JOURNAL OF
EMERGING INVESTIGATORS

Article

Impact of length of audio on music classification with deep learning

Srinithi Rajan[1], **Chaya Ravindra**[2]
[1] Notre Dame High School, San Jose, California
[2] AIClub Research Institute, Mountain View, California

SUMMARY

Music genre classification is a challenging task within music information retrieval. Genre plays a crucial role in music recommendation systems, influencing the quality of track suggestions. An automated approach to classify different genres of music would help with creating a high-quality recommendation system. In this study, we proposed an approach to accomplish such automation using machine learning models. The dataset consisted of 1,000 samples and 10 different genre categories. Our approach leveraged digital signal processing for feature extraction from music clips which were subsequently employed for genre classification using machine learning techniques. We hypothesized that a 30-second audio clip enables a higher accuracy for music genre classification using machine learning algorithms compared to a 3-second audio clip. We tested this hypothesis by analyzing clips of 30 seconds and 3 seconds. The experiments were conducted by dividing the dataset into a disjoint set of train and test for evaluating model performance for both cohorts. The highest accuracy for the thirty seconds audio clip dataset was 97% from K-Nearest Neighbors (KNN) and Random Forest, while the highest for the three seconds audio clips was 92% from KNN.

INTRODUCTION

The first computer-generated music was created in 1957 (1). The widespread use of the internet has brought significant changes to the music industry, and online music streaming platforms represent a major shift in how people consume music. The convenience of downloading and purchasing music has led to an exponential increase in the size of music libraries, resulting in a need for more efficient organizational, searching, retrieval, and recommendation systems. While users often express their musical preferences in terms of genres such as hip-hop, classical, and disco, a challenge arises from the vast majority of available tracks lacking automatic genre classification (2). The qualities that a piece of music possesses identify its genre (3). Many factors, including frequency, decibel level, and bandwidth, are used to define sound. A combination of these characteristics yields a sound spectrogram, a tool for analyzing the properties of the audio clips. Its attributes often include harmonic content, rhythmic structure, and instrumentation (4).

The classification of music plays a crucial role in the music industry and marketing, where it assists in genre identification to target specific audiences. For instance, categorizing music helps streaming platforms curate personalized playlists, introducing listeners to new artists and genres. In radio programming, classification aids in generating playlists that align with the station's theme. But previously, in the 1980s, music genre classification heavily relied on the expertise of musicologists and professional listeners, who applied their trained hearing to categorize music based on factors like instrumentation, rhythm, melody, and lyrical themes. However, this manual approach was inherently labor-intensive, subjective, and prone to biases influenced by individual musical backgrounds and cultural contexts (5). The use of artificial intelligence (AI) to categorize music genres reduces the time-intensive work of conducting the classification of music. While helping locate and recommend music based on the genres of your picks, it also supports the music profession by allowing avid music learners to understand the concept of genre classification amongst different pieces of music. By incorporating automatic music-generating technologies into our daily lives, we can ultimately contribute to the development of audio-specific solutions (6).

For AI-based music classification to occur with ease, the audio itself needs to have high quality to be classified into a genre appropriately, necessitating alteration and processing of audio signals. Audio signals require specific considerations in their processing and the transformation of one-dimensional audio signals into two-dimensional time-frequency representations, while common, introduces non-homogeneous axes of time and frequency (7). Unlike images, audio signals must be studied sequentially in chronological order.

KNN classifier is used for classification datasets and uses a random point to find the nearest neighbors based on distance. To get the classifier's maximum accuracy value for the investigation, experiments were conducted with the K value. The K value determines the number of nearest neighbors the algorithm will take into consideration as it makes a precise prediction. A study on music classification with the KNN classifier used the George Tzanekis (GTZAN) dataset and achieved a mean accuracy of 81% (8).

A decision tree is made up of a data collection that has been divided up into nodes, or groupings. The root node, which is the highest node, is chosen by applying certain techniques for attribute selection (9). The same experiment on genre classification conducted with the decision tree algorithm using the GTZAN dataset and a binary tree produced an accuracy of 64% (8).

The random forest algorithm, a widely used machine learning method, combines the outputs of multiple decision trees to produce a single result. It is popular due to its versatility,

https://doi.org/10.59720/24-214

ease of use, and capability to handle both classification and regression tasks (10). The same experiment on music classification which utilized the Random Forest algorithm that incorporated multiple decision trees and the GTZAN dataset formed an accuracy of 81% (8).

Artificial neural networks (ANNs) are multi-layered models consisting of input, hidden, and output layers with weights connecting, and they process information which travels through interconnected nodes (11). Multi-layer perceptrons (MLPs) are a type of ANN with an additional hidden layer and non-linear activation functions, making them suitable for complex relationships between input and output (12). The same experiment on genre classification used an MLP AI model, producing an accuracy of 88% (8).

KNN has one hyper-parameter, which is the K-value. If the number of neighbors (K) is low (such as 3) and increases to a slightly higher number, KNN usually performs well. But if the value of K drastically increases to a very high number, 20 for example, KNN starts to perform worse overall (15). Random forest has two hyper-parameters, which are the maximum depth of the trees and the number of trees in the forest. As long as the computational cost is not too high, a greater number of trees usually produces better results. The same applies for the maximum depth, but similar to the K-value, if the value goes too high, the accuracy of the model can decrease (16). MLP also has two hyper-parameters that were tuned in this experimentation, which are the epochs and the learning rate. While it is common to decrease the learning rate with experimentation, the accuracy actually increases with a higher learning rate and higher epochs batch size (17). Previous work has used a mix of similar and different machine learning algorithms to classify music genre (13, 14). One example, used the GTZAN dataset, aiming to enhance machine learning algorithms specifically in the field. These studies highlight a comparative analysis of these algorithms, but they were methodologically limited as the results have not been proved to be accurate since one of the algorithms had an accuracy of 0%.

