Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 11, Issue. 2 : 2020
ISSN : **1906-9685**

USING NEURAL NETWORKS FOR MUSICAL GENRE CLASSIFICATION

[#1]Mr.CHADA SAMPATH REDDY, *Assistant Professor*
[#2]Mrs.VUMMENTHALA MAMATHA, *Assistant Professor*
Department of Computer Science and Engineering,
SREE CHAITANYA INSTITUTE OF TECHNOLOGICAL SCIENCES, KARIMNAGAR, TS.

Abstract. In this post, we'll look at how neural networks can be used to identify different musical genres. Spectrogram pictures generated from music snippets are used to train a neural network to categorize songs. The primary goal of the research is to examine the model's features. Using the NN approach on two separate data sets produced the greatest results. In this video, we utilize a Convolutional Neural Network model to classify ten different types of music.

INTRODUCTION

Music Genre Classification

Internet access allows you to discover new ways to enjoy live performances and remixes of old favorites. Instead of saving music locally, a user can stream it from an online song collection. It conserves space. To listen to the music whenever you want, "like" it or add it to a playlist. People are incredibly perceptive, and they can identify an artist, genre, album, and song from just a few seconds of video. To analyze this intelligence and expose material wealth, many NN (Neural Network) methodologies have been enhanced with cutting-edge technology.

Machine Learning and Neural Network

Machine learning has piqued the interest of many individuals. There are numerous machine learning (ML) algorithms available for diverse applications. Methods of machine learning include supervised, unsupervised, semi-supervised, and reinforcement learning. Labeled training data assists guided machine learning systems in producing results. This makes it simple to retrieve the model quickly. Unsupervised learning makes use of unlabeled data to locate relevant features and complete difficult tasks. Semi-supervised learning makes use of a dataset that has a large amount

of unlabeled data and a small amount of labeled data. Input is required for reinforcement learning. Reinforcement learning teaches the maximizing of rewards. Reinforcement learning is frequently used in online entertainment to award correct predictions. Neural networks (NNs) are machine learning algorithms that can extract meaningful features from massive amounts of data and utilize them to generate structures. With accessible data, the model is trained in a neural network. When the NN categorizes fresh or test data, the trained model predicts distributions.

Neural Network

The feature extractor and the classifier model are the two fundamental components of a conventional music genre categorization. Based on these two reasons, the feature extractor is critical in assessing the efficacy of the MGC.

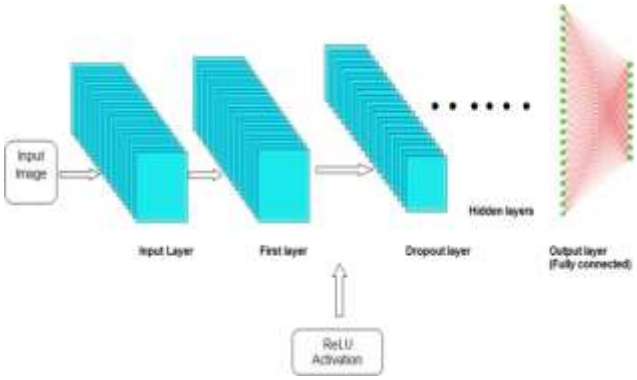


Fig. 1. Constructing a Cognitive Neural Network

CNNs, like pictures, handle grid-structured

multidimensional vectors. Figure 1 depicts a CNN layout concept. For the best results, use a CNN for binary and multiclass classification. Only the output type differentiates them. To train a model, use floral images. After preprocessing an image, CNN obtains pixel vectors. In the output class, pixel data are transformed into labels. Photographs of flowers will provide a variety of floral kinds to the final category. CNN models that have been trained on input images recognize new images. During classification, training set picture labels are matched to class names. Back propagation is used to change network weights during training in order to increase CNN accuracy. The number of rounds has an effect on the CNN structure. With training dataset features, a hierarchical model improves image categorization.

Feature extraction and genre recognition are two of our most important contributions.

Improving this model may benefit a variety of categorization tools.

This is how topics are categorized. Recent music genre classification advancements are listed in linked works. The document describes the concept and architecture in detail. In "Implementation work and findings," the techniques, motivation, and findings of the study are explained. Conclusions summarize studies and make recommendations for further measures.

