Automated Music Genre Classification through Deep Learning Techniques

Mamidi Kiran Kumar1*, K. Sujanasri1, B. Neha1, G. Akshara1, Pinki Chugh2, P. Haindavi3

1Department of CSBS, GRIET, Hyderabad, Telangana State, India
2Uttaranchal Institute of Technology, Uttaranchal University, Dehradun, 248007, India
3KG Reddy College of Engineering & Technology, Hyderabad, India

Abstract. Music Genre Classification (MGC) automatically categorizes music into different genres based on various musical attributes and features in a small number of music files. This is a crucial problem in the field of music information retrieval as it provides a way to organize and analyse large amounts of music files. MGC can be performed using conventional machine learning algorithms such as SVM, k-nearest neighbours, Decision trees, and neural networks. These algorithms learn to recognize different musical features and attributes to categorize the music files into different genres. The literature shows that the performance of conventional machine learning algorithms is inferior to deep learning algorithms such as CNN, RNN, etc., in various applications. Hence, the CNN algorithm is adapted to implement the classification of music files. This aims to classify music genres using CNN deep learning techniques. The performance of the algorithms for MGC can be evaluated using metrics such as accuracy, precision, recall, and F1-score. Additionally, the impact of different features and algorithms on the performance of MGC can be studied and compared. It has applications in areas such as automated music recommendation systems, music education, and music production. An accuracy of 83% is achieved by using CNN to accomplish the task of MGC.

1 Introduction

Music genre categorization is an open-ended method of classifying musical content into various categories depending on musical traits and elements. It is the primary objective of music information retrieval along with numerous applications such as music recommendation, playlist creation, and audio file analysis. This work explains the music genre categorization process, including the numerous strategies and approaches used to categorize a piece of music. Numerous musical characteristics are also studied that are commonly employed in genre classification, including melody, rhythm, harmony, pace, and timbre. Furthermore, the obstacles and disadvantages of musical genre classification are explored, such as building advanced algorithms capable of accurately classifying numerous music genres even with the addition of noise, distortion, and various other forms of signal.

* Corresponding author: mamidi.kirankumar09@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
deterioration. In the end, you will possess a comprehensive knowledge of music genre categorization and will be able to apply this information to your music categorization and analysis.

Music is a form of expression that combines melodies and sounds to generate aesthetic and emotive experiences for listeners. It is a global form of individual expression seen in all historical cultures and societies. It can be created with various techniques and instruments, such as stringed instruments, wind-based instruments, and digital instruments. Music is classified into genres depending on characteristics such as melody, timbre, tempo, rhythm, as well as harmony. Rock, blues, classical, pop, hip-hop, electronic, jazz, and country music are some well-known musical genres. Each genre has distinct characteristics that are typically associated with particular cultural, historical, and social contexts. Figure 1.1 shows the audio waveform.

![Audio waveform](image.png)

**Fig. 1.** Audio waveform (Courtesy: Source [1]).

## 2 Literature survey

In 2022, Tarannum Shaikh and Ashish Jadhav studied the performance of the CNN model for MGC [2]. Spectrogram images are given as input to the CNN model. The experiment was conducted using GTZAN and MusicNet datasets. Results show that the accuracy obtained for the GTZAN dataset is 92.65% and for MusicNet is 91.70%. In 2021, Ahmed A. Khamees et al. compared the performance between CNN and RNN to classify the genre of an audio file [3]. The GTZAN dataset is used to train the model. Results show that CNN outperformed all the other models with an accuracy of 83.74%. In 2021, Bharat Dave, Vinayak Chavan, Mujahid Khan, and Dr. Varsha Shah implemented a system [4] to extract features from the audio files and classify them into their respective genres using CNN. The GTZAN dataset is used to train the model. The accuracy achieved was 57.5%. In 2021, Ndiatenda Ndou, Ritesh Ajoodha, and Ashwini Jadhav reviewed the performance of several Machine Learning (ML) and Deep-Learning approaches in categorizing the genre of an audio file in [5].

