Literature Survey for Music Genre Classification Using Neural Network

Naman Kothari[1], Pawan Kumar[2]

[1]PG Student, Department of Computer Application, JAIN (Deemed-To-Be) University Bangalore, Karnataka, India
[2]Assistant Professor, Department of CS and IT, JAIN (Deemed-To-Be) University, Karnataka, India

Abstract - *Music classification is a field in the field of Music Recovery (MIR) and sound signal processing research. Neural Network is a modern way of classifying music. The classification of music using Neural Networks (NNs) has been very successful in recent years. Various song libraries, machine learning technologies, input formats, and the use of Neural Network are all successful to varying degrees. Spectrograms produced from time song servants are used as an entry in the Neural Network (NN) to separate songs into their genres. The generated spectrograms will be used as a Convolutional Neural Network (CNN) audio signal input system. The Deep Learning Approach is used for system training. The main method of extracting audio features is the Mel Frequency Cepstral Coefficient (MFCC). The Mel-Frequency Cepstral Coefficients MFCC is a collection of short-term audio spectrum signals that have been used in advanced vision and sound separation techniques.*

***Key Words*: Music Classification, CNN, MIR, spectrogram, GITZAN, MFCC….**

1.INTRODUCTION

In recent years, machine learning has grown in popularity. Some machine learning techniques are best suited depending on the type of application and the available data. Of the various applications, some are more efficient than others. In general, there are four types of machine learning algorithms. The most common types of learning algorithms are: Unsupported reading, supervised learning, and supervised reading and reinforced reading.

Unattended reading aims to remove useful features from a blank labelled data set in mind, while supervised reading creates a mathematical model using a fully labelled data set. For slightly monitored reading, on the other hand, use a data set that includes both labelled and non-labelled information. Learning to strengthen takes a different approach by using a feedback approach. Learning to reinforce occurs when a person is rewarded for taking appropriate treatment or making appropriate predictions.

1.1 Neural Networks

A neural network (NN) is a machine learning method that is often effective in extracting key elements from large data sets and producing a work or model that reflects those features. First, NN uses the training database to train the model. After model training, NN can be applied to new or previously unselected data points to separate data using a previously trained model.

A convolutional neural network (CNN) is a type of central network designed to process the same members with multiple sides as images. CNN can be used for dual classification and classification functions in multiple categories, the only difference being the number of output classes. An animal data set, for example, can be used to train an image separator. CNN is provided with a pixel value vector from the image and an outline segment defined by the vector (cat, dog, bird, etc.).

The Deep Neural Network (DNN) is the most widely used diagnostic tool, and it assists in the training of a large gene expression (MFCC) website. These later released structures are used as training neurons. Due to the selection and release of acceptable audio elements, the separation of music is considered a difficult task. The classification of music is made up of two basic steps:

Extraction: Various features are extracted from the waveform.
Classification: A classifier is built using the features extracted from the training data.

1.2 Music classification

Music classification is a type of music retrieval function (MIR) where labels are assigned to music features such as genre, heart rate, and instruments. It is also associated with concepts such as musical similarities and musical tastes. People have known about music since the beginning of time. The concept of classification of music allows us to distinguish between different genres based on their composition and frequency of hearing. With the increasing variety of genres around the world, the classification of genres has recently become quite popular. The classification of genres is an important step in building a useful commendation system for this project. The ultimate goal is to create a machine learning model that can more accurately classify music samples into different genres. These audio files will be categorized based on the minimum frequency and time zone characteristics.

