Covenant Journal of Informatics & Communication Technology. Vol. 11, No. 2, December 2023
ISSN: ***print*** **2354 – 3566** ***electronics*** **2354 – 3507 DOI: An Open Access Journal**

Mel-Frequency Cepstral Coefficients and Convolutional Neural Network for Genre Classification of Indigenous Nigerian Music

Akinosho Oluwadamilola [a], Abayomi-Alli Adebayo [a,*], Arogundade Oluwsefunmi 'Tale [a], Oladejo Rachel Adefunke [a, b], Adedapo Olufikayo A. [c]
[a] Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria.
[b] Department of Computer Science, Ogun State Institute of Technology, Igbesa, Nigeria.
[c] Department of Mathematical and Computer Sciences, Fountain University, Osogbo, Nigeria.
∗ Corresponding author: abayomiallia@funaab.edu.ng, 07030672420

Received: 12.08.2023 Accepted: 21.12.2023 ublication: December 2023

Abstract— **Music genre classification is a field of study within the broader domain of Music Information Retrieval (MIR) that is still an open problem. This study aims at classifying music by Nigerian artists into respective genres using Convolutional Neural Networks (CNNs) and audio features extracted from the songs. To achieve this, a dataset of 524 Nigerian songs was collected from different genres. Each downloaded music file was converted from standard MP3 to WAV format and then trimmed to 30 seconds. The Librosa sc library was used for the analysis, visualization and further pre-processing of the music file which includes converting the audio signals to Mel-frequency cepstral coefficients (MFCCs). The MFCCs were obtained by taking performing a Discrete Cosine Transform on the logarithm of the Mel-scale filtered power spectrum of the audio signals. CNN architecture with multiple convolutional and pooling layers was used to learn the relevant features and classify the genres. Six models were trained using a categorical cross-entropy loss function with different learning rates and optimizers. Performance of the models was evaluated using accuracy, precision, recall, and F1-score. The models returned varying results from the classification experiments but model 3 which was trained with an Adagrad optimizer and learning rate of 0.01 had accuracy and recall of 75.1% and 84%, respectively. The results from the study demonstrated the effectiveness of MFCC and CNNs in music genre classification particularly with indigenous Nigerian artists.**

Keywords/Index Terms— **CNN, Genre Classification, Librosa, MFCC, Music Information Retrieval (MIR), Nigerian.**

1. INTRODUCTION

Music genre classification (MGC) is a field of study within the broader domain of Music Information Retrieval (MIR). The goal of MGC is to automatically categorize pieces of music into predefined genres. This task is challenging due to the subjective nature of music and the wide variety of genres that exist. However, the ability to classify music into genres has many potential applications, such as music recommendation systems, music cataloguing, and music content analysis. MGC has been studied for several decades, and a wide range of methods and techniques have been proposed. Early approaches focused on hand-crafted features, such as tempo and melody, and rule-based systems. With the advent of traditional machine learning (ML) and deep learning (DL), more sophisticated approaches have been developed to automatically learn features from the audio data making it an active research area.[1] Deep learning is more and more used by the MIR community. This success can be explained by two reasons: the first one is that it avoids the more or less difficult extraction of carefully engineered audio features, the second one is that the deep learning hierarchical topology is beneficial for musical analysis because on one hand music is hierarchic in frequency and time and on the other hand relationships between musical events in the time domain which are important for human music perception can be analyzed by Convolutional Neural Networks (CNN). The most popular method is to use a spectrogram as an input to a CNN and to apply convolving filter kernels that extract patterns in 2D. However, as pointed out by J.Pons, T.Lidy and X.Serra [14], "a common criticism of deep learning relates to the difficulty in understanding the underlying relationships that the neural networks are learning, thus behaving like a black-box". It is why they experimented and showed, by playing with filter shapes, that musically motivated CNN may be beneficial. In the MIREX 2016 campaign, one of these authors showed that combining a CNN capturing temporal information and another one capturing timbral relations in the frequency domain is a promising approach for MGC .[15]

There are various musical preferences which make categorizing and recommending new songs in music listening apps and platforms is an essential and existing issue. One of the most effective methods for resolving this issue is to categorize music by their genre. There are several methods for music classification and recommendation. Many researchers have worked on this problem over the years, and different approaches have been proposed to achieve better accuracy. One main feature that separates one music from another is the genre. The classification of music genres is useful because it provides a framework for organizing and understanding different types of music. By categorizing music into different genres based on common musical characteristics such as rhythm, melody, and instrumentation, music can be easily identified and explored in a way that can be enjoyed and found meaningful.[2]

The listeners browse for and enjoy music has changed drastically. Storing or maintaining digital music is no longer an issue, but assisting customers to discover the music they enjoy by precisely analysing the music audio signals and creating suggestions based on the data collected from the signals has become an essential in practical applications such as music recommendation systems, music streaming services, and marketing of music products.[3] By understanding the genre of music that a listener enjoys, it is possible to make more accurate recommendations for other music that they may like. This can improve user experience for listeners and increase engagement with music products and services. Overall, MGC provides a valuable tool for both personal enjoyment and commercial applications in the music industry.

