Music Genre Classification Using Neural Network

Tarannum Shaikh[1*], *Ashish* Jadhav[1]

[1]Ramrao Adik Institute of Technology, D. Y. Patil Deemed to be University, Nerul, Navi Mumbai, India

Abstract. Music Genre classification on Neural Network is presented in this article. The research work uses spectrogram images generated from the songs timeslices and given as input to NN to do classification of songs to their respective musical genre. The research work focuses on analyzing the parameters of the model. Using two different datasets and implementing NN technique we have achieved an optimized result. The Convolutional Neural Network model presented in this article classifies 10 classes of Music Genres with the improved accuracy.

* Tarannum Shaikh: tanu99@gmail.com

1 Introduction

1.1 Music Genre Classification

On Internet web you could be connected to world of Music and experiencing it inclusive of live concerts, Remixes, Old Classic, etc. like never before [1]. Depending upon listener what he wants to listen, he simply downloads the music from songs database and listens without storing it in the device, saving the memory space. The music can be liked or formed a playlist so as to listen again whenever mood swings.

Humans are very intelligent to listen to the short samples of songs and differentiate the singer, beat, lyrics, the title, album, as well the genre of the song. Studying this intelligence, new technology is been applied on several NN (Neural Network) approaches, showing successful achievements [2].

1.2 Machine Learning and Neural Network

Today machine learning is very popular. There are many ML (machine learning) techniques for numerous applications. ML is divided into following techniques, those are supervised learning, unsupervised learning, semisupervised learning, and reinforcement learning [3].

Supervised Machine Learning algorithm is a technique which utilizes labelled training data to attain desired results. It helps the model to learn quickly. In unsupervised learning it uses unlabeled dataset, this technique extracts useful features and performs the complex processes. Semi supervised learning algorithm uses a dataset containing small number of labelled and large number of unlabeled data. Reinforcement learning technique uses a feedback mechanism. In reinforcement learning, it is about to take correct action to maximize reward. For instance, reinforcement learning is usually used in online entertainment when correct prediction is rewarded.

A NN is a technology of machine learning used explicitly to extract the essential features from complicated database and designing a framework that demonstrate the features [4]. The Neural Network uses the training dataset initially to train the designed model. Afterwards the NN is applied to new or test data and observed that the classification of the data is done correctly assumed by the trained model.

1.3 Neural Network

A standard Music Genre Classification usually includes two major elements, one is feature extractor and another is classifier model [5]. From both of these elements, feature extractor is very important parameter that gives the performance of MGC precisely.

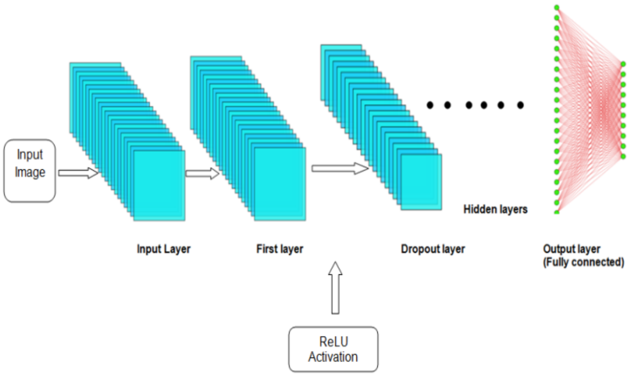


Fig. 1. Proposed CNN Architechture

Above figure 1 shows a proposed architechture of a Convolution Neural Network (CNN), CNN is a class of neural network which is specialized to perform certain process on multidimensional vectors having a grid structure for eg: images [4]. In order to attain optimized result, a CNN is applicable for binary classification as well as for multi classification. They are mostly similar but they vary in the output classes only. For instance, a data collection of floral photos may be used to train a model. The input applied to CNN is a vector of pixel values of an image processed from the dataset which is associated with the right label of vector defined as output class (eg. If floral images then output class is rose, tulip, lily, etc…). Once an input image is applied to CNN model for training purpose, the output will be mapped to particular class after classifying the images. While performing classification, every training dataset image is matched with the predicted output class label. During training the accuracy of the CNN is increased by back propagation technique in which the weights throughout the network are updated iteratively. Number of iterations helps to shape the CNN structure. This structured model does feature extraction of training datasets improving its capability to perform accurate classification of images.

