

Music Genre Classification using Artificial Neural Networks

Priya Shelke¹, Rushikesh Tanksale², Mrunali Gadekar³, Himanshu Naidu Akula⁴, Chaitali Shewale⁵ and Renuka kale⁶

¹⁻⁶Vishwakarma Institute of Information Technology, Pune, India

Email: priya.shelke@viit.ac.in, {tanksalerushikesh}@gmail.com; {mrunali.22010831, himanshu.22010394, chaitali.shewale, renuka.22010895}@viit.ac.in

Abstract—In the realm of music research, the analysis of musical genres holds paramount importance, and the application of artificial neural network models has garnered increasing interest. This comprehensive study delves into the examination of three pivotal models—multilayer-perceptron (MLP), convolutional-neural-network (CNN), and recurrent-neural-network with Long Short-term Memory (RNN-LSTM)—specifically for music genre classification. The motivation behind this exploration lies in the desire to unravel the potential of these neural network models in deciphering the intricate landscape of musical categorization. Employing a substantial audio dataset extracted from diverse musical compositions, our investigation meticulously compares the performance of the aforementioned models. Noteworthy is the finding that while MLPs exhibit utility in specific applications, they grapple with inherent limitations. Contrarily, RNN-LSTM and CNN emerge as frontrunners, surpassing MLPs in terms of accuracy and loss metrics when classifying music genres. Particularly, CNN showcases significant promise in its capacity to enhance precision in the nuanced domain of music analysis.

Index Terms— classification, neural networks, machine learning, signal processing, convoluted neural networks, multilayer perceptron, recurrent neural networks.

I. INTRODUCTION

A music genre is a phrase used to group distinct musical styles together based on a custom or set of guidelines. Although in fact both terms are used synonymously, the difference between musical form and musical style must be made. Music can be divided into distinct genres using a variety of techniques, such as dividing it into popular and art music or religious and secular music. Due to the artistic nature of music, these distinctions are frequently arbitrary and contentious, and some genres may overlap. Blues, pop, jazz, rock, metal, classical, country, disco, hip hop, and reggae are just a few of the ten genres that are being studied for the purpose of classifying music. Artificial neural networks are used in the classification process. Artificial neural network models such as Recurrent Neural Networks (RNN), Multilayer Perceptrons (MLP), and Convolutional Neural Networks (CNN) are used to test the model's accuracy.

II. LITERATURE SURVEY

In Nizamettin Aydin's study, the acoustic characteristics of the music were extracted, and then music genres were identified and recommendations for music were made using machine learning techniques. Convolutional

neural networks, another deep learning technique, was utilized to categorize the data by genre, recommend music, and rate the performance of the music [1]. The accuracy of the Vaibhavi Macharla model was 69.49 percent for a four-layer 2D CNN model and 77.92 percent for an unaugmented RCNN model. By performing data augmentation, the suggested method produced an accuracy of 81.55 percent for a four-layer 2D CNN model and 82.05 percent for an RCNN model [2].

In comparison to earlier studies, the suggested MusicRecNet model by A. Elbir and N. Aydin has demonstrated enhanced performance with regard to music genre classification, similarity analysis, and music recommendation. In example, CNN's results have improved significantly after the addition of some dropout layers. The MusicRecNet is a suitable performance model in terms of music genre classification, musical recommendation, and musical resemblance, because it only has three layers. When performance results are analyzed, some genres of music, such as jazz and classical, can result in incorrect classification and incorrect recommendations [3].

The study aims to compare the effectiveness of two machine learning models - the Convolutional-Neural-Network (CNN) and Long-Short-Term Memory (LSTM) model - in terms of their architectures, features, and classification accuracy. To achieve this, the researchers utilized two popular music datasets, namely GTZAN and FMA. After analyzing the results, it was found that the CNN model outperformed the LSTM model in both datasets. Specifically, the CNN model achieved an accuracy rate of 42.0% for GTZAN and 33.5% for FMA, while the LSTM model achieved 33.5% and 27.8% accuracy, respectively. [4]

In Hindavi's essay on the classification of musical genres, a more precise characterization of music is proposed. The essay suggests breaking down music into local MIDI passages, analyzing playing style, extracting passage features, and creating a sequence of features. This involves using a note feature matrix to segment the music and identifying the best features based on the segment theme[5]. The literature review by Lam Hoang depicts how music genres are generally categorized using the neural network, or NN. He has talked about the kinds of datasets that the majority of researchers chose for their papers. As can be observed, because of its popularity and accessibility, GTZAN is used for genre classification in the bulk of those studies. [6].