This study aims to develop MGC models to classify Nigerian music into respective genres.

2. EXPERIMENTAL AND RELATED WORK

There are numerous important steps in the research process for a study on MGC. Gathering a representative dataset of music from diverse genres is the first stage. This may entail obtaining or scraping audio files from different sources, converting the files to a standard format (for example, MP3 to WAV), and shortening the audio files to a predetermined duration (e.g., 30 seconds). Next, take key elements that can be utilized to train a ML model out of the audio samples. Mel-Frequency Cepstral Coefficients (MFCCs), a collection of coefficients that capture the spectral information of an audio signal, are often calculated to accomplish this.

After that, use the dataset to build a ML model. Select a suitable evaluation metric, such as accuracy or F1 score, to evaluate the effectiveness of the model. After the model has been trained, assess its performance on a different test set to gauge its accuracy and pinpoint any potential improvement areas. Utilizing strategies such as a confusion matrix may be necessary.

The study's approach framework, which is broken down into the two crucial stages of feature extraction and classification, is shown in Figure 1. Six steps altogether can be formed from these two phases:

1. The collection of the music files that were the source of the features that made up our dataset were collected.
2. The first thirty seconds and the mid thirty seconds of each audio files were extracted from the music files acquired in stage one,
3. The low-level features were extracted from the audio files obtained in step 2 and saved in as a JSON file.
4. The data was pre-processed at this step.
5. The model was trained on the dataset and then evaluated.
6. The model's accuracy was determined to demonstrate how accurately it categorized the audio.

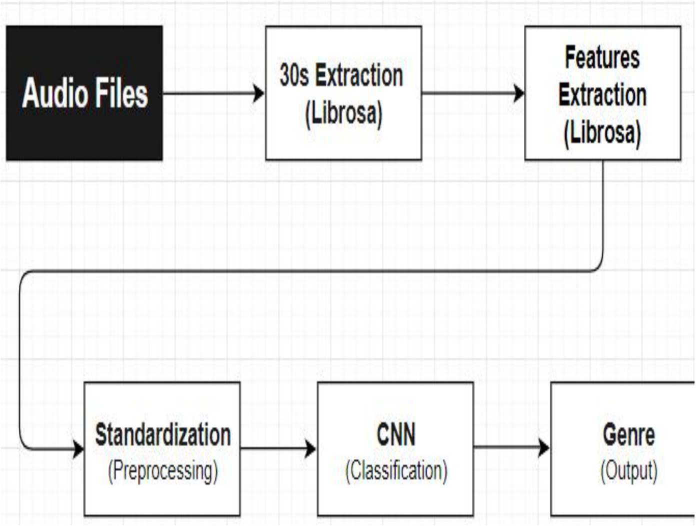

Figure 1: Methodology Framework

2.1 DATA COLLECTION

The data was downloaded from publicly posted songs on the Internet. The dataset consists of five hundred and twenty-four Nigerian songs of different genres. The songs were trimmed to 30 seconds from their initial lengths and then converted from .mp3 format to .wav format using the Librosa Python library. The song files were individually sampled at 22,050 Hz with a 32-bit mono resolution, allowing us to extract a significant number of features for the model. Except for highlife music, most of the downloaded songs were sung in Yoruba while others included traces of the language.

2.2 FEATURES EXTRACTION

This involves selecting and transforming a set of audio features that represent the musical content of the music in a compact and meaningful way. Common features used in MGC include (1) MFCCs, (2) Spectral Centroid, (3) Spectral Flux, (4) Chroma features, and others. The choice of the features depends on the specific task and the desired level of abstraction. The goal is to collect a set of characteristics that captures underlying musical structure, style, and genre information in the music. The primary characteristics that can be retrieved from the audio cues are content-based and text-based features. When describing music content, there are three primary characteristics that are typical, namely: (1) Timbre, (2) Melody/Harmony, and (3) Rhythm.[2]

2.2.1. TIMBRAL FEATURES

Timbral and textural features can distinguish similar beats or tunes. Each sound frame within a brief time span has timbral texture. Some timbre characteristics that are used to categorize music by genre are: (1) Temporal features, (2) Energy features, (3) Spectral shape features and (4) Perceptual features.[4]

For this study, the following timbre features were extracted:

a. Chroma_stft: The human perception of pitch is periodic in the sense that two pitches are thought to be similar in colour yet vary by an octave.[5]

b. Chromagram: They provide information about the harmonic content of a piece of music and can be useful for analysing the structure of musical pieces and understanding the relationship between different pitches. Chromas set consists of the twelve pitch spelling attributes: A, A♯, B, C, C ♯ D, D ♯, E, F, F♯ G, G♯ as used in Western music notation.

c. Root Mean Square Energy (RMSE): refers to the root mean square energy of a signal. Calculating the energy present in a signal is as follows.