The main contributions made in this literature is focusing both feature extraction and genre classification

The optimized model depicted here has good scope and growth in numerous technologies of classification.

Organization of topics in the paper is as follows. Related work section gives an overview of the current technology for MGC (Music Genre Classification). The Proposed architechture explains the designed model step by step. The next section is Implementation work and results in which complete implementation of the research work using various techniques and logic is explained in detail along with their metric results. At the last, the conclusion section comments over the research work done and mention future scope.

2 Related Works

In 2020 Nikki Pelchat and Craig M. Gelowitz presented a research work on music genre classification which classified the songs into their respective musical genres. He used two different music datasets consisting of 1880 different songs, these songs were splitted into 2.56 seconds section which eventually gathered with 132000 spectrograms snippets. This increased the number of training spectrogram slices per genre, typically 128X128 pixel slices were used for training the Neural Network. Activation functions used were softmax and Rectified Linear Unit (ReLU). Also used Hamming windows functions to convert mp3 files to spectrograms and increased the number of CNN layers to achieve 85% accuracy [6].

In 2021 Jaime Ramirez Castillo and M. Julia Flores presented a web based application in which retrieval of YouTube songs was done for music genres classification. Chunks of 10 seconds songs were classified into one or more musical genres. The models are made training several machine learning techniques, namely, Support Vector Machines (SVM), Naive Bayes classifiers, Feed forward deep neural networks and Recurrent neural networks and embedded those in a web application. These models let the user to envisage how each model "percept" the music in terms of music genre, at particular instant of song [7].

In 2020 Wing W. Y. NG Music genre consists of various features that are heterogeneous with abstractions in numerous levels. To eliminate same level feature extraction (FE), he proposed a combination of Convolutional Neural Network (CNN) with NetVLAD and self-attention to extract the information across different levels for learning long-term dependencies. Finally a meta classifier is utilized to make Music Genre Classifier.

His work shows that the proposed approach provides higher accuracies than other state-of-the-art models on GTZAN, ISMIR2004, and Extended Ballroom dataset [8].

Binary Pattern, which is the best result on this dataset In 2016 Costa et.al worked on Music genre recognition considering different datasets and their spectrograms. He trained the classifiers with various techniques like Gabor filters, Local Binary Patterns, Phase Quantization, to extract spectrograms and attained state-of-the-art outcome on numerous audio database. He has done the comparison of results produced by Convolutional Neural Network (CNN) with the outcomes procured by handcrafted features. His research was done on three different audio datasets with definite attributes, specifically a western music dataset (ISMIR 2004 Database), a Latin American music (LMD dataset), and African music dataset.

His research proved that the CNN is best alternative as it is favourable to all classifiers for music genre recognition. For the African database, the CNN excelled the handcrafted representations and showed the state-of-the-art by a margin. Another dataset i.e LMD database attained the recognition rate with 92% by combining CNN and Robust Local till then. On the ISMIR 2004 dataset, the CNN though did not improved the state of the art, but it gave comparatively better performance than the classifiers centered on other kind of attribute [3].

3 Proposed architecture of MGC

3.1 Overall Depiction of presented model

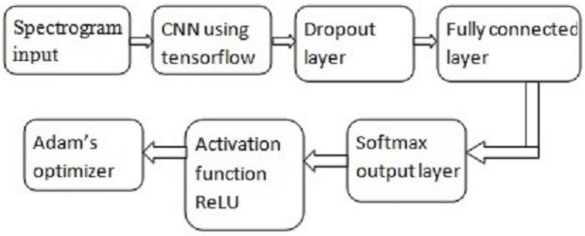


Fig. 2. Architechtural model

In the proposed architechture CNN model, a dense layer is used with 512 features, immediate after each dense layer a dropout layer is added by the factor of 0.2.