Our research built on this foundation by exploring different algorithms and assessing their effectiveness, particularly focusing on the random forest model. We aimed to test how accurately KNN, random forest, decision tree, and MLP algorithms would classify 10 music genres. We hypothesized that a 30-second audio clip provides more data for accurately classifying music genres with machine learning algorithms than a 3-second audio clip. We utilized the GTZAN dataset to train the algorithms and the Free Music Archive (FMA) dataset to test the final accuracy of the AI models. Our results showed that larger audio clips resulted in better accuracy for music genre classification. Random forest achieved the highest accuracy of 97.4% for 30-second clips and KNN achieved the highest accuracy 3-second clips (92%). Overall, KNN had high accuracy for both 3-second and 30-second clips. Our study points to a strategy to allow for accurate music recommendations on popular streaming services such as Spotify, aiding musicians in developing different music pieces for specific genres, and teaching music students the importance of identifying patterns in music that create unique styles and genres.

RESULTS

We tested three traditional algorithms: KNN, decision tree, and random forest, along with a neural network-based MLP, to test model accuracy in predicting music genres (18). We used the GTZAN dataset comprising 10 distinct music genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock (19). The dataset was split into training and test sets, with 80%, or 800 clips for the training sets while the test sets used the remaining 200 clips in the dataset. Algorithms were trained using the training set and their hyper-parameters were tuned using 5-fold cross-validation for optimization.

The best performing models on the testing dataset, as determined by standard parameters like F1-score and accuracy, were then tested for their performance on the test data from the GTZAN dataset (**Figure 1**). KNN achieved an accuracy of 97% for 30 seconds data and 92% for 3 seconds data. Random forest achieved an accuracy of 97.4% for 30 seconds data and 90.1% for 3 seconds data. Decision Tree achieved an accuracy of 91% for 30 seconds data and a 74% for 3 seconds data. MLP achieved 93% for 30 seconds data and 81% for 3 seconds data. Random forest and KNN performed the best overall on the testing dataset from the GTZAN database. Across all algorithms, the 30 seconds data produced better results and a higher accuracy compared to 3-second clips.

Since these results could only be validated with additional testing of the best models, we used the FMA dataset to evaluate the best models to verify the accuracies obtained

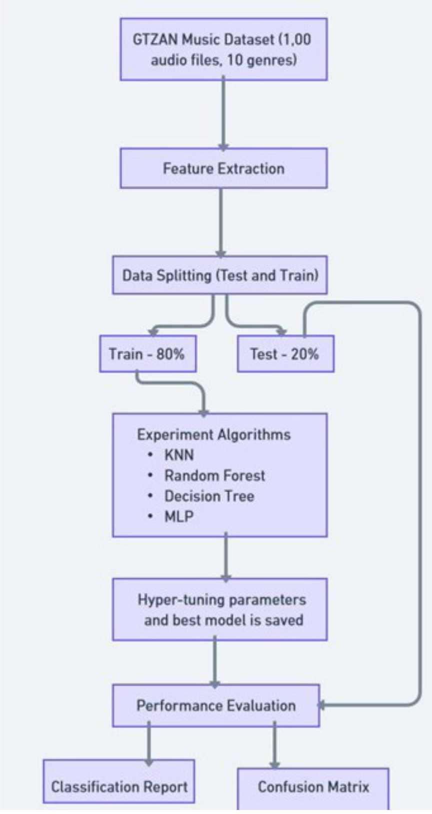


Figure 1. Flow chart of study methodology. A step-by- step process of the experimental setup with four algorithms, showing the progression from feature extraction to the performance evaluation of each algorithm and the final results.

by them. This dataset consisted of the same genres in the GTZAN database, except for reggae and hip-hop. The average time to make a single prediction per sample for 30 seconds data was 0.01272 seconds compared to 0.01011 seconds for the 3 seconds data. Even though they were very close, the time to make a single prediction with 30 seconds data was slower but provided higher accuracy. The reason for this was because the data that needed to be taken into the AI model for a 3 versus a 30 second clip was less.

The confusion matrices and classification reports generated for both the 30-second and 3-second clips in KNN experienced the greatest difficulty in recognizing rock music (96%) for 30 seconds and country music (87%) for 3 seconds. KNN worked the best in identifying disco music (100%) for 30 seconds and hip-hop music (98%) for 3 seconds (**Figure 2**). Results from hyper-parameter tuning of the KNN algorithm, where the value of K was varied between 1 and 20, showed accuracy values between 87% and 97% for 30-second clips and between 79% and 92% for 3-second clips. The best performance was obtained with K=1 for both clip lengths (**Figure 3**). The results demonstrated that increasing the K-value decreased the overall accuracy of the KNN classifier for both clip lengths. As the K-value increased, there were larger and broader clustering sections of data, creating an underfitted model that lacked an attention to detail.

Random forest generates its solutions through the accuracies of multiple decision tree algorithms that were trained with a random set of data and combined. While the number of trees refers to the amount of decision tree algorithms used, the maximum depth refers to the number of times the data can split for each tree (20). For the random forest algorithm, where the number of trees ranged between 10 and 100 and the maximum depth varied between 10 and 30, accuracy ranged from 87% to 97.2% for 30-second clips and from 80% to 90.1% for 3-second clips. The best performance was achieved with 100 trees and a maximum depth of 20 for both clip lengths (**Table 1**). Both KNN and random forest demonstrated a wide range in accuracy with a significant alteration of hyper-parameter values (**Figure 3**).

The MLP algorithm, with epochs varying from 20 to 100 and a learning rate ranging from 0.00001 to 0.005, produced accuracy values between 34% and 93% for 30-second clips and between 28% and 81% for 3-second clips. The best performance for MLP was obtained with 100 epochs and a learning rate of 0.005 for both clip lengths. The decision tree algorithm showed an accuracy of 91% for 30-second clips and 74% for 3-second clips (**Table 2**).

All four of these classifiers used for testing our datasets eventually obtained their highest accuracy after adjusting their respective hyper-parameters, which, when compared with each other demonstrate random forest's ability to perform better in classifying music genres. Even then, the accuracy decreased when the hyper-parameters were increased any further. All of these patterns are compared with a classifier accuracy comparison (**Table 3**).