2. LITERATURE SURVEY

In the research conducted by Pelchat and Craig M. [4], The GTZAN dataset's songs were categorised into seven genres and used in [4]. The stereo channels were then combined into one mono channel, and the music data was converted into a spectrogram using the SoX (Sound eXchange) command-line music application utility, which was then sliced into 128x128 pixel images, and the labelled spectrogram was used as inputs to the dataset, which was split into 70% training data, 20% validation data, and 10% test data. All of the weights were now initialised using the Xavier initialization method. The first four layers are convolutional layers with a kernel size of 2x2 and a stride of two, followed by a max pooling layer. Following the first four layers is a fully connected layer in which each output of the previous layer is fed into each input of the fully connected layer. This yields a vector of 1024 numbers. After that, a SoftMax layer is applied to generate seven outputs, one for each genre. The CNN implementation appeared to be overfit, as the accuracy for the training data was 97% versus 47% for the test data.

In the survey conducted by K. Meenakshi [5], They used to train the system by categorising the music database into different genres. After that, each song must go through a pre-processing stage. Feature Vector Extraction is performed in Python using the librosa package, also known as MFCC. The Mel Scale Filtering is then performed to obtain the Mel Frequency Spectrum by [5]. They obtained two types of feature vectors: Mel Spectrum with 128 coefficients and MFCC coefficients. ConvNet architectures are built using three types of layers: Convolutional Layer, Pooling Layer, and Fully Connected Layer. The database thus obtained is the MFCC, with a genre array size of 10 arrays. The input consists of 1000 songs with ten labels. The vector of features for MFCC The Anaconda Python package was used for the evaluation. The learning accuracy of the Mel Spec feature vector and the MFCC feature vector was found to be 76% and 47%, respectively.

Research conducted by Nirmal M R [2]. They did use a spectrogram, which is a visual representation of a signal's frequency spectrum as it varies over time. A spectrogram is a signal's Short Time Fourier Transform (STFT). Spectrograms are graphs that represent data in both the time and frequency domains. Two models are used to implement the convolutional neural network, and their results are compared in [2].

- User-defined sequential CNN model
- Pre-trained ConvNet (MobileNet)

The classification accuracy of the user-defined CNN model and MobileNet was 40% and 67%, respectively. Python 3 was used in the LINUX operating system. The deep learning model used in [2] was built-in Python Keras framework.

The methods used in [3] the stages of the proposed method were used in their research. Data collection and selection, pre-processing, feature extraction and selection, classification, evaluation, and measurement are the following methodologies used in [3]. This research makes use of the Spotify music dataset, which contains 228,159 songs across 26 genres and 18 features. To process data structures and perform data analysis, the Python programming language is used, along with the Python Data Analysis Library (PANDAS). In addition, Scikit Learn is used in this study, which is a package containing important modules for machine learning projects. Later on, this genre feature will be used to classify the target. This is done especially for the learning process that uses the SVM classifier. Using hyperparameter, the SVM-RBF classification was carried out by searching the grid for the best results. K-fold cross-validation with free random conditions and a comparison of 80% of training data and 20% of test data. The accuracy for the SVM, KNN, and NB was 80%, 77.18%, and 76.08%, respectively, in [3].

In the paper [6], Mingwen Dong states that they used deep learning, particularly convolutional networks (CNNs), which has lately been used successfully in computer vision and speech recognition. In the process of Musical Information Retrieval (MIR) one specific example is music genre classification. MFCC (Mel-frequency cepstral coefficients), texture, beat, and other human-engineered features have traditionally yielded 61% accuracy in the 10-genre classification task. They used a "divide-and-conquer" method to solve the problem: they split the spectrogram of the music signal into consecutive 3-second segments, made predictions for each segment, and then combined the predictions. [6] used the mel-spectrogram as the input to the CNN to further reduce the dimension of the spectrogram. For Prediction and Training 1000s of music tracks (converted to Mel-Spectrogram) are evenly divided into training, validation, and testing sets in a 5:2:3 ratio. During testing, all music (Mel-Spectrogram) is divided into 3-second segments with 50% overlap. The trained neural network then predicts the probabilities of each genre for each segment in [6]. Their model is the first to achieve human-level 70% accuracy in the 10-genre classification.