$$\sum_{n=1}^{N}|x(n)|^2 \qquad (1)$$

A signal's "root mean square energy" (RMSE) is determined as follows:

$$\sqrt{\frac{1}{N}\sum_{n}|x(n)|^2} \qquad (2)$$

d. Zero crossing rate: is a feature used in audio signal processing to estimate the frequency content and periodicity of a signal. A higher zero crossing rate indicates a more rapidly changing signal, while a lower zero crossing rate indicates a signal that changes more slowly.

e. Mel-Frequency Cepstral Coefficients: calculated for short-term spectral qualities. For audio segments of 10–100ms, MFCCs are small, short-time spectral envelope audio feature set descriptors.[6]

f. Spectral Rolloff: This correlates to the frequency below, which is a percentage of the total spectral energy lies. The average and standard deviation of the spectral roll off is calculated across all frames of the audio signal.[6]

g. Spectral-Bandwidth: The wavelength or frequency interval of radiation leaving the exit slit of a monochromator between limits set at a radiant power level halfway between the continuous background and the peak of an emission line or an absorption band of negligible intrinsic width.

$$\left(\sum_{k}s(k)(f(k)-f_c)^p\right)^{\frac{1}{p}} \qquad (3)$$

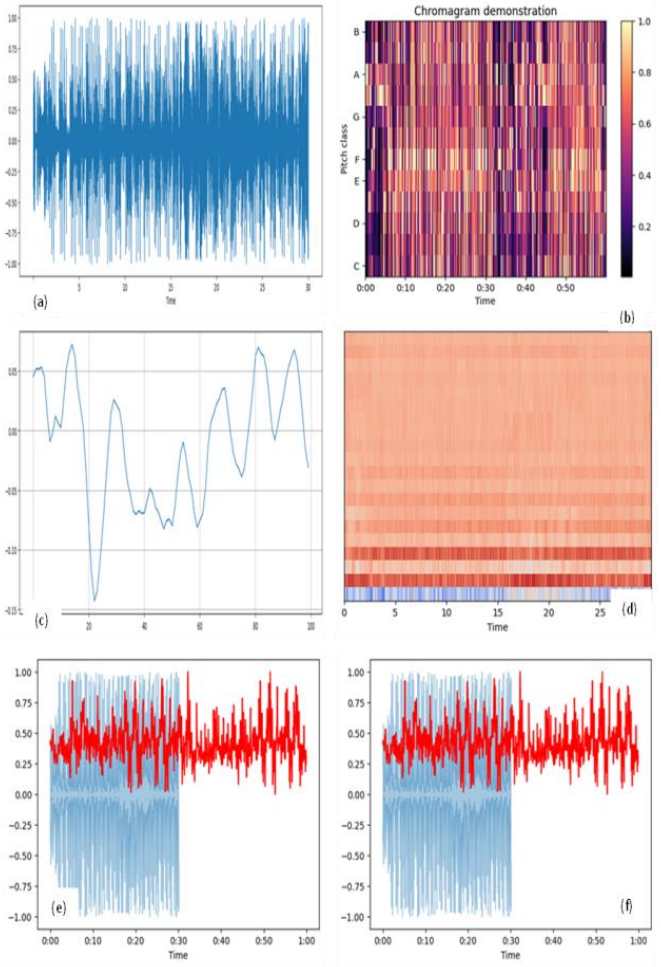

Figure 2: Visualization of (a) a waveform of a song signal (b) the computed chromagram of a song file (c) the computed ZCR of a song file (d) the computed MFCC of a song file (e) the computed spectral rolloff of a song file and, (f) the computed spectral centroid of a song file.

h. Spectral-Centroid: is a feature used in audio signal processing to estimate the "centre of mass" of the spectral content of a signal. It is a measure of the frequency content of a signal and provides information about the overall "brightness" or "darkness" of the signal.[6]

$$f_n = \frac{\sum_k S(k)f(k)}{\sum_k fk} \qquad (4)$$

2.2.2. RHYTHM

The time pattern of musical sounds and silences is rhythm. Rhythmic content refers to the characteristics of an audio stream that indicate its temporal movement, such as tempo. [7] The Beat Histogram (BH), which extracts the noteworthy periodicity from a song and is based on psycho-acoustic models that capture rhythmic and other fluctuations in frequency bands fundamental to the human aural system, is used to determine the rhythmic content's characteristics. Tempo is measured in beats per minute and denotes the musical speed (bmp).

2.2.3. MELODY/HARMONY

Harmony refers to the simultaneous sounding of different notes or chords, which creates a sense of vertical structure and progression within a piece of music. Melody, on the other hand, is the recognized sequence of pitched occurrences that occur in time. The vertical aspect of music is harmony, while the horizontal element is melody.[4]

2.3 DATA PREPROCESSING

Audio data pre-processing involves several steps to prepare audio data for ML models. The steps include (1) Sampling, (2) Normalization, (3) Resampling, (4) Windowing, (5) Spectrogram, (6) Mel Spectrogram, and (7) Data augmentation.

2.4 DATA EXPLORATORY

Data exploration is the process of analysing and understanding the characteristics of a dataset before building a model. This helps to identify patterns, relationships, and anomalies within the data and to ensure that the data is suitable for the intended use.