We verified the results obtained from using the GTZAN dataset by running the best performing model on the FMA-small dataset, which contained 30-second audio clips that

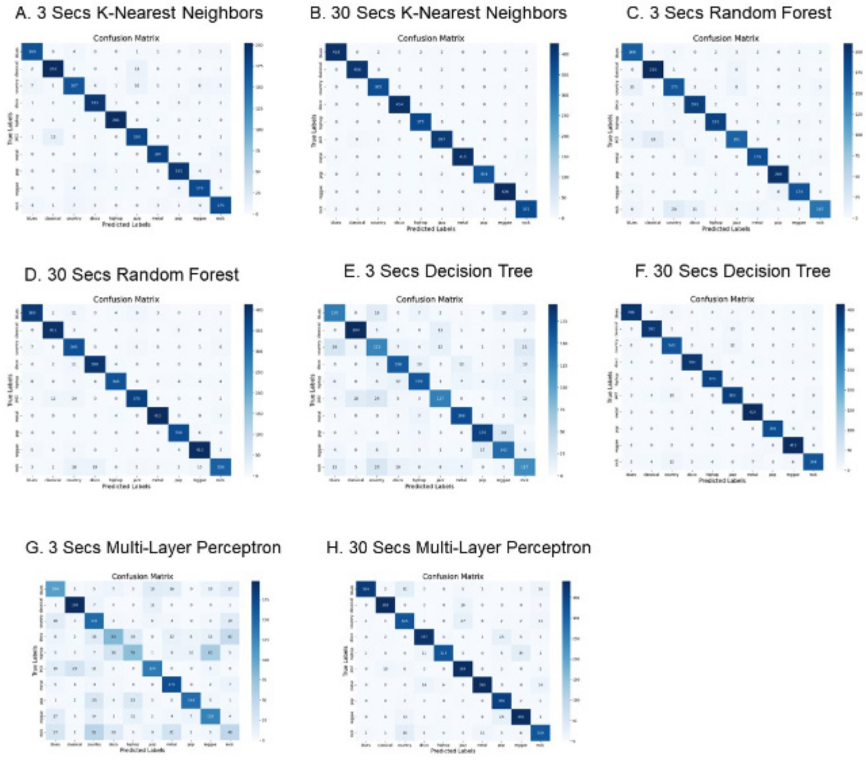
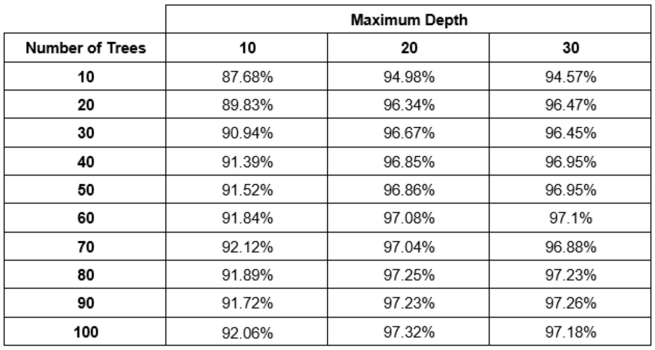


Figure 2. Confusion matrices from classifiers on the GTZAN dataset. A confusion matrix shows the performance of the models on the test data. These models were obtained by training the algorithms in the train subset and evaluating them on the validation subset of the data. The best models were saved and used to evaluate their performance on the test data. The KNN classifier for **A)** 3 second and **B)** 30 second clips. The random forest classifier for **C)** 3 second and **D)** 30 second clips. The decision tree classifier for **E)** 3 second and **F)** 30 second clips. The MLP classifier for **G)** 3 second and **H)** 30 second clips.

https://doi.org/10.59720/24-214



	Maximum Depth		
Number of Trees	10	20	30
10	87.68%	94.98%	94.57%
20	89.83%	96.34%	96.47%
30	90.94%	96.67%	96.45%
40	91.39%	96.85%	96.95%
50	91.52%	96.86%	96.95%
60	91.84%	97.08%	97.1%
70	92.12%	97.04%	96.88%
80	91.89%	97.25%	97.23%
90	91.72%	97.23%	97.26%
100	92.06%	97.32%	97.18%

Table 1. Random forest wider value range in the number of trees. The number of trees ranges from 10 to 100, and the maximum depth ranges from 10 to 30. The highest accuracy is 97.32% at 100 trees and a max_depth of 20. A visual depiction of the effect of the number of trees and the max_depth on the accuracy of random forest.

we split to also produce 3-second clips for additional testing. Therefore, we used the random forest model on the FMA dataset to look at accuracy on a completely different dataset (**Figure 4**). The accuracy percentage for 30 second clips was 16.7% while 3 second audio clips provided a 17.5% accuracy overall based on the confusion matrices.

DISCUSSION

We tested whether a 30-second audio clip provided significantly more information for music genre classification than a 3-second clip using four different machine learning algorithms. We first compared the performance of KNN, random forest, decision tree, and MLP classifiers on the GTZAN dataset. Across all models, classification performance was consistently better when using 30-second audio clips instead of 3-second segments, indicating that longer temporal context benefits genre recognition, supporting our hypothesis. Longer clips provided more comprehensive information about the music, such as melodic patterns, rhythmic structures, and harmonic progressions, which were crucial for accurate genre classification.

Among the models, random forest performed the best for 30-second clips, achieving 97.4% accuracy. KNN achieved a high accuracy overall for both clip lengths, obtaining an accuracy of 97% for 30-second clips and the highest accuracy of 92% for 3-second clips. The similar changes in accuracy for different audio clip sizes of the same algorithm can be interpreted visually using a comparison of sudden drops and growth in accuracy and how certain optimization features improve it further (**Figure 3**).

Our results showed how audio length impacts AI algorithms performance and indicating a higher accuracy can be achieved in the real world by considering longer clips for analysis. Since KNN had the highest accuracies for different audio clip lengths, it would most likely work in the most streamlined manner overall and would be a better machine learning algorithm to consider and incorporate into websites, such as Spotify, to suggest songs onto a person's playlist based on the type of genre they listen to. Currently, Spotify performs music genre classification using multiple machine and deep learning algorithms such as CNNs and natural language processing (NLP), but the streaming service utilizes two main models: audio model for analysis of each music track and a filtering model to distinguish each clip of music based on key characteristics such as tempo and instrumentals. For recommendation systems, Spotify uses a third filtering model to assess the type of music users of Spotify listen to. Spotify's accuracy in classifying music genres is 78.4% (21).