Figure 2 depicts the architechtural design of music Genre classification model, build using Keras and Tensorflow framework. The model consists of four hidden layers along with input layer.

These all layers are using the ReLU activation function. The output layer is fully connected network layer which uses the Softmax activation function. Evaluation metric considered here is loss and accuracy. Sparse categorical cross entropy function calculates the loss of the presented model. Adam's optimizer is used for updating the weights in neural network while training the model. The epoch chosen for training the proposed model is 600, for each epoch the model takes nearby 20ms (milli seconds).

Dropout is used to reduce the problem of overfitting by making all the parameters to zero. i.e making the neurons inactive. A dense layer along with the ReLu activation function is used to predict the data for classification. Evaluation metric can be improved more by adding more epochs, but there are certain limits to do so, hence the value for epochs must be considerate.

3.2 Music Database

In this article two different databases are utilized, a western database "GTZAN", and a Music Database "MusicNet", with different genres. [9]

GTZAN dataset consists of total 1,000 music excerpts each of 30 seconds long. The downloaded zip file consists of three different folders.

- First one is Genres Original: which consists of 1000 audio formats of .wav file of 30 seconds each equally divided in 10 genres. Therefore, each genre containing 100 songs.
- Second is Images original: which consist of graphical form or spectrograms of each 30 sec audio file. This is used as input image to the model.

- Third folder is CSV files having multiple features of audio files of above folders.

MusicNet is second dataset used in proposed research which has compilation of 330 freely-licensed classical music files [11]. This accumulation is compiled at three top-level files:

- musicnet.tar.gz - This file has the MusicNet dataset as it is, composed of PCM-encoded audio wave files (.wav) and related CSV-ciphered note labelled files (.csv). The data is arranged according to the train/test split required for training and validation.
- musicnet_metadata.csv – This metadata file is composed of music track-level details of recordings contained in MusicNet. The name of data and label files are done with MusicNet ids, hence it is used to cross-index the data and labels with this file.
- musicnet_midis.tar.gz - This file consists of reference Musical Instrument Digital Interface (MIDI) files which are useful to construct the MusicNet labels.

3.2.1 Data Analysis

Using GTZAN and Musicnet datasets, which are freely available on internet [10-11]. The audio files are all sampled with 22050Hz Mono 16-bit audio files in .wav format. The core of dataset is feature analysis for developers, including genres and images. The size of GTZAN dataset is 9990 by 60 where each row represents a single audio piece and each column signifying one feature.

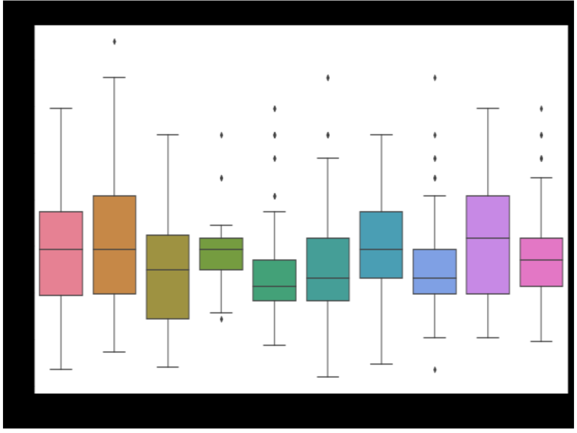


Fig. 3. Boxplot of Music Genre

Above figure illustrates Boxplot for genre and beats per minute (BPM) of the dataset, which shows detail analysis and distribution of the data in the dataset.

Using both the dataset the accuracy and loss is calculated. This shows the presented model can be used on any dataset for classification purpose which yields good results.

Principal Component Analysis (PCA) is a unsupervised technology which is used to reduce dimensionality of database. It is an algorithm applied to compress the high dimensional data into lower dimensional feature without any class label information. PCA reduces the data in two dimensions as shown in figure below (Principal Component 1(PC1), Principal Component 2(PC2)) or in three dimensions (PC1, PC2, PC3).