2.5 CLASSIFICATION MODEL

2D CNNs was the classifier used in this research to categorize the songs into their respective genres. A CNN is a form of deep learning neural network that is extensively used for image classification.[8] CNNs have made considerable advancements in recent years and are widely used for face recognition, object detection, audio recognition, and other computer vision applications.[9] In this study, CNNs were designed to extract features from input data using many layers of convolutional and pooling operations, and then classify the data using fully connected layers.

2.5.1 CONVOLUTION LAYER

The typical structure of CNN consists of the layers depicted in Figure 3. In the first few layers of a CNN, input features are subjected to convolutional procedures. The convolutional kernels are learned and used to extract local features from the input information. The network's depth is determined by the number of filters present in each of these layers.

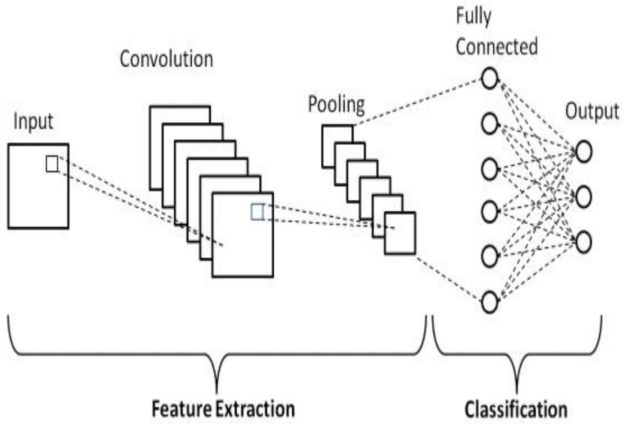

Figure 3: Convolutional Neural Network layers

2.5.2 POOLING LAYER

The pooling layers are used to reduce the spatial dimensions of the feature maps produced by the convolutional layers, while retaining the most important information. The pooling layer reduces the number of parameters and calculations in the network. This improves the efficiency of the network and avoids over-learning.

2.5.3 FLATTEN LAYER

A flatten layer is used in a CNN to convert a multi-dimensional array (such as a feature map generated by a Conv2D layer) into a one-dimensional array so that it can be fed into a Dense layer for classification or prediction. The flatten layer takes the high-dimensional feature maps from the previous layer and "flattens" them into a single 1D array of values, which is then fed into the next layer of the network.

2.5.4 MULTILAYER FEED FORWARD NETWORK (MLFF)

Multilayer Feed-Forward Network (MLFN) is an interconnected Artificial Neural Network with multiple layers that has neurons with weights associated with them and they compute the result using activation functions.[10] In a standard MLFF neural network, the neurons within each layer are fully connected to the neurons in the previous and next layers, meaning that each neuron receives input from every neuron in the previous layer and sends output to every neuron in the next layer. This allows the network to learn complex non-linear relationships between the input and output data, and to make predictions or classifications based on new input data.

2.6 ACTIVATION FUNCTIONS

In CNNs, activation functions are used to introduce non-linearity in the model. Some commonly used activation functions in CNNs are:

(1) ReLU (Rectified Linear Unit): This is the most used activation function in CNNs. It replaces all negative values with 0 and leaves positive values unchanged.
(2) Sigmoid: It compresses the input between 0 and 1, making it ideal for binary classification problems.
(3) Tanh (Hyperbolic tangent): Like sigmoid, it compresses the input between -1 and 1, and is often used in the output layer for multi-class classification problems.
(4) Softmax: It computes the probability distribution over multiple classes and is commonly used in the output layer for multi-class classification problems.

2.7.1 DROPOUTS

Dropout is a regularization technique in CNNs and fully connected networks (such as Multi-layer Perceptron). The idea behind dropout is to randomly "drop out" or "turn off" certain neurons during the training process, forcing the network to learn redundant representations and reduce overfitting.[11]

2.8 LOSS FUNCTION

A loss function, also known as a cost function, is a mathematical function that is used to evaluate the performance of a model by measuring the difference between its predicted output and the actual output for a given input. The goal of the loss function is to provide a quantitative measure of how well the model is performing, and to provide guidance for updating the model's parameters to improve its performance.

(1) Binary Cross Entropy (BCE): This is a binary classification loss function that measures the difference between the predicted probabilities and the true binary labels.

$$BCE\ loss =- y * log(p) \; - \; (1-y) * log(1-p) \qquad (5)$$

(2) Categorical Cross Entropy (CCE): This is a multi-class classification loss function that measures the difference between the predicted class probabilities and the true class labels.

$$\text{CCE} =- \sum_{i=1}^{c} y_i \log{(p_i)} \qquad (6)$$

(3) Softmax Cross Entropy (SCE): This is a multi-class classification loss function that uses the softmax activation function to convert the predicted class scores into class probabilities before computing the cross-entropy.

$$\text{SCE} =- \sum_{i=1}^{c} y_i \log{(softmax(z)_i)} \qquad (7)$$

2.9 TRAINING, TESTING AND VALIDATION

The dataset consists of 5,240 instances of song records, which was split into a training set and a test set of 80% (4,192) and 20% (1,048), respectively. The validation set will also be the same size as the test set. Splitting the dataset into these subsets will allow the ML model to learn patterns and relationships from the training data and then evaluate its performance on the test data, giving an estimate of how well it will generalize to new and/or unseen data.