To increase the reliability of the results of the completed GTZAN experiments, we added wider ranges of values for our model optimization features to visibly recognize the

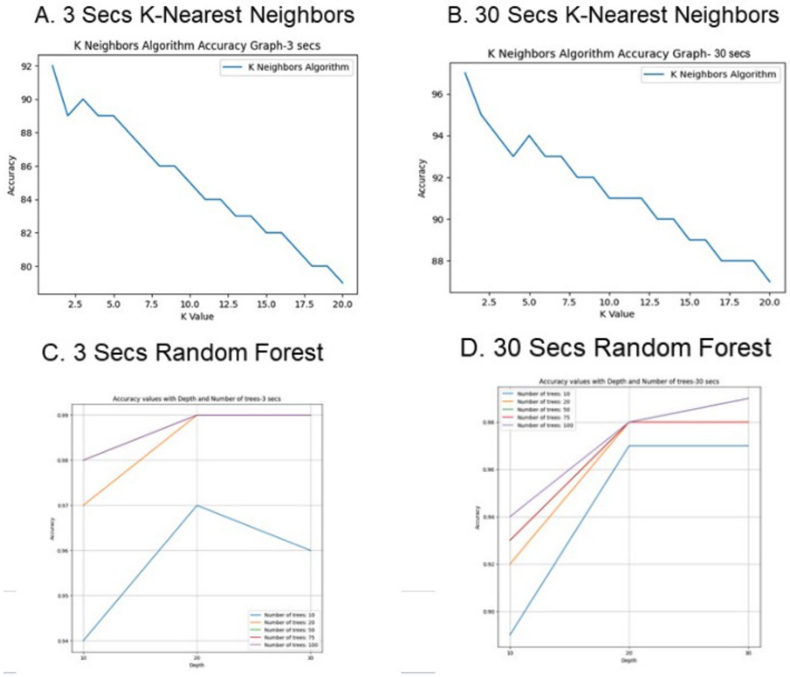
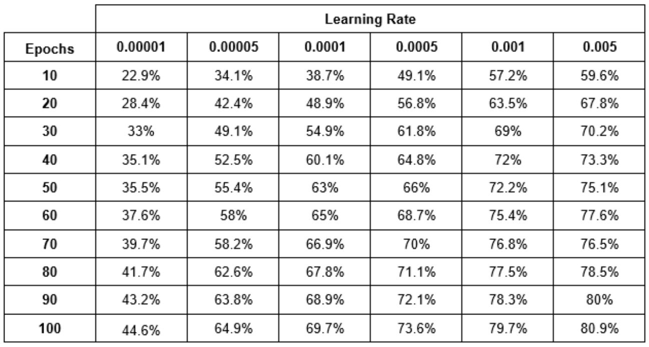


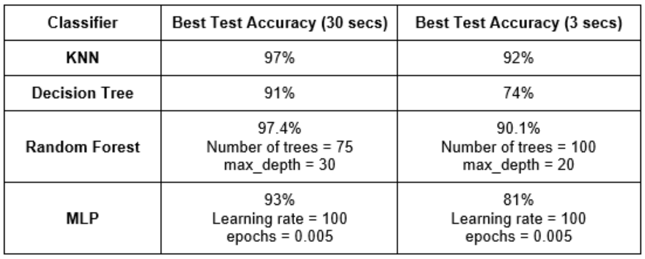
Figure 3. Overall accuracy is affected by changing different parameters of the classifier. The KNN classifier for **A)** 3 seconds and **B)** 30 seconds. The random forest classifier for **C)** 3 seconds and **D)** 30 seconds.

https://doi.org/10.59720/24-214



	Learning Rate					
Epochs	0.00001	0.00005	0.0001	0.0005	0.001	0.005
10	22.9%	34.1%	38.7%	49.1%	57.2%	59.6%
20	28.4%	42.4%	48.9%	56.8%	63.5%	67.8%
30	33%	49.1%	54.9%	61.8%	69%	70.2%
40	35.1%	52.5%	60.1%	64.8%	72%	73.3%
50	35.5%	55.4%	63%	66%	72.2%	75.1%
60	37.6%	58%	65%	68.7%	75.4%	77.6%
70	39.7%	58.2%	66.9%	70%	76.8%	76.5%
80	41.7%	62.6%	67.8%	71.1%	77.5%	78.5%
90	43.2%	63.8%	68.9%	72.1%	78.3%	80%
100	44.6%	64.9%	69.7%	73.6%	79.7%	80.9%

Table 2. Multi-layer perceptron has a wider value range in the learning rate. Represents the learning rate ranging from 10 - 100 and the epochs ranging from 0.00001 - 0.005. Highest accuracy is 80.9% at a learning rate of 100 and 0.005 epochs. Shows the effect of increasing the learning rate and epochs on the accuracy, increasing the accuracy from 22% to 80%.



Classifier	Best Test Accuracy (30 secs)	Best Test Accuracy (3 secs)
KNN	97%	92%
Decision Tree	91%	74%
Random Forest	97.4% Number of trees = 75 max_depth = 30	90.1% Number of trees = 100 max_depth = 20
MLP	93% Learning rate = 100 epochs = 0.005	81% Learning rate = 100 epochs = 0.005

Table 3. Variance of accuracy in machine learning and neural networks performance. KNN performed the best overall, gaining the highest accuracy for the 3 second clips and the second highest for 30 second audio clips.

changes in accuracy to obtain the best performance by the AI but acknowledged that further exploration could potentially enhance the model's performance. For the MLP algorithm, we explored a range of learning rates (from 0.00001 to 0.005) and epochs (from 20 to 100). For random forest, we tested a variety of tree counts (ranging from 10 to 100) and maximum depths (from 10 to 30). While the current hyper-parameter ranges produced competitive results, expanding the search space for random forest and MLP could potentially lead to more robust model optimization.

The significant improvement in accuracy for all algorithms with longer clips underscored the importance of audio length in automated music genre classification. For example, KNN's higher accuracy of 97% for 30 second clips instead of 92% occurred due to a longer audio length. Even then, the value of K (number of neighbors) affected the accuracy for both the 30 and 3 seconds data for KNN in the same ways. As the value of K increased, the accuracy decreased, so the highest accuracy was provided with a value of K as 1, while the lowest accuracy came from the value of K being equal to 20. We recognized the opposite pattern in the MLP classifier. In this case, the accuracy relied on the number of epochs and the learning rate. If the learning rate and the number of epochs were higher, the accuracy also increased, showing the direct relationship between the accuracy of MLP and its hyper-parameters (**Table 2**).