1. RELATED WORKS

For a 2020 study, Craig M. Gelowitz and Nikki Pelchat successfully categorised recordings into musical genres. In two audio files, he reassembled 132,000 spectrogram fragments from 1880 recordings. Each clip was separated by 2.56 seconds. Thus, 128x128 pixel slices and additional spectrogram slices per genre were used to train the Neural Network. For activation, we used softmax and ReLU. Hamming Windows generated spectrograms from mp3 data, and CNN layers improved accuracy to 85%.

In 2021, M. and Jaime Ramrez Castillo M. Mercado formed a partnership. Julia Flores demonstrated a YouTube music video platform based on genre. These ten-second musical samples fall into three categories. SVM, Naive Bayes, Feed Forward Deep Neural Networks, and Recurrent Neural Networks were used to train the models. The models are web-based services. The user can utilize these models to see how each model "hears" the music based on the kind at a specific point in the song.

Wing W. Y. NG's music in 2020 covers a wide range of sounds and levels of abstraction. To avoid same-level feature extraction (FE) and learn long-term dependencies from many layers, he proposed combining a CNN with NetVLAD and self-attention. Meta classifier is used by Music Genre Classifier. On the GTZAN, ISMIR2004, and Extended Ballroom datasets, he shows how the proposed technique outperforms state-of-the-art models.

Costa et al. (2016) examined music genre recognition using spectrograms and multiple datasets. On this data set, the Binary Pattern performed the best. To train the models and obtain cutting-edge results across numerous audio databases, he used Gabor filters, Local Binary Patterns, Phase Quantization, and spectrogram extraction. He compared CNN performance to features created by hand. For his research, he examined the ISMIR 2004 Database and listened to African, Latin American, and Western music.

CNN is ideal since it can be combined with any musical genre classifier, he discovered. CNN may have the most recent and greatest African photographs. The LMD database was created using CNN and Robust Local and has a 92% identification rate. On ISMIR 2004, CNN outperformed other attribute-focused classifiers but not the best modern models.

2. PROPOSED ARCHITECTURE OF MGC

Real-World Application of the Model

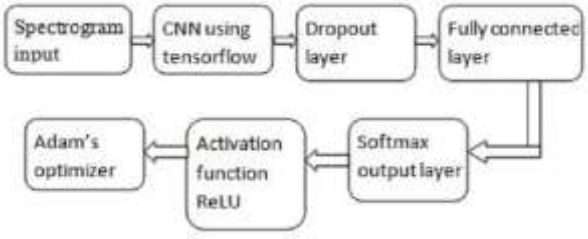


Fig. 2. Building design

A dense layer with 512 characteristics is used in the suggested CNN model. Each dense layer is immediately followed by a 0.2-factor dropout layer.

Figure 2 displays the overall framework of the

music genre categorization model constructed with Keras and Tensorflow. The model has four separate layers plus the input.

All of these layers rely on the activation function of the ReLU. The Softmax activation technique is used by the fully connected output network layer. This is rated according to its loss and accuracy. The suggested model's loss is determined using the sparse categorical cross entropy function. While the model is being trained, Adam's planner is utilized to make adjustments to the neural network weights. To train the proposed model, we run it through 600 epochs, each of which takes approximately 20 ms.

Simply set all variables to zero to eliminate overfitting. To put it another way, it is destroying brain cells. We may make educated assumptions about the category information using a dense layer and the ReLu activation function. It is not always the case that adding epochs results in a more accurate assessment. As a result, the value of epochs must be considered.

Music Database

- The music database "MusicNet" and the western database "GTZAN" are used in this work.
- The GTZAN archive contains a thousand 30-second music clips. There are three subfolders in the downloaded zip file.
- Genres Original includes 1,000 stereo.wav recordings ranging in length from 30 seconds to one minute, evenly spread over ten distinct musical genres. As a result, each category has 100 songs.
- The following picture depicts spectrograms for each 30-second audio clip. Display this image to the model.
- The third folder contains CSV files containing data from the prior two directories' audio recordings. Music Net, which contains 330 free classical MP3s, is the second dataset used in the suggested study. There are three documents in the set:
- This package contains the entire Music Net dataset. PCM audio wave files (.wav) and textual note label lead files (.csv) are included. Data is divided into train and test sets to train and test the model.
- The musicnet_metadata.csv file contains a list of Music Net-connected recordings. Using Music Net IDs, this file connects music data and label files.
- Music Net_midis.tar.gz Sample MIDI files for Music Net tags are included.