2.10 OPTIMIZER

Optimizers are algorithms used in deep learning to update the model parameters based on the gradients calculated during backpropagation. The purpose of the optimizer is to minimize the loss function, which measures the difference between the predicted outputs and the true labels.

2.11 EPOCH

An epoch is a term used in ML and refers to one complete iteration over the entire training dataset during training of a model. One epoch means that each sample in the training dataset has been used once for updating the model's parameters. The number of epochs is a hyperparameter that can be set by the user, and can impact the model's performance, with more epochs potentially leading to better results but also a longer training time.

2.12 BATCH SIZE

Batch size takes in several input samples at each epoch of training. More batch size can increase the accuracy as well as training time. It is a hyperparameter in deep learning that specifies the number of samples to work through before updating the model's weights.[12] In general, larger batch sizes result in faster progress in training, but don't always converge as quickly as smaller batch sizes. On the other hand, smaller batch sizes can converge faster, but can be slower to make progress in terms of actual training time.[13]

2.13 EVALUATION METRICS

Accuracy, Precision, Recall, and F1-score are the metrics used to assess the model that was built.

(1) $Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$ (8)

(2) $Precision = \frac{True\ positives}{(True\ positive + False\ positive)}$ (9)

(3) $Recall = \frac{True\ positives}{(True\ positive + False\ negative)}$ (10)

(4) $F1\ score = \frac{2*Precision*Recall}{(Precision+Recall)}$ (11)

(5) Training and validation loss are the measures of how well a ML model is fit to the training and validation data. The values are used to evaluate the performance of the model and to tune its hyperparameters. Training loss is the loss calculated on the training data after each iteration of the training process. It is used to monitor the progress of the training process and to check if the model is overfitting or underfitting the data. Validation loss is the loss calculated on the validation data after each iteration of the training process. It is used to monitor the performance of the model on unseen data and to check if the model is overfitting or underfitting the data.

2.9 RELATED WORK

Mermelstein used Mel-frequency cepstrum (MFCC) to measure the distance of the source if sound for speech recognition.[17] In MIR tasks information retrieval from images is also vital as information from music signal can be represented as images. Ojala et al. introduced Local Binary pattern (LBP) as an efficient texture descriptor that labeled the pixels of images by neighbor pixel thresholding and considered the result as a binary number.[18] In 2002, mixtures of Gaussians model and k-nearest neighbor (KNN) was used along with three hand-selected features (timbral texture, rhythmic content, and pitch content) for music genre classification (MGC) and achieved an accuracy of 61% , which in comparison to average human accuracy of 70% was a remarkable success.[19] However, the popularity of the use of hand-crafted feature extraction has decreased in recent times because this process imposes a significant blockade as expertise in the relevant field is required to obtain hand-crafted features. This requirement limits the generalization of MGC, as in different environments, the considered features change. In 2011, from spectrograms or images which were generated from audio signals using short-term Fourier transform (STFT), textural features were extracted.[20] This and the rise in popularity of a parallel processing architecture named Graphics Processing Unit (GPU) made the use of deep learning models for feature extraction and classification tasks, hence MIR tasks of different music tracks feasible. Deep learning is such a technique that enables systems to learn by example and has proven to enable systems to understand complex perception tasks with maximum precision. The deep learning process includes two phases, training and inception. Labeling of large data and identification of matching characteristics takes place during the training phase. On the other hand, in the inception phase, a concluding decision is made, and new unexposed data are labeled using previously learned knowledge. For long deep learning, models are in use with great success for different Computer Vision (CV) tasks, such as image classification [21], object detection[22,23], image caption[24],facial expression recognition[25], image recognition[26] and so on. Being inspired by the success of the application of deep learning models for various CV tasks, researchers being able to represent the audio signal as a spectrogram have applied different deep learning models such as CNN, RNN, and CRNN for various MIR tasks and have achieved state-of-the-art performance. CNN has been widely used for various MIR tasks such as music recommendations[27], automatic tagging[28], and feature learning[29], and so on. A popular approach of using CNN for MIR tasks involves using a spectrogram as an input to the CNN and extract patterns in 2D by applying convolving filter kernels. In 2010, for music genre prediction Li et al. developed a CNN using raw Mel-frequency cepstral coefficients (MFCC) as input[30]. For MGC, a CNN was used to capture temporal information, and another to capture timbral relations in the frequency domain[31]. The use of CNN for MGC has inspired many researchers such as Senac et al.[32] who used the filter dimensions of CNN in such a way so that it is interpretable time and frequency. In their experiment, they used eight features chosen along with dynamics, timbre, and tonality dimensions as inputs of CNN and achieved global accuracy of 89.6% against 87.8% for 513 frequency bins of a spectrogram. The experiment proved that music features are more efficient for MGC than a huge number of spectrogram frequency bins. Bahuleyan conducted a comparative study on the performance of deep learning models requiring a spectrogram as input, specifically VGG-16 (a robust CNN architecture) and machine learning classifiers that need is trained with hand-selected features for music genre classification.[33] In his experiments, the CNN architecture outperformed the feature-engineered models. Yang et al. used a Mel-scale spectrogram as input to his proposed CNN architecture, which applied the output of duplicated convolutional layers to different pooling layers to produce information for music genre classification.[34] Though the CNN architecture obtained a remarkable accuracy of 90.7% , the performance suffered in a significant amount when it came to correct classification of the country genre, and it suggested that the use of 3 seconds of raw audio as input have caused the loss of performance. But a major setback for most of the CNN-based music classification models is the requirement of data-augmentation and large datasets for the training of the models, as the models often require learning large parameters. A study in 2019 solved this problem to a certain extent by the introduction of a novel CNN architecture named Bottom- up Broadcast Neural Network (BBNN).[16]

