3%.

Model: "sequential"

| Layer (type) | Output Shape | Param # |
|---|---|---|
| dense (Dense) | (None, 512) | 30208 |
| dropout (Dropout) | (None, 512) | 0 |
| dense_1 (Dense) | (None, 256) | 131328 |
| dropout_1 (Dropout) | (None, 256) | 0 |
| dense_2 (Dense) | (None, 128) | 32896 |
| dropout_2 (Dropout) | (None, 128) | 0 |
| dense_3 (Dense) | (None, 64) | 8256 |
| dropout_3 (Dropout) | (None, 64) | 0 |
| dense_4 (Dense) | (None, 10) | 650 |

Total params: 203,338
Trainable params: 203,338
Non-trainable params: 0

Fig 3.1 model summary

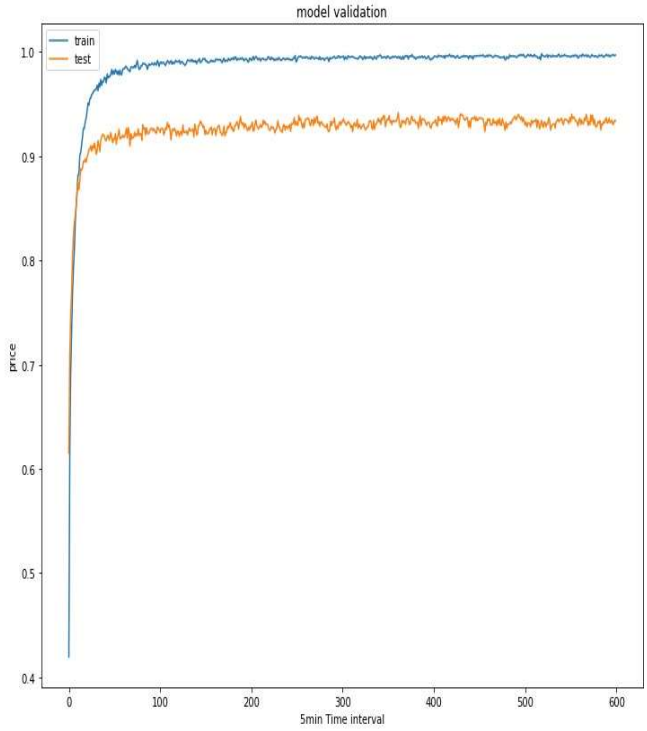


Fig 3.2 model validation graph

**CONCLUSION:**

Music genre classification can help users find the music they are interested, especially specific musicians and beginners. Because they're unaccustomed music and comparatively unfamiliar with various music styles, it takes plenty of your time to find a selected form of music from streaming media and this has caused in efficiency. For a particular musician, when looking for the music of the required genre, if he listens for a protracted time and judges the genre, the sound will grow tired and therefore the judgment will fail, and that they also spend plenty of your time searching for music. Therefore, an expressive style classification tool may be a time-saving method for these people. The best accuracy of our proposed Convolutional neural network for the expressive style classification is 93.3%, it'll help the longer-term work of music genre classification. Within the future, we'll continue the model and integration of streaming media and web crawlers to mix our CNN architecture to create it more complete, help music beginners and specific musicians shorten their time and increase efficiency.

**REFERENCES:**

[1].Data is available on tensorflow.org in catalog.

[2].De rosal Ignatius moses setiadi, Dewangga satriya rahardwika, eko hari rachmawanto, “comparison of SVM, KNN and NB classifier for genre music classification based on metadata”, In proceedings of the IEEE 2020 International seminar on application for technology of information and communication.

[3]. E. Chaudary, S. Aziz, M. U. Khan and P. Gretschmann, "Music Genre Classification using Support Vector Machine and Empirical Mode Decomposition," 2021 Mohammad Ali Jinnah University International Conference on Computing (MAJICC), 2021.

[4]. Changsheng Xu, N. C. Maddage, Xi Shao, Fang Cao and Qi Tian, "Musical genre classification using support vector machines," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003.

[5]. N. Ndou, R. Ajoodha and A. Jadhav, "Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches," 2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 2021.