Data Analysis

Using files from GTZAN and Musicnet that are publicly available. The.wav files were all 16-bit and sampled at 22050 Hz in mono. The collection is mostly geared for developers and focuses on visual and aural feature analysis. Each column in the GTZAN collection represents a feature, and each row is an audio recording. The dataset's dimensions are 9990 by 60.

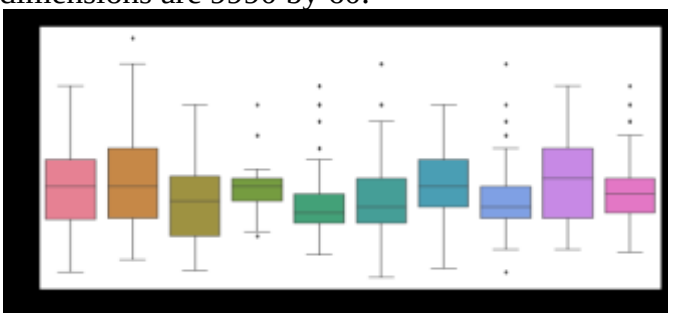


Fig.3. A Venn diagram of musical genres

The previous image is a box plot that depicts the distribution of genre and BPM data across the dataset.

Both types of data are used to compute accuracy and loss. This demonstrates the proposed classification methodology's general applicability and accuracy.

Principal Component Analysis (PCA) can be used to reduce the number of unsupervised dimensions in a database. An algorithm converts high-dimensional data into low-dimensional features without revealing the original category labels. PCA can decrease data in two or three directions, as seen in the diagrams below (PC1, PC2, and PC3).

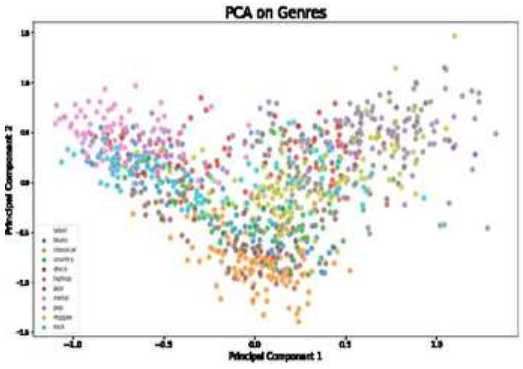


Fig.4. A Look at the Important Elements

Image 4. illustrates the PCA of the music

dataset, which highlights the important aspects of the data and reduces the dataset's dimensionality. No data is lost when a three-dimensional collection is reduced to a two-dimensional one. The data is more spread along PC1 than PC2, as seen in the scatterplot. This means that any classifier can separate the various types and subcategories correctly.

Preprocessing

Before feeding input into CNN for machine learning models, it is critical to "preprocess" the data. It is supposed to shorten the time required to train models and make the network more efficient. Pre-emphasis, frame, and widowing are all data preparation techniques used to extract meaningful information from raw data.

Scaling the features

Standard scalar is the best technique for normalizing audio features before processing since it eliminates the mean and normalizes the output to a variance of one.

Consider the example of x. You can get to it by

$z = (x - u) / s$

If you wish to employ machine learning or a neural network for classification, you must first standardize the dataset. The average value of the training samples is represented by u in this formula, while the standard deviation is represented by STD.

Visualizing Audio Files

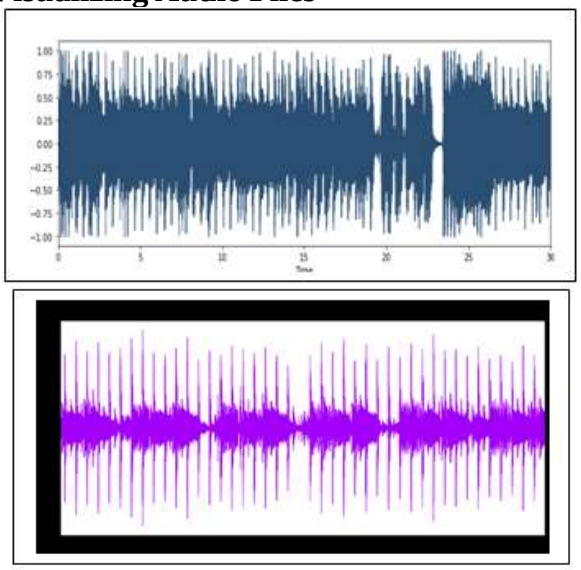


Fig. 5. The raw waveform representations of (a) hip-hop 14 and (b) reggae

Liburosa was used to plot the raw wave file data. Figure 5 depicts a sound waveform. The time and intensity axes are represented by the X and Y axes, respectively. Because these graphic representations enable for rapid processing of auditory information, the similarity score may be determined quickly.

