

Tamil Filmy Music Genre Classifier using Deep Learning Algorithms

Christina Gnanamani¹ and Shanthini Pandiaraj²

¹Department of ECE, Coimbatore Institute of Technology, TN, India
<https://orcid.org/0000-0003-3797-6165>

²Department of ECE, Karunya Institute of Technology and Sciences, TN, India
<https://orcid.org/0000-0003-1743-9661>

Abstract—The objective of this work is to classify the genre of the given Tamil music using CNN algorithm. Music plays an important role in human's life. Due to human evolution, the way of listening music and music catalogues also get advanced. So, this classifier technique helps the user to find the frequently used genre of that particular Tamil music. There are billions of songs in this world, so it is very difficult to classify the Tamil music genre of the songs by manual process for music streaming companies. Thus an accurate and predefined classifier is required, which can be succeed with deep learning using convolution neural network (CNN) algorithms. In this project, Jupyter notebook is used to compile the python code. The Machine is trained using GTZAN dataset which consist of 1000 audio tracks with 10 different types of Tamil music genres. To predict the specific Tamil music genre of the given audio file the feature of both the train and test data will be extracted and classification will be done with Convolution neural network algorithm and the output will be displayed.

Index Terms— CNN algorithms, Jupyter notebook, GTZAN datasets.

I. INTRODUCTION

Music plays an important role in every human's life. Earlier, people used to listen music only for their pleasure and to express their feelings, but nowadays people listen to songs for pleasure and they also analyze the technical terms behind that song, including the instruments used and the singer. Music is universal and composers' worldwide plagiarize or use foreign instruments to enrich their music and provide variety for listeners. We all are familiar with the elements of the music such as Melody, Folk, and Western. But Tamil music genre has songs from genres such as Jazz, Blues, Hip hop, Metal and so on. People have loved and listened to these types of songs without knowing their music genre. There is also several music streaming companies which act as a best entertaining platform for music lovers. As a listener listens to a particular song on repeat mode, the app identifies the genre and suggests songs in the same genre. This can be suggested with our Tamil music genre Classifier. Tamil music genre classifier is a model which is able to classify frequently used genre of the particular song. This can be done with the help of Convolution Neural network algorithm (CNN) which is a deep learning approach. Deep learning is a subset of Machine learning which deals with large amount of data set to train the machine. The 2 GTZAN data set consist of train and test data. There are 1000 audio files with 10 different types of genre each genre consist of 100 audio files. Totally 1000 audio files are there, each audio is 30 seconds in duration and they are in '.WAV' format. The features of the training data set are extracted and they are trained

with help of convolution neural network model to the machine. To test the model we give an input song, the features of the input song are then extracted, and according to the features of input song that matches with the training data the output will be displayed. The different types of genre identified are Blues, Jazz, Pop, Hip hop, Classical, Reggae, Metal, Country, Disco, Rock.

II. CONVOLUTION NEURAL NETWORK

Machine learning uses algorithms to parse data, learn from that data, and make informed decisions based on what it has learned. Deep learning structures algorithms in layers to create an artificial neural network that can learn and make intelligent decisions on its own. In this work, the audio is converted into image as image spectrogram and the features of this data will be extracted in terms of image itself. So, the image process will be done with deep leading model. There are some differences between the artificial neural network (ANN) and convolution neural network (CNN). Both are same in terms of their classification, but while going with accuracy, convolution neural network will win. In the case of ANN, while flattening the input image (i.e. converting the pixels in to the vector) there will be loss of data and there reduction in spatial property of the image due to flattening and accuracy will be affected during the image translation. To avoid these issues, CNN algorithm will be more effective. The convolution, Relu layer and pooling layer are combined with each other forms the convolution neural network.

Hence the CNN plays an important role in classification. It is a fully connected layer to predict the actual output and the accuracy of CNN algorithm is comparatively high for classification of genre from others.