Some past research on music genre classification focused solely on combinations of neural networks. The experimenters recommended hybrid algorithms for the FMA dataset, such as the convolutional recurrent neural network (CRNN), as it brought out the highest accuracy of 90% among the others, including CNN, ANN, support vector classifier (SVC), and parallel convolutional recurrent neural network (PCRNN) (22). In another experiment, the GTZAN dataset paired with a preprocessed algorithm by Mel-Spectrogram gave the accuracy of 90% (23). The highest accuracy in a paper was 95.2% with MFCCs and short-time Fourier transform (24). A new bidirectional memory method was tested in this new experiment on two datasets, and results showed 94% for the balanced GTZAN dataset (25). According to a machine learning study in music classification, although the success order for each algorithm was similar to our results, it was also surprising how the accuracies were far apart from each other, which differed from our results (13). Our findings do not directly match this research since we did not combine single algorithms to create hybrid ones but still produced a decent and applicable result with the models KNN, MLP, random forest, and decision tree. Our best models managed to achieve an accuracy of 97% for 30-second clips, which is higher than all of the others, and an accuracy of 92% for 3-second clips, which is a middle value compared to the other studies but still beneficial since with 3-second clips, there is access to less data.

Based on the confusion matrices and classification reports on the algorithms, we also determined the accuracy of each AI algorithm in identifying each of our 10 tested genres to find the AI's area of confusion and error. A common pattern in the error analysis when classifying genres of music was when the correct genre was rock, and the AI predicted the label to be country (**Figure 2**). Although this was not always the case, it was a prominent error found even in KNN. There were other errors but they did not necessarily apply to all of the classifiers. Both genres use similar rhythmic patterns, tempos, vocal styles, and sometimes overlapping instrumentation that usually helps the AI models distinguish between different genres.

Inspecting the performance of the KNN classifier could reveal patterns that improve music genre classification accuracy since it was the only algorithm that maintained the highest average of accuracy between the different clip lengths even though it did not achieve the highest accuracy for 30-second clips. In analyzing the KNN algorithm's performance on the 30-second audio clips, the confusion matrix revealed a high level of accuracy across all ten music genres. Precision, recall, and F1-scores for most genres hovered around 0.99, indicating the model's strong capacity to accurately classify the majority of audio signals. However, subtle misclassifications arose between similar genres, such as rock and metal, or jazz and country, where overlapping acoustic features made precise differentiation more difficult (**Figure 2**). In the case of the 3-second clips, the KNN classifier performed well for genres like hip-hop and metal, with both achieving 96% recall. Yet, there were notable misclassifications, such as jazz being confused with classical music, likely due to the shorter clip lengths providing fewer distinguishing features. Rock and county shared similar instrumentation. The overall precision and recall rates remained high, with the lowest F1-score at 0.85 for country, meaning that country was the most misclassified genre for

https://doi.org/10.59720/24-214

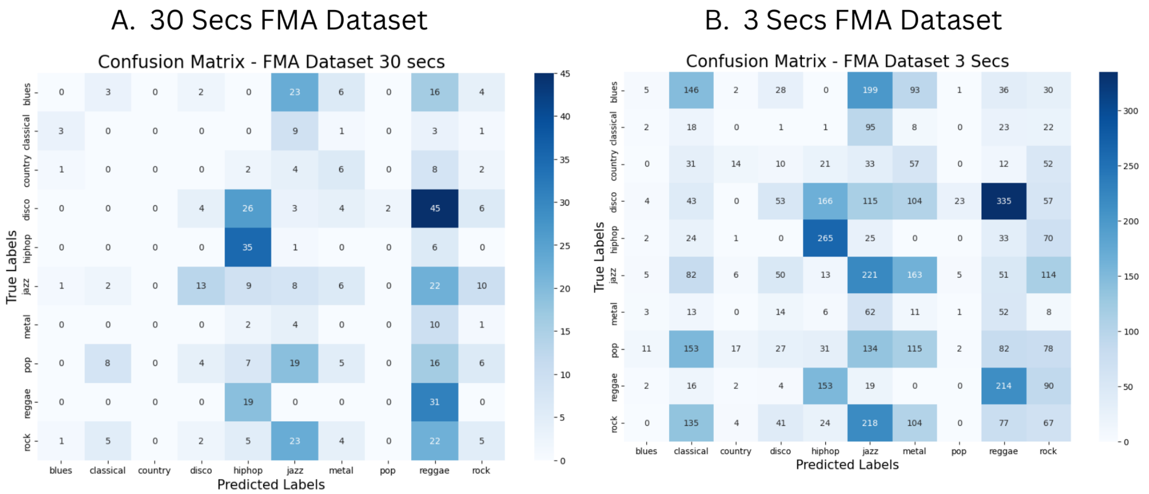


Figure 4. Confusion matrices from random forest through the FMA dataset. Confusion matrix of the accuracy of the Random Forest classifier at **A)** 30 seconds and **B)** 3 seconds.

all cases. The 3-second clips highlighted the challenges of classifying closely related genres with limited data. This suggested that longer audio clips or additional features might be necessary to improve classification accuracy, particularly for genres with overlapping characteristics (26).

Various factors may have influenced the results. Hyper-parameters significantly impacted final accuracies, with longer audio clips (30 seconds) consistently yielding higher accuracies than shorter ones (3 seconds), confirming our hypothesis. The number of samples also played a role; the dataset contained 1,000 clips split equally across 10 genres with 100 clips per genre, which might have been insufficient for optimal training. More data typically enhances accuracy (27). Additionally, expanding hyper-parameter ranges, such as increasing the number of trees in random forest beyond 100, could have improved reliability. We tested the dataset switching the 3 seconds and 30 seconds data in terms of training and testing. The accuracy was about 98% for 3 seconds and 99% for 30 seconds data, showing the improvement in accuracy in our best model, the random forest algorithm.