The work of N. D. Nivetha offers a thorough machine learning strategy for the problem of automatically classifying styles from audio signals. Convolutional Neural Network (CNN) technology is used in the system's development. to recognize different genres. Here, a CNN model is completely trained to forecast the genre of an audio stream. The GTZAN data set, a popular public data set for music recognition (MGR) research, is used in the experiment. [7]

The goal of Prajwal R's project is to create a classification and organization system for audio music that can be used in a variety of genres. The spectrogram and acoustic properties of the audio music serve as the basis for classification. This model was developed with the intention of improving the classification accuracy of musical genres. [8]

The study by Hareesh Bahuleyan compares two different types of model performance. The first approach employs deep learning to completely train a CNN model to identify the genre of an audio source based just on its spectrogram. The second approach uses time- and frequency-domain properties that were especially designed. [9] The goal of the research project is to use supervised learning methods, such as artificial neural networks, to categorize musical genres. evaluating the accuracy of music classifiers for datasets of various types. Additionally, the classification accuracy of genre classifiers is examined using a variety of modifiers and layers. [10]

The effectiveness of each of these models is evaluated in Snigdha Chillara's article, and the findings are logged in terms of prediction accuracy. A small number of these models are trained using both the mel-spectrograms and the audio properties of the songs, whilst a smaller number are trained solely using the spectrograms of the songs. When just providing the spectrograms as the dataset, it was discovered that one of the models, a convolutional neural network, had the highest accuracy of all the models. [11].

A hybrid design put up by Lin Feng comprises parallel CNN and bi-RNN blocks. They each concentrate on extracting temporal frame ordering and spatial characteristics, respectively. The two outputs are then combined into one potent musical signal representation and supplied to the softmax function for categorization. The extracted characteristics are guaranteed to be strong enough to represent music by the paralleling network. Additionally, experiments show that the design we suggest improves the performance of classifying music genres, and the additional Bi-RNN block is a complement for CNNs [12].

A number of categorization algorithms with two different forms of input were put into practice by Derek A. Huan. k-nearest neighbors, a straightforward feed-forward network, an advanced convolutional neural network, and an RBF kernel support vector machine were all employed in the trials. They experimented with both raw amplitude data and converted mel-spectrograms of that raw amplitude data as the input to our algorithms. Out of a total of ten popular music genres, a forecast genre is generated. All models performed better when raw audio

was transformed into mel-spectrograms, with our convolutional neural network outperforming human accuracy [13].

Neural networks and feature extraction were used as methods in the study by Xiabin Li and Jin Li. To execute feature extraction and classification as well as learn the features of the image, the convolutional neural network underwent training. With the use of traditional gradient descent methods, the model parameters were trained. With the same number of cycles, the study indicated that this strategy is 12 percent more effective than conventional procedures, making it appropriate for general implementation [14].

Two methods to improve convolutional neural network-based music genre classification are suggested in article number 15. First, peak- and average pooling is combined using a technique inspired by residual learning, providing higher level neural networks with more statistical information. Second, shortcut connections that skip one or more layers are added. The output of the KNN is then used to classify data in a different deep neural network [15].

III. MOTIVATION

The music genre classification model uses an audio stream as its input and seeks to guess the genre. Making the selection of songs easier and quicker is the goal of automating music classification. If one must manually categorize the songs or music, they must first be listened to a great deal before the genre is chosen. The system reduces human effort by automating the classification.

IV. PROPOSED SYSTEM

This paper proposes a classification model to segregate the given audio file input in WAV format. The proposed model includes preprocessing of the data, after preprocessing, we again load the data to further split it into training datasets, testing datasets, and validation datasets, followed by designing the model, compiling the model, training the model, and classification. During classification, the input would be assigned to one of the ten music genres. GTZAN dataset for music genre prediction is used in this model. The dataset has three parts. Genre original contains 10 genres with 100 audio files each. Length of audio files is 30 secs each. Images originally contain the visual representation of each audio file. There are 2 csv files, one containing mean and variance for audio length 30 seconds and the other contains mean and variance for audio length 3 seconds [16].