3.0 RESULTS AND DISCUSSION

This section presents the discussion of the results obtained from this study. The programming language used in this study was Python. The experiment was performed in Jupyter notebook using Scikit learn with the CNN classifier on the dataset.

3.1 DATA COLLATION

For this study, thirty seconds audio were collected from the start and middle of 524 songs across nine different genres namely: Afro-beat (38), Afrobeats (92), Apala (68), Fuji (62), Highlife (54), Juju (60), Reggae (54), Waka (46), and YorubaRap (50). The Python Librosa library was used to trim and extract the audio clips for further processing and analysis.

3.2 FEATURE EXTRACTION AND DATA PREPROCESSING

Librosa, one of the most well-known Python libraries for audio analysis, was utilized for data processing and feature extraction. The Mel Frequency Cepstral Coefficients (MFCCs) are a set of characteristics that describe the spectral envelope. The set has between 10 and 20 elements used to model the human voice qualities. Thirteen MFCCs were retrieved in the dataset collected for this study, and feature scaling was also performed. The dataset's labels were additionally pre-processed utilizing an array encoder. The target variables were converted to numeric target labels in alphabetical order as follows: Afro-Beat =0, Afrobeats =1, Apala =2, Fuji =3, Highlife =4, Juju =5, Reggae, Waka =7, YorubaRap =8.

3.3 CLASSIFICATION

Following the pre-processing of the dataset, a multiclass classifier was developed to categorize the songs appropriately using a 2D CNN as in the model summary and hyperparameters in Table 1.

Table 1: Summary and Hyperparameters for the 2D CNN Model

Layer (type)	**Output Shape**
Convo2d_33 (Conv2D)	(None, 128, 11, 32)
Max_pooling2d_22 (MaxPooling2d)	(None, 64, 6, 32)
Batch_normalization_12 (Batch Normalization)	(None, 64, 6, 32)
Convo2d_34 (Convo2D)	(None, 62, 4, 32)
Max_pooling2d_23 (MaxPooling2D)	(None, 31, 2, 32)
Batch_normalization_13 (Batch Normalization)	(None, 31, 2, 32)
Conv2d_35 (Convo2d)	(None, 30, 1, 32)
Max_pooling2d_24 (MaxPooling2d)	(None, 15, 1, 32)
Training data	80%
Test data	20%
Validation data	Same as test data
Batch size	32
Epochs	50
Kernel size	3
Dropout	0.5
Pool size	2
Activation functions	Relu and SoftMax
Loss function	Sparse categorical cross entropy

The values for training, validation accuracy and loss were obtained after the model had been trained and evaluated. Figure 4 depicts a comparison of the performance of all learning rate schedules and adaptive learning rate approaches.

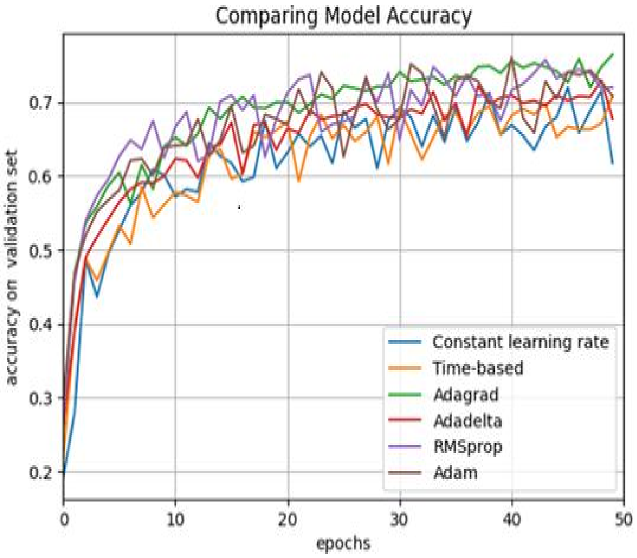

Figure 4: Comparing Performances of Different Learning Rate Schedules and Adaptive Learning Algorithms over 50 epochs.