Scaling the model for commercial use presents certain challenges. First, as the dataset size and diversity increase, KNN's computational costs can grow, as KNN requires comparing each new audio clip to every other clip in the dataset. Moreover, generalizing the model to a wider variety of music genres, languages, and cultural styles may impact its accuracy. Additionally, handling diverse audio qualities is difficult when moving from an experimental setting to real-world environments.

There are numerous ways to refine this experiment to relate it to more demographics and to solve more real-world problems. To pursue this project further, we would test different algorithms to determine the highest accuracy in identifying the artist and place of origin for the song and expand the dataset to produce more reliable results. Some scientific questions that still remain in regards to this experiment are why the time length of the audio clips affected the final accuracy and what are the largest factors of music that cause some genres to have a higher accuracy. This research can help newcomers classify music and improve AI-driven music recommendations by accurately identifying genres and suggesting similar music (28). Since the benefits of using 30 second audio resulted from all of the algorithms used in this project, our results strongly suggest that accuracy may be dependent on length, and with more information as to what the genre of music the audio may be, the AI can predict it better. This may be exactly how the AI's accuracy is directly correlated with the audio length, with the accuracy increasing as the length is increasing and vice versa.

MATERIALS AND METHODS

To prevent overfitting, where the model learns irrelevant patterns specific to the dataset, we employed a common data-splitting strategy (29). Cross-validation methods, such as k-fold cross-validation, ensure that the model was tested on various subsets of the data, improving its generalizability. Cross-validation and data-splitting were applied to prevent the model from fitting noise or irrelevant patterns in the data. We implemented the open dataset GTZAN, a widely recognized and popular benchmark for evaluating music classification methods, to build an AI-based music genre classification system (30). The dataset incorporated 1,000 tracks of music arranged equally in a hierarchical taxonomy of 100 clips per genre for 10 genres with each song as 30 and 3-second clips stored as wav. Each track contained 518 attributes categorized in audio features, which were obtained by data preprocessing of GTZAN dataset music tracks. The GTZAN dataset was randomly divided into two non-overlapping sets: 80% for training, and 20% for testing. This prevented the model from learning artist or album information, ensuring it focused solely on genre classification.

The dataset featured 100 audio samples per genre, each lasting 30 seconds and sampled at a rate of 22,050 Hz (29). We used python code to transform raw MP3 music clips into Mel-frequency cepstral coefficients (MFCC) features suitable for AI algorithms to process audio data. We used Librosa, a popular Python library specializing in music and audio analysis to function in experiments such as those involving music genre classification and speech recognition (31). Librosa simplified the process of building music information retrieval systems by offering tools for audio feature extraction.

https://doi.org/10.59720/24-214

By leveraging Librosa, we efficiently converted the audio clips into numerical representations which are MFCC features capturing characteristics like pitch, rhythm, and tempo for each song. This data was then saved in a comma-separated values (CSV) format, which was readily interpretable by AI algorithms.

CNNs have shown their powerful capabilities in categorizing complex audio patterns (32). Different audio representations were applied to the Visual Geometry Group of 16 layers (VGG-16) on spectrograms of audio signals, resulting in a 65% accuracy developed using data extracted from two formats of Librosa, chromagram and Mel-Frequency Cepstral Coefficients (MFCCs). VGG-16 is a CNN model with 16 convolutional and pooling layers that allow for accurate representations of visual features that create strong predictions in regards to image classification (30). A past study on utilizing machine learning techniques to enhance music classification utilized a duplicate convolution layer on mel-spectrograms, a type of audio representation that converts a sound wave into a visual format, incorporating statistical analysis for genre classification. These statistical methods interpret and analyze the data extracted from audio features (33).

Our AI model incorporated Mel-Spectrogram and MFCC's for the extraction process of the data and neural networks including KNN, random forest, decision tree, and MLP. Mel-Spectrogram was used as a visual image of a frequency scale of audio signals similar to a human's depiction of audio, and MFCC's were features from the Mel-Spectrogram utilized for music classification tasks. This was different compared to the usage of residual neural networks and the use of convolutional neural networks (34).

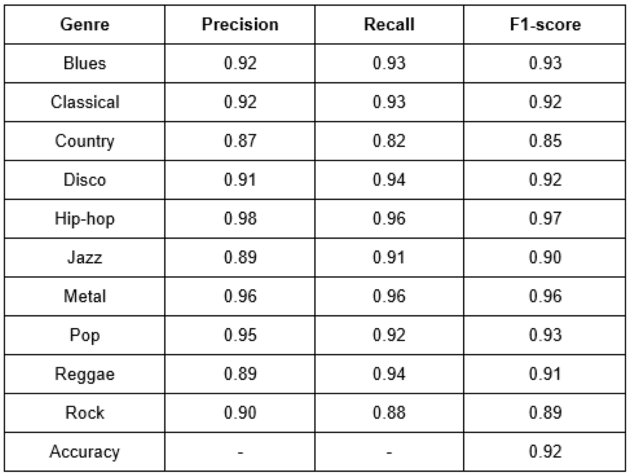
In our research, we employed a range of audio features that were also utilized in past studies regarding music genre classification, such as zero-crossing rate (ZCR), spectral centroid, spectral roll-off, spectral bandwidth, chroma frequency, root mean square energy (RMSE), delta, Mel-Spectrogram, tempo, and MFCC. The features were drawn from the music clips and stored in a CSV format. This formatted data was employed to train our proposed model.

After data-splitting and a five-fold cross-validation was conducted, KNN, decision tree, random forest, and MLP algorithms were trained on the training set of data, and the performance of the suggested model was assessed using the testing dataset.

Evaluation based on the GTZAN dataset focused on standard supervised learning metrics: accuracy, confusion matrix, precision, recall, and F1-score (**Table 4**). Accuracy reflects the models' ability to generate correct predictions, while the confusion matrix provides deeper insights (**Figure 2**). A confusion matrix helps determine an AI model's accuracy by comparing its predictions to actual labels, which also aids in locating patterns in genre prediction and calculating an algorithm's accuracy (35). Precision measures how well the models identify relevant cases, and recall measures the ability to capture important real-life cases. While precision focuses on the number of predictions by the AI that were positive over the total number of positive outcomes, recall is produced from the algorithm's accuracy in identifying a positive outcome correctly. The F1-score is the harmonic mean of precision and recall, ultimately creating an accuracy value that keeps track of all cases from the three other metrics. The parameters are provided in percentages to appropriately analyze the ability of a model to classify a music genre as columns are provided for each of the ten genres (**Table 4**).