A. MFCC-Feature-Extraction

MFCC stands for Mel-frequency cepstral coefficients (MFCC). They are a largely used attribute retrieval procedure in audio and speech processing. It captures relevant spectral and timbral characteristics of audio signals in a compact and perceptually meaningful way. MFCC features are useful for music genre classification due to their compact representation, robustness to noise, alignment with human perception, and ability to capture the timbral characteristics of music.

B. Data-preprocessing

First we define a dictionary with three keys mapping , mfcc, label. The data set used to train our model has ten music genres and each music genre having almost hundred music samples per genre. We make use of os.walk from the os module to iterate through each sub folder, in each subfolder we process each audio file in the given folder, we make use of librosa python module to extract key features like the given number of mfcc vector per segment and append them to the list associated to mfcc key in our dictionary, while iterating through the sub-folders we keep a track of the sub-folder path from which we extract the sub-folder name and append it to the list associated to mapping key of our dictionary. We also keep track of the folder number while walking through the sub-folders and append the number to the list associated with the key label of our dictionary. Make note that the values appended for labels would lie in the range of [0,10) as our model is trained to classify music into ten distinct categories. After walking over all the subfolders is concluded, we publish this dictionary into a .json file for further processing. For publishing our dictionary in the .file we make use of the json module in python

C. MLPN models for genre classification:

The training data is first loaded from a JSON file and transformed into numpy arrays for the inputs (X) and targets (y) in the function. Using the train_test_split function from the sklearn library, the data is divided into training and testing sets, with 70% of the data going to training and 30% to testing. After that, the neural network model is defined using the Sequential class from Keras. It has a Flatten input layer, three dense layers with various amounts of units and ReLU activation functions, an output layer with a softmax activation function for

multi-class classification, and so on. Since the labels are integers, the loss function is the sparse categorical cross entropy, and accuracy is used as the evaluation metric. The model is built using the Adam optimizer. The model is then trained using the fit function, which is supplied with the training data, validation data, batch size, and number of iterations. Once trained, the model is capable of making predictions for an audio input.

D. CNN model for genre classification

MFCC, labels, and mapping are first loaded from a JSON file containing training data. Following that, the data is divided into train, validation, and test sets. The architecture of the CNN model is then defined by the code. Multiple convolutional layers, which are frequently employed in CNNs to extract features from input data, max pooling layers for downsizing the sample, and batch normalization layers are used to build the model. This in some ways speeds up the training of our model. In order to do multi-class classification, the output of the final convolutional layer is flattened and then fed via a dense output layer with softmax activation. The function that makes predictions using the trained model is then invoked. Using preprocessed input sound data, the trained model is fed into this function to produce predicted class probabilities. We find the maximum of the given probability when classes are assigned after making our classification.

E. RNN-LSTM model for genre classification

We utilized a dataset in a JSON file to train a recurrent neural network (RNN) with layers for long short-term memory (LSTM). The dataset was divided into train, validation, and test sets using the train_test_split function from the scikit-learn library. Our RNN-LSTM model was created using the Keras package and consisted of two LSTM layers, a dense layer with ReLU activation, and an output layer with softmax activation. We built the model using the Adam-optimizer with a learning rate of 0.0001 and a loss function of sparse categorical cross-entropy. We trained the model using the training set for 30 epochs with a batch size of 32, while keeping track of training progress with the history object for further analysis.