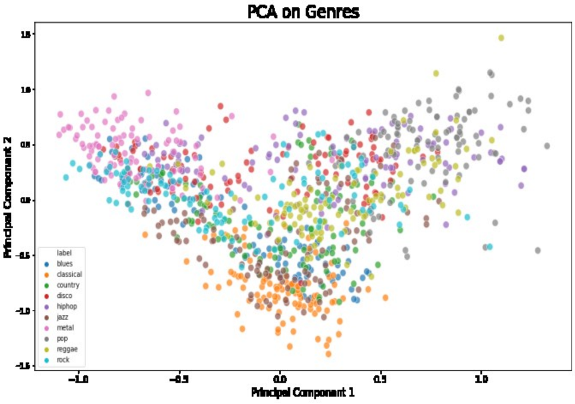


Fig. 4. Principal Component Analysis

Figure 4. Shows Principal Components Analysis (PCA) of the music dataset, indicating the principal components of the data, which reduces the dimensionality of dataset. (i.e) dataset can be reduced to 2 dimensions without incurring any loss in data. We see here in the plot that the data is scattered more along PC1 than PC2 that means any classifier can separate or classify the classes or genres perfectly.

3.3 Preprocessing

Preprocessing of data is an important step done on the dataset before applying the data to the input of CNN for machine learning models. It assures to improve the performance and efficiency of the network also saves time during training the model. Data preprocessing actually involves pre-emphasis, framing and widowing techniques on the given data doing so useful information can be retrieved from the data.

3.3.1 Scaling the features

To standardize the audio features Standard scalar is best preprocessing technique which removes the mean and scaling of audio signal to unit variance.

Consider sample x, which is calculated as

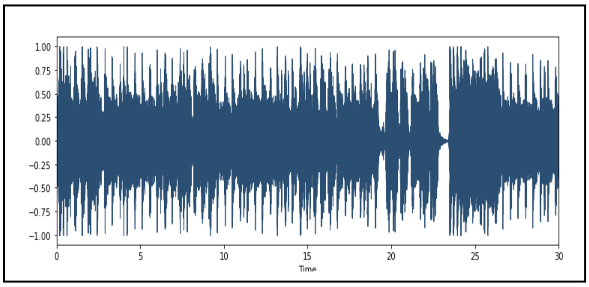
$$z = (x - u) / s$$

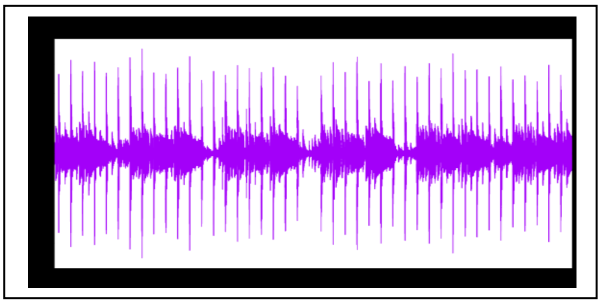
Where, u is training samples mean value

s stands for training samples standard deviation

Standardization of a dataset is a widely recommended for machine learning or Neural Network classification else the dataset response will be unfavourable.

3.3.2 Visualizing Audio Files





(b)

Fig. 5. Raw wave plot of audio file (a)hiphop 14, (b) reggae 36

We have used librosa to plot raw wave files. Figure 5 illustrates a waveform that is a visual representation of sound showing amplitude Vs time where the Y axis is amplitude and X axis is time. Such graphical images enable to process the audio data instantly and relate the similarity index.

3.4 Feature Extraction

Feature Extraction is done to analyse and find relation between different things. Using feature extraction techniques like PCA, Independent Components Analysis (ICA), and Linear Discriminant Analysis (LDA) dataset dimensionality can be reduced efficiently.

The dataset to be understood by models directly is converted to understandable format. FE is essential analysis for classification, prediction and recommendation algorithm.