The GTZAN dataset is used to train the model. The study concluded that KNN classifies the genre more accurately, which is 92.69%. In 2021, Macharla Vaibhavi and P. Radha Krishna used CNN and Recurrent Convolutional Neural Networks (RCNN) Deep Learning techniques for the classification of the genre [6]. The GTZAN dataset is used to train the model. Results show that comparatively the accuracy for the RCNN model with data augmentation is more, which is 82.05%. In 2021, Seethal V and Dr. A. Vijayakumar compared K Nearest Neighbours (KNN) and Support Vector Machine (SVM) in [7] for the classification of music into different genres. This experiment was conducted using the GTZAN dataset. Results show that KNN gave better accuracy than SVM which is 75%. In 2021, Navneet Parab, Shikta Das, Gunj Goda, and Ameya Naik used Neural Networks to classify the genre of an audio file [8]. This experiment was conducted using the GTZAN dataset. Results show that model 3 is reliable and consistent in predicting the genre with an accuracy of 91.76%. In 2022, Shivanshu Garg and Anshu Varshney performed Music Genre Classification (MGC) using K-Nearest Neighbours (KNN) [9]. This experiment was conducted using the GTZAN dataset. Results show that using the KNN algorithm the accuracy obtained is 70%. In 2022, Jessica Dias, Vaishak Pillai, Hruthvik Deshmukh, and
Ashok Shah worked on MGC using CNN [10]. Their objective was to determine the music genre by extracting the features through MFCC and classifying them by using CNN. The GTZAN dataset is used to train the model. Results show that the accuracy obtained by CNN is 76%. In 2022, Tarannum Shaikh and Ashish Jadhav studied the performance of the CNN model for MGC [11].

The experiment was conducted using GTZAN and MusicNet datasets. Results show that the accuracy obtained for the GTZAN dataset is 92.65% and for MusicNet is 91.70%. In 2022, Aza Kamala and Hossein Hassani proposed CNN and DNN methods [11] to classify music genres [12]. CNN achieved 92% accuracy while DNN, 90%. Hence it is concluded that CNN outperformed DNN in classifying the Kurdish music genre. In 2021, Jingwen Zhang performed “Music Feature Extraction and Classification Algorithm using Deep Learning” in [13]. The objective was to build a model for extracting and classifying music features based on a convolutional neural network. This experiment was conducted on the GTZAN dataset. An accuracy of 87% is achieved. In 2020, Dhevan Lau and Ritesh Ajoodha compared the Deep Learning Convolutional Neural Network (CNN) with other traditional classifiers concerning MGC [14]. They used spectrograms and content-based features for MGC. The GTZAN dataset is used in this experiment.

In 2020, Rajeeva Shreedhara Bhat, Rohith. B. R. and Mamata K. R. used Machine Learning techniques using log-based MEL spectrograms for MGC [15]. Results showed 98% accuracy for the training set and 73% accuracy for validation. In 2021, Prajwal R, Subham Sharma, Prasanna Naik worked on MGC using Machine Learning (ML) [16]. This experiment was conducted using the GTZAN dataset. Results show that SVM gave a better accuracy of 73.2% before using hyper-parameter tuning and 76.4% after using hyper-parameter tuning.

Authors [17] highlighted the significance of ML in prediction, pattern recognition and error reduction across diverse fields, emphasizing the impact of AI in broad domain. Author [18] presented text classification algorithms for various applications and explores the use of machine learning in detecting phishing attacks. Image restoration is to enhance images by removing noise and restoring them to their original quality. The present approach explored various methods in both frequency and spatial domains, followed by analysing their performance using simulations [19].

Authors [20] discussed the use of machine learning and neural networks, especially CNN, for recognizing handwriting patterns, with a focus on Telugu film industry names, achieving high accuracy (98.3%). Authors [21] emphasized the significance of feature selection in classification for accuracy and efficiency. It investigates combining features from different methods, demonstrating improved precision, contingent on dataset, algorithm, and metrics used. Author [22] explored the application of Transfer Learning (TL) in automated medical image analysis, highlighting its effectiveness in various tasks. TL models like AlexNet, ResNet, VGGNet, and GoogleNet prove valuable for enhancing medical image analysis.