V. SYSTEM DESIGN AND ARCHITECTURE

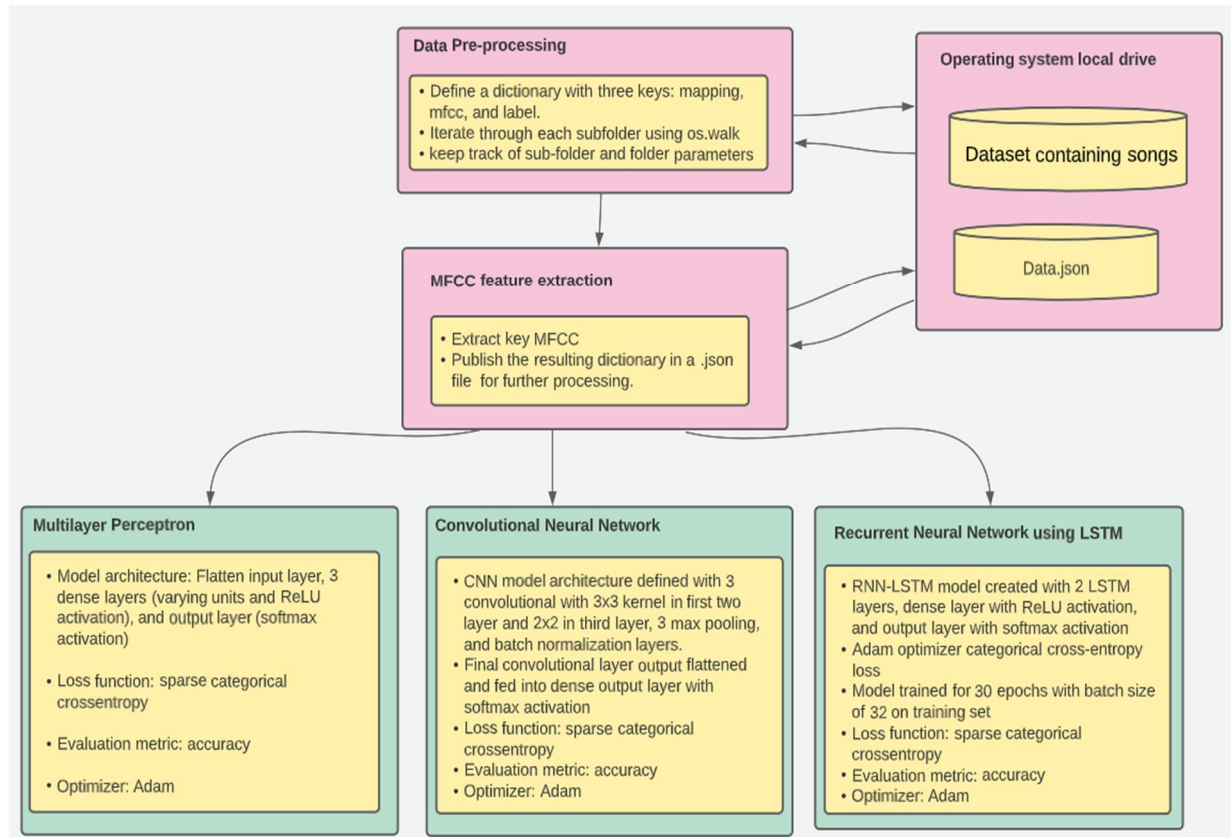


Figure1. System Architecture

A. MLP

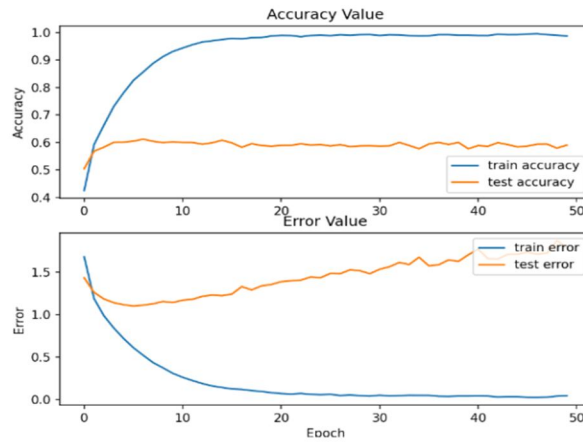


Figure 2. MLPN train-test accuracy v/s epochs and train-test error v/s epochs

The graph provided illustrates the relationship between train and test accuracy, as well as train and test error, over several epochs. It is evident from the graph that there is a notable difference between the train and test accuracy, indicating that the MLPN model is overfitting for music genre classification. Additionally, there is a considerable gap between train and test error, further supporting the notion of overfitting.