III. METHODOLOGY

Classification of genre is not an easy task without undergoing several processes. Those process will make the data as much as comfortable to train the model using particular method of classification, performance implications of various alternatives is central to this process. It is also extremely challenging. In the case of existing systems, measurement data is available. In the case of evolving systems, contemplated modifications often are straight forward, and limited iteration may be possible in validating a baseline model. In the case of proposed systems, these advantages do not exist. For this reason, it is tempting to stay on performance projections, which all too often prove to be significantly in error. The consequences can be serious, for performance, like reliability, is best designed in, rather than added on. Here our proposed model involves several process such as, splitting of dataset into training and testing. Both training and testing data are subjected to extracting their features. After extracting, training data is used to train the model using CNN algorithm and test data is used to test the trained model whether their prediction is appropriate to the expectation, this parameter can be measured using accuracy. A well versed model must have its accuracy rate comparatively higher than other model .To achieve this CNN plays a major role in our proposed system model.

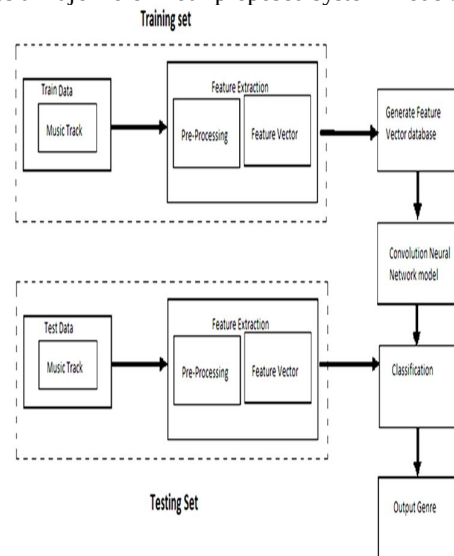


Fig. 1. Workflow of the model

A. Feature Extraction

Feature extraction plays an important role in the genre classification. Based on the different genre, their features are varied. Music are differentiated with three main parameters - beat, pitch, scale. We can extract these parameters into a valuable data in terms of spectrum. To visualize the spectrum of the feature, they were converted into image spectrogram. According to the variation in features of the Tamil music genre, the image spectrogram will be changed then they subjected to convolution process and output prediction. The feature that extracted are explained below.

Zero crossing rate indicates the number of times the wave crosses the value zero. We can also say the rate of sign changes in the given frame of time. Zero crossing rate is often employed in speech, music and audio classification process, with consideration of main property of the audio file. The taken audio file is 30 sec in duration. Training is stopped when the higher value where the maximum changes that satisfies the audio clip is reached.

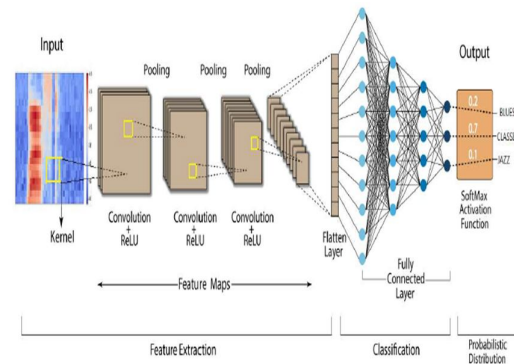


Fig 2 Complete process of Convolution neural network

Chroma is a main feature that differentiates the song or music from an audio. Chroma is technical based harmonic analysis of a music which can recognize the chord of the particular song. It is a main task that has advantage in music information retrieval. Chroma feature can identify the vast changes in the chord or scale between those ten different types of Tamil music genre. Each genre should have various changes in their scale. In this work, we consider the case of audio representations where a piece of music is given in the form of an audio performance. Here, the chord recognition task consists in first splitting up the audio into segments and then assigning a chord label to each segment. The segmentation specifies the start time and end time of a chord, and the chord label specifies which chord is played during this time period. The spectrogram usually displays time in the x-axis (horizontal) and frequency in the y-axis (vertical). In the chromagram, notes are displayed in the vertical axis. Usually, windows (or time slots, or time frames) with some amount of time samples (e.g., a window of 1 ms of a signal sampled at 44.1 kHz has roughly 44 samples) are FFT and the spectrum (or spectral frame, i.e. the FFT result) is displayed in the chromagram in vertical position. The frequently used pitch at the particular time for the given audio spectrogram and the unused area of pitch at the particular period of time is noted.