### Table 1. Summary of the existing approaches.

<table>
<thead>
<tr>
<th>Ref No.</th>
<th>Methodology</th>
<th>Accuracy</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>[2]</td>
<td>Transfer Learning</td>
<td>79%, 75.9%, 71.3%, 76.9% respectively</td>
<td>Hyper parameters are optimized</td>
<td>Inferences need to be improved for overlapping genres</td>
</tr>
<tr>
<td>[3]</td>
<td>CNN and RNN with LSTM</td>
<td>74% and 63.44% respectively</td>
<td>RNN is suitable for time series data.</td>
<td>RNN+LSTM is 64% slower than CNN</td>
</tr>
<tr>
<td>[4]</td>
<td>CNN</td>
<td>57.5%</td>
<td>Commodity</td>
<td>Low accuracy and no</td>
</tr>
<tr>
<td>Reference</td>
<td>Models Used</td>
<td>Hyperparameters/Features</td>
<td>Performance/Details</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>[5]</td>
<td>Linear LR, RF, SVM, Multilayer Perceptron, KNN, NB, Logistic Regression, CNN</td>
<td>KNN – 92.69% (3 sec feature set)</td>
<td>Information Gain Ranking algorithm to determine contributing features. Overlap between rock and country genres</td>
<td></td>
</tr>
<tr>
<td>[6]</td>
<td>CNN and RCNN</td>
<td>With and Without Data Augmentation, 81.55% 69.49% and 82.05%, 77.92% Respectively</td>
<td>Accuracy of both the models is improved by DA. Performance can be improved by further applying class conditional augmentation.</td>
<td></td>
</tr>
<tr>
<td>[7]</td>
<td>KNN and SVM</td>
<td>75% and 65% respectively</td>
<td>KNN is easy to implement. Accuracy of REGGAE is much lower in SVM than KNN.</td>
<td></td>
</tr>
<tr>
<td>[8]</td>
<td>Neural Networks. Features are extracted from time and frequency domains</td>
<td>91.76%</td>
<td>Reliable and consistent. Further improvement by curating the datasets.</td>
<td></td>
</tr>
<tr>
<td>[9]</td>
<td>KNN, MFCC is used for feature extraction</td>
<td>70%</td>
<td>KNN is easy to implement. Accuracy can be improved.</td>
<td></td>
</tr>
<tr>
<td>[10]</td>
<td>CNN and MFCC is used for feature extraction</td>
<td>76%</td>
<td>Works better than SVM, KNN and Feed-forward Neural Network. Further experiments need to be done to improve the accuracy.</td>
<td></td>
</tr>
<tr>
<td>[12]</td>
<td>CNN, DNN</td>
<td>92% and 90% respectively</td>
<td>Accuracy is good compared to previous models. Dataset can be extended for further improving the accuracy.</td>
<td></td>
</tr>
<tr>
<td>[13]</td>
<td>CNN</td>
<td>87%</td>
<td>Better performance than traditional ML and manual models. Further experiments on improving the accuracy should be conducted.</td>
<td></td>
</tr>
<tr>
<td>[14]</td>
<td>CNN, LR, KNN, SVM, RF, Simple MLP</td>
<td>92.66% for KNN</td>
<td>Extended dataset by cutting the audio samples. Utilizing more content-based feature representation.</td>
<td></td>
</tr>
<tr>
<td>[15]</td>
<td>CNN</td>
<td>98% for training set and 73% for validation set</td>
<td>Log-based MEL spectrograms enhance low frequency details and reduce dimensionality. Can be further improved by using bigger datasets with different formats.</td>
<td></td>
</tr>
<tr>
<td>[16]</td>
<td>KNN, SVM, LR, RF</td>
<td>with and without hyper-para tuning 60.4% 66.4% 73.2% 76.4% 62.4% 67.2% 66.8% 69.6% respectively</td>
<td>Selection of beat hyper-parameters. Experiments with other deep learning approaches can be done for optimizing the parameters.</td>
<td></td>
</tr>
</tbody>
</table>
3 Problem statement and objectives

3.1 Problem statement

Music has an important role in everyone’s life. People have their preferences in music. It is highly variable and subjective in nature. It is important to classify music according to their genre this can be achieved using automated music genre classification using deep learning techniques. There are potential overlaps between genres which make it difficult to identify them. Manual genre classification is a hectic task. Therefore, developing an accurate and robust music genre classification model requires thorough domain knowledge. Hence, automated process of MGC is done using CNN.

3.2 Objectives

- To understand various audio representations.
- To process the audio files and extract Mel-Frequency Cepstral Coefficients (MFCCs).
- To classify the music data into different genres effectively using CNN.

4 Proposed method

Music genre classification is a procedure in which a machine learning system undergoes training to forecast the genre of a specific music track. Music genre categorization is an active topic of research, with several approaches being created and tested to increase classification accuracy. In this work, CNN, which is a deep learning approach, is used to achieve the goal of MGC.

4.1 Architecture of the proposed method

![Architecture diagram]

Figure 2 shows the architecture diagram of the proposed work. The dataset contains audio files from 10 different genres. MFCCs (Mel-Frequency Cepstral Coefficients) are extracted
from the audio files which will be saved in a JSON file. These MFCCs are divided into training sets and test sets. The training set is used to train the CNN model. The trained model is then used for the classification of new audio files. The test set is given to the trained model for evaluation.

4.2 Modularization

The proposed method consists of 4 modules. They are audio visualization, data processing, classification, and evaluation. In this section, we discuss the inputs provided to each module, the processing that happens, and finally, the output of the module.

4.2.1 Audio visualization

Audio visualization is the display of audio data or its representation in a visual form. It is a method of converting audio signals into visual elements.

- Visualizing audio file as a waveform: Plotting an audio in the waveform includes representing the amplitude of the audio as a function of time. Figure 3 shows the waveform.

![Fig. 3. Waveform visualization.](image)

- Visualizing audio file as a spectrogram: A spectrogram is a visual representation of the frequency spectrum of a signal over time. It is a two-dimensional graph with time on x-axis and frequency on y-axis. The plot is shown in Figure 4.