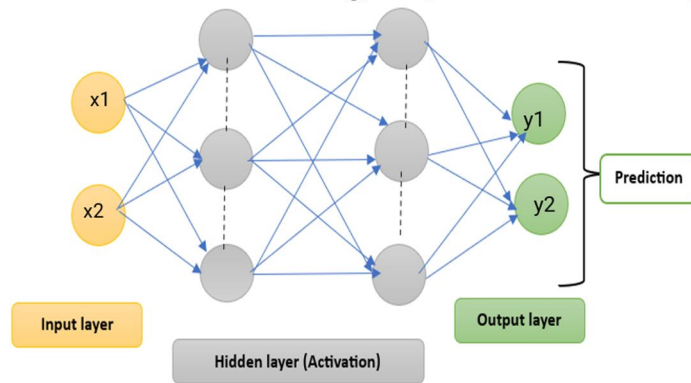


Figure 3. MLPN Architecture

Figure 3 depicts the MLPN model's system architecture. In this case, x1 and x2 are input neurons that are routed via a hidden layer that conducts matrix multiplications, followed by a Rectified Linear Unit, and finally output neurons that use the softmax activation function.

B. CNN

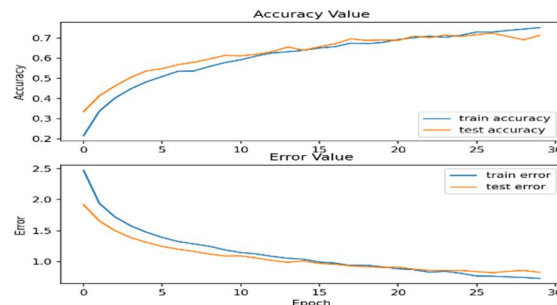


Figure 4. CNN train-test accuracy v/s epochs and train-test error v/s epochs

Again, for the CNN model, figure 4 is a graph of train and test accuracy v/s epochs, as well as train and test error v/s epochs. We can easily observe that the difference between train and test accuracy is very small, as is the difference between train and test error. As a result, CNN is an excellent fit for music genre classification.

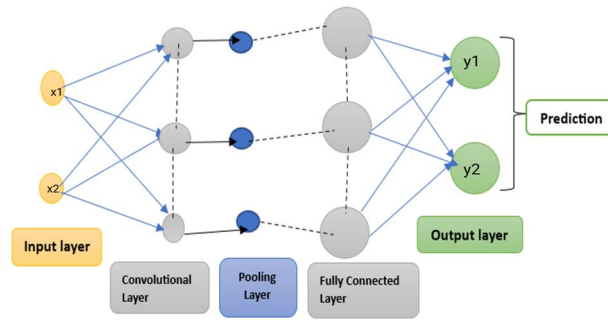


Figure 5 CNN Architecture

Figure 5 provides a visual representation of the functioning of a CNN. The input neurons x_1 and x_2 are fed into the convolution layer, which includes learnable filters that generate feature maps. The resulting output is then passed through several pooling layers that reduce the dimensionality of the feature map. The final pooling layer's output is then fed into a fully connected layer, which generates the final outputs y_1 and y_2 .

C. RNN-LSTM

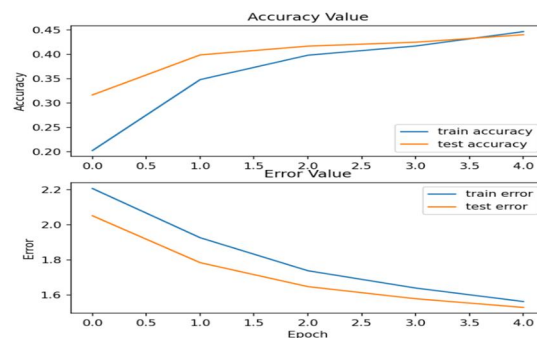


Figure 6 : RNN-LSTM train-test accuracy v/s epochs and train-test error v/s epochs

Figure 6 displays the relationship between train and test accuracy and train and test error over several epochs. The graph indicates that there is relatively little difference between the train and test accuracy and train and test error. Therefore, we can infer that the RNN with LSTM model is a suitable fit for our project.