We brief the features extracted in the research work [12].

3.4.1 Zero-crossing rate

Zero crossing rate (ZCR) [13] is the rate in which the audio signal passes number of times from upper half cycle to lower half cycle or from lower half cycle to upper half cycle through zero. Application of this feature is in audio or speech classification, recognition and music information retrieval.

$$\text{ZCR} = \frac{\sum_{n=1}^{N} |sign\ (An) - sign\ (An-1)|}{2N} \qquad (1)$$

Where, N is the number of samples in the frame and
An is the amplitude of the n-th sample.
sign (·) represents the sign function.

3.4.2 Spectral Centroid

Spectral Centroid [13] feature gives essential facts about the frequency and energy distribution of the audio signal, we evaluate the center of gravity of the spectrum as

$$\text{Centroid} = \frac{\sum_{k=1}^{N} P(fk)fk}{\sum_{k=1}^{N} P(fk)} \qquad (2)$$

Where, fk is the k-th frequency,
N is the number of frequency bins.
P (fk) is the spectral amplitude on the k-th frequency [8].

3.4.3 Chroma Feature

Chroma Feature is a robust tool for semantic analysis of the musical audio features. It's a 12 element feature representing spectral energy level in the signal. It's a powerful way for finding out similarity index of audio samples.

Short Time Fourier Transform (STFT) are used to extract the Chroma feature from audio signal. Better the extraction quality, best are the results.

3.4.4 Spectral Roll-Off

Spectral roll-off [14] is the specified amount of frequency lying below the total spectral frequency.
The spectral rolloff point is calculated as described in
rolloffPoint = i

$$\text{Such that } \sum_{k=b1}^{i} sk = k \sum_{k=b1}^{b2} sk \qquad (3)$$

Where, s_k is the spectral value at bin k.
b_1 and b_2 are the band edges, in bins, over which to calculate the spectral spread.
k is the percentage of total energy contained between b_1 and i.
k can be set using Threshold.

3.4.5 Mel-Spectrogram

Mel-spectrogram is a graphical representation of the frequencies spectrum of an audio signal with respect to time. Mel spectrogram is the magnitude spectrogram of the input audio data on the log scaled frequencies on the Y-axis. The expression of mel-frequency is as follows [13]

$$\text{Mel}(f) = 2595\text{log10}\ \ 1 + f/700 \qquad (4)$$

3.4.6 Mel-Frequency Cepstral Coefficient (MFCC)

Mel-Frequency Cepstral Coefficient (MFCC) [12] feature is broadly used in automatic audio and speech classification. The cepstral coefficient obtained on the mel-frequency is called (MFCC) mel-frequency cepstrum coefficient.

Above features in the form of vectors are used to build the input feature vector.

4 Implementation

The proposed Music Genre Classification model is implemented in Python Programming and the corresponding output were achieved. For training and testing purpose, we have split the dataset into 67% and 33% respectively. All the images are preprocessed using Librosa and then given as the input to the proposed CNN model.

Using Deep Convolutional network the NN model was implemented applying Tensorflow framework. The batch size given here is 128. The first four layers are convolutional layers and dropout layers with 0.2

probabilities, which avoid a curb overfitting and is implemented after each successive layer. The activation function used here at input and hidden layers is ReLU and Softmax for the output layer. The last layer is fully connected layer to which each and every output of the previous layer is given. This process gives a lengthy array. A softmax layer is applied to regulate the 10 outputs, these 10 outputs is nothing but 10 Genre.

Adopting “sparse categorical cross entropy” with the standard initial learning rate the loss function is measured. The Adam’s optimizer is used to converge towards the minima fast. It is computationally efficient and is used for huge data or maximum parameters.

To see that the designed neural network has the same accuracy throughout different datasets, the same NN architechture was used to train another data set. The two datasets used in this research has different genre, but number of songs or music piece in each genre of both dataset is equal. This is done intentionally for consistency and comparing test accuracy.

Table 1. Table of GTZAN[10].