Many researchers have worked on researching musical parameters and methods for classifying them into different genres. They have used a variety of NN approaches to try to replicate these skills, with varying degrees of success. Shazam is most renowned for being able to identify a song's title and artist after just a few seconds of listening. Shazam uses a mechanism known as a song's "signature," according to [2]. Shazam describes a song's trademark as the song's spectrogram's big peaks in amplitude. The use of deep learning, particularly convolutional networks (CNNs), has lately been used successfully in computer vision and speech recognition. In Music Information Retrieval (MIR), Using audio spectrogram and MFCC, a Deep Neural Network (DNN) was developed to improve music genre classification performance.

Spectrograms are Short Time Fourier Transforms (STFT) of a signal in the time and frequency domain. The spectrogram image is a good representation of the audio clip. Extracting features from such a large collection of music is a time-consuming procedure. The GTZAN dataset is a widely used musical genre database for model feeding. The GTZAN dataset was gathered by Tzanetakis and Cook [7] [14]. There are thousands of song snippets in ten different genres, relatively evenly distributed: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, and Rock.

TensorFlow is an implementation of a neural network as a deep convolutional neural network using TensorFlow [7]. All the weights can be initialized for the input vector, which will be a 128x128 pixel spectrogram. The first four layers are convolutional layers with a kernel size of 2x2 with a stride of two and a max-pooling layer after each successive layer. After the first four layers, there is a fully connected layer where each output of the last layer is fed into each input of the fully connected layer. A SoftMax layer is then applied to get different outputs that represent each genre [4].

3. SYSTEM DESIGN & ANALYSIS

In my research, the major categories of music genre to achieve the goal are: Dataset, Data Pre-Processing, Feature Extraction from Dataset and Training & Testing Model.

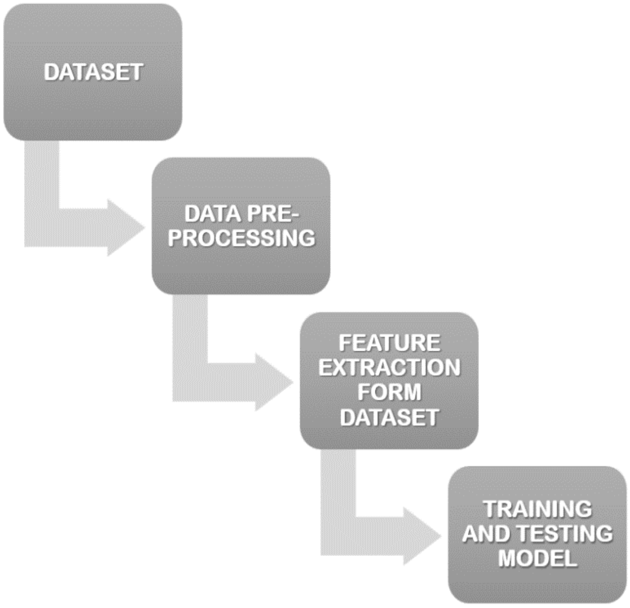

Fig -1: System Design Process

3.1 Dataset

The GTZAN Database is used to enter data into the system because it is a collection of free accessible songs from many genres. The collection is made up of thousands of audio tracks divided into 10 different genres. This database includes blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock music. Music data on the GTZAN database is taken at 22050 Hz and lasts for about 30 seconds, a total of 22020 x 30 = 661500 samples. For each a smooth window of 2048 samples, with a change of 1024 samples, as calculated during the study, all the results provided below are rated at more than ten runs, and the accuracy of the sections was selected as metric performance metrics.

MUSIC GENRE	NUMBER OF SONGS
Blues	1000
Classical	1000
Country	1000
Folk	1000
Hip-Hop	1000
Jazz	1000
Metal	1000
Electronic	1000
Reggae	1000
Rock	1000
TOTAL	**10000**

3.2 Data Pre-processing

At this stage, each music signal is first converted from a waveform to a mel-spectrogram with a time window of 23ms and passed through the Librosa package. The mel-spectrogram is then converted to a log to set values from different mel scales to the same range. The mel-spectrogram has a simpler understanding of the PCA-whitening method because it is a biologically inspired image [6]. Training and testing data were distributed, 80% of training data and 20% of test data distributed.