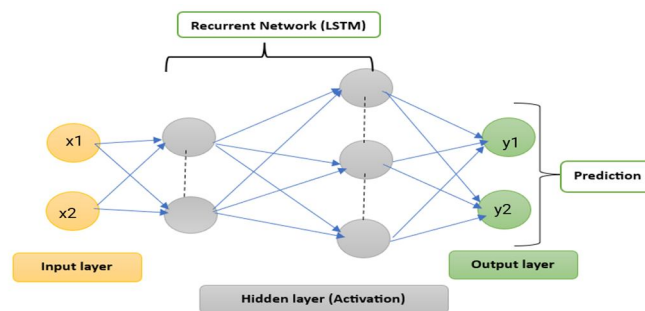


Figure 7 :RNN-LSTM Architecture

The figure presented above depicts the RNN model. It involves passing the input neurons x_1 and x_2 through a recurrent neural network that follows the LSTM mechanism. LSTM utilizes the current input, hidden state, and cell state to create a new hidden state. The output from the LSTM layer is then fed into a fully connected layer, which generates the output neurons y_1 and y_2 .

VI. EXPERIMENTAL DETAILS -IMPLEMENTATION OF ALGORITHMS USED

A subfield of artificial intelligence called artificial neural networks is based on the brain and is biologically inspired. Artificial neural networks are often referred to as computational networks based on biological neural networks, which form the framework of the human brain. Just like neurons in a real brain, neurons in artificial

neural networks also have connections between them in various layers of the networks. These neuronal structures are known as nodes. To create a model for categorizing musical genres, the following ANN sub-techniques are employed:

Batch normalization formula: Batch normalization is a technique to enhance the development of deep neural networks by normalizing the inputs of each of the layers. The batch normalization formula is:

$$x^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{\text{Var}[x^{(k)}] + \epsilon}} \quad (1)$$

where

$x^{(k)}$ is the input to the k^{th} layer

$E[x^{(k)}]$ is the mean of $x^{(k)}$ overcurrent.

branch

$\text{Var}[x^{(k)}]$ is variance of $x^{(k)}$ over current

branch

ϵ is a small constant.

Adam optimization formula: Adam is an optimization strategy that is used during training to change the weights of a neural network. Adam's formula is as follows:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \quad (2)$$

where

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \quad (3)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \quad (4)$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t} \quad (5)$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t} \quad (6)$$

where

g_t = Gradient at iteration t

β_1, β_2 = Hyperparameters that control the decay rate.

$\beta_1 = 0.90$,

$\beta_2 = 0.99$

Sparse categorical cross-entropy loss formula: Sparse categorical Cross-Entropy is a loss function that is used in Multi-Class classification problems with integer targets representing class labels. The formula for sparse categorical cross-entropy is:

$$\text{loss}(y, \hat{y}) = \sum_{i=1}^N y_i \log \hat{y}_i \quad (7)$$

Convolution is a mathematical operation that combines two functions to produce a third function. Here we use it to process the multivariate time feature of MFCC.

$$y_{i,j} = \alpha \left(\sum_{k=1}^m \sum_{l=1}^n w_{k,l} x_{i+k-1,j+l-1} + b \right) \quad (8)$$

where

w is the filter or kernel

x is the input

b is the bias

m, n are the dimensions of the kernel.

$\{m, n\} = \{3, 3\}$

α is the activation function

y is the output.

Max pooling is a down-sampling operation commonly used in convolutional neural networks to reduce the feature map dimensionality. It divides the input into non-overlapping regions and outputs the maximum value in each region, thus preserving the most salient features while reducing computation and overfitting. The formula for the maximum pooling operation is:

$$y_{i,j} = \max_{k,l} x_{i+k-1,j+l-1} \quad (9)$$

where

Here input variable is x

Variables used for dimensions of the pooling filter are k and i respectively
 Y is the output.

The proposed model's effectiveness was evaluated using performance measures, which included accuracy, recall, and specificity.

$$Recall = \frac{TP}{TP+FN} \quad (10)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

where

TP stands for true-positives.
 FN stands for false negatives.
 TN stands for true negatives.
 FP stands for false positives.

$$F1 \text{ score} = 2 * \frac{p*r}{p+r} \quad (12)$$

where

$$p = \frac{TP}{TP+FP}$$

$$r = \frac{TP}{TP+F}$$

VII. RESULT AND DISCUSSION

TABLE I. COMPARISON TABLE OF ACCURACY AND LOSS PARAMETERS

Algorithms	Parameters				
	Accuracy	Loss	Recall	F1 score	Training time (s)
MLP	59.91%	2.6	0.602	0.599	202.15
CNN	70.97%	0.82	0.709	0.708	210.06
RNN-LSTM	63.60%	1.05	0.633	0.627	595.67