3.4 EVALUATION

The metrics used to evaluate the performance of the CNN classification model are accuracy, confusion matrix, precision, recall and f1-score.

3.5 RESULTS AND DISCUSSION

Table 2 shows the model performances using different adaptive learning rate methods and optimizers, Model 3, which used Adagrad returned the best model accuracy with 75%.

Table 2: Classification accuracy on the CNN models

CNN models	**Optimizers**	**Learning rate (Lr)**	**Momentum**	**Decay**	**Test Accuracy**
Model 1	SDG	Constant lr = 0.1	0.0	0.0	63%
Model 2	SDG	Time based decay lr = 0.1	0.8	Lr/epoch	71%
Model 3	Adagrad	0.01		0.0	75%
Model 4	Adadelta	1.0		0.0	69%
Model 5	RMSprop	0.001		0.0	69%
Model 6	Adam	0.001		0.0	70%

A summary of the result is given in Table 3 shows the classification performance in form of a confusion matrix while the precision, recall and f1-score of each music genre is presented below. The column in Table 3 indicates the correct genres while the rows indicate the predicted genres. The portion of the songs rightly classified lies in the diagonal of the table highlighted in bold. The retuned classification accuracy of Afro-beats, Afrobeats, Apala, Fuji, Highlife, Juju, Reggae, Waka, and Yoruba Rap are 87.50%, 80.89%, 98.86%, 77.95%, 75.74%, 86.46%, 69.34%, 100% and 95.88%, respectively. Waka music had a perfect genre classification with no instance incorrectly classified as also reflected in its recall and F1-score. It was closely followed by Apala that had one misclassification wrongly classified as Fuji music while the model performs least with Reggae music having 42 instances misclassified as Juju (13), Highlife (12), Afrobeats (8), Afrobeat (4), Apala (2), Yoruba rap (2) and Fuji (1). The reason why some genres had low performance rate might be because the segmented pieces do not contain enough information for these genres or there are similar sounds in these genre types. Afro-beat, Afrobeats, Fuji and Highlife also had poor genre classification results, which can be attributed to the seemly similarity in their percussions. Reggae's poor result could be attributed to the similarity its shared with Highlife, Juju, and afrobeat(s).

4.0 CONCLUSION

Using MFCC features extracted from song time slices and a 2D CNN as the model, the Nigerian songs were successfully categorized into different genres for this research. Based on the study results, it was determined that the CNN model with the AdaGrad optimizer performed the best for music genre classification. The model's accuracy was 75.19%, and its recall was 84.00%. Variable misclassifications of the genres included in the dataset were determined using the model's confusion matrix output. The results of the experiment also suggest that incorrect genre classifications were consistent with human observations.

This was expected, as some songs span multiple genres, making it difficult for humans to determine the genre. To expand the dataset to include Nigerian-English contemporary music, artist recognition, and mood detection, additional study is required. By building and utilizing datasets containing more genres and songs in future research, models that categorize music genres more precisely and bring more benefits to the learning and teaching of music can be uncovered. In addition, if the established models are incorporated into music recommendation systems, the effectiveness and precision of these systems can be enhanced. For, future studies, experiment with alternative deep learning methodologies should be considered to examine how parameter optimization will improve the predictions of the deep learning models.

Table 3: Results of confusion matrix with CNN model 3 classifier

Genre	Afro-beats	Afro beats	Apala	Fuji	Highlife	Juju	Reggae	Waka	Yoruba rap
Afro-beat	**63**	11	0	1	4	1	4	0	0
Afrobeats	4	**165**	0	6	5	0	8	0	0
Apala	1	4	**87**	17	10	5	2	2	4
Fuji	0	8	1	**99**	5	1	1	0	0
Highlife	0	4	0	3	**103**	5	12	0	0
Juju	0	1	0	0	6	**83**	13	0	0
Reggae	0	0	0	0	3	1	**95**	0	0
Waka	0	0	0	1	0	0	0	**89**	0
Yoruba rap	4	11	0	0	0	0	2	0	**93**
Precision, Recall and F1-scores for Model 3									
Precision	0.875	0.8088	**0.9886**	0.7795	0.7574	0.8646	0.6935	0.9780	0.9588
Recall	0.7500	0.8777	0.6591	0.8609	0.8110	0.8058	0.9596	**0.9889**	0.8455
F1-score	0.8076	0.8418	0.7909	0.8182	0.7833	0.8342	0.8051	**0.9834**	0.8986
Correctly classified	87.50	80.89	98.86	77.95	75.74	86.46	69.34	**100**	95.88

REFERENCES

[1] Li, T., Ogihara, M., Li, Q. (2003). A Comparative Study on Content-Based Music Genre Classification. SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, July28-August 1, 2003, Toronto, Canada, pp. 282–289. Doi:10.1145/860435.860487

[2] Tzanetakis G., Cook P. (2002). Musical genre classification of audio signals," in IEEE Transactions on Speech and Audio Processing, 10(5), 293-302, July 2002, doi: 10.1109/TSA.2002.800560.