Feature Extraction

Two common uses of feature extraction are data analysis and the finding of causal linkages. The use of feature extraction techniques such as Independent Components Analysis, Principal Components Analysis, and Linear Discriminant Analysis can minimize the complexity of a set.

Models can read the information right away because it has been transformed to their preferred format. For categorizing, guessing, and recommending algorithms, FE is a critical evaluation.

We go over the study's findings once more.

Zero-crossing rate

The zero crossing rate (ZCR) represents how frequently an audio wave passes from positive to negative or vice versa [13]. This ability to organize, identify, and recall facts connected with spoken or verbal communication is important to people.

Liburosa was used to plot the raw wave file data. Figure 5 depicts a sound waveform. The time and intensity axes are represented by the X and Y axes, respectively. Because these graphic representations enable for rapid processing of auditory information, the similarity score may be determined quickly.

Feature Extraction

Two common uses of feature extraction are data analysis and the finding of causal linkages. The use of feature extraction techniques such as Independent Components Analysis, Principal Components Analysis, and Linear Discriminant Analysis can minimize the complexity of a set.

Models can read the information right away because it has been transformed to their preferred format. For categorizing, guessing, and recommending algorithms, FE is a critical evaluation.

We go over the study's findings once more.

Zero-crossing rate

The zero crossing rate (ZCR) represents how frequently an audio wave passes from positive to negative or vice versa [13]. This ability to

organize, identify, and recall facts connected with spoken or verbal communication is important to people.

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

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.

Chroma Feature

The Chroma Feature is an excellent tool for determining the significance of musical sound features. The spectrum energy of the signal is displayed throughout its 12 segments. To safely determine the degree of resemblance between two audio samples.

We use a method called a Short Time Fourier Transform (STFT) to extract color information from sound. When the extraction quality is high, the results are excellent.

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

Where, fk is the k-th frequency,
N is the number of frequency bins.

The frequency-dependent spectral amplitude is denoted by P(fk).

Spectral Roll-Off

The portion of the frequency spectrum that is lower than the complete spectral frequency is referred to as spectral roll-off.

Solving for roll off Point = i yields the spectral roll off point.

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

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.

Mel-Spectrogram

A mel-spectrogram plots an audio source's frequency content against time. A logarithmically scaled magnitude spectrogram (or mel) displays sound frequency information along the Y-axis. Mel-frequency is written as follows:

Mel (f) is calculated as f/700 + 2595log10.

(4)

Mel-Frequency Cepstral Coefficient (MFCC)

The Mel-Frequency Cepstral Coefficient (MFCC) is commonly used to perform automatic sound and speech categorization. The cepstral coefficient determined using the mel-frequency is known as the mel-frequency cepstrum coefficient (MFCC). The vectors of the preceding characteristics are combined to form the input feature vector.

3. IMPLEMENTATION

The proposed model for categorizing music categories was implemented in Python, and the anticipated results were obtained. As a result, 67% of the sample was used for training and 33% for testing. Librosa, as a preprocessor, prepares each image for the specified CNN model. The NN model was built using Deep Convolutional Networks and the Tensorflow framework. In this case, the number of rounds is limited to 128. The first four layers are made up of dropout and convolutional layers with a 0.2 probability. To prevent the model from fitting too well, these layers are placed after each additional layer. In this scenario, we use ReLU to activate the output layer's input layer and Soft max to activate its hidden layer. The outputs of all levels are pooled and transferred to the final, fully connected layer. This approach can produce a variety of unique results. A max layer gently governs the outputs of ten simple kinds. The "sparse categorical cross entropy" and the mean initialization learning rate are used to calculate the loss function. With Adam's optimizer, it is simple to find limits. It comes in handy when you need to quickly compute a factor or work with a large dataset. We trained a second neural network using the identical NN design to confirm consistency in performance across data sets. Despite their genre differences, the two datasets analyzed in this analysis have the same number of melodies or musical pieces. This is done on purpose in order to measure test accuracy and preserve fairness.

