VAWKUM Transactions on Computer Sciences
http://vfast.org/journals/index.php/VTCS@ 2024, ISSN(e): 2308-8168, ISSN(p): 2411-6335
Volume 12, Number 2, July-December 2024 pp:237-252

Optimized Music Classification with a Hybrid VGG16-RNN Using Mel-Spectrogram and MFCC Features

Saima Ashraf [1] **and Mohsin Ashraf** [1*]

[1]Department of Computer Science, University of Central Punjab, Lahore 54700, Pakistan

Keywords: *visual geometry group; recurrent neural network; Mel-Spectrogram; MFCC.*

Journal Info:
Submitted:
November 04, 2024
Accepted:
December 15, 2024
Published:
December 26, 2024

Abstract

Music classification using deep neural networks has gained a lot of attention in recent years. This is due to the difficult task of capturing every essential aspect of music in features and interpretability of classifiers. There is limited research on the integration of VGG16 and RNNs, but the researchers found that few classifiers accurately capture intrinsic musical characteristics. Previous work in this field has primarily focused on spectral features, which has constrained overall performance. To address this issue, we proposed a novel hybrid neural architecture based on Visual Geometry Group 16 (VGG16), which is highly effective in extracting important features from musical variations. We combined VGG16 with several recurrent neural network (RNN) variants, including Gated Recurrent Unit (GRU), Bidirectional GRU (BiGRU), Long Short-Term Memory (LSTM), and Bidirectional LSTM (BiLSTM). Additionally, we compared their performance for the GTZAN dataset using both Mel-Spectrogram and Mel-Frequency Cepstral Coefficients (MFCC) features. Our results indicate that the VGG16+GRU model achieved the highest accuracy of 89. 60% with Mel spectrograms and 82. 70% with MFCC features. These findings demonstrate the effectiveness of combining advanced feature extraction techniques with deep learning models for music genre classification.

***Correspondence author email address:** mohsin.ashraf@ucp.edu.pk
DOI: 10.21015/vtcs.v12i2.1962

1 Introduction

Music Classification is a task in music information retrieval (MIR) to understand music semantics. Music has many applications in the industry, such as recommendation systems, personalized playlists, and content-based retrieval. The evolution of music is linked to changes in dynasties, the development of the market economy, and changes in humanistic spirit, leading to the prosperous stage today driven by rapid economic growth [1]. The field of music genre classification has roots both in musicology and in the development of computational techniques. Although the concept of categorizing music into genres has existed for centuries based on human intuition and cultural

conventions, the application of computational methods to automatically classify music genres is a more recent development. In the past decade, the rise of electronic music and technological advancements have prompted researchers to develop computational techniques for music classification. This gave rise to the field of Music Information Retrieval (MIR). This has led to the development of more advanced algorithms for classifying music. In [2], hand-crafted features from various sources (audio, chords, lyrics, and visual spectrograms) were evaluated and compared to the learned features using various fusion algorithms. The hand-crafted features used are the Local Binary Pattern (LBP), the Robust Local Binary Pattern (RLBP), Statistical Spectrum Descriptors (SSD), Simplified Chord Sequences (SCS) [3, 4], and the Mel-frequency Cepstral Coefficient (MFCC) [5]. Several DNN models have achieved impressive results in traditional music classification fields such as VGGNet [6, 7], ResNet [8], DenseNet [9], NANSNet [10, 11], MobileNet [12].

Deep learning significantly improves music classification [13, 14] by automatically extracting complex, high-level features from raw audio data that traditional methods struggle to capture. In contrast to hand-crafted characteristics, deep learning models such as VGG16 (Visual Geometry Group) and RNNs (Recurrent Neural Networks) learn to recognize patterns in musical aspects such as rhythm, timbre, melody, and harmony.

1.1 VGG16

VGG16 [15] is well regarded for its simplicity and effectiveness in image classification tasks. Convolution operations on the audio waveform enable the automatic extraction of features from raw audio signals. The architecture is notable for using very small convolutional filters, shown in Figure 1.

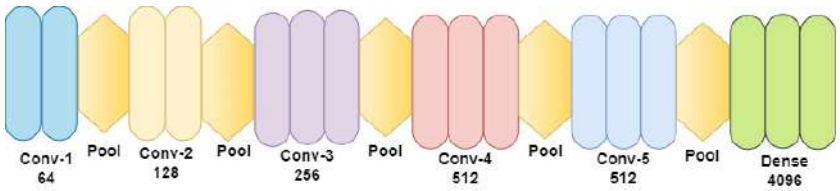

Figure 1. Architecture of VGG16

1.2 RNNs

Recurrent Neural Networks (RNNs) [16] are a type of artificial neural network that handles sequence data by maintaining context from earlier data points. Despite typical feedforward neural networks that process each input independently, RNNs leverage the temporal or sequential structure of the data by passing information from one step to the next through hidden states. This makes them particularly useful for tasks where context or previous inputs are important, as shown in Figure 2.

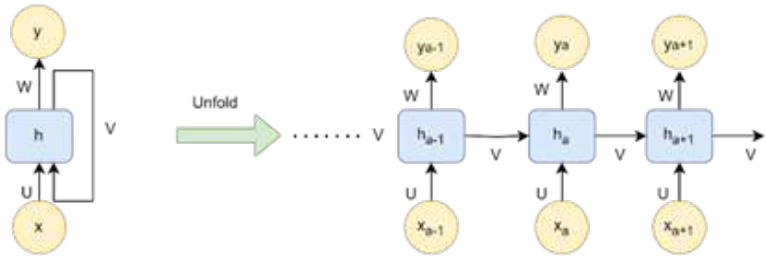

Figure 2. Architecture of RNNs

Previous studies have extensively used the GTZAN [17–19] dataset for music classification, establishing it as

a standard benchmark for musical analysis. However, there is still a need for a robust analytical framework that is capable of automatically analyzing the vast and growing digital music libraries. The purpose of combining VGG16 and RNN variations is to employ feature learning with estimated parameters while simultaneously utilizing sequential modeling, which is required for neural networks to successfully interpret the intrinsically sequential structure of music files.

The primary contributions of this paper are as follows:

- MFCC and Mel-Spectrograms are used to identify the most effective neural architecture for music classification.
- For music classification, various hybrid models such as VGG16+LSTM, VGG16+BiLSTM, VGG16+GRU, and VGG16+BiGRU are considered.
- Finally, the performance of the different hybrid models was evaluated across various extracted features utilizing a similar music dataset.

The remaining sections of this paper are structured as follows: Section 2 includes a literature review. Section 3 describes the proposed hybrid architecture, while Section 4 covers dataset overview, the experimental design, and the result and analysis. Section 5 discusses the results and process of feature extraction. The conclusion is in Section 6.

2 Literature Review

In Music Information Retrieval (MIR) tasks, extracting information from images is crucial because of the increasing challenges in feature extraction, it is more difficult to directly identify the genre of music from the audio input and use classifiers to categorize the music into specific genres. Many researchers used machine learning algorithms like Support Vector Machines (SVM) [20] and K-Nearest Neighbors (K-NN) [21]. Feature extraction from the audio file is performed with Mel-Frequency Cepstral Coefficients (MFCC) to obtain feature vectors. The feature vectors were classified using supervised learning methods [22], namely K-NN, Linear SVM, and Polynomial SVM, achieving overall accuracies of 64.4%, 60%, and 77.78%, respectively.

Ahmed et al. [23] utilized digital signal processing techniques followed by music genre classification and recommendations by machine learning techniques to extract the acoustic features of music. SVM outperformed other models with mixed characteristics, reaching 72% accuracy. Furthermore, convolutional neural networks, a deep learning approach, were also employed, and they performed their task in three steps: creating raw data, utilizing STFT with a hop size of 1024 and a window length of 2048, and MFCC with 13 coefficients. The results showed a 66% accuracy. Similarly, Suo et al. [24] compared different strategies for recognizing and classifying the GTZAN music dataset and employed a CNN model for the genre classification by using STFT and MFCC but were unable to perform a predictable result.

To classify music, Ashuman et al. used a model approach [25] collected eight features from audio files. These features were utilized to train neural networks for classical and Sufi genres. The model attained an overall test accuracy of 85%. Similarly, Heakl et al. [26] utilized the GTZAN dataset with constrained parameters and a global pooling strategy to evaluate a CNN-based network but only achieved a test accuracy of 70.60%. Li et al. proposed [27] a spectrogram-based approach for evaluating the models' performance in a deep neural network (DNN). They created a balanced, trusted model, ResNet50-trust. The FMA dataset has the highest classification accuracy of 80.14%.

The sequential nature of audio and RNN variants like LSTM and GRU are proposed for music classification in [28]. Mohsin et al. [29] introduced a Globally Regularized hybrid architecture that combines the convolutional and recurrent neural networks for music classification, addressing feature biases and training complexity caused by a fixed batch size, which hampers consistent training performance. Fulzele et al. [30] proposed a combined model

of LSTM and SVM for music classification, which enhanced the accuracy compared to using the individual methods. Similarly, Choi et al. [31] employed a CRNN (Convolutional Recurrent Neural Network) for music auto-tagging and evaluated its performance with three different CNN architectures, focusing on the result and training time per sample while controlling for the number of parameters. Furthermore, Mohsin et al. [32] successfully achieved outstanding outcomes and proved to have been highly effective in music classification by utilizing MFCC and Mel-Spectrogram features and used four models: CNN+GRU, CNN+LSTM, CNN+Bi-GRU, and CNN+Bi-LSTM, reaching a high accuracy of 89.30%. Similarly, Noopur et al.[33] utilized MFCCs for audio representation and applied two CRNN models on the GTZAN dataset, CNN-GRU, and CNN-LSTM, achieving an accuracy of 87.5%.

Although several studies have demonstrated consistent performance, they suffered from limited feature representation, inadequate training, and lack of compactness, which led to performance overhead. To enhance the effectiveness and accuracy of neural networks in classification tasks, we proposed a hybrid approach combining VGG16 with RNN variants models. By using two main features, Mel-Spectrograms and MFCC, in a new joint design. These features are first processed using VGG16 layers before being given to optimized RNN variants like LSTM, Bi-LSTM, GRU, and Bi-GRU, which were used to evaluate and compare the performance on the GTZAN dataset.

3 Proposed Hybrid VGG16-RNN Model

The proposed hybrid model for music genre classification has been implemented with MFCC and Mel-Spectrogram features. The workflow begins by generating Mel-Spectrograms and MFCCs through the use of the librosa library and Python package for music analysis. These Mel-Spectrograms and MFCCs are then fed into the joint architecture, which integrates VGG16 with RNN variants, followed by a comparison of accuracy between Mel-Spectrograms and MFCCs as shown in the below Figure 3.

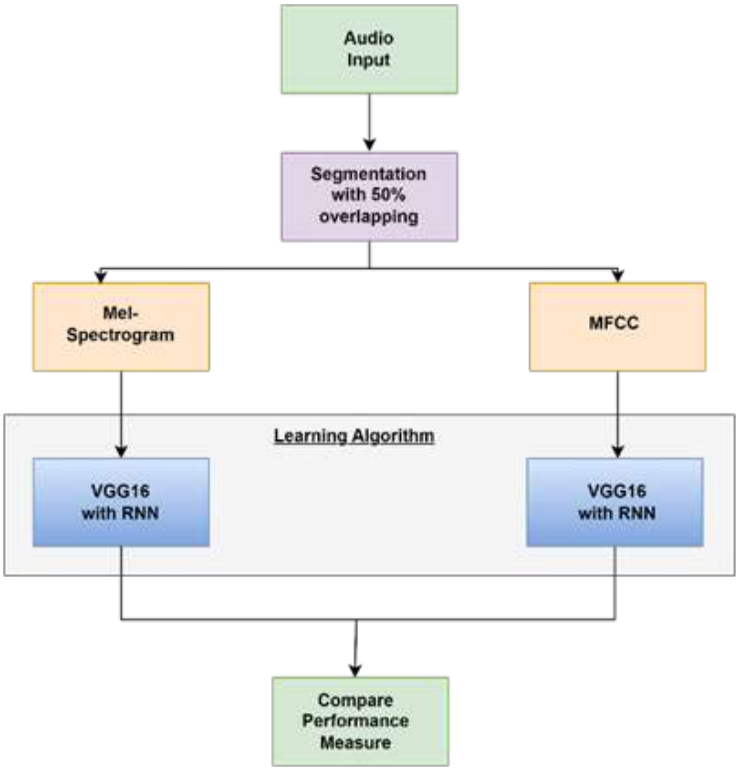

Figure 3. Proposed Hybrid VRNN Model

The subsequent steps of our proposed architecture are as follows:

- Segmentation with 50% overlapping
- Creating Mel-Spectrogram and MFCC
- Learning Algorithm (VGG16-RNN)

3.1 Segmentation

After preprocessing an audio input, an audio-splitting process is applied to convert a 30-second audio clip into 3-second segments, allowing for a more detailed evaluation of the proposed model.

3.2 Creating Mel-Spectrogram and MFCC

Each music clip underwent evaluation using different feature extraction techniques. The methods used for this comparison are Mel-Spectrograms and MFCCs. A spectrogram provides a two-dimensional visualization of data on frequencies across time and is also associated with bandpass filters, where $X(t)$ is any input data, n is the count of analysis filter banks, and $a_n(f)$ represents the analytical filters. By applying these analysis filter banks, a signal is decomposed into a set of sub-components signals $f_n(t)$ as shown in Figure 4, with each sub-component signal having a chunk of the original frequency spectrum. It closely resembles the analysis of digital filter banks versus Mel filter banks in Figure 5

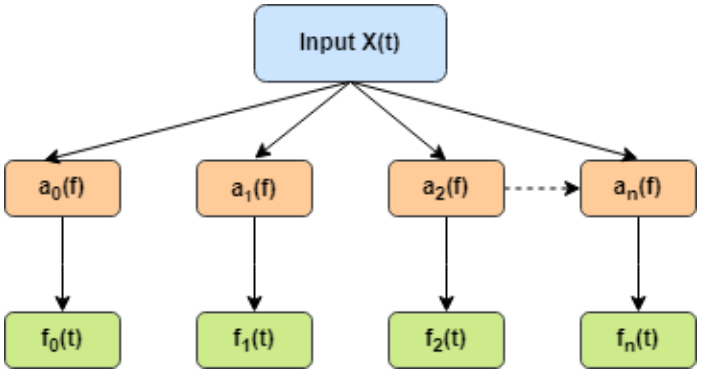

Figure 4. Digital filter bank analysis

During Mel-Spectrogram processing, we initiate by taking the audio clip as input and applying a Hann window process. Then, FFT is performed on each block, transforming the time-domain signal into a frequency-domain signal, that is alike to the Short-Time Fourier Transform (STFT). It is also utilized to process each frequency-domain signal, as illustrated in Figure 5.

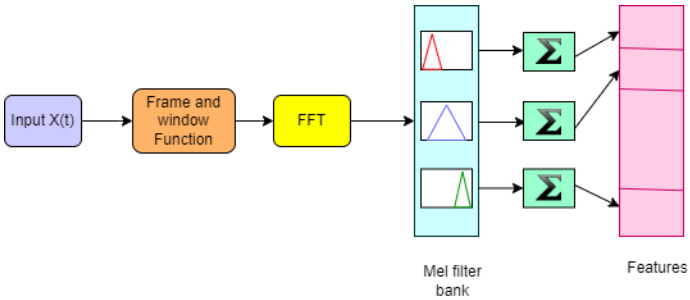

Figure 5. Evaluation process of Mel-Spectrogram

To evaluate further, the filter bank energy is calculated by multiplying the signal by the filter bank and summing the coefficients. Mel-Spectrogram vectors are formed by aggregating the coefficients across *n* filters.

$$h_{a,k} = f(x) = \begin{cases} \frac{f_k - f_{a-1}}{f_a - f_{a-1}}, & \text{if } f_{a-1} \leq f_k < f_a \\ \frac{f_{a+1} - f_k}{f_{a+1} - f_a}, & \text{if } f_a \leq f_k < f_{a+1} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

In this context, f_k represents the frequency bin obtained from the FFT, while f_a, f_{a-1}, and f_{a+1} denote the frequencies of the adjacent Mel bands. When this equation is applied, the resulting values create a triangular graph, as illustrated in Figure 6. Each triangle is colored differently to represent various analysis filter banks, which facilitates the comparison of these filter bands with the Mel-based filter bank. Triangle filters are effective for smoothing signals and are fundamental to the Mel scale, which mimics the way humans perceive sound.

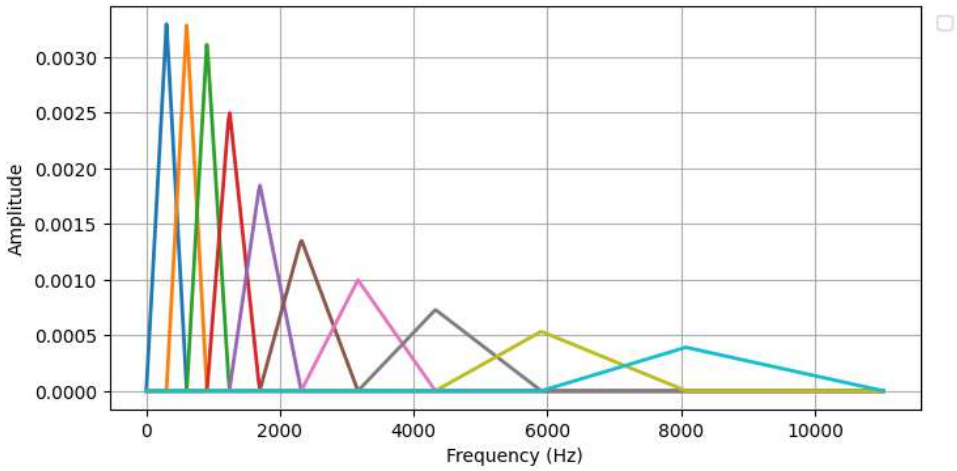

Figure 6. Mel-based filter bank for n=10

Each clip has dimensions of (128,129) when the Mel-Spectrogram is generated using an FFT window length and hop size of 1024 and 512, respectively as depicted in Figure 7.

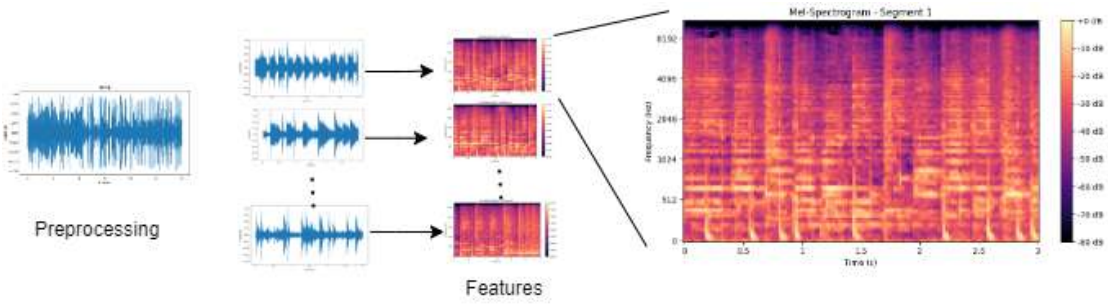

Figure 7. Generation of a Mel-Spectrogram for a Segment of Hip-hop Music (3-Sec duration).

.

Mel-Frequency Cepstral Coefficients (MFCC) is a representation of audio signals that capture the perceptually significant features of sound. The process we used involves taking the logarithm of the power spectrum and applying the Discrete Cosine Transform (DCT) to extract a compact set of 13 selected coefficients. These coefficients

effectively summarize the spectral characteristics of the audio. We chose to use the Discrete Cosine Transform (DCT) instead of the inverse Fast Fourier Transform (IFFT) because the DCT is expected to perform similarly to the FFT while being easier to compute and implement, as shown in Figure. 8.

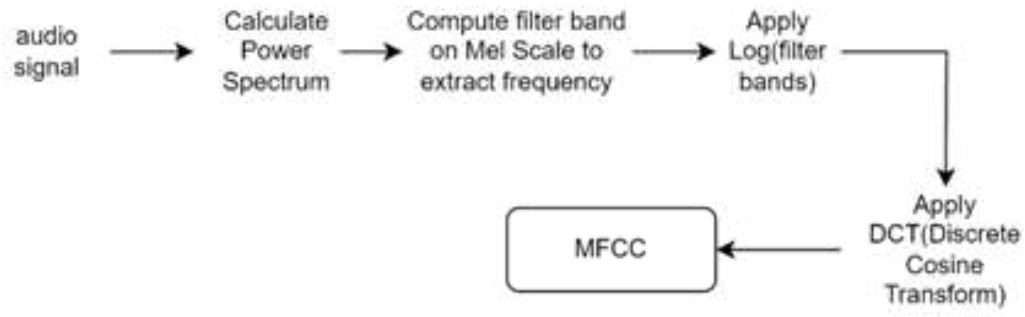

Figure 8. Process of MFCC evaluation

The MFCC for the Hip-Hop genre is illustrated in Figure 9.

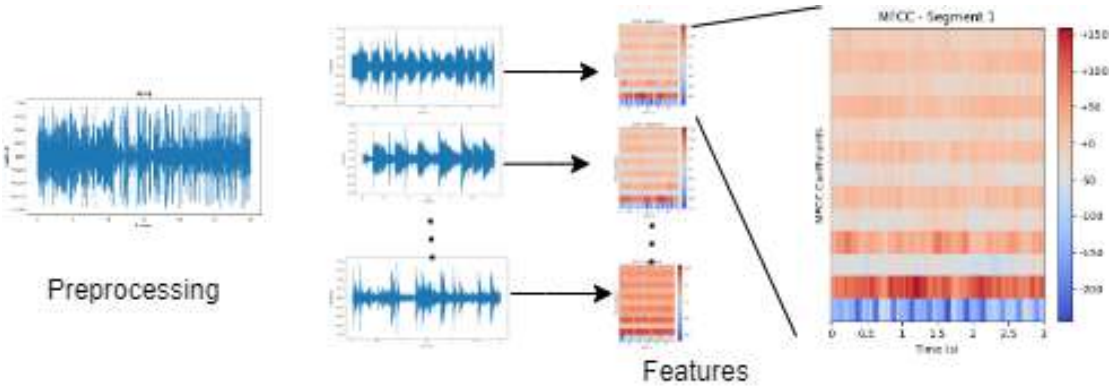

Figure 9. Generation of an MFCC for a Segment of Hip-hop Music (3-Sec duration).

3.3 Learning Algorithm (VGG16-RNN)

We segmented the dataset into three sections: training, testing, and validation, using an 8:1:1 ratio. To maintain continuity in the sequence, we preserved 50% of the original data, allowing for shuffling within each genre without losing information. After training, the model's performance on validation data is evaluated after each epoch to assess its ability to generalize to unseen data. The performance is measured using test samples following the training and validation data assessment. VGG16 has achieved remarkable outcomes in the analysis of image data. Music classification demonstrates its ability to capture complex patterns and features within audio signals. The methodology used [34] involves an architecture comprising 16 convolutional layers, which allows it to extract high-level features from Mel-Spectrograms and Mel Frequency Cepstral Coefficients (MFCC). While VGG16 excels at feature extraction from static images, it does not account for changes over time. To address this limitation, RNN variants play a significant role in storing information for future use. Additionally, during the training process, we made various adjustments to the hyperparameters to optimize performance.

4 Experiments

This section covers the experiments conducted with our proposed model, including an overview of the dataset, baseline models, the design of the experiments, and results and detailed analysis.

4.1 Dataset Overview

We utilized the publicly available GTZAN dataset to evaluate our proposed model. This dataset consists of 1,000 audio tracks, each lasting 30 seconds. It includes 10 distinct genres: blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock, with 100 tracks per genre. The audio files are in .WAV format, sampled at 22.05 kHz with a 16-bit resolution.

4.2 Baseline Model

For our assessment of CNN and RNN models, we begin by showcasing the baseline model using various features applied to the GTZAN dataset, as displayed in Table 1. Following this, we experiment with the baseline model by incorporating the VGG16 architecture alongside RNNs, as also presented in the table 2.

Table 1. Base model utilizes the GTZAN dataset

Standard Model	Feature
CNN-RNN [32]	Mel-Spectrogram and MFCC

Table 2. Proposed model utilizes the GTZAN dataset

Standard Model	Feature
VGG16-RNN	Mel-Spectrogram and MFCC

4.3 Experimental Design

We utilized Jupyter Notebook for our experiment, as it is compatible with various packages and software. The Librosa package transforms music samples into Mel-Spectrograms and MFCCs during the architectural development process. Using the logarithmic function on the music files, we achieve a scalable result when the window length is 2048 and the hop length is 512. This method leverages the human perception of intensity, which is measured in decibels. It's important to note that using the same hyperparameters across all datasets is ineffective, as different datasets can uniquely influence various architectures. Therefore, selecting the right hyperparameter values and network size is crucial for training the neural network model. We conducted numerous experiments to identify the optimal attributes, including the VGG16 layers, kernel size, kernel count, hidden layers in RNN variants, and the learning rate, as shown in the below Table 3.

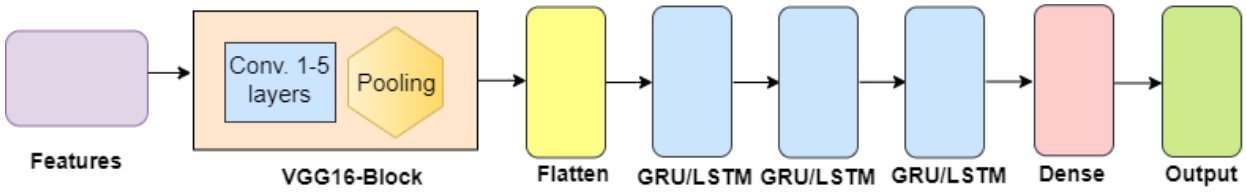

Figure 10. Proposed VGG16 and RNN variants (LSTM/GRU)

Table 3. Configuration of attributes for the proposed model

Attributes	Options value	Heightened
Window Size	-	2048
Hop Size	-	512
VGG16	-	5
Filter Size	-	3
Number of Filters	[64, 128, 256]	128
LSTM/GRU/BiGRU/BiLSTM	-	3
Dense	[16, 32, 64]	32
Dropout	[0.3, 0.5, 0.7]	0.5
Epochs Counts	[15, 20, 25, 30]	20, 30
Learning-Rate	[0.1, 0.01, 0.001]	0.001

In our proposed architecture, we employed a VGG16 block (with its initial five layers frozen) to extract features from the input data. Each layer has the same filter size. We then flattened the output into a 1D array and reshaped the features. The reshaped data is subsequently processed through three stacked GRU (Gated Recurrent Unit) or LSTM (Long Short-Term Memory) layers, each with 128 units, followed by a 50% dropout, which helps prevent overfitting. Lastly, use dense layers followed by an output layer as shown in Figure 10. We utilized the Adam optimizer for 15 to 30 epochs with a learning-rate of 0.001, using categorical cross-entropy as the loss function.

We used a similar approach by applying a single LSTM or GRU layer with 128 units to the VGG16 model. This was combined with a Bi-LSTM or Bi-GRU layer, which was then followed by another LSTM or GRU layer, also consisting of 128 units. These layers were then connected to a dense layer and finally an output layer.

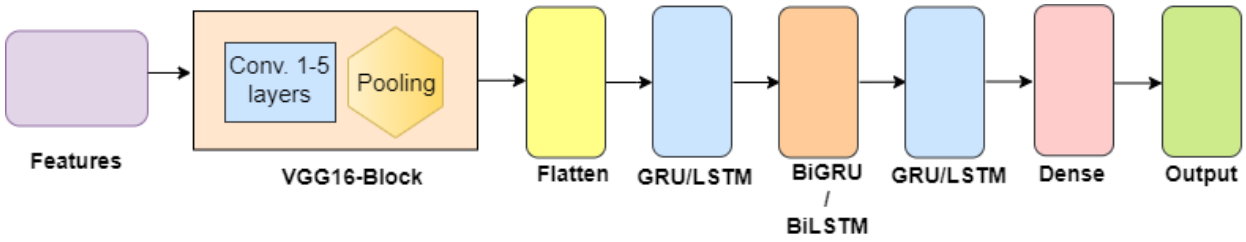

Figure 11. Proposed VGG16 and RNN variants (BiLSTM/BiGRU))

4.4 Results and Analysis

To evaluate the performance of our proposed model, we compared the accuracies of both the base model and the proposed model in a Table.4 and **??**.In the base model for the GTZAN dataset and achieved accuracies of 87.78% for the CNN-RNN model using Mel-Spectrogram features and 71.50% using MFCC features. In contrast, the VGG16-RNN model achieved accuracy rates of 89.60% with Mel-Spectrograms and 82.70% with MFCCs.

Table 4. Accuracy% of base model utilizes the GTZAN dataset

Standard Model	Feature	Result
CNN-RNN	Mel-Spectrogram	87.78
CNN-RNN	MFCC	71.50

Furthermore, we compared the accuracies of four distinct variants: VGG16+LSTM, VGG16+GRU, VGG16+BiLSTM, and VGG16+BiGRU using both Mel-Spectrogram and MFCC features as shown below in Table 5 and Figure 12 to assess the proposed hybrid model's performance.

Table 5. Outcomes of feature extraction using the proposed hybrid model

Audio-Feature	Architecture	Accuracy%
Mel-Spectrogram	VGG16-GRU	89.60%
Mel-Spectrogram	VGG16-LSTM	87.00%
Mel-Spectrogram	VGG16-BiGRU	89.20%
Mel-Spectrogram	VGG16-BiLSTM	86.70%
MFCC	VGG16+GRU	82.70%
MFCC	VGG16+LSTM	81.70%
MFCC	VGG16+BiGRU	80.60%
MFCC	VGG16+BiLSTM	81.70%

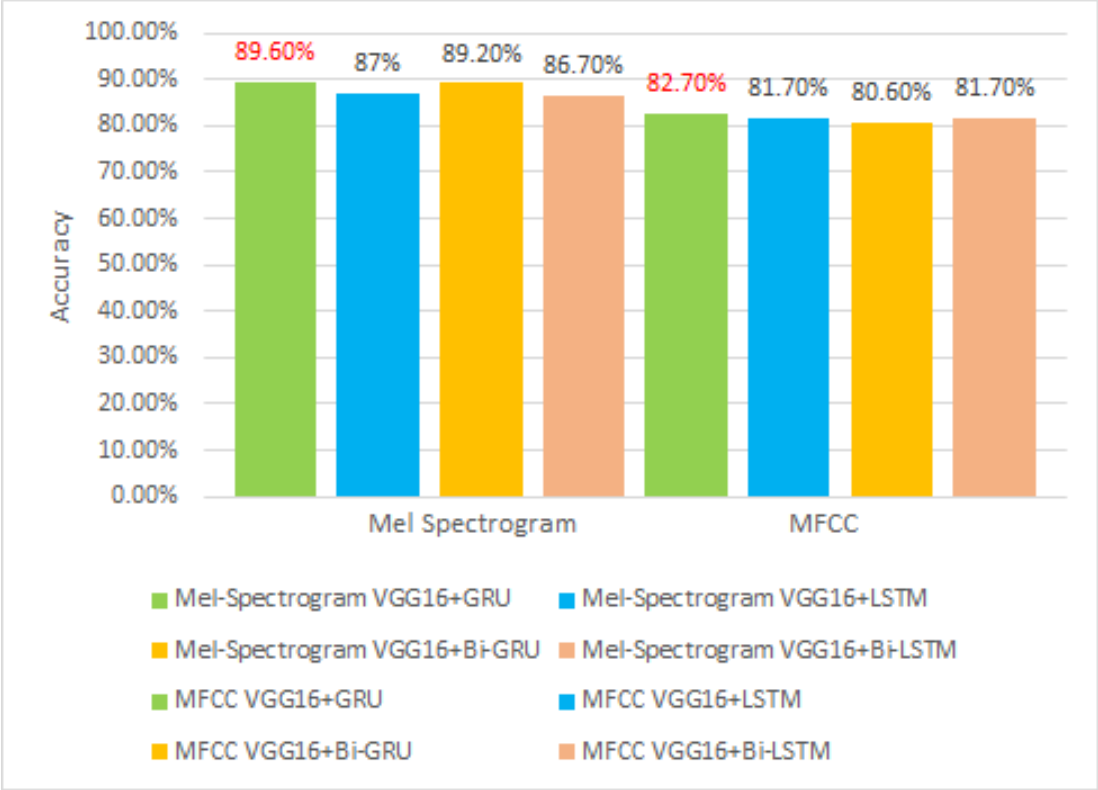

Figure 12. Extracted features' outcomes with proposed joint model

The VGG16+GRU combination, as shown in Figure 12, achieved the highest performance on Mel-Spectrograms with an accuracy of 89.60%, surpassing other models. The VGG16+BiGRU architecture followed closely, securing an accuracy of 89.20%. In contrast, when we evaluated the proposed model using MFCC features, the VGG16+GRU combination achieved the highest accuracy of 82.70%. A detailed comparison of our proposed architecture with

state-of-the-art models can be found in Table 6 and Figure 13.

Table 6. Results of feature processing using the proposed hybrid structure

Methodology	Accuracy%
Patil et al. [22]	64.00%
Ahmed et al. [23]	66.00%
Suo et al. [24]	68.70 %
Heakl et al. [26]	70.00%
Li et al. [27]	80.14%
Ashuman et al. [25]	85.00%
Noopur et al. [33]	87.50%
Jukubik. [28]	87.70%
Mohsin et al. [32]	89.30%
Proposed Model Work(VRNN)	89.60%

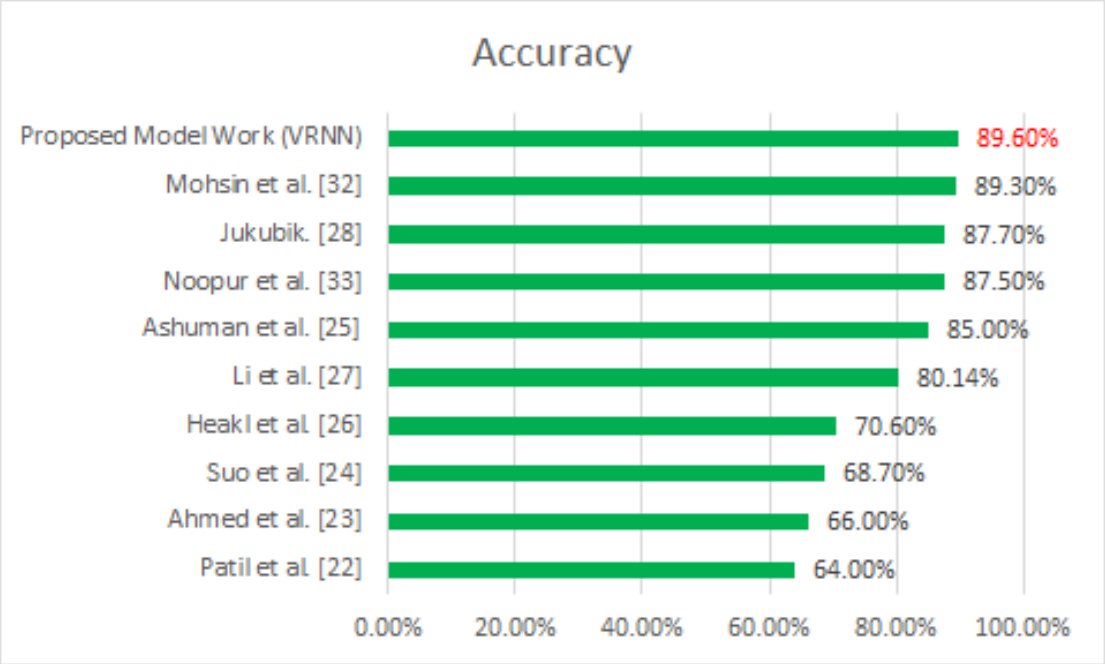

Figure 13. Comparison between the Innovative architecture and proposed architecture

We computed precision, recall, and F1-score for each genre in the GTZAN dataset to assess model performance, as shown in 7. In the GTZAN dataset, the reggae genre performs best, while classical shows the lowest results. This difference is due to some music samples being highly distinct, leading to high precision, recall, and F1-Score, while others have similar beats and rhythms, resulting in lower performance metrics.

Moreover, the outcomes are influenced by image dimensions, layer count, the number of filters, and kernel dimensions. The proposed model achieves an accuracy of 90%.

Table 7. Evaluation Metrics

Genre	Precision	Recall	F1-Score
Blues	0.92	0.93	0.92
Classical	0.85	0.73	0.78
Country	0.88	0.91	0.90
Disco	0.96	0.86	0.91
Hiphop	0.84	0.92	0.88
Jazz	0.96	0.89	0.92
Metal	0.94	0.91	0.92
Raggae	0.93	0.94	0.93
Rock	0.79	0.92	0.85

5 Discussion

After analyzing the proposed VGG16-RNN model, we compared it with different RNN variants for music classification tasks discussed in this section. Music features include zero-cross rate, spectral centroid, spectral contrast, and other techniques for classification tasks that restrict model performance as mentioned in [35, 36]. In comparison, this approach preserves spatiotemporal dependencies while normalizing the input feature map, resulting in improved training complexity and accuracy performance. By utilizing the Mel-Scale, spectrograms produce Mel-spectrograms and visualize samples as points evenly separated by time (t) and frequency (f). Our work leverages the advantages of the spatiotemporal domain by utilizing Mel-Spectrograms and MFCCs for improved music analysis. The GTZAN [17–19] dataset, frequently used in music information retrieval (MIR), has several limitations, including class imbalance and data overlap. Class imbalance can lead models to favor popular genres, reducing their ability to generalize to underrepresented ones. Additionally, overlapping data can cause overfitting, which may distort performance metrics. The GTZAN dataset is widely used in music information retrieval (MIR), with over 90% of studies depending on it. Our model produced strong results, but testing on alternative datasets could further confirm these findings. VGG16 combined with various RNN variants delivers impressive performance. Moreover, identifying the information patterns allows the model to more accurately assess the significance of features extracted by the VGG16 network at a defined period. Moreover, recurrent neural networks (RNNs) effectively capture long-term dependencies through temporal aggregation, enabling efficient management of sequential data. In the initial stages of the hybrid VGG16 and RNN, lower accuracy may arise due to RNNs' sensitivity to longer sequences, where information can degrade. Moreover, tuning issues with hyperparameter optimization. It was conducted to balance model accuracy and computational efficiency more effectively: the learning rate was initially set to 0.01 and gradually reduced to 0.001 using a learning rate scheduler for stable convergence, and selected the dropout rate to avoid overfitting, I experimented with numbers ranging from 0.2 to 0.5, and 0.5 produced the best results. Various kernel sizes (3x3, 5x5, 7x7) were tested for convolutional layers with 3x3 providing an optimal balance of feature extraction and speed.

The following 14 and 15 illustrate the model's accuracy during training and validation along the vertical axis increases as epochs along the horizontal axis.

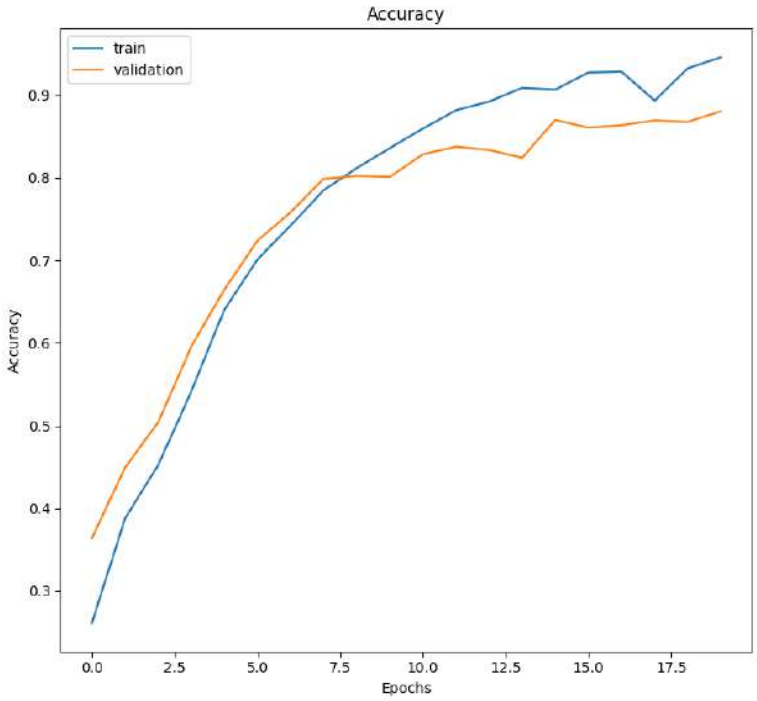

Figure 14. Accuracy between training and validation

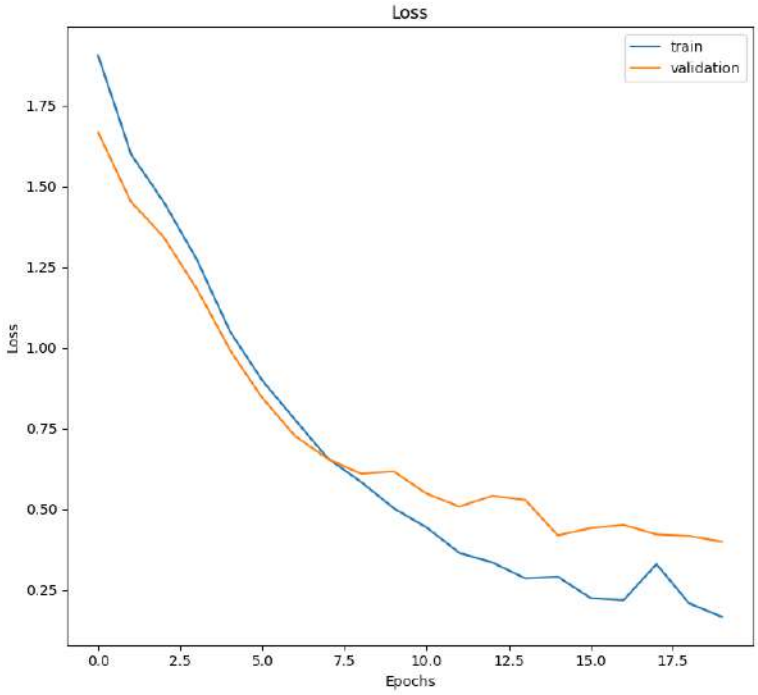

Figure 15. Loss between training and validation

The validation accuracy stabilizes after 15 epochs and remains consistent through the 20-epoch Mel-Spectrogram feature, while the training accuracy stabilizes after 25 epochs and continues to be consistent

through the 30-epoch MFCC feature. Our tests revealed that using Mel-Spectrograms and MFCC with VGG16-RNN with distinct variants accuracy results shown in TABLES 5 We further examined how the VGG16-RNN hybrid architecture is better optimized for Mel-Spectrograms, which aligns effectively with the spatial feature extraction capabilities of VGG16 and the temporal modeling strength of RNNs.

6 Conclusion

This research has contributed to the field of music genre classification by implementing and evaluating a hybrid deep-learning approach on the GTZAN dataset. We achieved notable classification accuracy by leveraging the strength of VGG16 combined with RNN variants and utilizing Mel-Spectrogram and MFCC techniques for transforming audio data. Our findings show that the VGG16+GRU model yielded the highest accuracy of 83% with MFCC features and an impressive 89.60% with Mel-Spectrograms, outperforming other tested configurations. This approach not only demonstrated competitive results with current state-of-the-art methods but also emphasized the benefits of combining advanced feature extraction with deep learning techniques. This work has important real-world applications. Music recommendation systems can use genre classification models to understand user preferences and create personalized playlists. Automated music tagging systems can efficiently manage large audio libraries with these models. Moreover, this research can support advancements in audio-based emotion recognition, music education tools, and content moderation, where genre classification is key. Future research could build on these results by experimenting with more complex feature extraction methods and diverse model architectures to further refine and enhance classification accuracy.

Author Contributions

Saima Ashraf: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation. **Mohsin Ashraf** : Supervision, Visualization, Investigation, Validation, Writing- Reviewing and Editing

Compliance with Ethical Standards

It is declare that all authors don't have any conflict of interest. It is also declare that this article does not contain any studies with human participants or animals performed by any of the authors. Furthermore, informed consent was obtained from all individual participants included in the study.

References

[1] Y. Li, Q. Zhang, and T. Gong, "Quantitative Influence Analysis of the Development Scale of Market Economy on the Level of Music Innovation," *Discrete Dynamics in Nature and Society*, vol. 2022, pp. 1–13, 2022. [Online]. Available: [CrossRef].

[2] R. M. Pereira, Y. M. G. Costa, R. L. Aguiar, A. S. Britto, L. E. S. Oliveira, and C. N. Silla, "Representation learning vs. handcrafted features for music genre classification," in *2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2019, pp. 1–8. [Online]. Available: [CrossRef].

[3] O. K. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Transactions on Multimedia*, vol. 23, pp. 3978–3985, 2020. [Online]. Available: [CrossRef].

[4] R. M. Pereira and C. N. Silla, "Using simplified chords sequences to classify songs genres," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 1446–1451. [Online]. Available: [CrossRef].

[5] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4th International Conference on Signal Processing and Communication Systems*, 2010, pp. 1–5. [Online]. Available: [CrossRef].

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: [CrossRef].

[7] M. F. Haque, H.-Y. Lim, and D.-S. Kang, "Object detection based on VGG with ResNet network," in *2019 International Conference on Electronics, Information, and Communication (ICEIC)*, 2019, pp. 1–3. [Online]. Available: [CrossRef].

[8] I. C. Duta, L. Liu, F. Zhu, and L. Shao, "Improved residual networks for image and video recognition," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 9415–9422. [Online]. Available: [CrossRef].

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708. [Online]. Available: [CrossRef].

[10] D. Baymurzina, E. Golikov, and M. Burtsev, "A review of neural architecture search," *Neurocomputing*, vol. 474, pp. 82–93, 2022. [Online]. Available: [CrossRef].

[11] Y. Kim, Y. Li, H. Park, Y. Venkatesha, and P. Panda, "Neural architecture search for spiking neural networks," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, Springer, 2022, pp. 36–56. [Online]. Available: [CrossRef].

[12] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. [Online]. Available: [CrossRef].

[13] D. Bisharad and R. H. Laskar, "Music genre recognition using convolutional recurrent neural network architecture," Expert Syst., vol.36, no.4 ,pp. 1-13, Aug.2019. [Online]. Available: [CrossRef].

[14] A.Huang and R.Wu, "Deep learning for music," 2016, [Online]. Available: [CrossRef]

[15] P. P. Das and A. Acharjee, "Double coated VGG16 architecture: An enhanced approach for genre classification of spectrographic representation of musical pieces," in *2019 22nd International Conference on Computer and Information Technology (ICCIT)*, 2019, pp. 1–5. IEEE. [Online]. Available: [CrossRef].

[16] S.-Y. Yin, Y. Huang, T.-Y. Chang, S.-F. Chang, and V. S. Tseng, "Continual learning with attentive recurrent neural networks for temporal data classification," *Neural Networks*, vol. 158, pp. 171–187, 2023. Elsevier. [Online]. Available: [CrossRef].

[17] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013. [Online]. Available: [CrossRef].

[18] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*vol. 10, no. 5, p. 293, 2002, [Online]. Available:[CrossRef].

[19] S.Sigtia and S.Dixon, "Improved music feature learning with deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2014, pp. 6959–6963. [Online]. Available: [CrossRef].

[20] M. I. Mandel and D. P. W. Ellis, "Song-level features and support vector machines for music classification," *Journal Name*, vol. X, no. Y, pp. Z–W, 2005. [Online]. Available: [CrossRef]

[21] D. Kostrzewa, R. Brzeski, and M. Kubanski, "The classification of music by the genre using the KNN classifier," in *Beyond Databases, Architectures and Structures. Facing the Challenges of Data Proliferation and Growing Variety: 14th International Conference, BDAS 2018, Held at the 24th IFIP World Computer Congress, WCC 2018, Poznan, Poland, September 18-20, 2018, Proceedings 14*, Springer, 2018, pp. 233–242.[Online]. Available: [CrossRef]

[22] N. M. Patil and M. U. Nemade, "Music genre classification using MFCC, K-NN and SVM classifier," *International Journal of Computer Engineering In Research Trends*, vol. 4, no. 2, pp. 43–47, 2017.[Online]. Available: [CrossRef]

[23] A. Elbir, H. B. Çam, M. E. Iyican, B. Öztürk, and N. Aydin, "Music genre classification and recommendation by using machine learning techniques," in *2018 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2018, pp. 1–5. [Online]. Available: [CrossRef]

[24] W. Suo, "Efficient Music Genre Classification with Deep Convolutional Neural Networks," in *2022 5th International Conference on Data Science and Information Technology (DSIT)*, 2022, pp. 01–05. [Online]. Available: [CrossRef]

[25] G. Ashuman, M. Sheezan, S. Masood, and A. Saleem, "Genre Classification of Songs Using Neural Network," *Department of Computer Engg, New Delhi*, 2016. [Online]. Available: [CrossRef]

[26] A. Heakl, A. Abdelgawad, and V. Parque, "A study on broadcast networks for music genre classification," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8. [Online]. Available: [CrossRef]

[27] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, "An evaluation of deep neural network models for music classification using spectrograms," *Multimedia Tools and Applications*, vol. 81, pp. 1–27, 2022. [Online]. Available: [CrossRef]

[28] J. Jakubik, "Evaluation of gated recurrent neural networks in music classification tasks," in *Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology–ISAT 2017: Part I*, 2018, pp. 27–37. [Online]. Available: [CrossRef]

[29] M. Ashraf, G. Geng, X. Wang, F. Ahmad, and F. Abid, "A globally regularized joint neural architecture for music classification," *IEEE Access*, vol. 8, pp. 220980–220989, 2020. [Online]. Available: CrossRef

[30] P. Fulzele, R. Singh, N. Kaushik, and K. Pandey, "A hybrid model for music genre classification using LSTM and SVM," *Proc. 2018 11th Int. Conf. Contemporary Comput. (IC3)*, 2018, pp. 1–3, [Online]. Available: [CrossRef].

[31] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, New Orleans, LA, 2017, pp. 2392–2396, [Online]. Available: [CrossRef].

[32] Ashraf, M.; Abid, F.; Din, I.U.; Rasheed, J.; Yesiltepe, M.; Yeo, S.F.; Ersoy, M.T. A hybrid CNN and RNN variant model for music classification. *Appl. Sci.* **2023**, *13*, 1476. Available online: [CrossRef].

[33] N. Srivastava, S. Ruhil, and G. Kaushal, "Music Genre Classification using Convolutional Recurrent Neural Networks," in *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, 2022, pp. 1–5. [Online]. Available: CrossRef

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: CrossRef

[35] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia, vol. 13, no. 2, pp. 303–319*, Apr. 2011. [Online]. Available: CrossRef

[36] Z. Fu, G. Lu, K.-M. Ting, and D. Zhang, "On feature combination for music classification," *in Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern Recognit. (SPR), Struct. Syntactic Pattern Recognit. (SSPR)*, 2010, pp. 453–462. [Online]. Available: CrossRef