Mel-frequency cepstral coefficient is a scale that compares the perceived frequency of the tone with the actual frequency. The MFCC gives a discrete cosine transform (DCT) of a real logarithm of the short-term energy displayed on the Mel frequency scale. In simple terms, MFCC is used in detection and identification the pitch of the song. The MFCC feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. To obtain back the hidden time information, the modified Fourier transform STFT is used. This transform is a compromise between a time and a frequency based representation by determining the sinusoidal frequency and phase content of local sections of a signal as it changes over time. In this way, the STFT extracts data about which frequencies are contained in the signal and also at which points of times or, to be more precise, in which time intervals these frequencies appear.

Spectral rolloff means in every song there will be sudden changes or sudden reduction in frequency that differs for each genre in different frame of an audio. Spectral rolloff is the frequency below which a specified percentage of the total spectral energy or 85%. Time function that is relatively smooth tends to have higher

frequency rolloff rate and narrower bandwidth. Time function that discontinuous in signal will have lower frequency roll of rate and wider bandwidth.

Spectral centroid is the one of the characteristic of the audio spectrum. It indicates that where the center of mass of the spectrum is located. The spectral centroid of a signal is the curve whose value at any given time is the centroid of the corresponding constant time cross section of the audio's spectrogram. A spectral centroid provides a noise robust estimate of how the dominant frequency of a signal changes over time.

B. Preprocessing

Audio preprocessing, feature extraction and feature analyzing are steps we take to extract information from the present data and information that in a machine learning context should be useful for predicting the genre of a sample or the value of some targeted attributes. In audio analysis this process is largely based on the components of an audio signal that can help the model to differentiate one audio from the other. Meanwhile they should have two different features. By the above feature extraction mentioned, the zero crossing, chroma, MFCC are the some of the tools that hopefully classifies and differentiates the audios. In preprocessing the extracted features are arranged in a comprehensible manner by removing the unwanted noise from the data.

C. Dataset

The main component in machine and deep learning to train the model is dataset. Here we have used GTZAN dataset to train the model.

Genre	Number of Records	Genre	Number of Records
blues	100		
classical	100		
country	100		
disco	100		
Hip-hop	100		
jazz	100		
metal	100		
pop	100		
reggae	100		
rock	100		
TOTAL	1000		

Fig. 3 Split-up of GTZAN dataset

The GTZAN dataset appears in at least 100 published works, and is the most-used public dataset for evaluation in machine listening research for music genre recognition (MGR). In this dataset, there are 10 different music genre. Each genre consist of 100 audio files and totally 1000 audio files. Each audio consist 30 seconds of duration. The tracks are all 22050Hz Mono 16-bit audio files in .wav format. The memory is 1.2GB.

The entire real dataset is not completely subjected to the model. It is separated into two sets, training and testing dataset. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing. Analysis Services randomly samples the data to help ensure that the testing and training sets are similar. By using similar data for training and testing, one can minimize the effects of data discrepancies and better understand the characteristics of the model. After a model has been processed by using the training set, we test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that we want to predict, it is easy to determine whether the model's guesses are correct. By this method, the model can define the accuracy of the proposed system model. In real time model, the accuracy may differ according to the valid input to be predicted, because the test data will not be same for every time. A best model will give the best accuracy for the given input data.

D. Software tools

The Python Kernel runs the computations and the results are obtained But in music genre classification it is difficult to train the model in jupyter IDE. Because the dataset size is large and is about 1.2 GB. So, Google Colab is used to train the datasets. In this domain, we upload the datasets in drive and access that datasets using clouds to Google Colab, it will be very easier to train the datasets. We have used lots of python library packages. They are Tensorflow, Keras, Scikit-Learn, Tkinter, Librosa, Numpy, Pandas and Matplotlib.

IV. OUTPUT

After the completion of training the machine using the proposed system model it starts to predict the genre of the given input song. The input audio is taken from YouTube, as will be done by a user. While playing the song it automatically recognizes the song and can be seen in audio visualization. And the genre prediction in 2 different cases is explained as follows.

A. Case 1

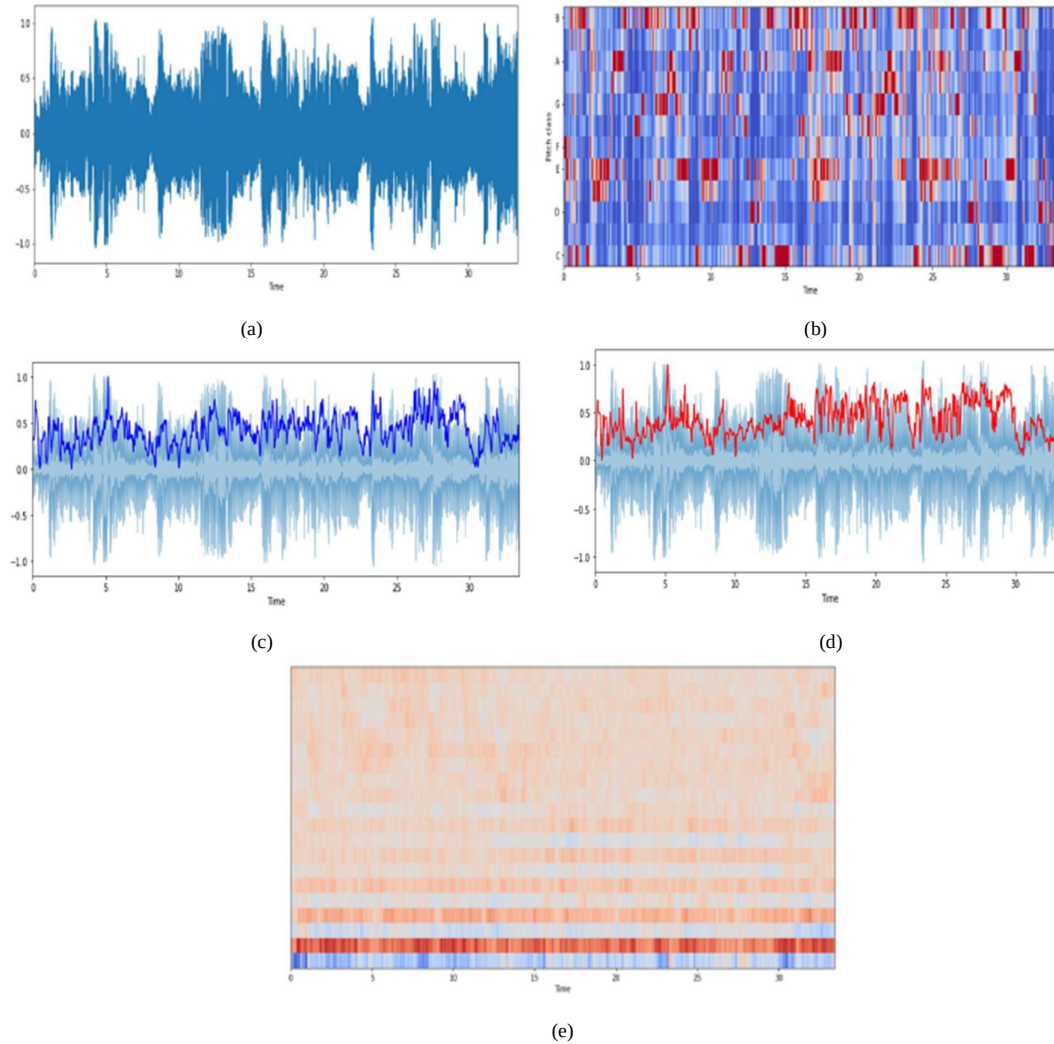


Fig 4. Spectrogram image (a) zero crossing rate of the particulate time period of the song
 (b) the chroma feature of this song , the region that filled with red in colour is indicates that the maximum used note for this song hence the chord used are Am , Em (E minor) which are suitable for classical song.
 (c) is spectral roll-off of the audio clip.
 (d) is spectral centroid.
 (e) shows the MFCC for this songs

Once the song starts to play, the model starts to predict the output in the form of value horizontal bar graph. Here the graph shows the 92% of classical, hence the song that played is a classical Tamil music genre.

B. Case 2

Similarly, this time by playing another song, the model shows the output as 74% of Reggae. Hence the frequently used genre of this particular song is reggae.

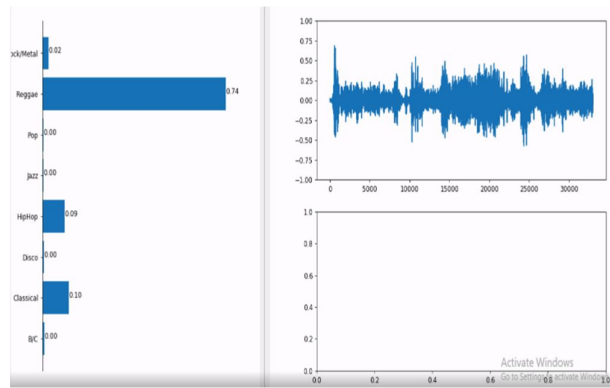


Fig.4.Output for case 2

V. RESULTS AND DISCUSSION

Confusion matrix is another simple way of analyzing the accuracy. The predicted output is in x-axis and actual (true output) is in y-axis. The all are done with train and test data, since test data is subset of train data, we could create this matrix. The matrix explains that in x-axis the model predicts that 15% the given song is blues but when actually it is a song with rock Tamil music genre. And it predicts 57% the input song is blues when actually the given song is blues. This continues for the further prediction. The highest value taken as the actual predicted output. This matrix will also defines the behavior of the trained mode.

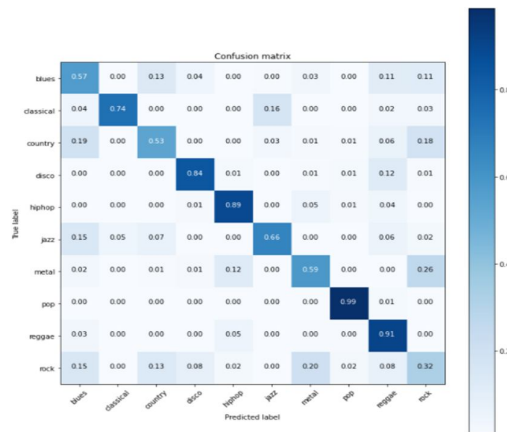


Fig.5. Confusion matrix

Performing results clearly explains that the prediction capabilities of a model in terms of accuracy, precision, recall and F1 score. Where F1 score depicts how well the model is performing. It is basically harmonic mean of precision and recall. To get the high F1 score, both the precision and recall value must be high. After training the model with the multiple iteration and increased its accuracy by passing the value with multiple epochs, we have achieved 73% of the model accuracy, which is higher than that of the conventional model which trained with artificial neural network. Hence we overcome the drawback that present in the ANN using CNN algorithm.

TABLE I. RESULT TABULATION

Models	Accuracy	Precision	Recall	F1 score
CNN	0.73	0.77	0.69	0.73
ANN-Spectrogram model	0.68	0.66	0.52	0.68
KNN	0.65	0.62	0.52	0.65
SVM	0.63	0.59	0.50	0.64