Genre			
Name	**No. of files**	**Name**	**No. of files**
Blues Country Hiphop Metal reggae	100 100 100 100 100	Classical Disco Jazz Pop rock	100 100 100 100 100

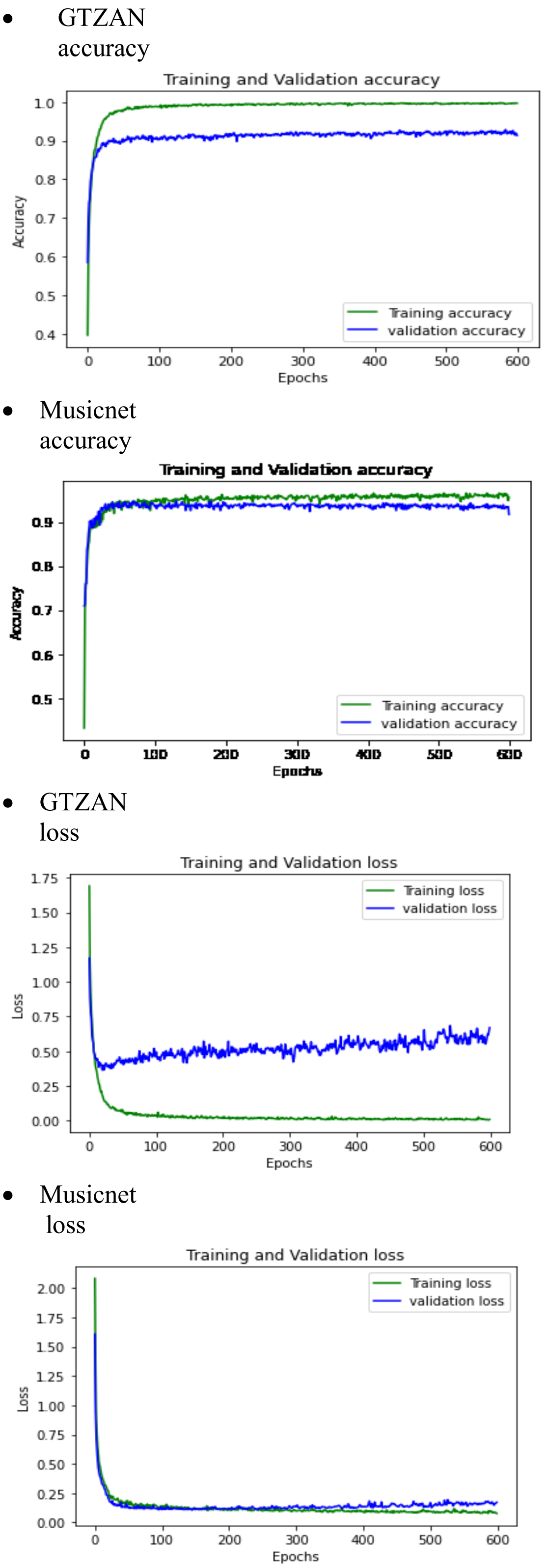
[a] Genre details, (files are 30 sec long .wav format).

Table 1 shows respective number of files of each Genre. The other dataset had different genre with same number of files in each genre. Equalizing number of songs per genre in two different dataset, a test accuracy attained was 92.65%. Hence using different dataset on the same model is possible that too with higher accuracy.

Following are the comparison results of accuracy and loss for both the datasets. Table 2 gives the metric evaluation of the results obtained by the proposed MGC model.

Table **2.** Accuracy And loss result.

Dataset	**Accuracy**	**Loss**
GTZAN	92.65%	57.37%
Musicnet	91.70%	25.46%



- GTZAN accuracy
- Musicnet accuracy
- GTZAN loss
- Musicnet loss

Fig. 6. Training accuracy and training loss of two datasets.

Figure 6 illustrates graphical representation of table 2. The plot shows both the training and validation data acuuracy and loss curves.