3.3 Feature Extraction from Dataset

The librosa module in Python is used for Feature Vector Extraction. This software is used for audio analysis only. Each audio file is extracted and the vector feature is calculated. By recording approximately, the type of log energy spectrum on the Mel-frequency scale, MFCCs incorporate timbral features of the music signal.

Mel Spectrogram: Mel-Spectrogram mimics human genetic makeup by producing a representation of the time frequency of sound. The magnitude spectrum is computerized from time series data and mapped to mel scale.

3.4 Training & Testing of Model

10,000 music tracks (converted to mel-spectrogram) are evenly divided into training, validation, and testing sets in a 5:2:3 ratio. The following is the training procedure:

a) Selection of a subset of track from the training set.

b) Take 3-second continuous segments from all selected tunes and sample a starting point at random.

c) Calculate the gradients using the back-propagation algorithm with the segments as input and the original music labels as target genres.

d) Update the weights using the gradients.

e) Repeat the procedure until the classification accuracy on the cross-validation data set no longer improves.

During testing, all music (mel-spectrogram) is divided into 3-second chunks with 50% overlap. The trained neural network then forecasts the odds of each genre for each section. The genre with the highest averaged probability is the genre predicted for each piece of music.

3.5 Convolutional Neural Network Model

Convolutional Neural Network (CNN) is made up of one or more dynamic layers followed by one or more fully connected layers, such as a normal neural multilayer network. This step involves passing a matrix filter (say, 3x3) over the inserted image, which is a size (image_width x image_height). The filter is first applied to the image matrix, and then the intelligent repetition of the element between the filter and the image area is calculated, followed by a summary to give the element value. The model is made up of four convolutional layers. Each layer is made up of a convolutional filter, a ReLU activation function, and a mass integration layer to reduce the size. There is a flat layer and a stop layer before entering the neural network. The flat layer converts the image tensor into a vector. This vector is a neural network input. To prevent overcrowding, a stopping layer is applied. The neural network is made up of a dense layer of 512 nodes and an outgoing layer with nodes equal to the number of classes to be separated.

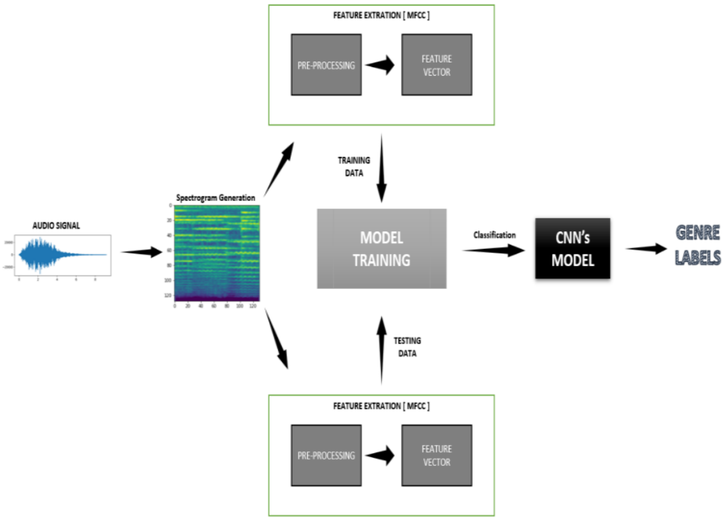

Fig. 2: *System Work flow Diagram*

4. CONCLUSIONS

This project presents a Music Genre Classification application that employs neural network techniques. There are a number of different audio feature extraction techniques, and it was decided that MFCC would be the best fit for this project. To train our model and perform classification on our dataset, we used KNN and CNN algorithms. This paper demonstrates a music genre classification system based on Neural Networks. Music classification is a type of music information retrieval (MIR) activity in which labels are assigned to musical elements such as genre, mood, and instrumentation. The librosa library, which is based on Python, assists in extracting features and hence in supplying acceptable parameters for network training. As a result, this system appears to be promising for categorizing a large database of music into the appropriate genre.