According to the aforementioned findings, CNN is by far the most accurate algorithm among the three methods used to train music genre classification. The highest accuracy of the three algorithms is 70.97% for CNN. The accuracy rates for MLP and RNN-LSTM are 59.9% and 63.60%, respectively. CNN had the least loss, at 0.82 units. MLP and RNN-LSTM have respective losses of 2.6 & 1.05. MLP, CNN, and RNN-LSTM have recall-value of 0.602, 0.709, and 0.633, respectively. MLP and RNN-LSTM training take 202.15 and 210.06 seconds, respectively. However, compared to MLP and RNN-LSTM, CNN requires a training period of 595.67 seconds, almost twice as long.

VIII. CONCLUSIONS AND FUTURE SCOPE

Three artificial neural network models were investigated in this research for the creation of music genre classification: the Multilayer Perceptron, the Convolved Neural Network, and the Recurrent Neural Network followed by a mechanism called as long short-term memory. Even though MLPs are still useful for some purposes, they have limitations because they can't handle sequential input and can't pick up spatial information. Although they fall short of CNN or RNN-LSTM on more difficult problems, MLPs are nevertheless used for classification and regression issues. To conclude, this study explored the effectiveness of three artificial neural network models, namely multilayer-perceptron-network (MLPN), convoluted-neural-network (CNN), and recurrent-neural-network followed by long short-term memory (RNN-LSTM), for music genre categorization. The outcomes of the study indicated that, while MLPs are still useful for some purposes, they have limitations

and fall short on more challenging problems. On the other hand, CNN and RNN-LSTM showed superior performance in music genre classification, with CNN demonstrating increased accuracy and less loss. Therefore, it can be concluded that CNN and RNN-LSTM are more suitable for music genre classification tasks compared to MLPs. These findings are valuable for researchers and developers interested in using artificial neural networks for music genre classification and can inform the development of more accurate and efficient music classification models in the future.

REFERENCES

- [1] Nizamettin Aydin, Music Genre Classification and Music Recommendation by Using Deep Learning (oct 2018).
- [2] Vaibhavi Macharla, Music Genre Classification using Neural Networks with Data Augmentation A Make in India Creation (Jan 12, 2021).
- [3] Elbir, A., Aydin, N.: ‘ *Music genre classification and recommendation by using machine learning and deep learning*’. 2018 Innovations in Intelligent Systems and Applications Conf. (ASYU), 2018, Adana, Turkey, 2018, pp. 1– 5
- [4] Xining Luo, Automatic Music Genre Classification based on CNN and LSTM (2023).
- [5] Hindawi, A Music Genre Classification Method Based on Deep Learning (2022).
- [6] Lam Hoang, Literature Review about Music Genre Classification (June 2018).
- [7] D.Nevetha, A.Nithyasree, A.Parveenbanu, Mrs.Jetlin CP, Music Genre Classification Using Machine Learning (2021).
- [8] Prajwal R, Shubham Sharma, Prasanna Naik, Mrs. Sugna Mk, MUSIC GENRE CLASSIFICATION USING MACHINE LEARNING (July 2021).
- [9] Hareesh Bahuleyan, Music Genre Classification using Machine Learning Techniques.
- [10] Navneet Parab, Shikta Das, Gunj Goda, Ameya Naik, Music Genre Classification Using Deep Learning (oct 2021).
- [11] Snigdha Chillara, Kavitha A S, Shwetha A Neginhal, Shreya Haldia, Vidyullatha K S, Music Genre Classification using Machine Learning Algorithms: A comparison (May 2019).
- [12] Lin Feng, Shenlan Liu, Jianing Yao, Music Genre Classification with Paralleling Recurrent Convolutional Neural Network (Dec 2017).
- [13] Derek A. Huang, Arianna A. Serafini, Eli J. Pugh, Music Genre Classification.
- [14] Xiabin Li and Jin Li, Music Classification Method Using Big Data Feature Extraction and Neural Networks (2022).
- [15] Sheeba Fathima, Sagar U S, Shreya C K, Music Genre Classification using Deep Learning (2021)
- [16] <https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification>