International Journal of Computer Techniques – IJCT Volume 12 Issue 5, October 2025
Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

A Hybrid Deep Learning System for Automatic Music Genre Classification

[1] Adebanjo, Adedoyin S., [2] Odesanya, Konyisola A., [3] Bello, Sheriff K., [4] Oyerinde Emmanuel, [5] Adeoti Babajide Ebenezer, [6] Mgbeahuruike Emmanuel

[1] [5] [6] Department of Software Engineering, School of Computing, Babcock University, Ilishan-Remo, Ogun-Sate, Nigeria.
[2] [3] Department of Computer Science, School of Computing, Babcock University, Ilishan-Remo, Ogun-Sate, Nigeria.
[4] Department of Information Technology, School of Computing, Babcock University, Ilishan-Remo, Ogun-Sate, Nigeria.
[1] adebanjoa@babcock.edu.ng, [2] odesanya2822@student.babcock.edu.ng, [3] bello7571@student.babcock.edu.ng, [4] oyerindee@babcock.edu.ng,, [5] adeotib@babcock.edu.ng, [6] mgbeahuruikee@babcock.edu.ng

Abstract*— The rapid growth of digital music libraries has made manual genre classification inefficient and inconsistent. This study proposes an automatic music genre classification system using a hybrid deep learning model. The model combines Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) for recognizing sequential patterns. It uses Mel-spectrograms and MFCCs extracted with Librosa from a Kaggle dataset of 10 genres. The implementation is done with Python, TensorFlow, Keras, and the system is deployed via Streamlit for real-time predictions. Experimental results show a classification accuracy of 90%. This outperforms standalone models by 3% and demonstrates the system's effectiveness, scalability, and potential to improve music recommendation and retrieval platforms.

Index Terms— *Audio Feature Extraction, CNN-RNN Hybrid Model, Deep Learning,Music Information Retreieval, Music Genre Classification.*

I. INTRODUCTION

In recent years, the rapid growth of digital streaming platforms like Spotify, Apple Music, and YouTube Music has changed how people consume music. These platforms give audiences worldwide unprecedented access to large music libraries, leading to a huge increase in the amount of available content. By February 2024, for instance, YouTube Music reported over 100 million subscribers [1]. Users in the Middle East and Africa (MEA) region spent more than two hours daily on Apple Music [2]. In Nigeria, music streaming services are expected to generate revenues of about USD 87.91 million [3]. With more time and engagement spent on these platforms, the large amount of data being processed highlights the need for effective music classification systems.

Music genre classification is crucial in the field of Music Information Retrieval (MIR). This field focuses on organizing, recommending, and discovering music in large digital repositories [4]. With streaming platforms now offering millions of tracks, traditional manual classification methods have become impractical. These methods, once used by musicologists and curators, relied on analyzing musical features like rhythm, harmony, instrumentation, and cultural context. Such analysis provided valuable insights into the structure and beauty of music [5]. However, these methods were labor-intensive, took a lot of time, and often relied on personal opinions. Limited exposure to different genres could introduce biases, leading to inconsistencies and less reliable classification results.[6]-[10].

The rise of machine learning, and more recently, deep learning has brought a significant change in how we classify music genres. These methods use improved feature extraction techniques to examine audio signals and identify patterns specific to different genres [11]. Common features like Mel-spectrograms and Mel-frequency cepstral coefficients (MFCCs) offer numerical representations of sound, making analysis easier. By using these representations, machine learning models can automatically learn and classify musical patterns more accurately and efficiently than traditional manual methods or rules [12], [13].

Despite significant progress, music genre classification still encounters major challenges. Ambiguity in genre boundaries is a key issue. Many songs share overlapping traits, which makes accurate categorization difficult. Additionally, multicultural and underrepresented genres often get little attention in current systems. This leads to misclassification and decreased visibility for these types

of music [14]. To tackle these problems, we need to develop strong classification models that can handle diverse datasets and reflect the complexity of modern music. Such improvements would improve user experiences by providing better recommendations and personalized playlists. They would also help discover new and diverse music, enriching the global music scene.

The aim of the study is to develop an automatic music genre classification system using a hybrid CNN-RNN-based deep learning technique, capable of accurately identifying and categorizing a wide range of music genres.
The specific objectives of the study are:

i. To identify the various music genres that would be used to build the classification model.
ii. To build a hybrid CNN-RNN-based music genre classification system.
iii. To evaluate the model accuracy and performance.

II. LITERATURE REVIEW

2.1 Music Information Retrieval (MIR) and Music Genre Classification

Music Information Retrieval (MIR) is a key field that focuses on extracting and organizing information from music. This includes tasks like genre recognition, pitch detection, and recommendation [4]. Its importance has increased with the rise of streaming platforms, where MIR helps with search, personalized recommendations, and managing large databases [12]. Beyond commercial uses, MIR also supports musicology, emotional analysis, and therapeutic studies [15].

In MIR, genre classification is essential for organizing large music libraries, improving retrieval, and enabling personalized recommendations [7], [10]. Accurate classification not only enhances playlist generation and user engagement but also encourages the discovery of new styles and aids research into cultural and stylistic evolution. Therefore, genre classification is crucial for both practical uses and broader music research [7], [9]-[10].

2.2 Traditional Methods for Music Genre Classification

Early music genre classification relied on manual labeling, where experts tagged songs based on rhythm, melody, instrumentation, and style. While important, these methods were subjective, inconsistent, and not suitable for large digital collections due to cultural biases and high resource demands. To solve these problems, computational methods introduced features like temporal, cepstral, and spectral attributes, which were classified using algorithms such as k-nearest neighbors (KNN), support vector machines (SVM), and decision trees [5]-[9]. Although these methods offered greater scalability, they were still limited by predefined rules and often missed capturing overlapping or hybrid genres [10].

Machine learning improved the field by allowing models to learn patterns from data instead of relying only on expert-designed features [9]. However, traditional algorithms like SVMs and decision trees depended heavily on the quality of features and struggled to capture the complex and temporal dynamics of music signals [16]. These challenges pointed to the need for more advanced, data-driven methods that could model both spectral and sequential structures, leading to the rise of deep learning techniques.

2.3 Deep Learning Approaches

Deep learning has changed how we classify music genres. It now allows for direct learning from raw audio or spectrograms. This is different from traditional machine learning, which depended on handcrafted features. Models like CNNs and RNNs automatically extract representations, increasing accuracy across various genres These improvements are backed by large datasets, such as GTZAN and the Million Song Dataset, along with greater computational power. New methods using multimodal inputs and Transformer-based architectures further boost performance [11]-[12].

CNNs are commonly used because they capture local spectral and temporal features from spectrograms or MFCCs. Deeper layers learn higher-level abstractions [13], [17]. Adding recurrent layers like Long Short-Term Memory (LSTM) networks enhances temporal modeling. Meanwhile, attention mechanisms improve feature selection [18]-[19]. Techniques such as transfer learning, data augmentation, and multimodal integration help deal with overfitting and data scarcity [20]-[21].

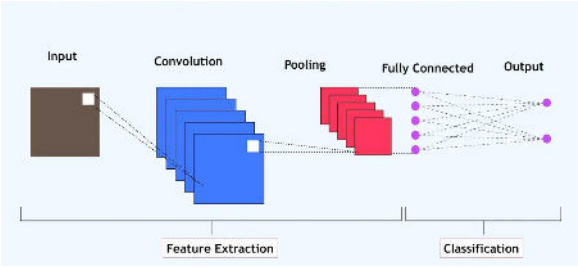

Figure 1: CNN Architecture

RNNs, including LSTM and Gated Recurrent Units (GRU) work well with CNNs by modeling sequential dependencies and keeping temporal context. They effectively capture rhythm and melody [13], [21]. LSTMs handle vanishing gradients for long-term dependencies, while GRUs provide similar accuracy with lower complexity. Bidirectional versions like Bi-LSTM and

Bi-GRU improve performance by considering both past and future context [18], [22].

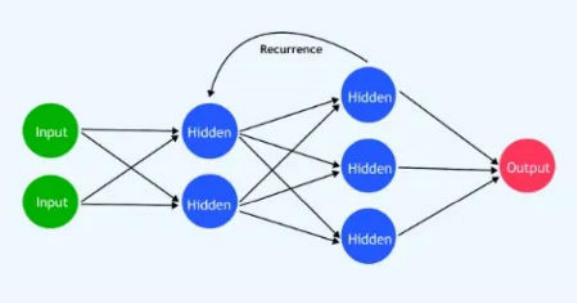

Figure 2: RNN Architecture

Hybrid CNN-RNN architectures merge spatial and temporal modeling for better classification accuracy. CNN layers reduce dimensionality, and recurrent layers capture sequential dynamics. Models like CNN-BiGRU and CNN-BiLSTM outperform standalone networks in benchmarks.

Despite challenges, such as computational costs and gradient instability, approaches like GRU variants, bidirectional designs, dropout regularization, and optimized learning rates enhance robustness. Studies indicate that Mel-spectrograms often work better than MFCCs as input features because they offer richer signal representation [13], [14].

2.4 Key Datasets and Benchmarks

Datasets and benchmarks play a crucial role in advancing music genre classification by offering standard resources for training and evaluation. The GTZAN dataset [23], which contains 1,000 clips across ten genres, is the most widely used due to its easy access and balanced representation. It supports models that range from CNNs to hybrid CNN-RNNs [13]. Other datasets, like the Extended Ballroom with rhythm annotations and the large-scale Million Song Dataset (MSD), expand research into tempo-focused studies and large-scale retrieval tasks. Additional resources such as Ballroom, Emotify, and LastFM offer different perspectives: rhythm-based, emotion-based, and social-metadata. Performance is typically evaluated using metrics like accuracy, precision, recall, and F1-score [13]-[14].

Despite their usefulness, current datasets have several drawbacks. For instance, GTZAN is small, contains mislabeled and duplicated tracks, and lacks diversity outside Western genres. This raises concerns about overfitting and cultural bias. Limited audio quality and incomplete metadata also hinder multimodal analysis. These issues impact the reliability of benchmarks, often leading to inflated accuracy scores and restricted generalizability [13], [14].

Having diverse datasets is therefore vital for strong classification. A broader range of genres, cultures, and production styles helps reduce bias and improves model flexibility. This is especially important for real-world applications like streaming recommendations. Using multimodal inputs, which combine audio with metadata such as artist background or listener context, further enriches classification and supports sophisticated deep learning frameworks [13], [24].

2.5 Motivation for Hybrid CNN-RNN Models

Recent studies show that using hybrid CNN-RNN architectures is effective in classifying music genres. These models combine CNN layers, which extract spectral and timbral features from Mel-spectrograms or MFCCs, with RNN layers that capture temporal dynamics. This combination utilizes both spatial and sequential structures in music. Bidirectional variants like BiLSTM and BiGRU improve performance by modeling past and future context. Additionally, CNN-based preprocessing helps solve gradient problems in deep RNNs [13].

This integration reflects the dual nature of music signals. It has consistently outperformed standalone CNNs or RNNs on various datasets. Beyond accuracy, hybrid frameworks also support improvements in explainability and classification. They offer richer feature representations that can incorporate metadata or multimodal inputs for systems that are more context-aware and culturally inclusive [13], [14].

III. METHODOLOGY

This study followed a structured plan for building the MGCS, ensuring that each component is scientifically sound and systematically implemented. The design includes data collection, feature extraction, model development, evaluation, and deployment.

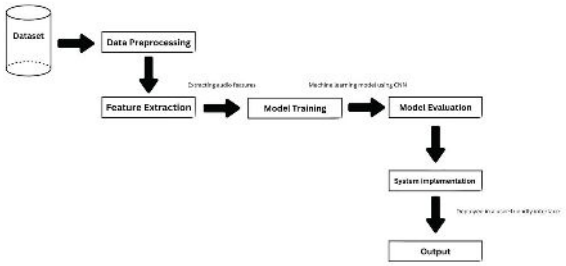

Figure 3: Music Genre Classification Model System Architecture

Each stage is explained in detail below.

3.1 Data Collection

A reliable dataset is crucial for effective model training. This study used the GTZAN dataset as the main source because it's widely recognized in music genre classification. GTZAN has 1,000 audio clips evenly spread across 10 genres, such as jazz, rock, classical, and

blues. This provides a balanced representation for evaluation.

To address the limited size of labeled data and improve model robustness, data augmentation techniques were applied. These include:

i. Pitch Shifting: Adjusting the pitch without changing the tempo.

ii. Time Stretching: Modifying the playback speed while maintaining pitch.

iii. Adding Noise: Introducing synthetic noise to simulate real-world recording conditions.

These techniques create synthetic variations of the dataset, increasing diversity and enhancing generalization.

3.2 Feature Extraction

Feature extraction bridges the gap between raw audio and machine learning models by converting audio signals into structured representations. Using Python's Librosa library, the following features were extracted:

i. Mel-Frequency Cepstral Coefficients (MFCC): Capture timbral information by modeling how humans perceive sound frequencies, making them essential for distinguishing between genres.

ii. Mel-Spectrograms: Provide a time-frequency representation of audio, capturing rhythmic patterns and tonal variations. These spectrograms are visualized and processed directly by the CNN model, making them particularly effective.

These extracted features reduce the complexity of raw audio while preserving essential genre-specific information.

3.3 Model Training

The CNN-RNN hybrid model is trained on extracted feature sets, with the dataset divided into training and testing subsets in an 80-20 split. Model parameters were optimized using backpropagation and gradient descent. To improve convergence and avoid overfitting, hyperparameters such as learning rate, number of layers, and batch size were tuned. Regularization strategies, including dropout and batch normalization, were applied to enhance generalization.

3.4 Model Evaluation

The performance of the trained models was assessed using standard metrics like accuracy, and the classification report, which provides details on precision, recall, and F1-score for each classification.

3.4 Model Deployment

The The final phase involved deploying the final model through a user interface website created using a Python library Streamlit which is capable of providing frontend features to allow users to upload audio files while also allowing to integrate the data preprocessing and model prediction to enable real-time classification to the end user.

IV. RESULTS

Table 1 and figure 4 present the results of the study. Table 1 shows the performance summary of the model. Fig. 4 shows the confusion matrix classification results of the hybrid CNN-RNN model. CNN achieved an accuracy of 89% on the validation set, performing particularly well on Metal, Hip-hop, and Classical genres. However, it struggled slightly with Country and Blues, where genre characteristics overlap. RNN achieved an accuracy of 87% on the validation set, demonstrating strong results in detecting temporal variations. However, it had lower precision for Rock and Disco. The Hybrid CNN-RNN model outperformed both CNN and RNN, achieving 90% accuracy on the validation set. It successfully captured both spatial and temporal features, making it the most robust model.

Table 1: Model Performance Summary

Model	Accuracy	Best Genres	Weak Genres
CNN	89%	Metal, Classical	Country, Blues
RNN	87%	Metal, Blues, Classical	Rock, Disco
Hybrid	90%	Metal, Classical	Rock, Country

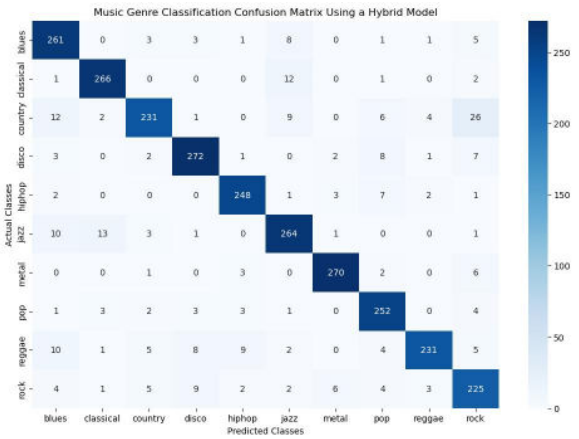

Figure 4: Hybrid CNN-RNN confusion matrix showing classification results.

CNN performed well for spatial feature extraction but struggled with temporal variations. RNN captured temporal dependencies but had slightly lower accuracy. Hybrid CNN-RNN achieved the highest accuracy (90%) by leveraging both spatial and sequential information. It is thus the most effective approach for music genre classification, balancing accuracy, robustness, and generalization.

V. CONCLUSION

This study explored the potential of deep learning models in accurately identifying different music genres. The study implemented three distinct architectures: Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and a hybrid CNN-RNN model. Through extensive data preprocessing, including noise reduction, feature extraction, and augmentation, the system was optimized for high performance. The results demonstrated that the hybrid CNN-RNN model provided the best accuracy, effectively capturing both spatial and temporal dependencies in music signals.

By leveraging CNNs and RNNs, the system effectively classified music genres with high accuracy, demonstrating the effectiveness of combining spatial and temporal feature extraction techniques. The findings confirm that deep learning approaches can significantly improve automated music classification, with implications for applications in streaming services, recommendation systems, and digital music libraries.

REFERENCES

[1] Statista, “Number of paying YouTube Music and YouTube Premium subscribers worldwide from 2020 to 2024,” 2024. [Online]. Available: https://www.statista.com/statistics/1344265/youtube-paying-subscribers/.

[2] Statista, “Average daily time spent streaming music on Spotify and Apple Music among internet users worldwide as of the 4th quarter of 2018, by region,” 2024. [Online]. Available: https://www.statista.com/statistics/1100724/time-spent-streaming-music-by-region/.

[3] Statista, “Music streaming – Nigeria,” 2024. [Online]. Available: https://www.statista.com/outlook/dmo/digital-media/digital-music/music-streaming/nigeria/.

[4] J. S. Downie, D. Byrd, and T. Crawford, “Ten years of ISMIR: Reflections on challenges and opportunities,” in Proc. 10th Int. Soc. Music Inf. Retrieval Conf., Kobe, Japan, 2009, pp. 13–18.

[5] C. Palisca, “Marc Scacchi’s defense of new music,” Muzyka, vol. 43, pp. 131–132, 1998.

[6] A. North and D. Hargreaves, The Social and Applied Psychology of Music. Oxford, U.K.: Oxford Univ. Press, 2008.

[7] G. Sun, “Research on architecture for long-tailed genre computer intelligent classification with music information retrieval and deep learning,” J. Phys.: Conf. Ser., vol. 2033, p. 012008, 2021.

[8] W. Wang, Y. Huang, Y. Wang, and L. Wang, “Generalized autoencoder: A neural network framework for dimensionality reduction,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2014, pp. 490–497.

[9] A. Dorochowicz, A. Kurowski, and B. Kostek, “Employing subjective tests and deep learning for discovering the relationship between personality types and preferred music genres,” Electronics, vol. 9, no. 12, p. 2016, 2020. [Online]. Available: https://doi.org/10.3390/electronics9122016.

[10] Y. Zhang and T. Li, “Music genre classification with parallel convolutional neural networks and capuchin search algorithm,” Sci. Rep., vol. 15, p. 9580, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-90619-7.

[11] L. Liu, “The implementation of a proposed deep-learning algorithm to classify music genres,” Open Comput. Sci., vol. 14, no. 1, p. 20230106, 2024. [Online]. Available: https://doi.org/10.1515/comp-2023-0106.

[12] N. Ndou, R. Ajoodha, and A. Jadhav, “Music genre classification: A review of deep-learning and traditional machine-learning approaches,” in Proc. 2021 IEEE Int. IoT, Electron. and Mechatronics Conf. (IEMTRONICS), 2021, pp. 1–6. doi: 10.1109/IEMTRONICS52119.2021.9422487.

[13] M. Ashraf, F. Abid, I. U. Din, J. Rasheed, M. Yesiltepe, S. F. Yeo, and M. T. Ersoy, “A hybrid CNN and RNN variant model for music classification,” Applied Sciences, vol. 13, no. 3, p. 1476, 2023. doi: 10.3390/app13031476.

[14] H. Foroughmand and G. Peeters, “Extending deep rhythm for tempo and genre estimation using complex convolutions, multitask learning and multi-input network,” in Proc. 2020 Joint Conf. AI Music Creativity, Stockholm, Sweden, Oct. 2020. [Online]. Available: https://hal.science/hal-03127155.

[15] M. Schedl, E. Gómez, and J. Urbano, Music Information Retrieval: Recent Developments and Applications. Boston, MA, USA: Now Publishers, 2014. [Online]. Available: https://doi.org/10.1561/1500000042.

[16] K. Choi, G. Fazekas, M. B. Sandler, and K. Cho, “Convolutional recurrent neural networks for music classification,” arXiv preprint arXiv:1609.04243, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1609.04243.

[17] L. Qiu, S. Li, and Y. Sung, “DBTMPE: Deep bidirectional transformers-based masked predictive encoder approach for music genre classification,” Mathematics, vol. 9, no. 5, p. 530, 2021. [Online]. Available: https://doi.org/10.3390/math9050530.

[18] Y. Tian, "Deep neural networks and fractional grey lag goose optimization for music genre identification," Sci. Rep., vol. 15, p. 6702, 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-91203-9.

[19] Y. Ding, H. Zhang, W. Huang, X. Zhou, and Z. Shi, "Efficient music genre recognition using ECAS-CNN: A novel channel-aware neural network architecture," Sensors, vol. 24, no. 21, p. 7021, 2024. [Online]. Available: https://doi.org/10.3390/s24217021.

[20] M. K. Kumar, K. Sujanasri, B. Neha, G. Akshara, P. Chugh, and P. Haindavi, "Automated music genre classification through deep learning techniques," E3S Web Conf., vol. 430, p. 01033, 2023. [Online]. Available: https://doi.org/10.1051/e3sconf/202343001033.

[21] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," J. Big Data, vol. 8, p. 53, 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00444-8.

[22] Y.-J. Mon, "LSTM-based music generation technologies," Computers, vol. 14, no. 6, p. 229, 2025. [Online]. Available: https://doi.org/10.3390/computers14060229.

[23] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," IEEE Trans. Speech Audio Process., vol. 10, no. 5, pp. 293–302, Jul. 2002. doi: 10.1109/TSA.2002.800560.

[24] J. Perianez-Pascual, J. D. Gutiérrez, L. Escobar-Encinas, Á. Rubio-Largo, and R. Rodriguez-Echeverria, "Beyond spectrograms: Rethinking audio classification from EnCodec's latent space," Algorithms, vol. 18, no. 2, p. 108, 2025. [Online]. Available: https://doi.org/10.3390/a18020108.