The CNN module was employed as the feature extractor to learn mid-level and high-level features from the spectrograms. CNNs are biologically inspired multilayer perceptron variations. In this work, we make use of Audio Set, which is a large-scale human annotated database of sounds. A music collection of songs was used, divided into several genres. Then, slicing up the larger spectrograms into 128-pixel wide PNGs, which represent 2.56 seconds of a given song. The GTZAN dataset will be used here. There are ten classes (10 music genres) with 1000 audio tracks each. The tracks are all in.wav format. It includes audio tracks from the ten genres listed below. Blues, classical, country, disco, hip hop, jazz, metal, pop, reggae, and rock are some of the genres represented. The GTZAN dataset has long been used as a standard for determining music genre classification, and MFCC spectrograms are also used to preprocess the tracks. MFCC spectrograms were used.

REFERENCES

[1] https://www.kaggle.com/andradaolteanu/gtzan-dataset-music-genre-classification

[2] Nirmal M R, "Music Genre Classification using Spectrograms", International Conference on Power, Instrumentation, Control and Computing (PICC) – 2020.

[3] De Rosal Ignatius Moses Setiadi, Dewangga Satriya Rahardwika, Candra Irawan, Desi Purwanti Kusumaningrum, "Comparison of SVM, KNN, and NB Classifier for Genre Music Classification based on Metadata", International Seminar on Application for Technology of Information and Communication (iSemantic) – 2020.

[4] Nikki Pelchat and Craig M. Gelowitz, "Neural Network Music Genre Classification", IEEE Canadian Conference of Electrical and Computer Engineering (CCECE) – 2019.

[5] K. Meenakshi, "Automatic Music Genre Classification using Convolution Neural Network", International Conference on Computer Communication and Informatics (ICCCI)– 2018.

[6] Mingwen Dong, "Convolutional Neural Network Achieves Human-level Accuracy in Music Genre Classification", arXiv:1802.09697v1 [cs.SD] – 2018.

[7] Hareesh Bahuleyan, “Music Genre Classification using Machine Learning Techniques”, arXiv:1804.01149v1 [cs.SD] – 2018.

[8] Weibin Zhang, Wenkang Lei, Xiangmin Xu, Xiaofeng Xing, “Improved Music Genre Classification with Convolutional Neural Networks”, Interspeech – 2016.

[9] Balachandra K, Neha Kumari, Tushar Shukla, Kumar Satyam, “Music Genre Classification for Indian Music Genre”, IJRASET – 2021.

[10] Nicolas Scaringella, Giorgio Zoia, Daniel Miynek, “Automatic Genre Classification of Music Content”, IEEE Signal Processing Magazine – March 2006.

[11] Deepanway Ghosal, Maheshkumar H. Kolekar, “Music Genre Recognition using Deep Neural Network and Transfer Learning”, in Interspeech vol. 28, no. 24, pp. 2087-2097, September 2018.

[12] Jose J. Valero-Mas, Antonio Pertusa,” End-to-End Optical Music Recognition Using Neural Networks”, 18th International Society for Music Information Retrieval Conference – 2017.

[13] Prasenjeet Fulzele1, Rajat Singh2, Naman Kaushik3,” A Hybrid Model for Music Genre Classification Using LSTM & SVM”, ICCC – 2018.

[14] George Tzanetakis, Perry Cook,” Musical Genre Classification of Audio Signals”, IEEE Transactions on Speech & Audio Processing, vol. 10, No. 5, July 2002*.*