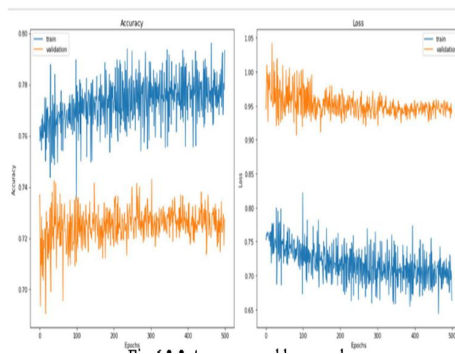


Fig 6 Accuracy and loss graph

VI. CONCLUSION AND FUTURE WORK

Here, a deep learning model based on ANN is designed to predict the genre of the songs that we are listening. By analyzing the literature survey it is conformed that Our CNN model is more accurate in classification of audio data. Obviously it took lot of time to train the machine but it is more efficient while using in real time. The main advantage of this model compared to KNN and SVM is that CNN is an unsupervised one, whereas others are supervised model i.e. user should label each data that is given as an input. As we are using large amount of data to achieve an effective classification of Tamil music genre, we need unsupervised model. The future scope of this algorithm is a vast. Music streaming company like Jio saavn, Spotify can apply this algorithm to analyze the frequently listened song by the user in their streams. According to the song one is listening, the frequently used genre can be identified and can be displayed in their suggestion feed. It will help the user to listen different songs in same genre that they love to listen. Correlating and listening to music of the same genre in different languages and cultures enhances the listener's experience.

REFERENCES

- [1] Wei Zhang, Xiaohui Chen, Yueqi Liu, Qian Xi(2020), "A Distributed Storage and Computation k-Nearest Neighbor Algorithm Based cloudedge Computing for Cyber-Physical-Social Systems", IEEE Access, Vol.8, pp. 50118-50130.
- [2] Kristina P. Sinaga, Miin-Shen Yang, "Unsupervised kmeans Clustering Algorithm", IEEE Access, Vol.8, pp.1-1, 2020
- [3] Xiaodan Lin, Jingxian Liu, Xiangui Kang, "Audio Recapture Detection With Convolutional Neural Networks", IEEE Transactions on Multimedia, Vol.18, pp.1-8, 2016.
- [4] G.Christina, A.Rajeswari, S.Mathivanan, "Real Time Analysis of a 24 ghz Planar Microstrip Antenna for Vehicular Communications", Springer-Wireless Personal Communications-Springer ,Volume 97 Issue:1, pp-1129-1139, May 2017.
- [5] Linyuan Fan, "Audio Example Recognition and Retrieval Based on Geometric Incremental Learning Support Vector Machine System", IEEE Access, Vol.8, pp. 78630-78638.
- [6] Christina Gnanamani, A.Rajeswari et al., "Design and Development of Wearable Antennas for Tele-Medicene Applications", IEEE International Conference on Communication and Signal Processing, April 6-8, 2016, Chennai, India
- [7] E.Wold, T. Blum, D.J.Keislar, "Wheaten, Content-based classification, search, and retrieval of audio," IEEE multimedia 3 (3), 27- 36, 1996.
- [8] Christina Gnanamani, S. Aishwarya, "RF Energy harvesting using 2 x 4 Circular Microstrip Array Antenna for 5G-iot Zero Power Wireless Sensors" at 2019 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications , Indonesian Science Expo, Indonesia, Oct 23-24, 2019.
- [9] F. Weninger, B. Schuller B, "Audio recognition in the wild: Static and dynamic classification on a real-world database of animal vocalizations," in: acoustics, speech and signal processing (ICASSP), IEEE international conference, IEEE, pp. 337-340, 2011.
- [10] M. V.Ghiurcau, C.Rusu, R.C.Bilcu, J.Astola, "Audio based solutions for detecting intruders in wild areas," Signal Processing 92 (3), 829-840, 2012.
- [11] Anitha, N. Dharshini, B. Raga Ravali, Shriya Chaurasia, G. Christina, "Low Cost Head Gesture Controlled Wheelchair for Quadriplegic Patients" Springer-Lecture Notes on Data Engineering and Communications Technologies book series, Volume 26, pp-1425-1435, Dec 2018
- [12] G.Christina, A.Rajeswari "Design of Compact Circularly polarized Antenna for GPS Application" in the International Conference on Evolving Trends in Engineering Economics, Health Systems Management and Business Communication , 1-2 March, 2018, CIT, Coimbatore.