Table 1. The GTZAN table is shown below.

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

The total number of records for each study topic is shown in Table 1. All of the fields

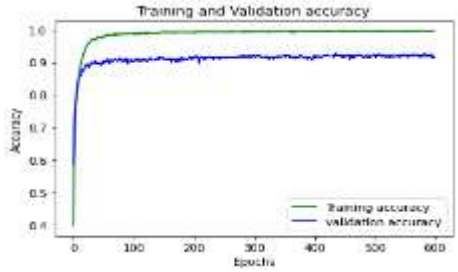
represented in the other collection had a consistent number of files. The test's 92.65% accuracy was obtained by comparing two datasets with the same number of songs in each genre. As a result, the same model can be used to a range of datasets with promising results.

The accuracy-to-loss ratios for the two datasets are shown below. Table 2 displays the outcomes of the suggested MGC model.

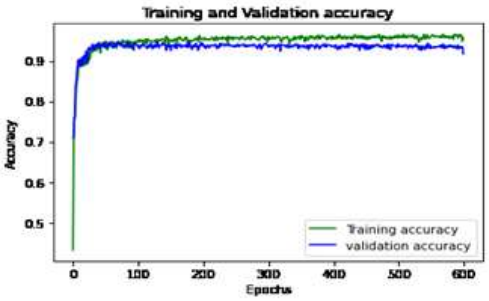
Table 2. As a result, precision and suffering.

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

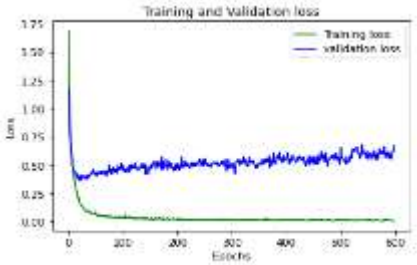
GTZAN's Precision



The Precision of Music Net



completing GTZAN



There is an issue with MusicNet.

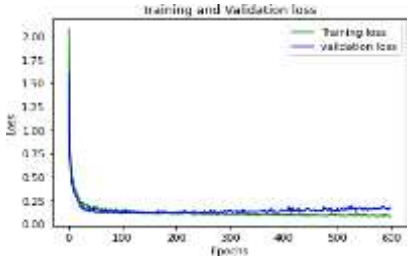


Fig. 6. By training two datasets, we were able to calculate training loss and accuracy.

Figure 6 illustrates a table outline.

Loss and accuracy curves for both the training and testing datasets are shown in the graph.

The third image. Compare and contrast pre-trained models

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

Table 3 provides a historical view on pre-trained model performance. This shows that simply adding more layers does not increase performance. On the other hand, increasing the model's size will enhance its precision.

4. CONCLUSION

This study gives a framework for categorizing various musical styles. The process of preparing materials for processing and obtaining musical characteristics. Data extracted from spectrograms is fed into a convolutional neural network. The neural network is divided into four tiers. To calculate probabilities, the ReLU activation function is utilized in the input layer, while the 10 classes and Soft max activation are employed in the output layer. A MIDI music library was used to examine the datasets' successive portions. We also feel that neural networks are the most effective machine learning method accessible today. Tensorflow is a framework for efficiently constructing and running CNN models.

In the future, changes to the weights' initialization and the inclusion of the entire melody, rather than only the first thirty seconds, are possible. We shall merge our separate efforts in the future to eliminate inaccuracies. Binary algorithms are also being studied for detecting whether or not a user enjoys a song recommendation from 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. T. Kim, J. Lee, J. Nam, Comparison and analysis of sample CNN architectures for audio classification, IEEE J. Sel. Top. Sign. Proces. 13 (2) (2019) 285–297.
4. 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.
5. J. Thickstun, Z. Harchaoui, S. M. Kakade, November 30, 2016 ,15 August 2021 [https://zenodo.org/record/5120004#.YbnBRjNBzIV]
6. 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.
7. 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.
8. 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.
9. 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.
10. Scheirer, E., and M. Slaney, "Construction and Evaluation of a Robust Multi feature Speech/Music Discriminator," IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume **2**, 1997, pp. 1221–1224.