![Fig. 4. Spectrogram visualization.](image)

- Visualizing audio file as a Mel-spectrogram: Traditional spectrograms use a linear frequency scale. We can visualize the Mel spectrogram plot through matplotlib. The plot is shown in Figure 5.
4.2.2 Data processing

MFCC: It stands for Mel-Frequency Cepstral Coefficients. The MFCC algorithm converts the spectral information of an audio signal into a set of coefficients. These coefficients are used to characterize the signal’s spectral envelope. Figure 6 shows MFCC coefficients with time on x-axis and MFCC index on y-axis. The color represents the magnitude of the coefficient.

4.2.3 Classification

Convolutional Neural Networks (CNN) is used to serve the purpose of music genre classification. Several research papers concluded that using CNN is an effective way for MGC. Before building the CNN model, the MFCCs from the json file are split into training sets, validation sets and test sets. The ratios of the split are 55%, 20% and 25% for training, validation and testing respectively. After this process we define the CNN architecture. Figure 7 shows the CNN architecture of the implemented model for genre classification.

4.2.4 Evaluation

The trained model needs to be evaluated to analyse its performance. Various performance metrics such as accuracy, f1-score, precision, recall, and confusion matrix are implemented to study the model’s performance. The performance metrics are as follows:

- Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)
• F1 Score = \( \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \)
• Precision = \( \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \)
• Recall = \( \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \)

A confusion matrix for multiclass classification is a table that shows the number of correct and incorrect predictions for each class of the model. The diagonal elements represent the number of correct predictions for each class, while the off-diagonal elements represent the number of incorrect predictions.

**Fig. 7.** CNN architecture.

### 5 Results and discussions

#### 5.1 About the dataset

GTZAN dataset consists of 1000 audio clips, each lasting 30 seconds and derived from commercially available audio clips. Blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock are among the ten music genres represented in the GTZAN dataset. The GTZAN database comprises 100 audio clips for each genre. The GTZAN dataset contains audio clips in WAV format with a sampling rate of 22.05 kHz and a resolution of 16 bits.

#### 5.2 Experimental results

Mel-frequency cepstral coefficient (MFCC) is used to train the model and gain useful insights from it. The famous GTZAN dataset is used. MGC is a multi-label classification task. There are several metrics that can be evaluated for these kinds of tasks. The overall accuracy of the model is 83%.

The accuracy of the model against the number of epochs during the training phase is plotted. The confusion matrix depicts the extent to which the multi-label classification model performs correctly in terms of predicting the presence or absence of labels. Figure 10 shows the confusion matrix implemented for MGC using CNN.
5.3 Significance of the proposed method

To achieve the goal of automated music genre classification, CNN approach is adopted. They can be trained on huge datasets and learn complex patterns while also generalizing well to new music samples. CNNs can be used to classify music genres in real-time once they are trained, which makes them appropriate for online music applications. The proposed approach has satisfactory performance in the task of classifying the automated music genres. The accuracy is 83% which is acceptable. From Table 2, the model performs well when compared to other existing methods. Modification in the CNN architecture resulted in accuracy improvement.
6 Conclusion and future enhancements

The CNN technique is used to meet the automated MGC objectives. The audio files are divided into shorter segments. MFCCs are retrieved from audio segments and saved as JSON files. Each segment contains 13 MFCCs that represent each audio file uniquely. By encoding audio signals in a small feature space, MFCC decreases the dimensionality of the signals. This improves computing efficiency and reduces the complexity of machine learning methods used for audio processing tasks like automated music genre classification. As a result, it is well suited to capture genre-specific elements that are consistent across different pitches and timbres of music. We attained an accuracy of 83%, demonstrating that the model can classify the genre more accurately. The F1 rating is 83%. Plots for accuracy versus epochs and loss versus epochs are shown to provide a more visual understanding of the model's performance. Our model achieved an 83% precision and recall rate. The confusion matrix is presented to summarize the forecasts for each genre.

Future enhancements: Current work processes the audio if it is of 30-second duration. Further experiments should be done to handle audio of any length. Implementing music genre classification for various audio formats can be explored. The implemented model works well for the WAV format. There are many other formats such as MP3, FLAC, and others. Data Augmentation can be done to further improve the performance of the model. It helps to increase the diversity and quantity of the training data. Hyperparameter tuning assists in determining the best combination of hyperparameter values to achieve the best performance.
of the model. Research towards solving this issue should be done to increase the performance of the model.

References
10. Dias, Jessica, Pillai, Vaishak, Deshmukh, Hrutvik, Shah, Ashok, Music Genre Classification & Recommendation System using CNN, SSRN, 4111849 (2022)