[3] Cano, P., Koppenberger, M., Wack, N. (2005). Content-based music audio recommendation. Multimedia MM05 13th Annual ACM International Conference on Multimedia, Hilton, Singapore November 6-11, 2005.pp. 211-212. Doi:10.1145/1101149.1101181

[4] Scaringella, N., Zoia, G., Mlynek, D. (2006). Automatic genre classification of music content: a survey. IEEE Signal Processing Magazine, 23(2),133-141. March 2006, doi:10.1109/MSP.2006.1598089.

[5] Ellis, D. P. W. (2007). Classifying music audio with timbral and chroma features. In Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR) (pp. 81-86).

[6] Tjoa S. (2017). Music information retrieval. Available online at https://musicinformationretrieval.com/mfcc.html. Retrieved January 2023.

[7] Yang Y. H., Chen H. H. (2012). Machine Recognition of Music Emotion: A Review. ACM Transactions on Intelligent Systems and Technology, 3(3), 1-30. Doi:10.1145/2168752.2168754

[8] Jogin, M., Madhulika, M. S., Divya, G. D., Meghana, R. K., Apoorva, S. (2018). Feature extraction using convolution neural networks (CNN) and deep learning. In 2018 3rd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT) (pp. 2319-2323). IEEE. doi:10.1109/RTEICT42901.2018.9012507

[9] Lu, H., Zhang, Q. (2016). Applications of deep convolutional neural network in computer vision. 31. 1-17. Multimedia Tools and Applications, 80(22), 34355-34374. Doi:10.16337/j.1004-9037.2016.01.001

[10] Feedforward neural network. (2023). In Wikipedia. Retrieved March 31, 2023. https://en.wikipedia.org/wiki/Feedforward_neural_network

[11] Park, S., Kwak, N. (2016). Analysis on the Dropout Effect in Convolutional Neural Networks. Asian Conference on Computer Vision. pp. 189-204. Doi:10.1007/978-3-319-54184-6_12

[12] Sharma, S. (2017). Epoch vs Batch Size vs Iterations. Towards Data Science. Retrieved from https://towardsdatascience.com/epoch-vs-batch-size-vs-iterations-over-a-epoch-in-deep-learning-4ba5f066fcf9

[13] Shen, K. (2018). Effect of batch size on training dynamics. Medium. Retrieved April 28, 2023, Retrieved June 19, 2023 from

https://medium.com/mini-distill/effect-of-batch-size-on-training-dynamics-21c14f7a716e

[14] Pons, J., Lidy, T., & Serra, X. (2016, June). Experimenting with musically motivated convolutional neural networks. In *2016 14th international workshop on content-based multimedia indexing (CBMI)* (pp. 1-6). IEEE.

[15] Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (pp. 807-814).

[16] Liu, C., Feng, L., Liu, G., Wang, H., & Liu, S. (2021). Bottom-up broadcast neural network for music genre classification. *Multimedia Tools and Applications*, *80*, 7313-7331.

[17] Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 374-388.

[18] Ojala, T., Pietikainen, M., & Harwood, D. (1994, October). Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *Proceedings of 12th international conference on pattern recognition* (Vol. 1, pp. 582-585). IEEE.

[19] Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, *10*(5), 293-302.

[20] Costa, Y. M., Oliveira, L. S., Koericb, A. L., & Gouyon, F. (2011, June). Music genre recognition using spectrograms. In *2011 18th International conference on systems, signals and image processing* (pp. 1-4). IEEE.

[21] Mnih, V., Heess, N., & Graves, A. (2014). Recurrent models of visual attention. *Advances in neural information processing systems*, *27*.

[22] Ba, J., Mnih, V., & Kavukcuoglu, K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

[23] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

[24] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning* (pp. 2048-2057). PMLR.

[25] Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., & Dobaie, A. M. (2018). Facial expression recognition via learning deep sparse autoencoders. *Neurocomputing*, *273*, 643-649.

[26] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[27] Chiliguano, P., & Fazekas, G. (2016, March). Hybrid music recommender using content-based and social information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2618-2622). IEEE.

[28] Choi, K., Fazekas, G., & Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. *arXiv preprint arXiv:1606.00298*.

[29] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

[30] Li, T. L., Chan, A. B., & Chun, A. H. (2010). Automatic musical pattern feature extraction using convolutional neural network. *Genre*, *10*(2010), 1x1.

[31] Lidy, T., & Schindler, A. (2016). Parallel convolutional neural networks for music genre and mood classification. *MIREX2016*, *3*.

[32] Senac, C., Pellegrini, T., Mouret, F., & Pinquier, J. (2017, June). Music feature maps with convolutional neural networks for music genre classification. In *Proceedings of the 15th international workshop on content-based multimedia indexing* (pp. 1-5).

[33] Huang, D. A., Serafini, A. A., & Pugh, E. J. (2018). Music Genre Classification. *CS229 Stanford*.

[34] Yang, H., & Zhang, W. Q. (2019, September). Music Genre Classification Using Duplicated Convolutional Layers in Neural Networks. In *Interspeech* (pp. 3382-3386).