To fine-tune the machine learning algorithms KNN, MLP, decision tree, and random forest, we varied their hyper-parameters and evaluated each algorithm on the validation dataset to compare accuracy. We saved the model that demonstrated the best performance based on the evaluation criteria and further evaluated this top-performing model on the test subset of the dataset to validate its robustness and generalization capabilities. Subsequently, we trained AI models utilizing the four machine learning algorithms and employed them on the GTZAN dataset. During the training process, specific hyper-parameters were fine-tuned for each algorithm. For MLP, parameters such as maximum iterations, learning rate, and the number of hidden layers were optimized. In the case of random forest, the number of trees was adjusted. For KNN, the optimal value of the hyper-parameter "K" was determined. When the highest accuracies were achieved for all algorithms, the most successful models were tested using the FMA dataset as a form of verification of our original results.

The FMA dataset consists of 8,000 tracks of 30 seconds each, across 163 balanced genres among which we extracted the 8 genres listed here: rock, pop, metal, jazz, disco, country, classical, and blues. Our original dataset, the GTZAN dataset, included 10 genres, so we sourced the additional 2 genres, reggae and hip-hop, from the website called Pixabay to achieve the goal of matching the 10 genres from the GTZAN to the FMA dataset (36). We randomly downloaded few audio clips from the website and split them into 30-second clips and 3-second clips are generated using the 30-second clips. We combined the eight genres from the FMA-small dataset and the two genres from pixabay sourced data. We made sure the two pixabay sourced genres in the dataset had the same number of clips as the other eight to maintain the balance in the dataset. For testing the 3-second model, we split the 30-second samples into 3-second chunks and used it for testing (37).



Genre	Precision	Recall	F1-score
Blues	0.92	0.93	0.93
Classical	0.92	0.93	0.92
Country	0.87	0.82	0.85
Disco	0.91	0.94	0.92
Hip-hop	0.98	0.96	0.97
Jazz	0.89	0.91	0.90
Metal	0.96	0.96	0.96
Pop	0.95	0.92	0.93
Reggae	0.89	0.94	0.91
Rock	0.90	0.88	0.89
Accuracy	-	-	0.92

Table 4. Classification report of the KNN classifier for 3 second audio lengths. Shows the hyper-parameters used to optimize the algorithm through hyper-tuning to produce higher accuracies for each music genre.

https://doi.org/10.59720/24-214

Received: May 16, 2024
Accepted: December 2, 2024
Published: October 8, 2025

REFERENCES

1. "Restoring the first recording of computer music." *British Library*, 13 Sept. 2016, https://blogs.bl.uk/sound-and-vision/2016/09/restoring-the-first-recording-of-computer-music.html. Accessed 9 Nov. 2024
2. McFee, Brian, et al. "librosa: Audio and music signal analysis in python." *SciPy*, 2015. https://doi.org/10.25080/Majora-7b98e3ed-003.
3. Tzanetakis, George, and Perry Cook. "Musical Genre Classification of Audio Signals." *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, July 2002, pp. 293-302. https://www.doi.org/10.1109/TSA.2002.800560.
4. Joshi, Dipti, et al. "Comparative Study of Mfcc and Mel Spectrogram for Raga Classification Using CNN." *Indian Journal of Science and Technology*, vol. 16, no. 11, 20 Mar. 2023, pp. 816-822. https://doi.org/10.17485/ijst/v16i11.1809.
5. van Venrooij, Alex, and Rens Wilderom. "The dynamics of dance: An early history of electronic dance music." *The Bloomsbury Handbook of Popular Music and Youth Culture*, edited by Andy Bennett, e-book ed., Bloomsbury Publishing, 2022, pp. 257.
6. Bisharad, Dipjyoti, and Rabul Hussain Laskar. "Music Genre Recognition Using Residual Neural Networks." *TENCON 2019 - 2019 IEEE Region 10 Conference*, Oct. 2019, https://doi.org/10.1109/tencon.2019.8929406.
7. Mehrish, Ambuj, et al. "A review of deep learning techniques for speech processing." *Information Fusion*, vol. 99, 3 June 2023, Paper no. 101869. https://doi.org/10.1016/j.inffus.2023.101869.
8. Islam, Md Shofiqul, et al. "Machine learning-based music genre classification with pre-processed feature analysis." *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika,* vol. 7, no. 3, 12 Jan. 2021, pp. 491-502, https://doi.org/10.26555/jiteki.v7i3.22327.
9. Grandini, Margherita, et al. "Metrics for Multi-ClassClassification: an Overview." *arXiv,* Aug. 2020, https://doi.org/10.48550/ar756.Xiv.2008.05.
10. Navin, Maria, et al. "Performance analysis of text classification algorithms using confusion matrix." *International Journal of Engineering and Technical Research*, vol. 6, no. 4, Dec. 2016, pp. 2454-4698.
11. Guerra, Maria I. S., et al. "Assessing Maximum Power Point Tracking Intelligent Techniques on a PV System with a Buck–Boost Converter." *Energies*, vol. 14, no. 22, 9 Nov. 2021, Paper no. 7453, https://doi.org/10.3390/en14227453.
12. Hridoy, Rashidul Hasan, et al. "An Efficient Computer Vision Approach for Rapid Recognition of Poisonous Plants by Classifying Leaf Images using Transfer Learning." *12th International Conference on Computing Communication and Networking Technologies*, July 2021, https://doi.org/10.1109/ICCCNT51525.2021.9580011.
13. Arxiv: Mogonediwa, Keoikantse. "Music Genre Classification: Training an AI model." 23 May 2024, https://doi.org/10.48550/arXiv.2405.15096.
14. Kamuni, Navin, and Dheerendra Panwar. "Enhancing Music Genre Classification through Multi-Algorithm Analysis and User-Friendly Visualization." *Arxiv*, 27 May 2024, https://doi.org/10.48550/arXiv.2405.17413.
15. Nti, Isaac Kofi, et al. "Performance of machine learning algorithms with different K values in K-fold cross-validation." *International Journal of Information Technology and Computer Science* vol. 13, no. 6, 8 Dec. 2021, pp. 61-71, http://doi.org/10.5815/ijitcs.2021.06.05.
16. Duroux, Roxane, and Erwan Scornet. "Impact of subsampling and tree depth on random forests." *ESAIM: Probability and Statistics,* vol. 22, 14 Dec. 2018, pp. 96-128, https://doi.org/10.1051/ps/2018008.
17. Smith, Samuel L., et al. "Don't decay the learning rate, increase the batch size." *arXiv preprint,* 24 Feb. 2018, https://doi.org/10.48550/arXiv.1711.00489.
18. Aitkin, Murray and Rob Foxall. "Statistical modelling of artificial neural networks using the multi-layer perceptron." *Statistics and Computing*, vol. 13, pp. 227-239, Aug. 2003, https://doi.org/10.1023/A:1024218716736.
19. Liang, Beici, and Minwei Gu. "Music genre classification using transfer learning." *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Aug. 2020, https://doi.org/10.1109/mipr49039.2020.00085.
20. Belgiu, Mariana, and Lucian Drăguţ. "Random forest in remote sensing: A review of applications and future directions." *ISPRS Journal of Photogrammetry and Remote Sensing,* vol. 114, Apr. 2016, pp. 24-31, https://doi.org/10.1016/j.isprsjprs.2016.01.011.
21. Heidbreder, Eric. "Classifying Song Genre: Using Spotify's Built-in Features vs. Extracting My Own." Towards Data Science, www.towardsdatascience.com/classifying-song-genre-using-spotifys-built-in-features-vs-extracting-my-own-a4d5fe448948. Accessed 28 Jan. 2025.
22. Ghosh, Partha, et al. "A Study on Music Genre Classification using Machine Learning." *International Journal of Engineering Business and Social Science*, vol. 1, no. 04, 25 Apr. 2023, pp. 308-320. https://doi.org/10.58451/ijebss.v1i04.55.
23. Gokhman, Ruslan. "Machine Learning and Deep Learning methods for music genre Classification." *Yeshiva University*, May 2023.
24. Li, Teng. "Optimizing the configuration of deep learning models for music genre classification." *Heliyon*, vol. 10, no. 2, 30 Jan. 2024, https://doi.org/10.1016/j.heliyon.2024.e24892.
25. Wijaya, Nantalira Niar, et al. "Music-genre classification using Bidirectional long short-term memory and mel-frequency cepstral coefficients." *Journal of Computing Theories and Applications*, vol. 1, no. 3, Jan. 2024, pp. 243-256. https://doi.org/10.62411/jcta.9655.
26. Choi, Keunwoo, et al. "Convolutional recurrent neural networks for music classification." *2017 IEEE International conference on acoustics, speech and signal processing,* 19 June 2017, https://doi.org/10.48550/arXiv.1609.04243.
27. Sturm, Bob L. "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use." *arXiv,* 27 Nov. 2024, https://doi.org/10.48550/arXiv.1306.1461.

28. Schedl, Markus, et al. “Music information retrieval: Recent developments and applications.” *Foundations and Trends in Information Retrieval,* vol. 8, no. 2-3, 12 Sept. 2014, pp. 127-261, http://doi.org/10.1561/1500000042.
29. Paradzinets, Aliaksandr, et al. “Multiexpert system for automatic music genre classification.” *Teknik Rapor, Ecole Centrale de Lyon, Departement MathInfo*, Jan. 2009.
30. Zhang, Scott, et al. “Music Genre Classification: Near-Realtime Vs Sequential Approach,” CS 229 Projects, Stanford University, 2019. https://cs229.stanford.edu/proj2019spr/report/3.pdf.
31. McFee, B., et al. Librosa/librosa: 0.11.0. 0.11.0, Zenodo, 11 Mar. 2025, https://doi.org/10.5281/zenodo.15006942.
32. Sinha, Harsh, et al. “Audio classification using braided convolutional neural networks.” *IET Signal Processing* 14.7 (2020): 448-454. https://doi.org/10.1049/iet-spr.2019.0381.
33. Silla, Carlos N., et al. “A Machine Learning Approach to Automatic Music Genre Classification.” *Journal of the Brazilian Computer Society*, vol. 14, no. 3, Sept. 2008, pp. 7–18. https://doi.org/10.1007/bf03192561.
34. Qi, Yanjun. “Random Forest for Bioinformatics.” *Ensemble Machine Learning*, 19 Jan. 2012, pp. 307–323, https://doi.org/10.1007/978-1-4419-9326-7_11.
35. Villasís Giribets, Albert, et al. “User acceptance of two-sided shared vehicles: The cases of Barcelona and Leuven.” *Escola de Camins*, 30 Jan. 2024.
36. “Royalty Free Music Downloads.” *Pixabay,* www.pixabay.com/music/, Accessed 5 June 2025.
37. Defferrard, Michaël, et al. “FMA: A dataset for music analysis.” *arXiv*, 6 Dec. 2016, https://doi.org/10.48550/arXiv.1612.01840.
38. Golzari, Shahram, et al. “A Hybrid Approach to Traditional Malay Music Genre Classification: Combining Feature Selection and Artificial Immune Recognition System.” *2008 International Symposium on Information Technology*, 2008, https://doi.org/10.1109/ITSIM.2008.4631692.
39. Liuwanyue, Shi. “Course genres classification of music e-learning platform based on deep learning big data intelligent processing algorithm.” *Entertainment Computing*, vol. 50, 28 Apr. 2024, https://doi.org/10.1016/j.entcom.2024.100704.
40. Ndou, Ndiatenda, et al. “Music Genre Classification: A Review of Deep-Learning and Traditional Machine-Learning Approaches.” *IEEE International IOT, Electronics and Mechatronics Conference*, 14 May 2021, https://doi.org/10.1109/IEMTRONICS52119.2021.9422487.
41. “Music-Genre-Classification-AI” *Github*, www.github.com/AIML-engineer/Music-Genre-Classification-AI. Accessed 8 Mar. 2024