Table 3. Comparison of Pre-Trained Models [1]

Authors/Models	**Accuracy**
Stochastic Gradient Descent	65%
Pelchat, Gelowitz	85%
Naïve Bayes	51.9%
Costa *et.al.*	67%
Lopes *et al.*	60%
Despois	90%

Table 3 shows an overview of accuracy obtained by pre trained models in recent years, which indicates that to improve the performance only increasing the number of layers is not sufficient. On the contrary we can increase the accuracy by increasing the model depth.

5 Conclusion

This research work proposes an approach of implementation of a Music Genre classification model. The work involved of preprocessing and extracting the features of the music. This extracted feature like spectrograms are provided as an input to the CNN. The NN comprises of four layers including a input layer with ReLU, and finally a fully connected output layer with 10 classes including Softmax activation function predict the probability of each genre. In addition we have used a song library of MIDI data for contemplation of datasets sequential features. Concluding further, we consider that Neural Networks are most efficient in machine learning technology. Framework Tensorflow is very important in building the CNN model and to execute it effectively.

Future work may have some modification in initialization of weights and including full song instead of 30s slice classification. Implement pooling, for individual output to nullify false positives is a future task. Also contemplating binary classifiers to find out the user may like a predicted song, found on their personal music library.

References

[1] M. Goto and R. B. Dannenberg, "Music Interfaces Based on Automatic Music Signal Analysis: New Ways to Create and Listen to Music," in *IEEE Signal Processing Magazine*, vol. **36**, no. 1, pp. 74-81, Jan. 2019, doi: 10.1109/MSP.2018.2874360.

[2] Y. M.G. Costa, L. S. Oliveira, C. N. Silla, "An evaluation of Convolutional Neural Networks for music classification using spectrograms", in *Applied Soft Computing*, Volume **52**, 2017, Pages 28-38, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2016.12.024.

[3] S. Gollapudi, *Practial Machine Learning*. Birmingham, U.K.: Packt, 2016.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning.(Report)," *Nature*, vol. **521**, no. 7553, p. 436, May 2015, 2015.

[5] T. Kim, J. Lee, J. Nam, Comparison and analysis of sampleCNN architectures for audio classification, *IEEE* J. Sel. Top. Sign. Proces. 13 (2) (2019) 285–297.

[6] N. Pelchat and C. M. Gelowitz, "Neural Network Music Genre Classification," in *Canadian Journal of Electrical and Computer Engineering*, vol. **43**, no. 3, pp. 170-173, Summer 2020, doi: 10.1109/CJECE.2020.2970144.

[7] J. R. Castillo and M. J. Flores, "Web-Based Music Genre Classification for Timeline Song Visualization and Analysis," in IEEE Access, vol. **9**, pp. 18801-18816, 2021, doi: 10.1109/ACCESS.2021.3053864.

[8] W. W. Y. Ng, W. Zeng and T. Wang, "Multi-Level Local Feature Coding Fusion for Music Genre Recognition," in *IEEE Access*, vol. **8**, pp. 152713-152727, 2020, doi: 10.1109/ACCESS.2020.3017661.

[9] G. Peeters, 2021, 15 August 2021 [https://ismir.net/resources/datasets/]

[10] G. Tzanetakis, 2015, 15 August 2021, [http://marsyas.info/downloads/datasets.html]

[11] J. Thickstun, Z. Harchaoui, S. M. Kakade, November 30, 2016 ,15 August 2021 [https://zenodo.org/record/5120004#.YbnBRjNBzIV]

[12] D. Yu, H. Duan, J. Fang and B. Zeng, "Predominant Instrument Recognition Based on Deep Neural Network With Auxiliary Classification," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. **28**, pp. 852-861, 2020, doi: 10.1109/TASLP.2020.2971419.

[13] J. D. Deng, C. Simmermacher, and S. Cranefield, "A study on feature analysis for musical instrument classification," *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)*, vol. **38**, no. 2, pp. 429–438, Apr. 2008.

[14] Scheirer, E., and M. Slaney, "Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator," *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Volume **2**, 1997, pp. 1221–1224.